Data Mining

Table of Content

Suitable for people

✓ People with high school mathematics basis

✓ People who want to quickly get started with data mining

✓ People who want to develop in data analysis and machine learning

Pre class preparation

Environment	Windows 64 bit or Linux64 bit PC
Data	Exercise data
Tool	Download and install Raqsoft YModel Download and install Raqsoft esProc

Pre class preparation (exercise data and tool download address)

Exercise data

Exercise_1.csv
Exercise_2.csv
Titanic.csv
Houseprice.csv
meter_data.csv
catering_sale.csv

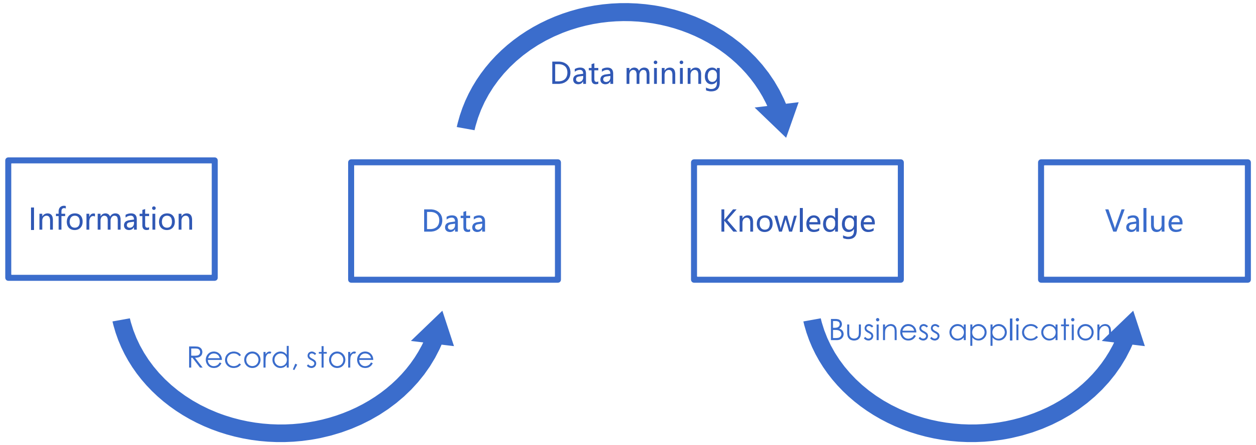
Download of Ymodel, esProc and the free license

http://www.raqsoft.com/ymodel-download

Chapter 1 The concept of data mining

The concept of data mining

Data mining is a process of extracting hidden, unknown and potentially useful information and knowledge from a large number of incomplete, noisy, fuzzy and random practical application data.



Generally, when we transform information into value, we have to go through four levels: information, data, knowledge and value. Data mining is an important part in the process of finding knowledge from data.

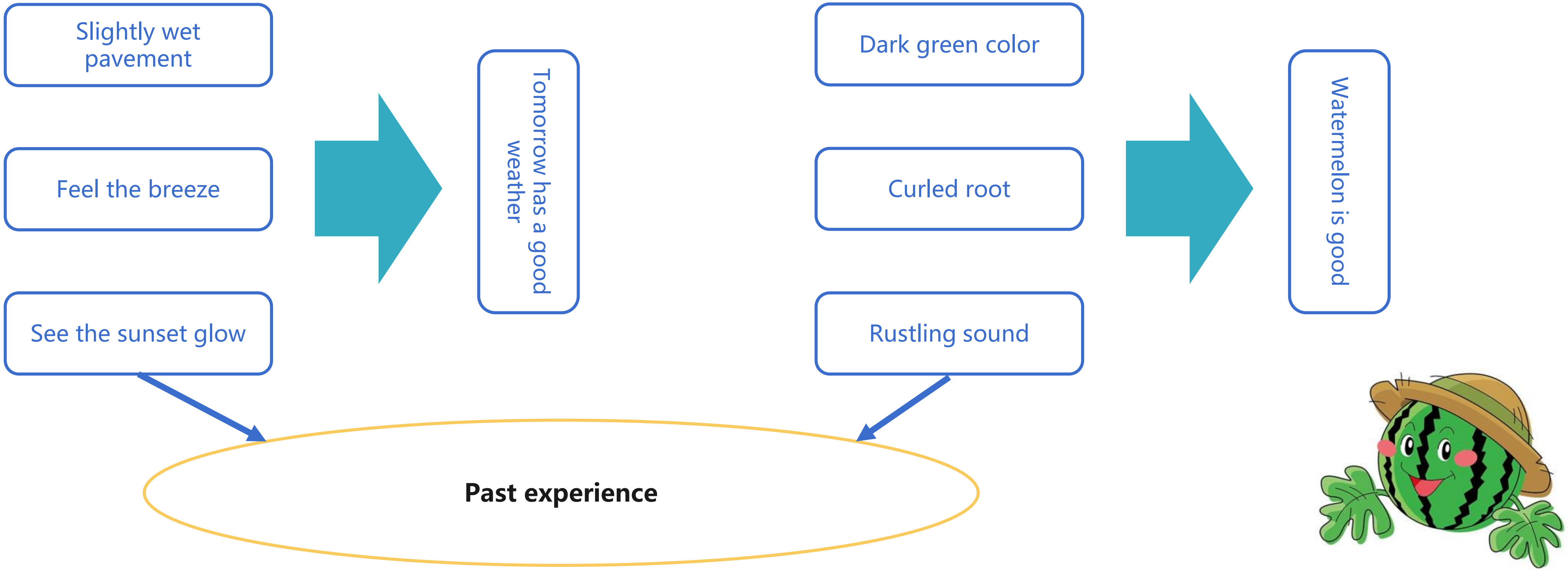
The concept of data mining

Synonyms similar to data mining include machine learning, artificial intelligence, business intelligence, pattern recognition, knowledge discovery, data analysis and decision support, etc.

The common methods of data analysis using data mining include classification, regression analysis, clustering, association rules, features, change and deviation analysis, web page mining, etc.

The concept of data mining

In the evening, the road surface of the street is wet after a little rain, and the gentle breeze blows. Look up at the sunset glow in the sky. Well, tomorrow has fine weather. Go to the fruit stand, pick up a dark green watermelon with curled root and rustling sound, and look forward to enjoying it.



Slightly wet pavement

Feel the breeze

See the sunset glow

Tomorrow has a good weather

Dark green color

Curled root

Rustling sound

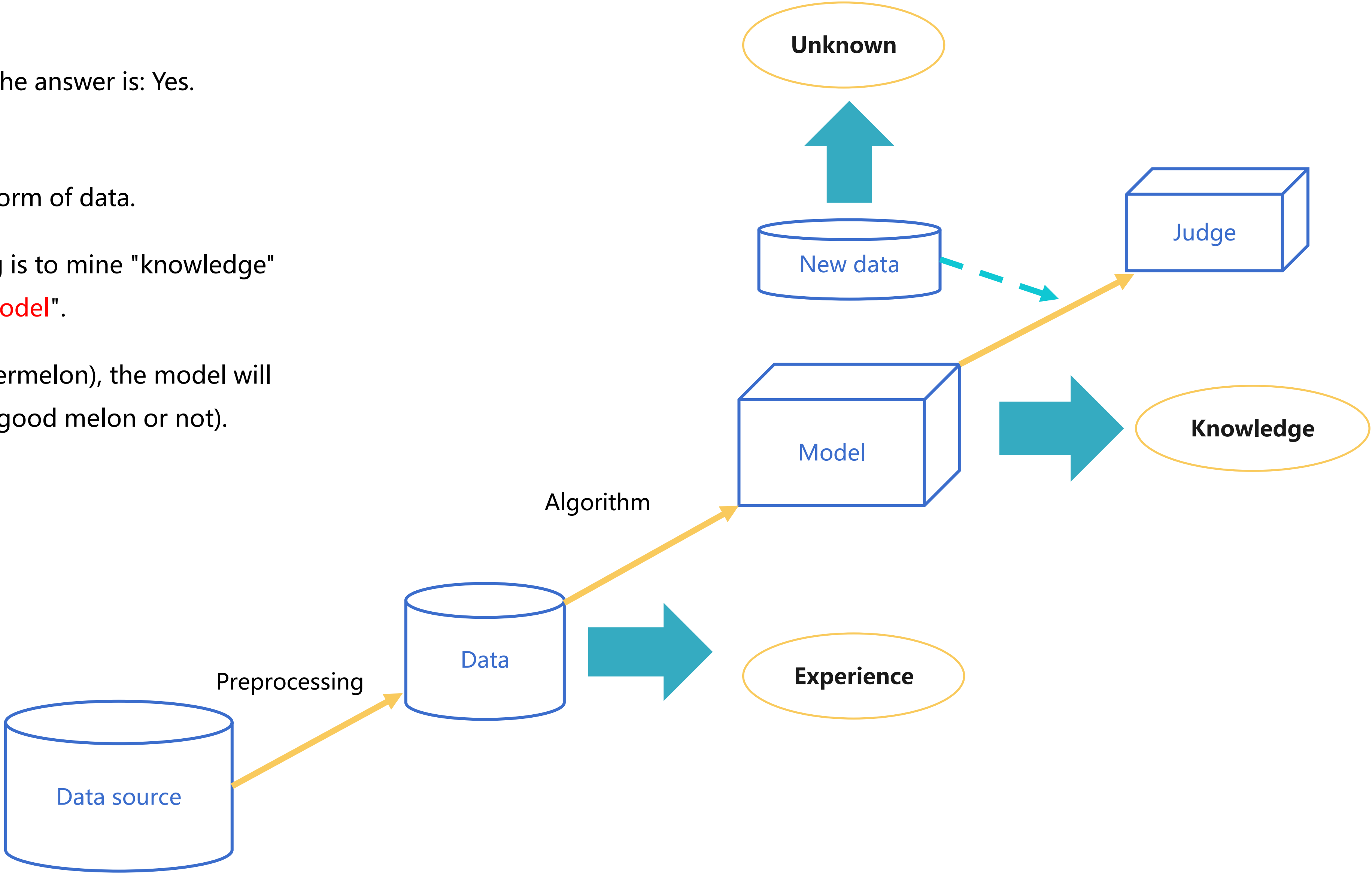
Watermelon is good

Past experience

The concept of data mining

Can the machine help us finish this? The answer is: Yes.

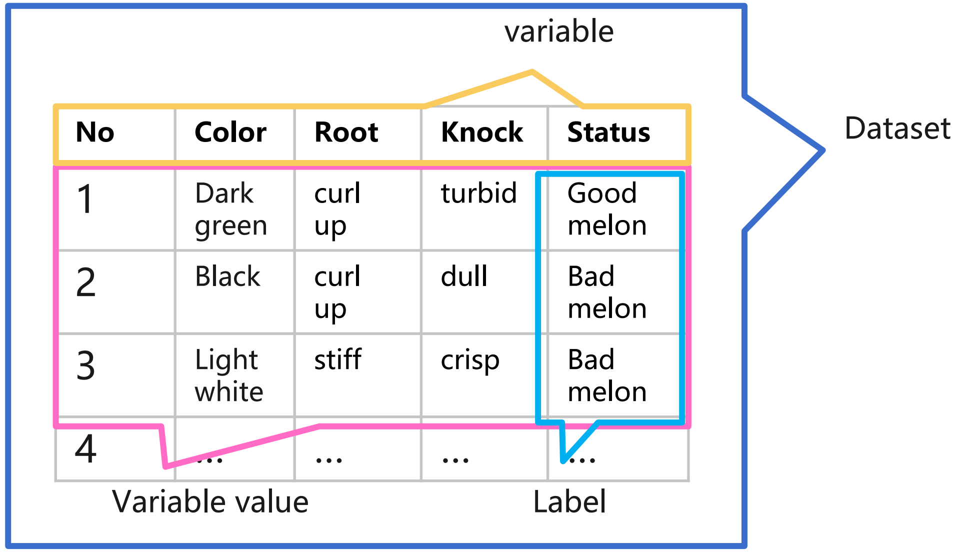
- Experience usually exists in the form of data.
- The main content of data mining is to mine "knowledge" from historical data to create "model".
- In the new situations (uncut watermelon), the model will help us to judge (whether it is a good melon or not).



The concept of data mining

In terms of mathematical language that high school students can understand, the essence of modeling task is:
According to some existing correspondence from input space X (such as {[color = dark green; root = curl up; knock = turbid sound], [color = black; root = curl up; knock = dull], [color = light white; root = stiff; knock = crisp]}) to output space Y (such as {good melon, bad melon, bad melon}), find a function f : $X \xrightarrow{f} Y$ to describe this correspondence, this function is the model we want.

With the model, it's easy to make predictions, it means, take a new set of x and use this function to calculate the y.



No	Color	Root	Knock	Status
1	Dark green	curl up	turbid	Good melon
2	Black	curl up	dull	Bad melon
3	Light white	stiff	crisp	Bad melon
4	...	...	...	...

Model ≠ function?

The reason why we are more accustomed to call the model as a model rather than a function is that it does not meet the certainty we usually expect from the function. Here, the same X may correspond to different Y (melons with the same color, root and knocking sound may be good or bad).

The concept of data mining

But how is the model build, in other words, how to find the function?

Think about how to make a person have the ability to judge whether a melon is good or bad? You need to practice with a batch of melons to get the characteristics (color, root, knocking, etc.) before you cut it, and then you can cut it to see whether it is good or not. Over time, this person will be able to learn to judge the quality of the melon by the characteristics of the melon before it is cut open.

Simply think that the more melons you use for practice, the more experience you can gain, and the more accurate your judgment will be in the future.

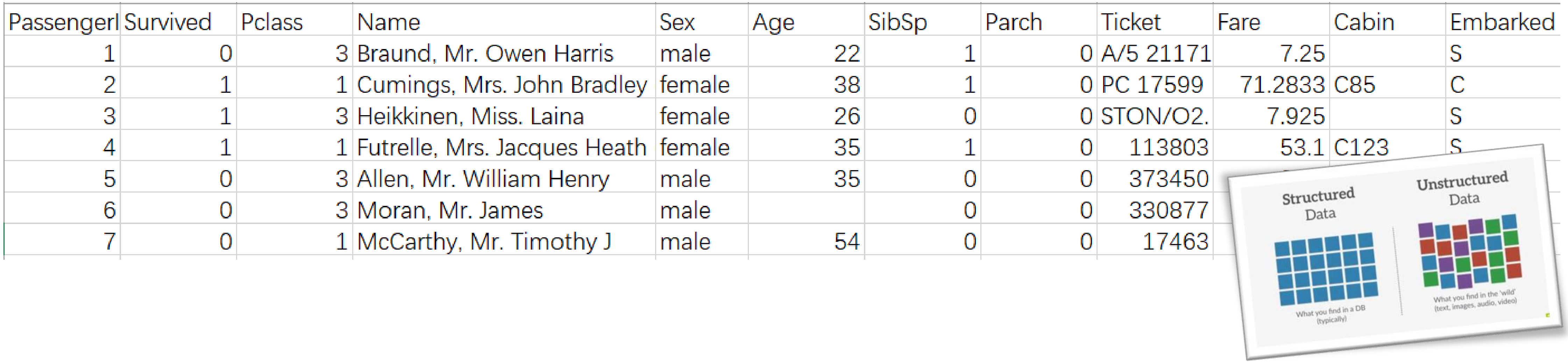
It's the same thing to do data mining with machines. We need to use historical data (melon used for practice) to build models, and the modeling process is also called training, and these historical data are called training datasets.

The concept of data mining

We usually say that training data should be organized into structured data before modeling, so what is structured data?

Structured data refers to data in two-dimensional form. The general feature is that data is in rows (also known as samples), one row of data represents the information of an entity, and the attributes (also known as fields) of each row of data are the same. It can come from databases, text, or file storage systems such as HDFS.

See the figure below for the data of predicting Titanic survivors:



PassengerI	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	[illegible]	[illegible]	[illegible]
6	0	3	Moran, Mr. James	male		0	0	330877	[illegible]	[illegible]	[illegible]
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	[illegible]	[illegible]	[illegible]

Structured Data
What you find in a DB (typically)
Unstructured Data
What you find in the 'wild' (text, images, audio, video)

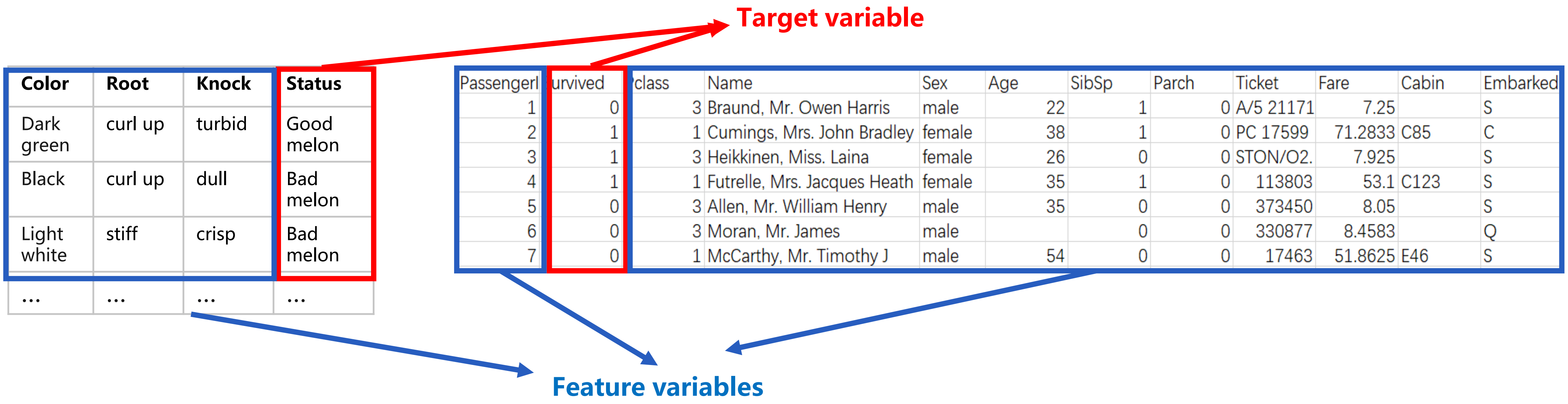
The concept of data mining

Obviously, the training dataset must have the target we care about (the quality of the melon), that is, the Y must have a value (the melon used for practice, its quality is known), which is called the target variable.

In the training data set, of course, there are also features to judge whether the melon is good or not, such as color, rooting, knocking sound, which are called feature variables.

In terms of structured data, target variables and feature variables are attributes or fields of data.

The target variables and feature variables of the predicted good melon and Titanic survivors are as follows:



Target variable

Color	**Root**	**Knock**	**Status**
Dark green	curl up	turbid	Good melon
Black	curl up	dull	Bad melon
Light white	stiff	crisp	Bad melon
…	…	…	…

PassengerI	urvived	class	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S

Feature variables

Class exercise

Consider a data mining task: build a model based on the following dataset to predict the type of flowers (setosa, versicolor, virginica) through their attributes. The dataset is as follows:

sepal length	sepal width	Petal length	Petal width	Type
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
…	…	…	…	…

Thinking:

What are the feature variables in this task?

What is the target variable?

Types of data mining

Machine learning can be divided into supervised learning, unsupervised learning, semi supervised learning (also known as reinforcement learning by Hinton), etc.

Here, we mainly understand supervised learning and unsupervised learning.

1. **Supervised learning**: know the relationship between input and output result according to the existing dataset. According to this known relationship, an optimal model is obtained by training.

In supervised learning, the training data has both features and label(target variable). Through training, the machine can find the relationship between features and label by itself. When facing the data with only features but no label, it can judge the label.

For example, the accuracy of the exercises with standard answers and then to the exam is higher than that of the exercises without answers and then to the exam

Another example: when we were young, we didn't know whether cattle and birds belonged to the same category, but as we grew up, we kept inputting all kinds of knowledge, and the models in our brains became more and more accurate, and the judgment of animals became more and more accurate.

Types of data mining

2. **Unsupervised learning**: we don't know the relationship between data and features in dataset, but we need to get the relationship between data according to clustering or certain model.

Compared with supervised learning, unsupervised learning is more like self-learning. There is no label (target variable) for machines to learn to do things by themselves.

For example, when we visit a painting exhibition, we know nothing about art, but after we appreciate many works, we can also divide them into different factions.

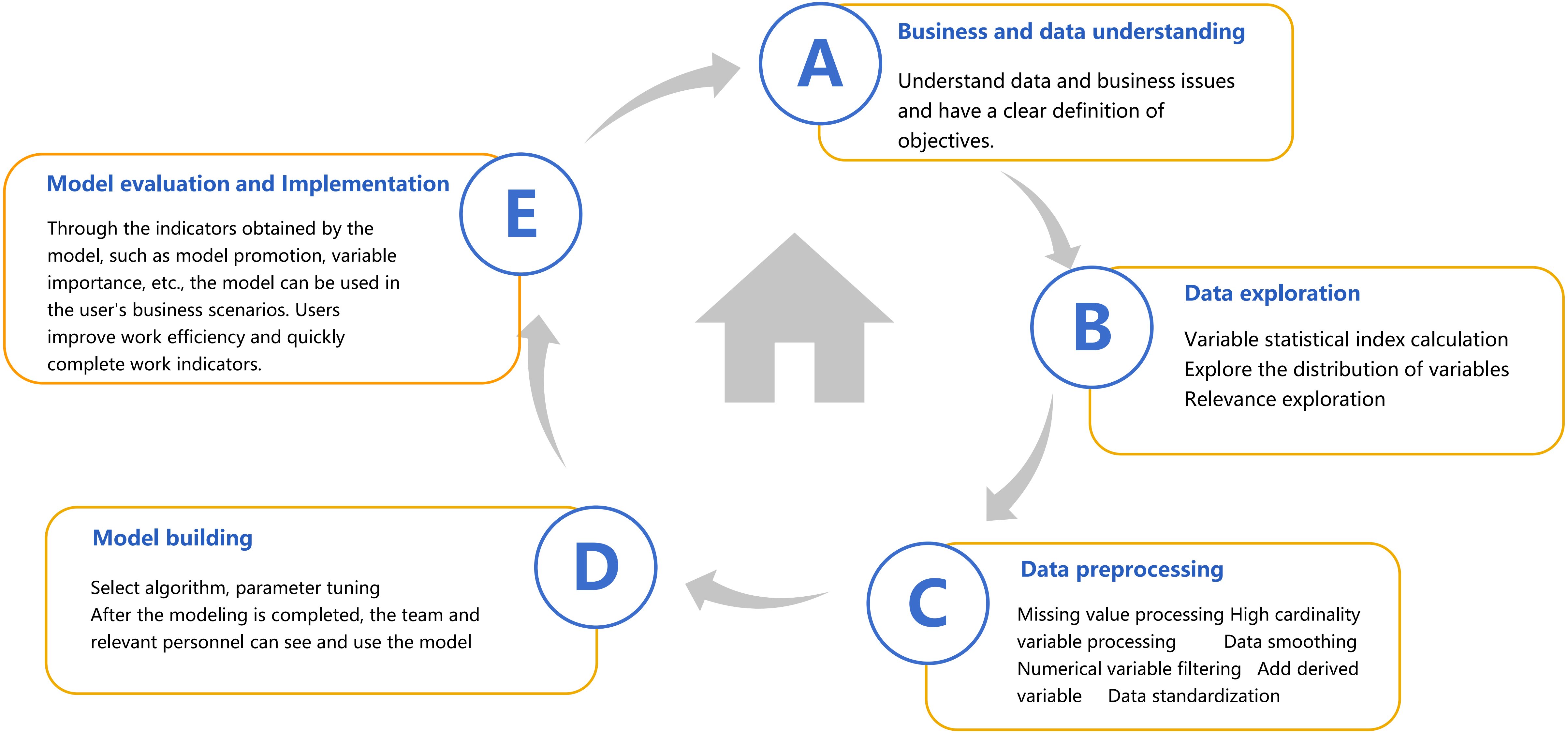
In short: the biggest difference between supervised and unsupervised learning is whether there is a label, i.e. target variable Y.

Class exercise

1. **Predict house price**: there are different house positions, sizes, orientations and prices in the dataset. Build a model to predict the prices of other houses.
2. **Estimate the nature of the tumor**: the dataset includes the patient's gender, age , tumor size, location, label of benign or malignant tumor. Use it to establish a model to judge the nature of the tumor of the new patient.
3. **Google News**: Google News collects a lot of news content on the Internet every day. It groups these news to form related news. These news events are all on the same theme, so they are presented together. For example, according to the different content structure, it can be divided into finance, entertainment, sports, etc.
4. **Group people according to a given gene**: for a different group of people, we measure the expression of a specific gene in their DNA. Then we can use clustering algorithm to divide them into different types according to the measurement results.

Thinking：Which are supervised learning and which are unsupervised learning?

Data mining process



A

Business and data understanding

Understand data and business issues
and have a clear definition of
objectives.

B

Data exploration

Variable statistical index calculation
Explore the distribution of variables
Relevance exploration

C

Data preprocessing

Missing value processing High cardinality
variable processing Data smoothing
Numerical variable filtering Add derived
variable Data standardization

D

Model building

Select algorithm, parameter tuning
After the modeling is completed, the team and
relevant personnel can see and use the model

E

Model evaluation and Implementation

Through the indicators obtained by the model, such as model promotion, variable importance, etc., the model can be used in the user's business scenarios. Users improve work efficiency and quickly complete work indicators.

Glossary - for reference

data mining

machine learning

model

sample

feature

label

train set

structured data

supervised learning

unsupervised learning

semi-supervised learning

Chapter 2 Data exploration

The significance of data exploration

View features

Using tools to view the characteristics of data

Perceive value

Understand the influence of featue variables on target variable and decide which variables to choose

Understand data

Understand the statistical characteristics of variables and the correlation between variables

Data type

A

Quantitative data

B

Qualitative data

Data type - A Quantitative data

Definition	the data can measured on a numerical scale.
Example	Value of sales volume, whose size represents its sales status.
Content	Quantitative data in this tutorial includes: ● **Numerical variable** ● **Count variable** ● **Time date variable**

Count variable

Count variable	Variable with integer value
Examples	Class size of a class [45,67,53， ...] The number of trains arriving at Beijing railway station every hour [2,5,6,7， ...]， but it can't be that the number of trains arriving every hour is 6.5
Thinking	1. The values of a set of data are: [-12,-5,-8,...] Is it a count variable? 2. User's ID number :[110000198003198182, 130000197407258697， ...] Is it a count variable?

Numerical variable

Numerical variable	The variable with floating-point value. It is a variable that can get any numerical value within a given range, that is, a variable with decimal point
Examples	The height of a class [175.5,180.4,165.3...] Sales volume of the enterprise in the first quarter: [2300.87,1098.8，...] Bank deposit of depositor：[4035.65,2053.89，...]

Time date variable

Time date variable	Variables representing time and date
Examples	Date of birth:[2009-01-01，1998-03-05，...] Login time of user:[2019/1/1 12:00:00,...] Corporate loan date：[2019/Sep,2015/May...]
Note	Time and date variables are important factors in many scenarios, such as the change of house price with time, the periodicity of sales volume of a product (more down jackets are purchased in winter), user consumption habits and payday, etc ... Note: due to different data sources, there may be many formats of date, such as separated by "-" or "/" , whether the date is in Chinese or English. For the convenience of calculation, unified format is required first.

Data type - B Qualitative data

Definition	Measurements that cannot be recorded on a numerical scale, which can only be classified into different categories.
Example	Gender, two values represent two groups
Content	Qualitative data in this tutorial includes: ● **Unary variable** ● **Binary variable** ● **Categorical variable** ● **Text string variable**

Count variable

Thinking

1. The values of a set of data are: [-12,-5,-8,...] Is it a count variable?

2. User's ID number :[110000198003198182, 130000197407258697， ...] Is it a count variable?

Single valued variable

Unary variable	Variable containing only one category (without missing values)
Examples	Household voltage :[220,220,,...] Sold or not (only the sold ones are recorded) :[1,,1,1,,,,1...]
Thinking	Is unary variable useless in data mining?

Binary variable Categorical variable

Binary variable	Variable with only two categories (without missing values), which is often the target variable
Examples	Gender: [male, female] Sold or not: [1,0,1,1,0,0,0,1...] Whether the user defaults：[Yes, No,...]
Categorical variable	Variables with more categories than two
Examples	Industry: [tourism, manufacturing, IT, ...] Education background: [doctor, master, bachelor, ...]
Thinking	Annual income: [1,2,3，...] (use numbers to show income levels, such as high, medium, low,...) What kind of variable is annual income?

Long text variable

Text string variable	A long string (usually more than 128 bytes) or a text variable with a large number of categories
Examples	Story introduction:[Harry potter says:” ....” ,He is ....,......] Home address: [a community in Haidian District, Beijing, a community in Wuhan City, Hubei Province, ...] Name：[Braund, Mr. Owen Harris, Cumings, Mrs. John Bradley (Florence Briggs Thayer),...]
Thinking	Text string variables generally can't be used directly. They need to be transformed again. There are no uniform rules for the transformation and it needs to be analyzed specifically according to data characteristics and business requirements. For example, the province, city can be extracted from the home address. Mr, Mrs, miss can be extracted from names.

Data type

	Variable type	Description	Example
Quantitative data	Count variable	Variable with integer value	Class size:[45,67,53...] Number of rooms :[2,5,6,7...]
	Numerical variable	Variable with floating point value	Height:[175.5,180.4,165.3...] Sales volume:[2300.87,1098,8...]
	Time date variable	Variable representing time and date	Birthday:[2009-01-01...] Login time:[2019/1/1 12:00:00,...]
Qualitative data	Unary variable	Variable containing only one category (without missing values)	Household voltage :[220,220,,...] Sold or not (only recorded sold):[1,,1,1,,,,1...]
	Binary variable	Variable with only two categories (without missing values), which is often the target variable	Gender:[male, female] Sold or not :[1,0,1,1,0,0,0,1...]
	Categorical variable	Variables with more categories than two	Industry: [tourism, manufacturing, IT, ...] Annual income:[1（High）,2（Medium）,3（Low）, ...]
	Text string variable	Variables with a length of more than 128 bytes and a very large number of classifications, which generally cannot be used directly and need to be transformed again	Story introduction:[Harry potter says:” ....” ,He is ....,......]
	ID	A unique identifier for each record, which is usually useless.	ID:[110000198003198182, 130000197407258697， ...]

Class exercise

Chemical and manufacturing plants sometimes discharge toxic substances, such as DDT, into nearby rivers. These toxic substances will harm the animals and plants in the river and on the bank. The U.S. Army Corps of engineers conducted a survey of fish contamination in three tributaries of the Tennessee River (Alabama) - Flint River, limestone River and spring river. A total of 144 fish were caught and the values of the following five variables were measured and recorded:

1. Rivers from which fish come
2. Species of fish (River catfish, broad mouth bass, small mouth bullfish)
3. Length (CM)
4. Weight (g)
5. DDT concentration in parts per million

Please confirm whether these five variables are qualitative data or quantitative data?

Class exercise

Body length, weight and DDT concentration are measured by numbers, so they are quantitative variables. Body length is in centimeters, weight is in grams, and DDT is in parts per million.

The types of rivers and species of fish can not be measured by numbers, but can only be classified into categories (for example, there are four kinds of fish), so they are qualitative variables.

The description and analysis methods of quantitative data and qualitative data are different. Therefore, it is very important to learn to distinguish data types in data analysis.

Class exercise

Exercise: using YModel to identify the data types of Titanic survival prediction data.

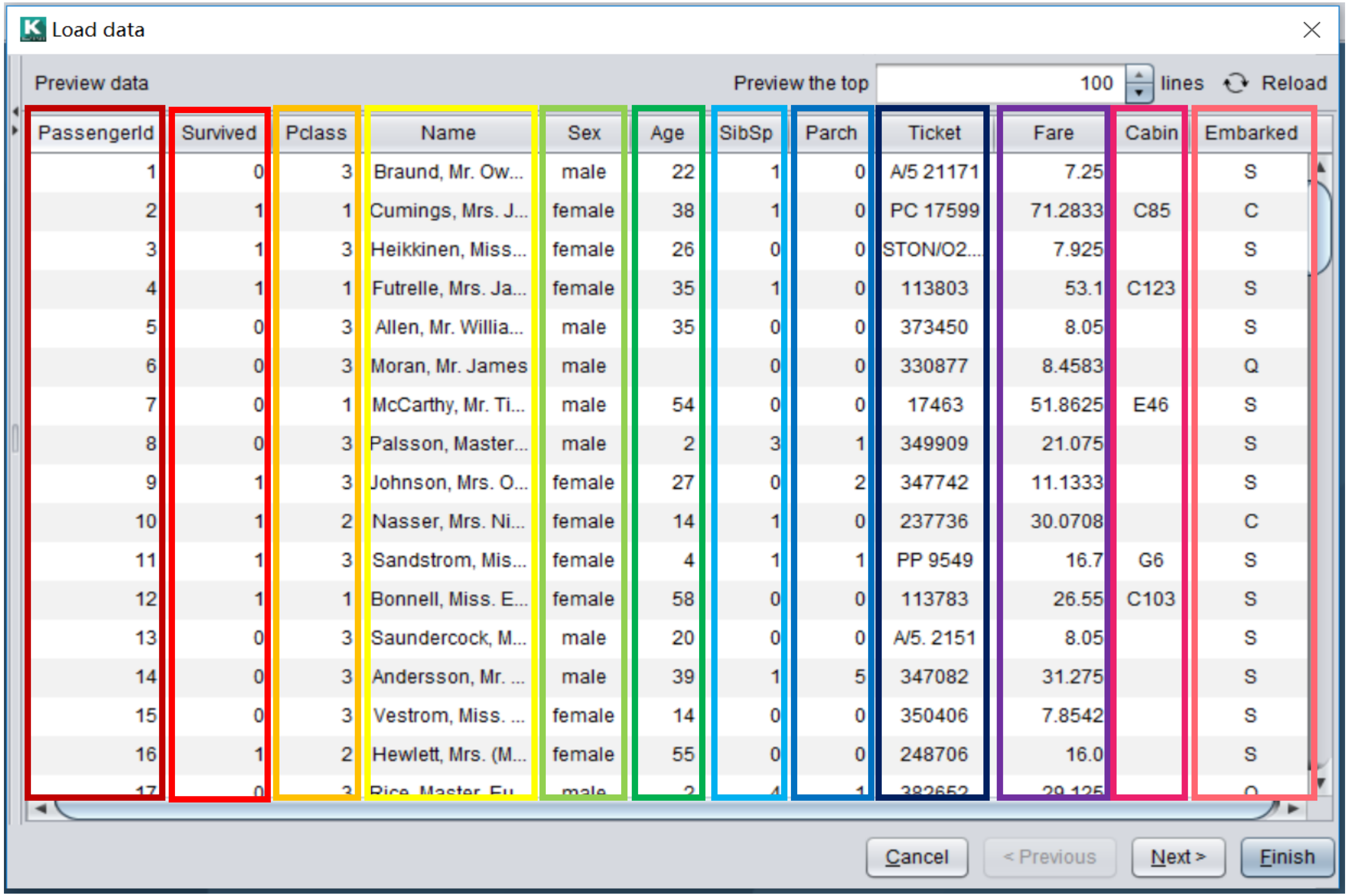
Data：Titanic.csv

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S

Class exercise

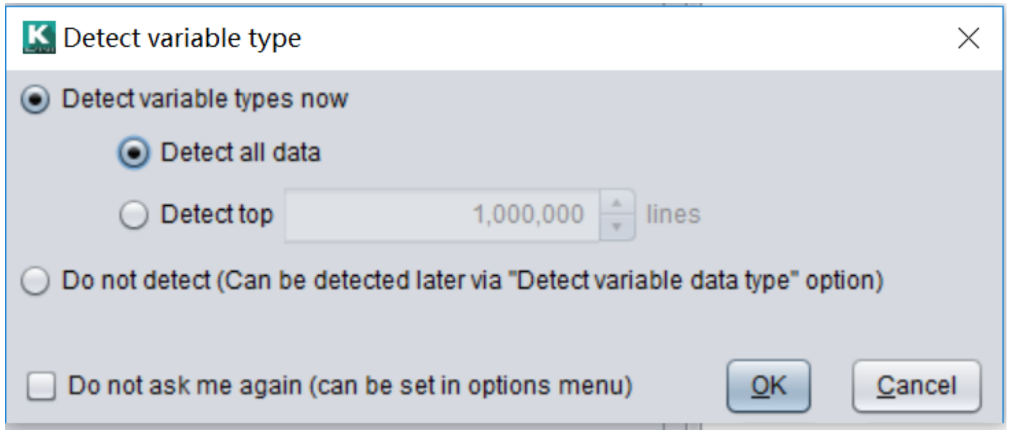
We take the Titanic survival prediction data on kaggle as an example, and use YModel to identify data types

1. Data preview



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Ow...	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. J...	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss...	female	26	0	0	STON/O2...	7.925		S
4	1	1	Futrelle, Mrs. Ja...	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. Willia...	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Ti...	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master...	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. O...	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Ni...	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Mis...	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. E...	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, M...	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. ...	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. ...	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (M...	female	55	0	0	248706	16.0		S
17	0	3	Rice, Master. Eu...	male	2	4	1	382652	29.125		Q

Class exercise



Detect variable type

● Detect variable types now

● Detect all data

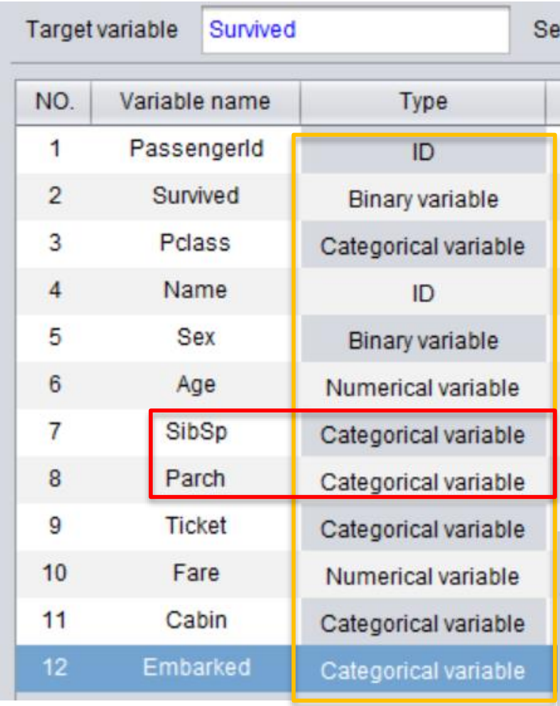
○ Detect top 1,000,000 lines

○ Do not detect (Can be detected later via "Detect variable data type" option)

☐ Do not ask me again (can be set in options menu)

2. Select detection data range

3. Automatic identification of data types by YModel tool

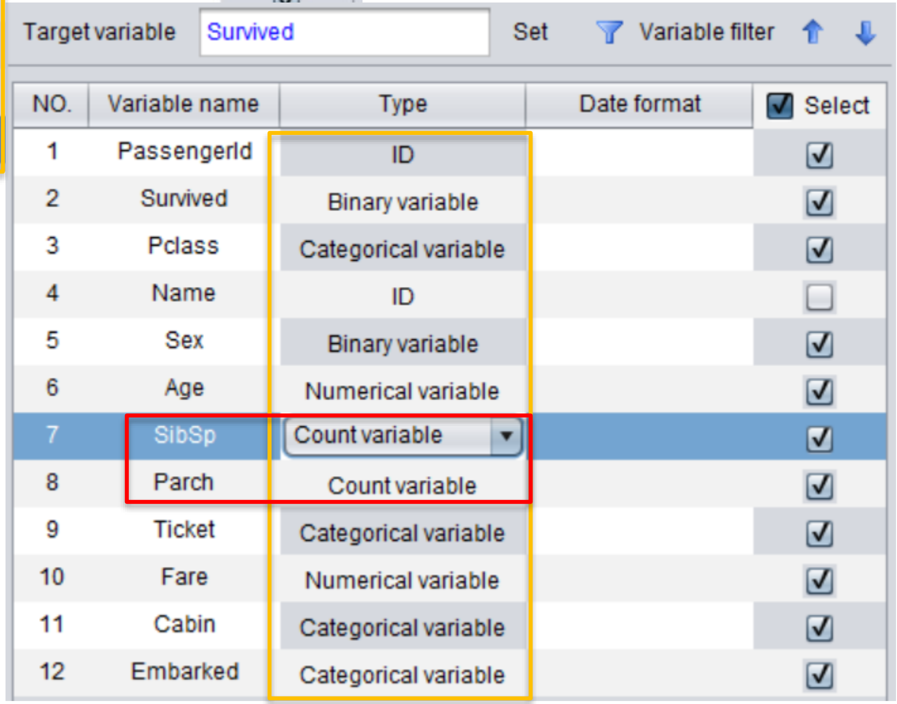


Target variable: Survived

NO.	Variable name	Type	Date format	Select
1	PassengerId	ID		
2	Survived	Binary variable		
3	Pclass	Categorical variable		
4	Name	ID		☐
5	Sex	Binary variable		☑
6	Age	Numerical variable		☑
7	SibSp	Categorical variable		☑
8	Parch	Categorical variable		☑
9	Ticket	Categorical variable		☑
10	Fare	Numerical variable		
11	Cabin	Categorical variable		
12	Embarked	Categorical variable		

Automatically identify data types

Two fields, SibSp and Parch, have only numbers 0-6, so they are recognized as categorical variables.
However, the fields are interpreted as the number of siblings and spouses and the number of parents and children, so they should be changed to count variables.



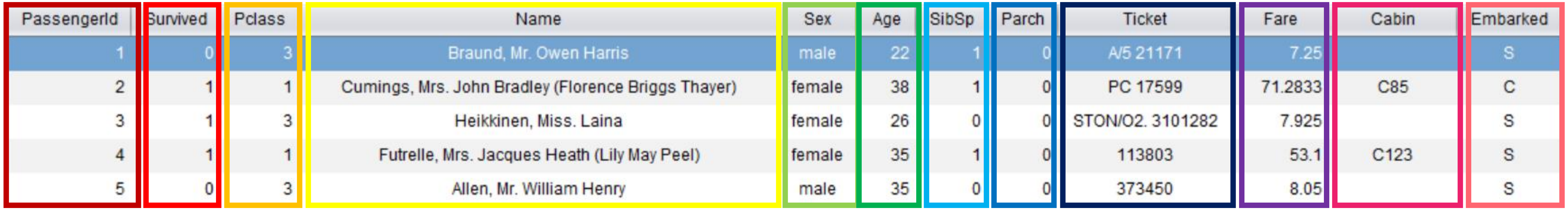
Target variable: Survived

NO.	Variable name	Type	Date format	☑ Select
1	PassengerId	ID		☑
2	Survived	Binary variable		☑
3	Pclass	Categorical variable		☑
4	Name	ID		☐
5	Sex	Binary variable		☑
6	Age	Numerical variable		☑
7	SibSp	Count variable		☑
8	Parch	Count variable		☑
9	Ticket	Categorical variable		☑
10	Fare	Numerical variable		☑
11	Cabin	Categorical variable		☑
12	Embarked	Categorical variable		☑

Class exercise

Variables of Titanic data

	No.	Variable	Description	Type
🟥	1	PassengerId	Passenger ID	ID，Unique ID
🟥	2	Survived	Survived or not	Binary variable, target variable
🟧	3	Pclass	Ticket class	Categorical variable
🟨	4	Name	Passenger name	ID，Unique ID
🟩	5	Sex	Passenger gender	Binary variable
🟩	6	Age	Passenger age	Numerical variable
🟦	7	SibSp	Number of siblings and spouses	Count variable
🟦	8	Parch	Number of parents and children	Count variable
🟦	9	Ticket	Ticket No	Categorical variable
🟪	10	Fare	Fare price	Numerical variable
🟥	11	Cabin	Cabin	Categorical variable
🟥	12	Embarked	Port of embarkation	Categorical variable



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

Data type

1

For different data types, in the model training of algorithm, the way of processing and treatment is different. Quantitative data are calculated directly.

2

Qualitative data may be converted to sparse matrix: each category is a new field, and then calculated according to its value "1" "0 ".

3

Make sure to correctly judge the data type based on the business meaning of the field.

Glossary - for reference

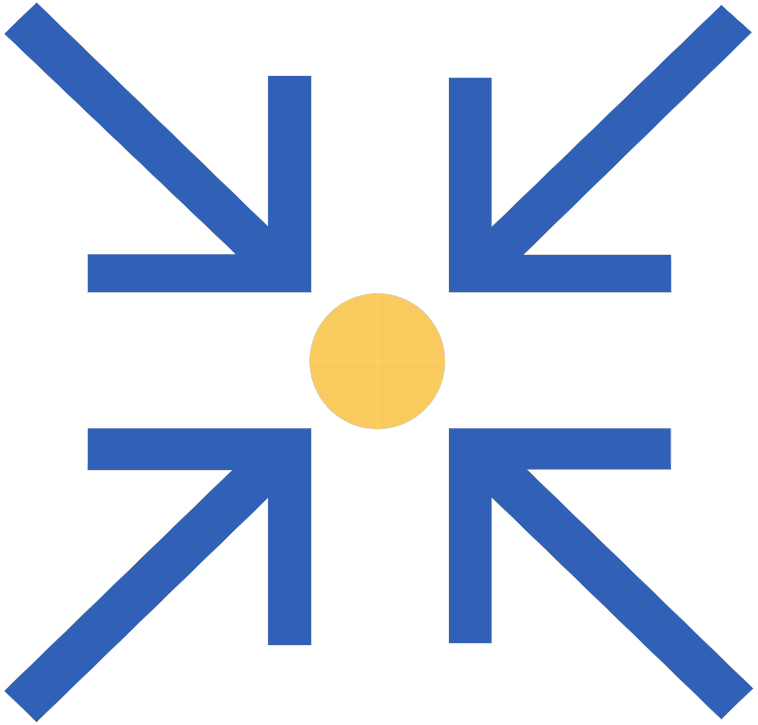
quantitative data

qualitative data

variable

sample

Exploration method of quantitative data



Measures of central tendency

Mean
Median
Mode

Measures of variability

Range
Variance
Standard deviation

Measures of relative standing

Upper quartile
lower quartile
Z-score

Measures of symmetry

Skewness

Numeric variable, count variable, time and date variable

Exploration method of quantitative data

Measures of central tendency

Mean

That is, the average value, whose size reflects the overall level

For example: dataset {5, 3, 8, 5, 6}, Mean value = (5 + 3 + 8 + 5 + 6) / 5 = 5.4

Median

The value in the middle after the dataset is arranged in ascending (or descending) order

Arrange n measurements from small to large:
If n is an odd number, the median is the middle number
If n is even, it's the average of the middle two numbers

For example, the median of data set {3,4,5,7,8} is 5,
The median of data set {2,4,5,7} should be (4 + 5) / 2 = 4.5

Mode

Is the most frequent number in the dataset

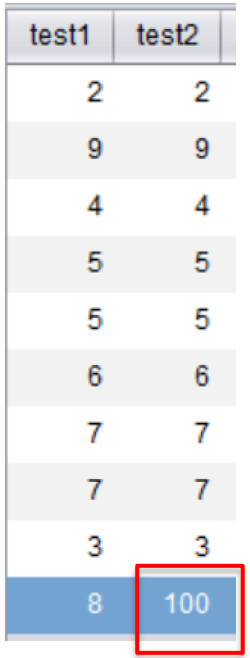
For example: the mode of data set {1,2,2,3,4,7,9} is 2

Exploration method of quantitative data

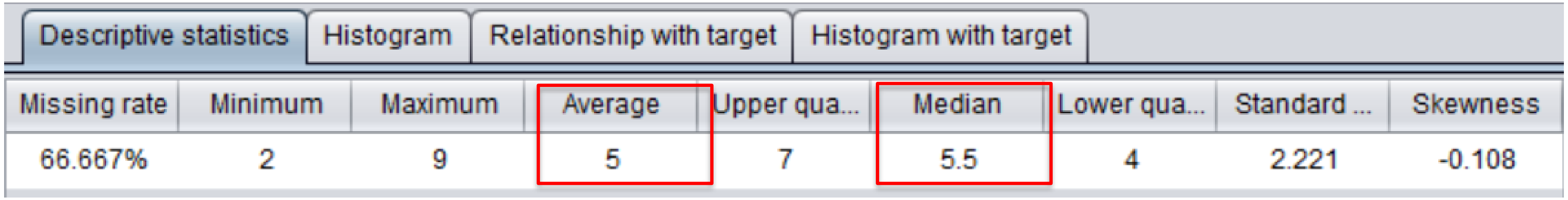
The average is the most commonly used and easily understood measure of the central trend. If the data distribution is **normal,** the **average(mean value)** is usually the best measure of the central trend.

If there are several **maxima / minima or skews** in the dataset, the **median** is the best measure of the central trend.

For example ,for test1 and test2 in exercise data Exercise_1.csv, open the data with YModel, and the statistical results are as follows:

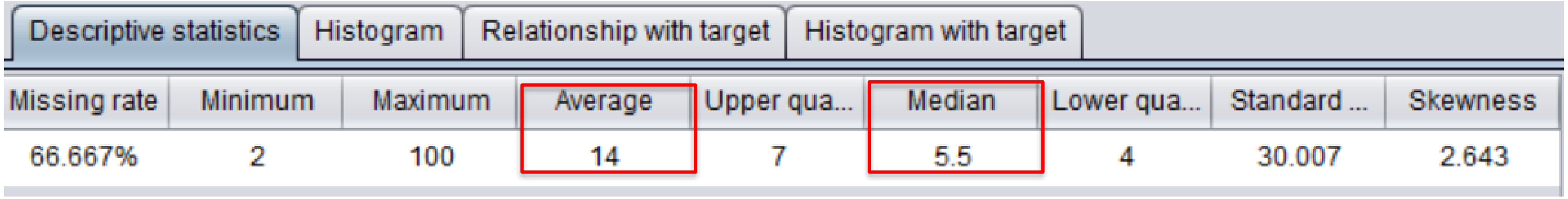


test1	test2
2	2
9	9
4	4
5	5
5	5
6	6
7	7
7	7
3	3
8	100



Missing rate	Minimum	Maximum	Average	Upper qua...	Median	Lower qua...	Standard ...	Skewness
66.667%	2	9	5	7	5.5	4	2.221	-0.108

test1



Missing rate	Minimum	Maximum	Average	Upper qua...	Median	Lower qua...	Standard ...	Skewness
66.667%	2	100	14	7	5.5	4	30.007	2.643

test2

Due to the existence of the extreme value 100, the mean value is affected but the median value remains unchanged.

Class exercise – Mean, Median

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Player income	4023	3274	227	129	154	269	463	196	190	232	90	84	65	90	36	35

This is the income data of a group of players (the player income variant of Exercise_1.csv)

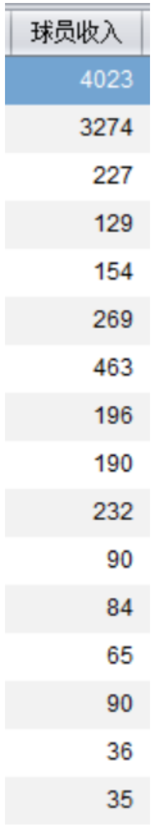
Calculate

What is the average value of the data? What's the median?

Which can accurately describe a player's typical income situation?

Class exercise – Mean, Median

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Player income	**4023**	**3274**	**227**	**129**	**154**	**269**	**463**	**196**	**190**	**232**	**90**	**84**	**65**	**90**	**36**	**35**

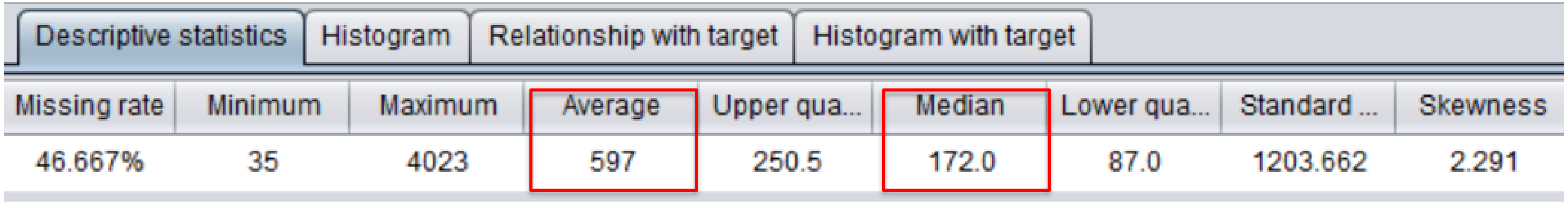


球员收入
4023
3274
227
129
154
269
463
196
190
232
90
84
65
90
36
35

The average is 597, and the median is 172

The average value exceeds the salary value of most players in the sample, thus misleading the measurement of concentration trend.

The median can more accurately describe the typical salary situation of professional athletes.



Missing rate	Minimum	Maximum	Average	Upper qua...	Median	Lower qua...	Standard ...	Skewness
46.667%	35	4023	597	250.5	172.0	87.0	1203.662	2.291

Class exercise – Mode

Mode is the most frequent number in the dataset.

A set of data may have multiple modes, or no mode.

1	2	3	4	5	6	7	8	9	10
13.5	**8.4**	**10.5**	**10.6**	**10.1**	**8.2**	**11.1**	**11.3**	**9.5**	**11.7**

1	2	3	4	5	6	7	8	9	10
13.5	**8.4**	**10.5**	**10.6**	**10.1**	**8.2**	**11.1**	**11.3**	**8.4**	**8.2**

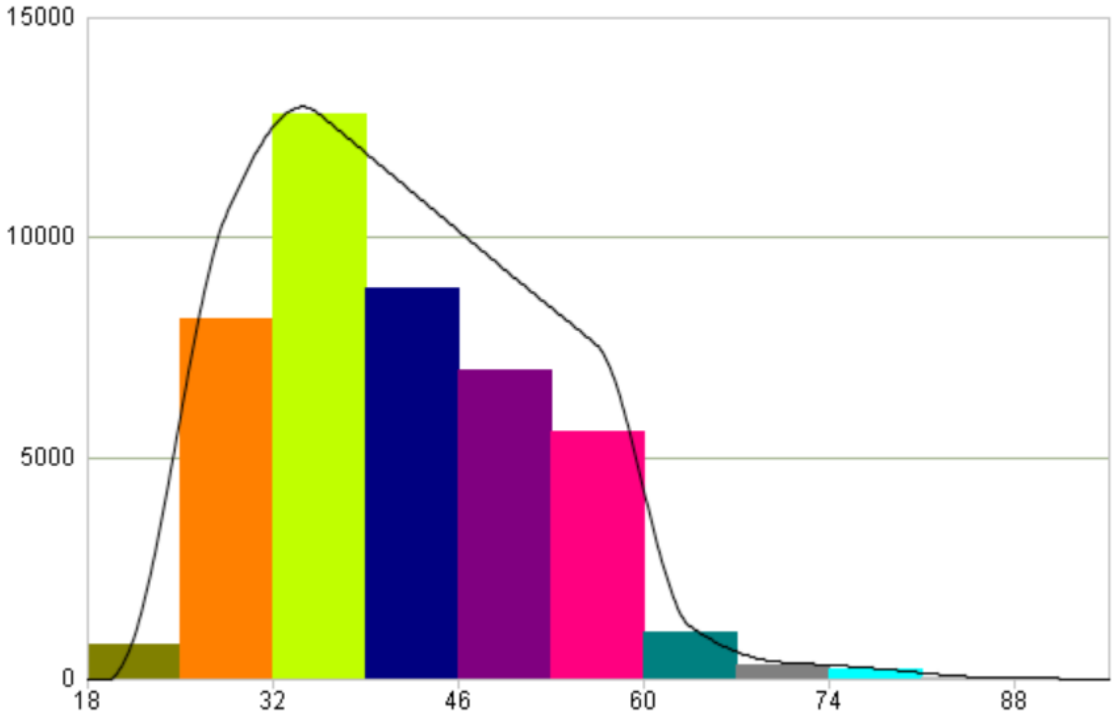
Do two sets of data have modes respectively?
What is it?

Exploration method of quantitative data

Mode may not be meaningful for quantitative data, but the grouping interval (modal class) containing the maximum frequency is more meaningful.

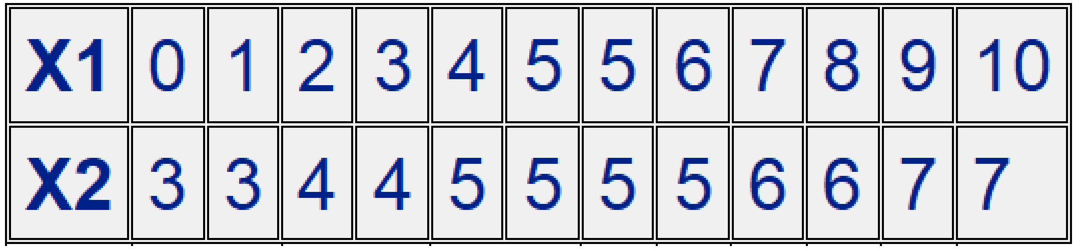
As shown in the figure, the age mode range of a product customer group is (32, 39) (age variant in Exercise_2.csv)

In applications involving quantitative data, the mean value and median usually provide more information than modes.



Exploration method of quantitative data

The measurement of the centralized trend only provides a part of the description of the dataset, which is incomplete. There are many other significant differences in the data that cannot be described by the centralized trend.



X1	0	1	2	3	4	5	5	6	7	8	9	10
X2	3	3	4	4	5	5	5	5	6	6	7	7

The mean and median of the two groups of data are the same, but the data distribution is different.

The dispersion measure (also known as variability) of datasets indicates the dispersion or dispersion degree of data, and the variability of data is also very important.

Exploration method of quantitative data

Measures of variability

Range

Range: max - min

Easy to calculate, but when the dataset is large, the response to data changes is quite insensitive.

Variance

Variance reflects the dispersion degree of each value from the average value.

The larger the variance value is, the greater the dispersion degree is .

Variance formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Standard deviation

Is the square root of variance, meaning similar to variance.

Class exercise - Range

Range: max - min

Easy to calculate, but when the dataset is large, the response to data changes is quite insensitive.

Age1

22 38 26 35 35 0 54 2 27 14 4 58 20 39 14 55 2 56 31 29 35 34 15 28 8 38 30 19 25 28

Age2

52 68 56 65 65 30 84 32 57 44 34 88 50 69 44 85 32 86 61 59 65 64 45 58 38 68 60 49 55 58

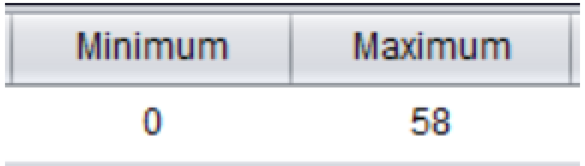
Check the range of two sets of data using the tool. (age1 and age2 variants in Exercise_1.csv)

Do two groups of data represent the same age group?

Class exercise - Range

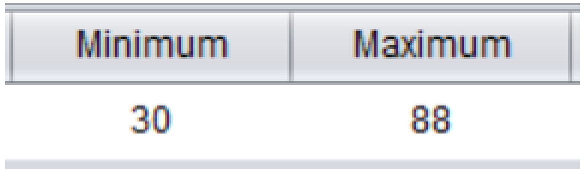
Open exercise data, import to YModel, view statistics

Age1:



Minimum	Maximum
0	58

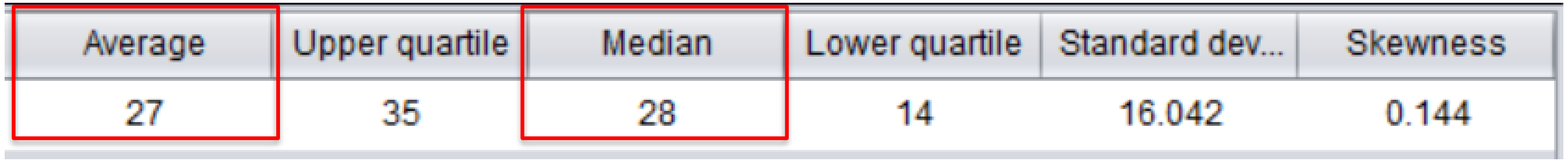
Age2:



Minimum	Maximum
30	88

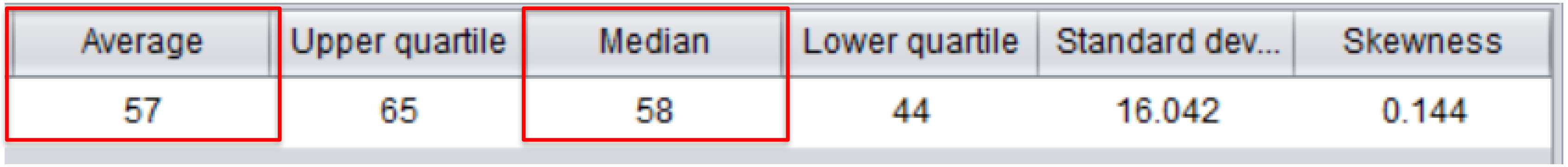
The range of the two groups of data is the same, but the characteristics of the data are totally different. Age1 indicates that the main customer group is young people, age2 indicates that the main customer group is middle-aged and old people.

Age1:

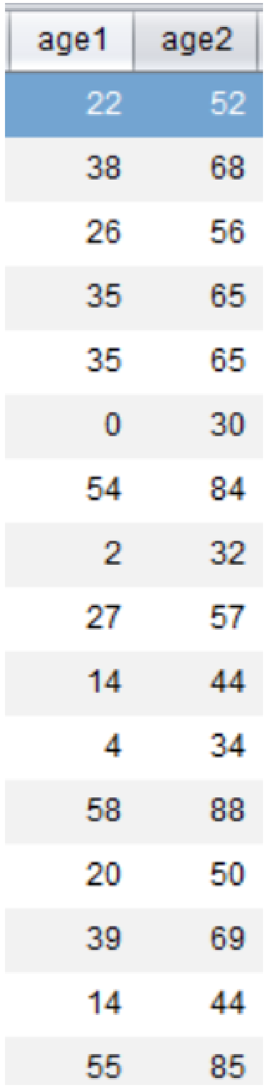


Average	Upper quartile	Median	Lower quartile	Standard dev...	Skewness
27	35	28	14	16.042	0.144

Age2:



Average	Upper quartile	Median	Lower quartile	Standard dev...	Skewness
57	65	58	44	16.042	0.144



age1	age2
22	52
38	68
26	56
35	65
35	65
0	30
54	84
2	32
27	57
14	44
4	34
58	88
20	50
39	69
14	44
55	85

Class exercise – Variance, Standard deviation

Calculate the variance and standard deviation of the following samples： 2， 3， 3， 3， 4

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Average $\overline{x}$ =(2+3+3+3+4)/5=3

Variance $s^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1} = \frac{(2-3)^2+(3-3)^2+(3-3)^2+(3-3)^2+(4-3)^2}{5-1}$=0.5

Standard deviation $s = \sqrt{0.5} = 0.71$

Class exercise – Variance, Standard deviation

Variance and standard deviation reflect the dispersion degree of each value from the average value.

The larger the value is, the greater the dispersion degree is .

Example

The math scores of the two groups are as follows:

Group 1：50，100，100，60，50

Group 2：73， 70，75，72，70

Use tools to calculate the mean and standard deviation of two sets of data （Math_score1 and Math_score2 variants in Exercise_1.csv ）

How are the two sets of data different?

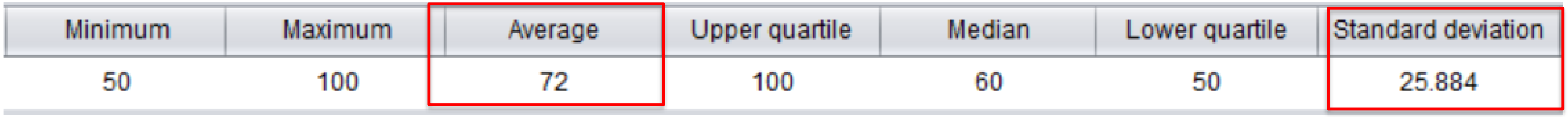
Class exercise – Variance, Standard deviation

Example

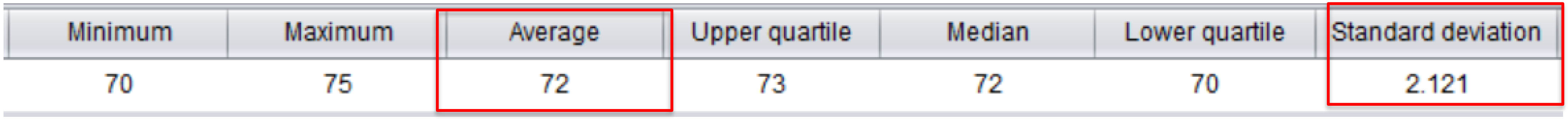
The math scores of the two groups are as follows:

Group 1：50，100，100，60，50

Group 2：73， 70，75，72，70



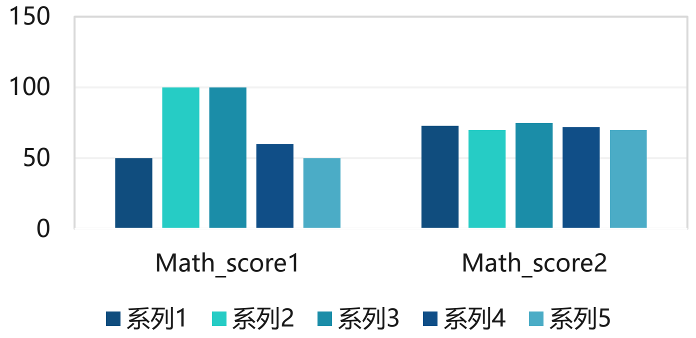
Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard deviation
50	100	72	100	60	50	25.884



Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard deviation
70	75	72	73	72	70	2.121

The average of the two groups is the same, but the standard deviation is different, and the dispersion is different.

The second group had more stable score.



Exploration method of quantitative data

Measures of relative standing

Quartile

Arrange all values from small to large and divide them into four equal parts,
The value at 25% is called the lower quartile,
The value at 75% is called the upper quartile.

The quartiles are usually represented by a box line chart.
Box line chart and Z-score are important indexes for outlier judgment, which will be explained later.

Z-score

It describes the distance between a given measurement value x and the average value, which is expressed in standard deviation.

z-score formula：

$$Z = \frac{X - \bar{X}}{s}$$

Class exercise – Quartile

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Player income	**4023**	**3274**	**227**	**129**	**154**	**269**	**463**	**196**	**190**	**232**	**90**	**84**	**65**	**90**	**36**	**35**

The income variant in Exercise_1.csv

Use the software to view the upper and lower quartiles of the data.

Calculate IQR (quartile difference)

$$\text{IQR} = Q_U - Q_L$$

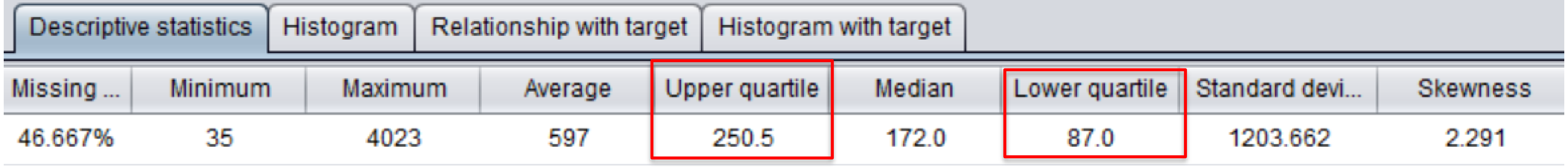
Class exercise – Quartile

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Player income	**4023**	**3274**	**227**	**129**	**154**	**269**	**463**	**196**	**190**	**232**	**90**	**84**	**65**	**90**	**36**	**35**

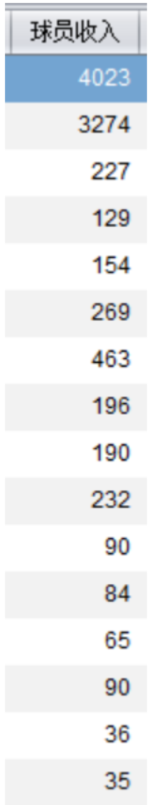
Upper quartile Q_U=250.5

Lower quartile Q_L=87

IQR=163.5



Missing ...	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard devi...	Skewness
46.667%	35	4023	597	250.5	172.0	87.0	1203.662	2.291



球员收入
4023
3274
227
129
154
269
463
196
190
232
90
84
65
90
36
35

Class exercise - Z-score

2000 students who took GMAT were selected as a random sample. In this sample, the average GMAT score is $\bar{x}$ =540, and the standard deviation is s = 100. Smith is one of the students in the sample, whose GMAT score is x = 440. What is Smith's Z-score?

$$Z=\frac{x-\bar{x}}{s}=\frac{440-540}{100}=-1.0$$

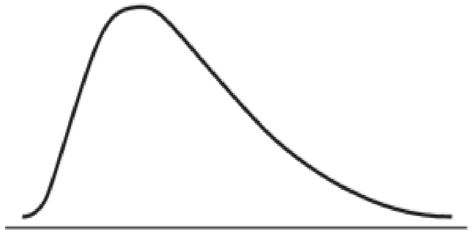
This score means that Smith's GMAT score is one standard deviation lower than the sample's average.

The value of Z-score reflects the relative position of the measured value. A large positive Z-value implies that the measurement is larger than almost all other measurements, while a large (in number) negative Z-value implies that the measurement is smaller than almost all other measurements. If a Z value is 0 or close to 0, the measurement will be located at or next to the mean of the sample.

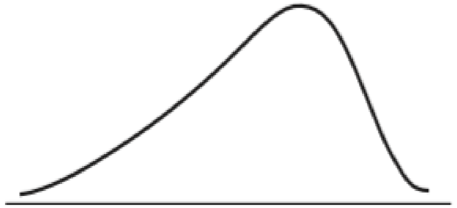
Exploration method of quantitative data

Skewness

If the value is 0, it means a symmetrical distribution; if the value is positive, it means the peak value of the distribution is to the left; if the value is negative, it means the peak value of the distribution is to the right.



Skewness是正值

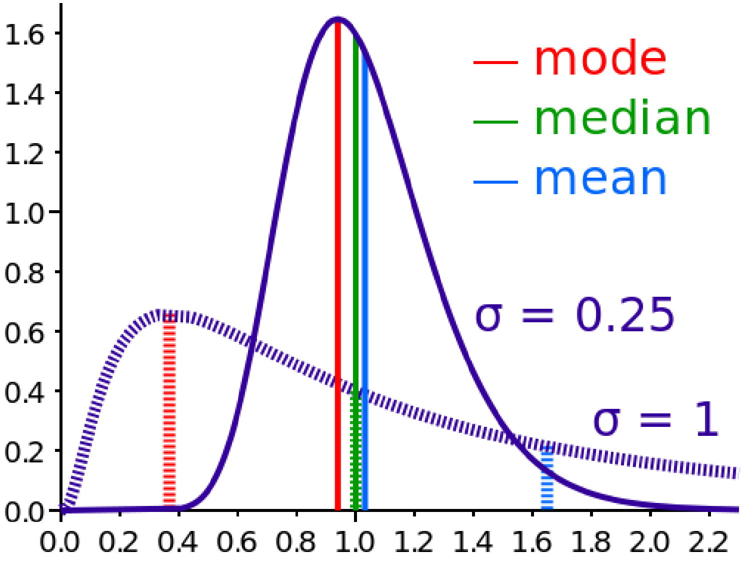


Skewness是负值

Measures of symmetry

Under different skewness, the values of mean, median and mode are very different. The skewness of distribution can be judged by the difference between median and mean.

- Median < mean: right deviation of data
- There is little difference between median and mean: symmetrical distribution
- Median > mean: left deviation of data



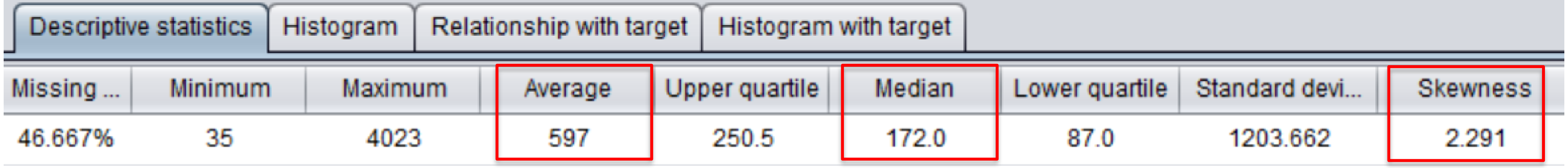
Class exercise - Skewness

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Player income	**4023**	**3274**	**227**	**129**	**154**	**269**	**463**	**196**	**190**	**232**	**90**	**84**	**65**	**90**	**36**	**35**

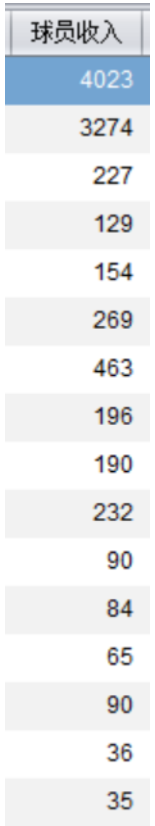
Use the tool to view the skewness of the data.（the player income variant in Exercise_1.csv)

Skewness=2.291， Median < mean: right deviation of data

Skewness greater than 1 is a very obvious signal, your data distribution has obvious asymmetry. It can only be used after correction, otherwise the value of the index will be greatly reduced. This is because many algorithms assume that the data is normally distributed (skewness is 0)



Missing ...	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard devi...	Skewness
46.667%	35	4023	597	250.5	172.0	87.0	1203.662	2.291



球员收入
4023
3274
227
129
154
269
463
196
190
232
90
84
65
90
36
35

Quantitative data

1

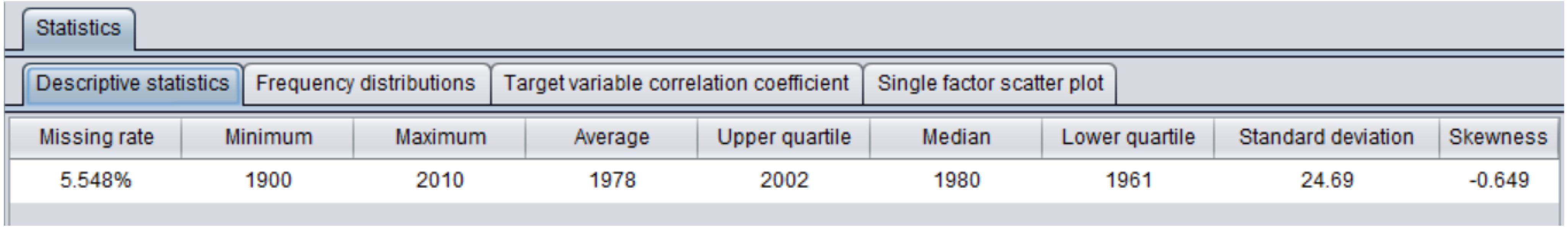
In the exploration of continuous data, the first index to be focused on is the missing rate.

2

Then there are the mean, median and other indexes, which can help data analysts to have a good understanding of the characteristics of the data.

3

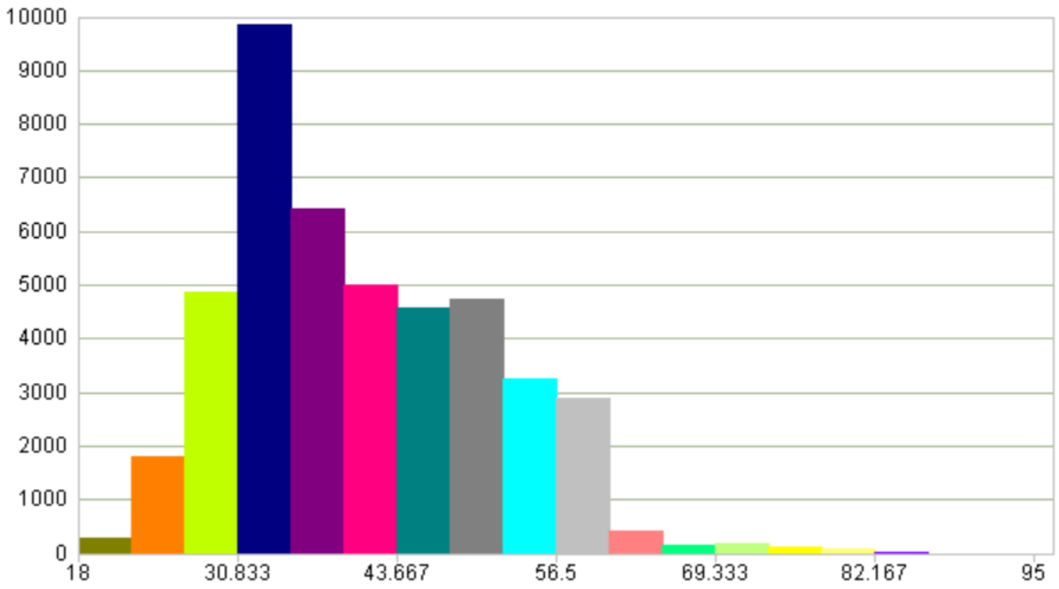
Skewness is another very important index, but when its absolute value is close to 1 or greater than 1, the data can only be used after mathematical conversion.

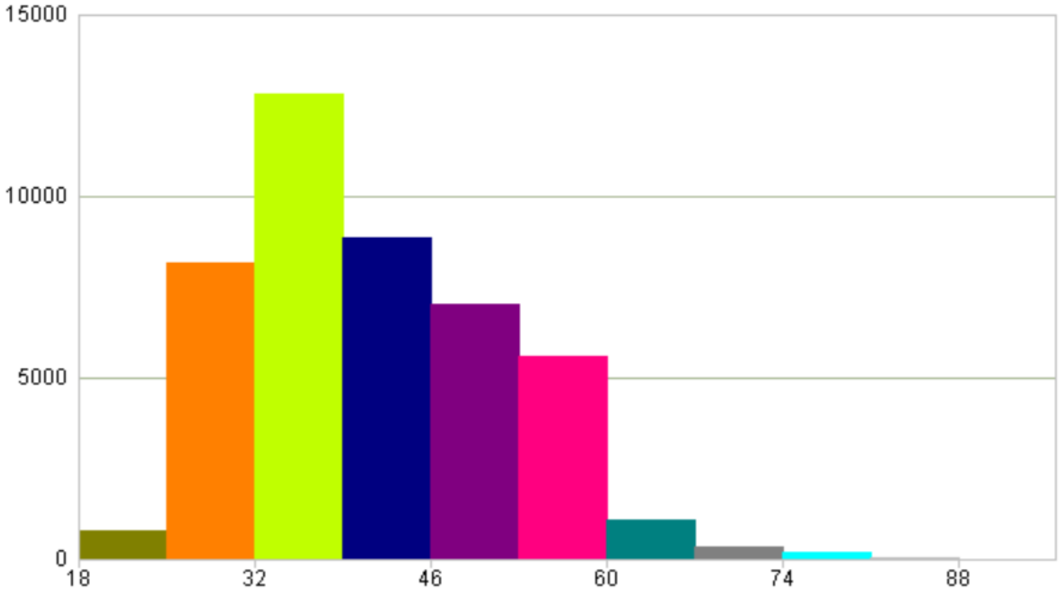

Statistics
Descriptive statistics
Frequency distributions
Target variable correlation coefficient
Single factor scatter plot

Missing rate	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard deviation	Skewness
5.548%	1900	2010	1978	2002	1980	1961	24.69	-0.649

Graphical method of quantitative data

Histogram: it can be used to represent the frequency or the number of related frequencies falling into each category interval.





Graphical method of quantitative data

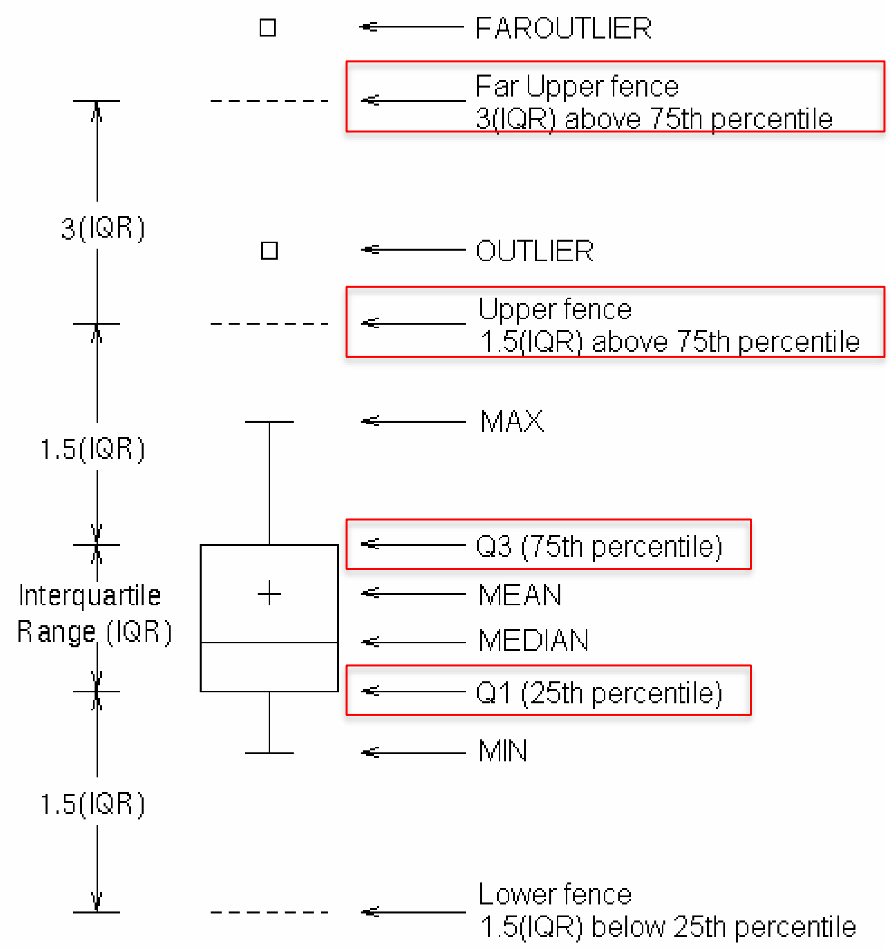
Box plots

Drawing method: draw a rectangle with two ends at the top and bottom quarter respectively. The median of the data is usually displayed in the box by a line in the box.

Draw two lines at Q_3+1.5IQR and Q_1 - 1.5IQR , which are the same as the median line, called the inner fences.

Draw two lines at Q_3+3IQR and Q_1 - 3IQR, which are called outer fences.

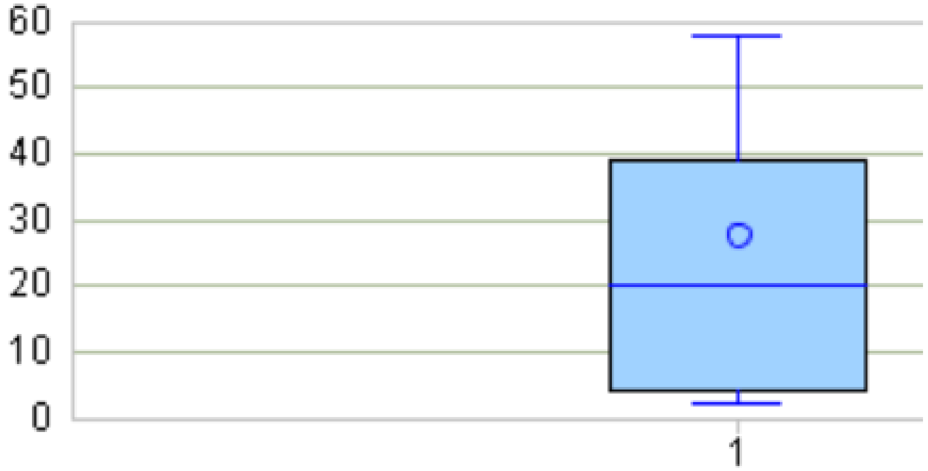
Interquartile range： IQR=Q_3-Q_1



Class exercise - Box plots

Consider the box line on the right （Exercise_1.csv， data with target = 1 in the age1 variable ）：

1. What is the median of the dataset (approximate)?
2. What are the upper and lower quartiles of the dataset (approximate)?
3. What is the interquartile range of the dataset (approximate)?
4. Is the average of this dataset the same as the median?



Class exercise - Box plots

1. What is the median of the dataset (approximate)?

   Median: 20

2. What are the upper and lower quartiles of the dataset (approximate)?

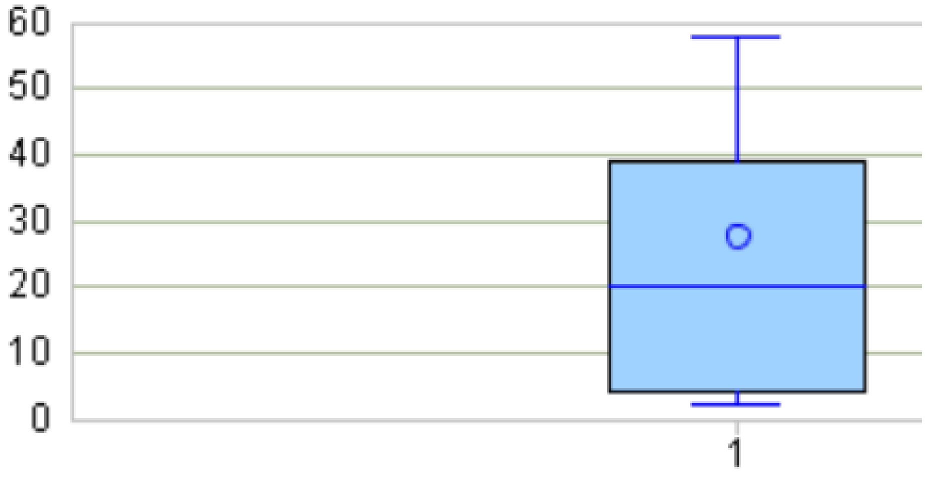
   Upper quartile:39 Lower quartile:4

3. What is the interquartile range of the dataset (approximate) ?

   IQR=35

4. Is the average of this dataset the same as the median?

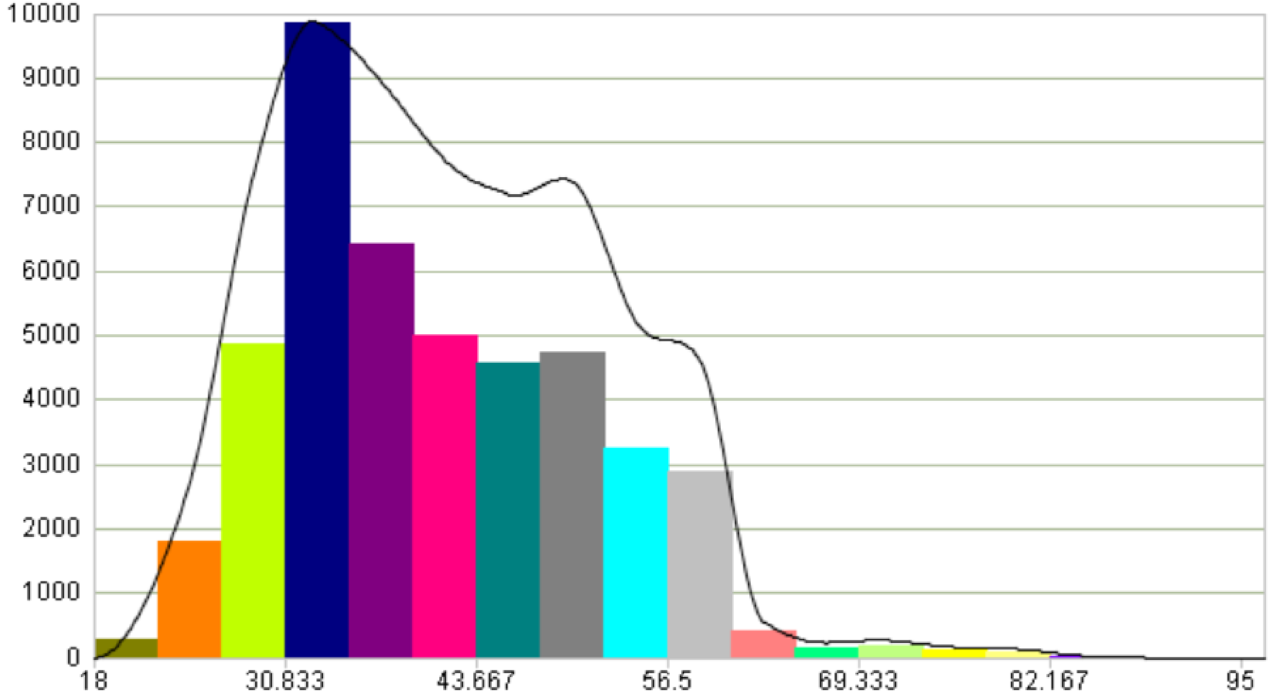
   Average=28， greater than median



Graphical method of quantitative data

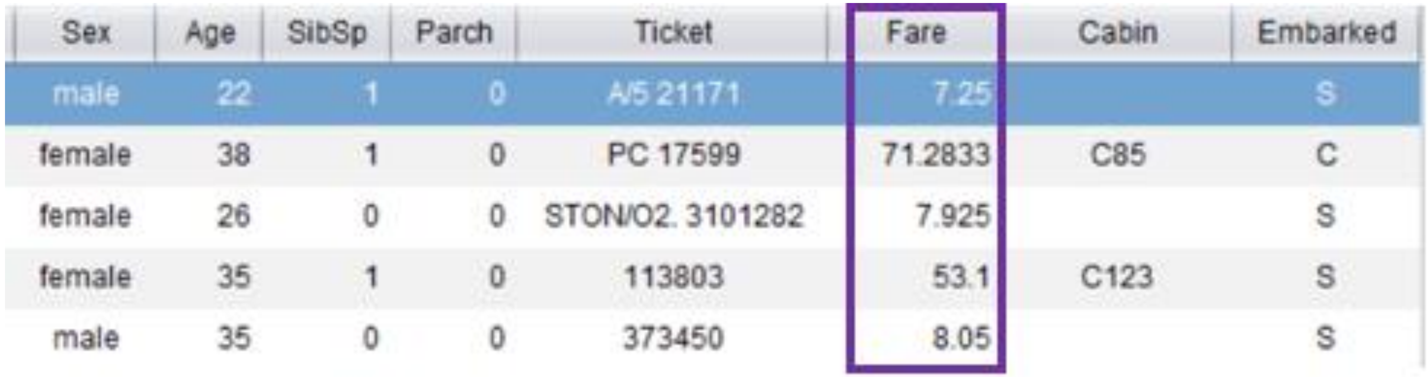
Density curve

One disadvantage of using histograms is that histograms depend on the number of intervals selected. One way to solve this problem is to smooth the data and draw the density estimation curve. A common smoothing method is kernel estimation.



The smoothing algorithm is complex.
Here is a concept .

Class exercise - Review



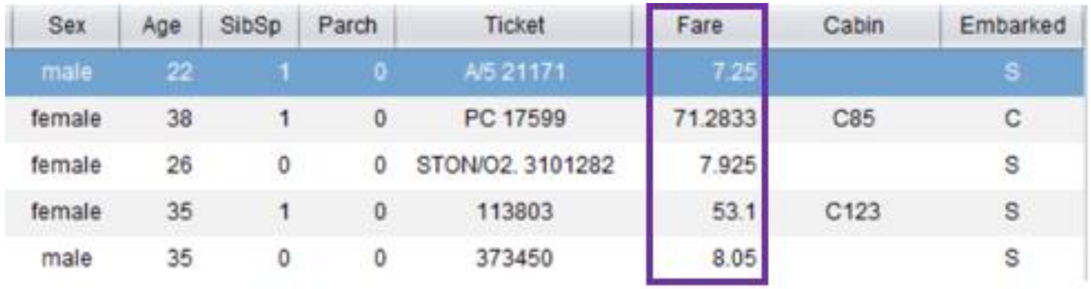
Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
male	22	1	0	A/5 21171	7.25		S
female	38	1	0	PC 17599	71.2833	C85	C
female	26	0	0	STON/O2. 3101282	7.925		S
female	35	1	0	113803	53.1	C123	S
male	35	0	0	373450	8.05		S

There is a variable "fare" in the Titanic data, and data exploration is carried out for it.

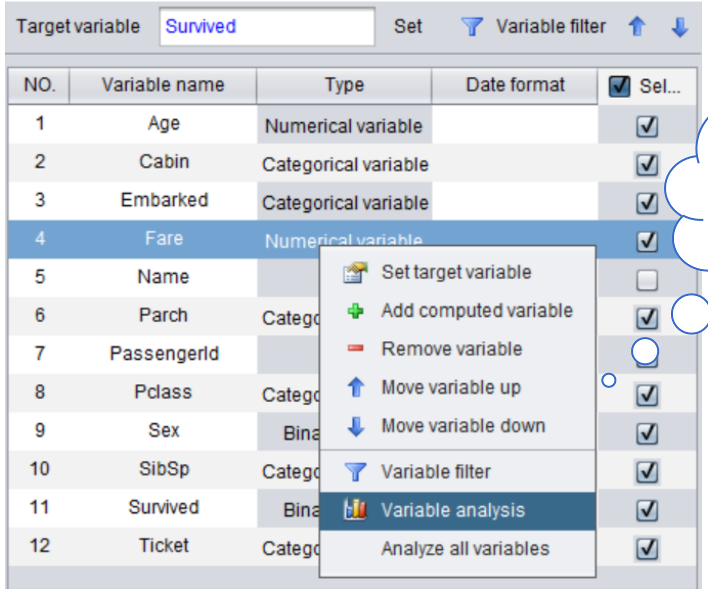
Calculate the maximum / minimum, average, median, skewness and other statistical indicators?

Analyze data distribution?

Class exercise - Review



Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
male	22	1	0	A/5 21171	7.25		S
female	38	1	0	PC 17599	71.2833	C85	C
female	26	0	0	STON/O2. 3101282	7.925		S
female	35	1	0	113803	53.1	C123	S
male	35	0	0	373450	8.05		S



Target variable: Survived

NO.	Variable name	Type	Date format	☑ Sel...
1	Age	Numerical variable		☑
2	Cabin	Categorical variable		☑
3	Embarked	Categorical variable		☑
4	Fare	Numerical variable		☑
5	Name			☐
6	Parch	Categ		☑
7	PassengerId			
8	Pclass	Categ		☑
9	Sex	Bina		☑
10	SibSp	Categ		☑
11	Survived	Bina		☑
12	Ticket	Categ		☑

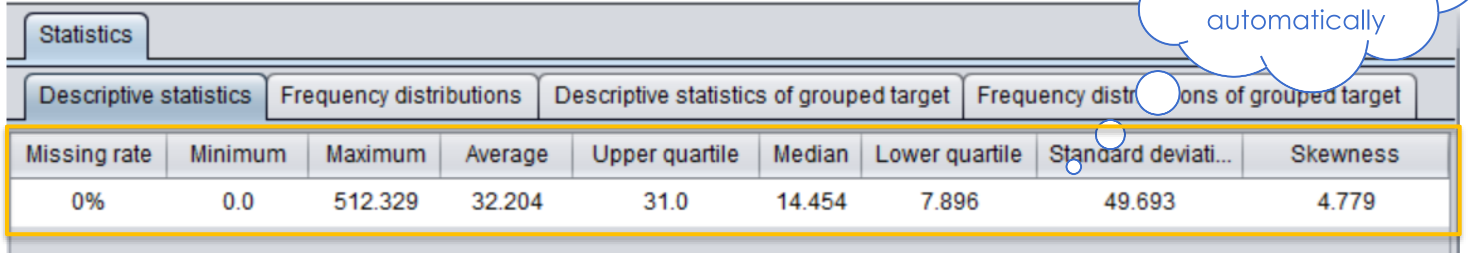
Explore variables automatically

There is a variable "fare" in the Titanic data, and data exploration is carried out for it.

The statistical values are shown in the figure below, and the graphical distribution is shown in the figure below on the right,

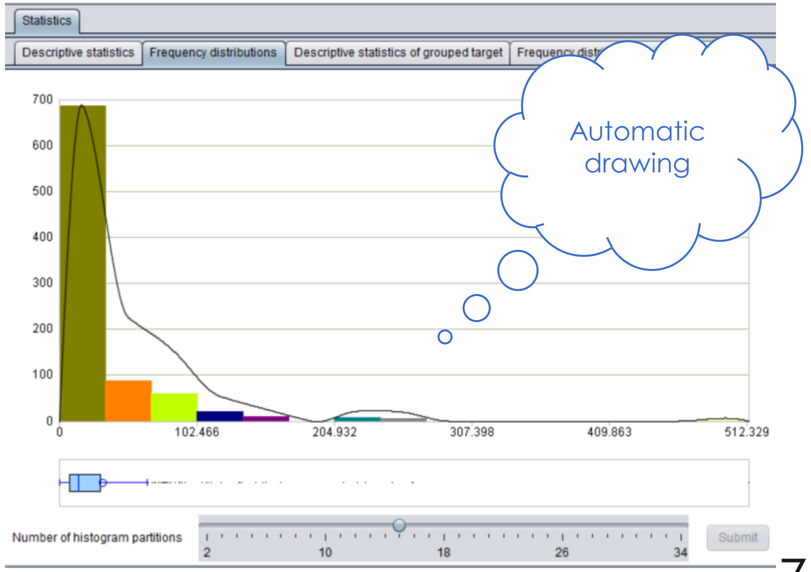
It can be seen that the data range is 0-512, the average value is greater than the median, and the data skewness is large, indicating that the rich are in the minority.

calculate various statistical values automatically



Statistics
Descriptive statistics
Frequency distributions
Descriptive statistics of grouped target
Frequency distributions of grouped target

Missing rate	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard deviati...	Skewness
0%	0.0	512.329	32.204	31.0	14.454	7.896	49.693	4.779



Automatic drawing

Glossary - for reference

measures of central tendency
measures of relative standing
measures of variability
measures of symmetry
mean
median
mode
modal class
range
variance
standard deviation
quartile
upper quartile
lower quartile
interquartile range, IQR
z-score
skewness
histogram
box plots
density curve
inner fences
outer fences
outliers

Exploration method of qualitative data

Class

Is that qualitative data can be divided into several categories

Class frequency

Refers to the number of samples belonging to a certain category

Class percentage

(Class frequency/n) *100%

n: the total number of samples (also known as base)

(Class frequency/n): class relative frequency

Mode

The most frequent data in the dataset.

For example, for a certain subtype, the value A, B, C and D have the most occurrence of C, then C is the mode.

Unary variable, binary variable, categorical variable, Text string variable

Class exercise

On the right is the education background of

Forbes ' top 30 CEOs with the highest incomes

(the CEO education variant in Exercise_1.csv)

Please calculate :

How many categories and What are they?

What is the sample size for each category?

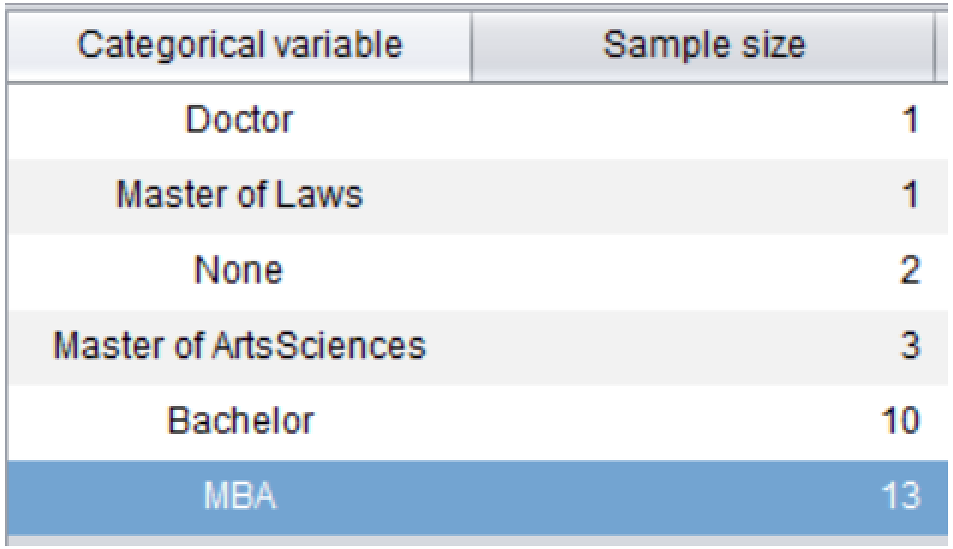
What is the percentage of each?

Which is the mode of the data?

ID	CEO education	ID	CEO education
1	Bachelor	16	Master of Arts and Sciences
2	MBA	17	Bachelor
3	Bachelor	18	No university degree
4	Bachelor	19	Bachelor
5	MBA	20	Bachelor
6	No university degree	21	MBA
7	Doctor	22	Bachelor
8	MBA	23	Bachelor
9	MBA	24	MBA
10	MBA	25	MBA
11	Master of Arts and Sciences	26	MBA
12	MBA	27	Master of law
13	MBA	28	Bachelor
14	Master of Arts and Sciences	29	MBA
15	MBA	30	Bachelor

Class exercise - Qualitative data indicators

Education background of the 30 CEOs:
Bachelor, MBA, master of Arts and Sciences, master of law, doctor and no university degree, 6 categories in total, and the sample size of each category as follows:



Categorical variable	Sample size
Doctor	1
Master of Laws	1
None	2
Master of ArtsSciences	3
Bachelor	10
MBA	13

ID	CEO education	ID	CEO education
1	Bachelor	16	Master of Arts and Sciences
2	MBA	17	Bachelor
3	Bachelor	18	No university degree
4	Bachelor	19	Bachelor
5	MBA	20	Bachelor
6	No university degree	21	MBA
7	Doctor	22	Bachelor
8	MBA	23	Bachelor
9	MBA	24	MBA
10	MBA	25	MBA
11	Master of Arts and Sciences	26	MBA
12	MBA	27	Master of law
13	MBA	28	Bachelor
14	Master of Arts and Sciences	29	MBA
15	MBA	30	Bachelor

Class exercise - Qualitative data indicators

The percentages for each category are:

Doctor: 1 / 30 = 0.033

LLM: 1 / 30 = 0.033

No university degree: 2 / 30 = 0.067

Master of Arts: 3 / 30 = 0.1

Bachelor: 10 / 30 = 0.333

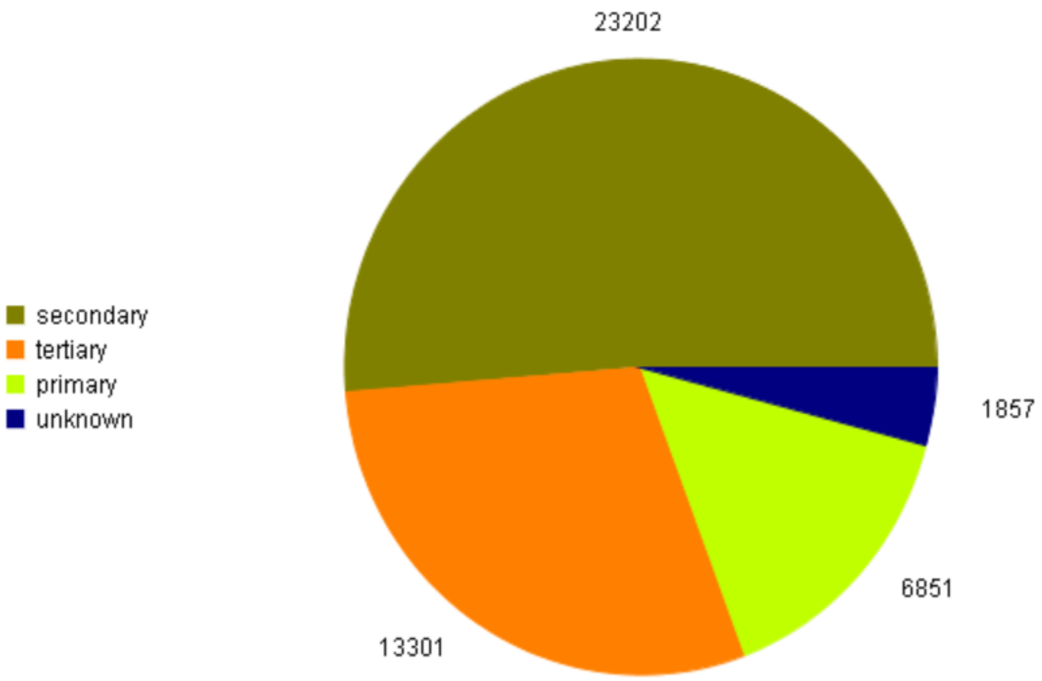
MBA=13/30=0.4333

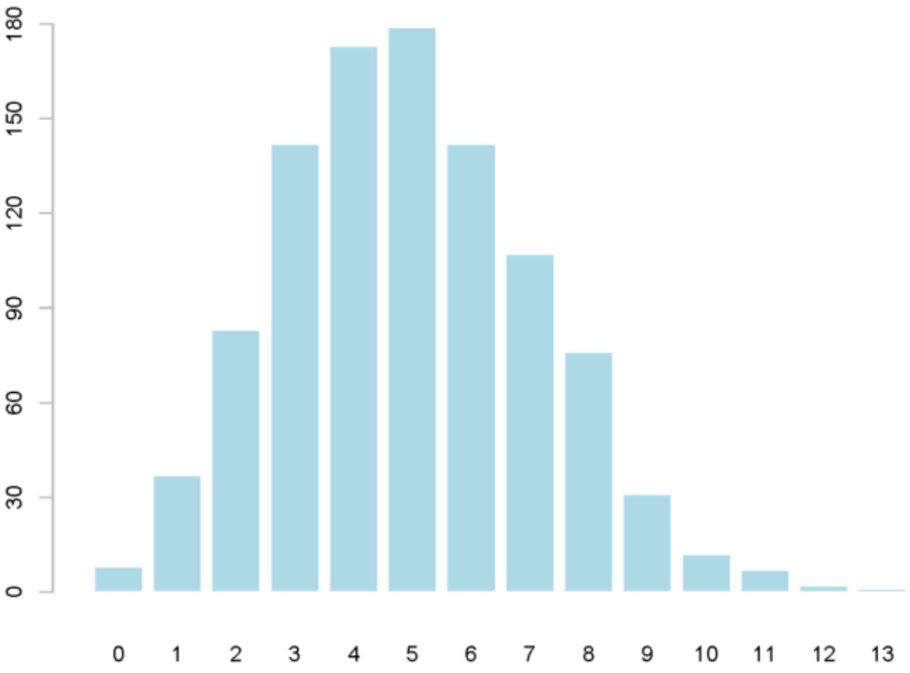
The mode of this variable is the most frequent category: MBA

ID	CEO education	ID	CEO education
1	Bachelor	16	Master of Arts and Sciences
2	MBA	17	Bachelor
3	Bachelor	18	No university degree
4	Bachelor	19	Bachelor
5	MBA	20	Bachelor
6	No university degree	21	MBA
7	Doctor	22	Bachelor
8	MBA	23	Bachelor
9	MBA	24	MBA
10	MBA	25	MBA
11	Master of Arts and Sciences	26	MBA
12	MBA	27	Master of law
13	MBA	28	Bachelor
14	Master of Arts and Sciences	29	MBA
15	MBA	30	Bachelor

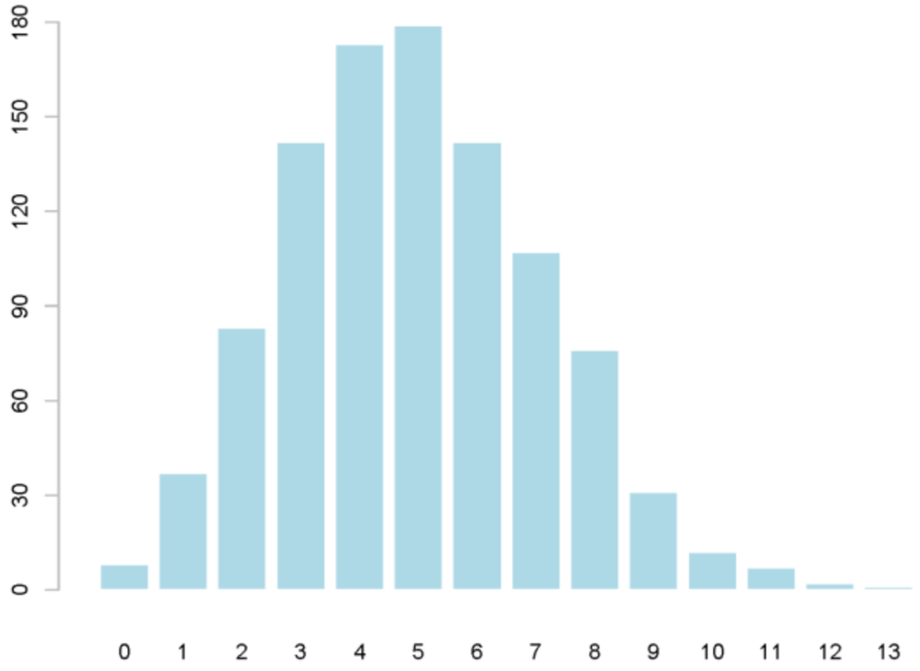
Graphical method of qualitative data

Graphic method of qualitative data: pie chart and bar graph

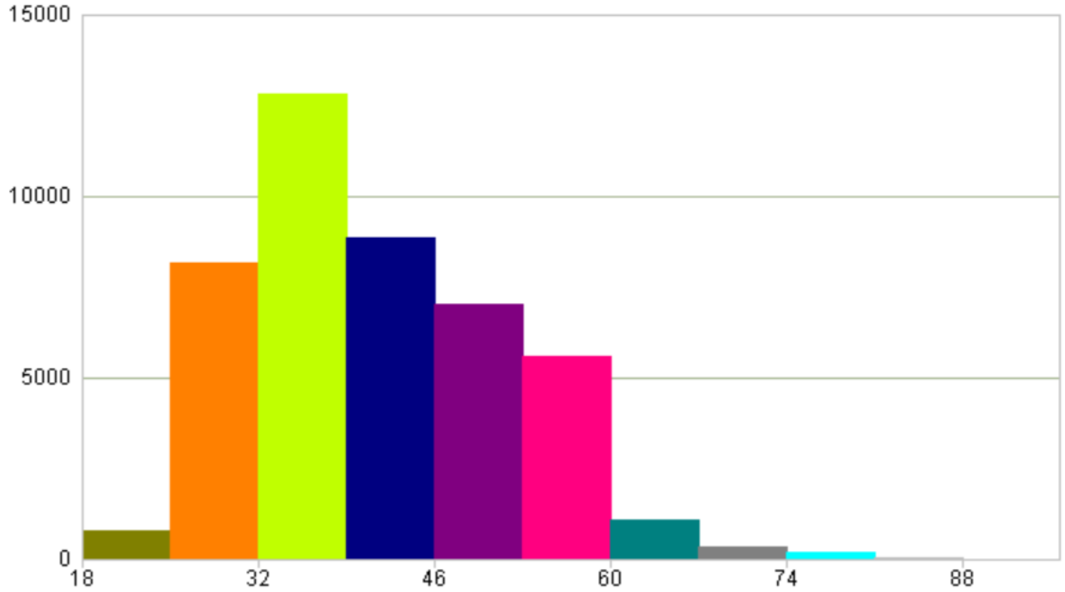




Bar graph and histogram



Bar graph

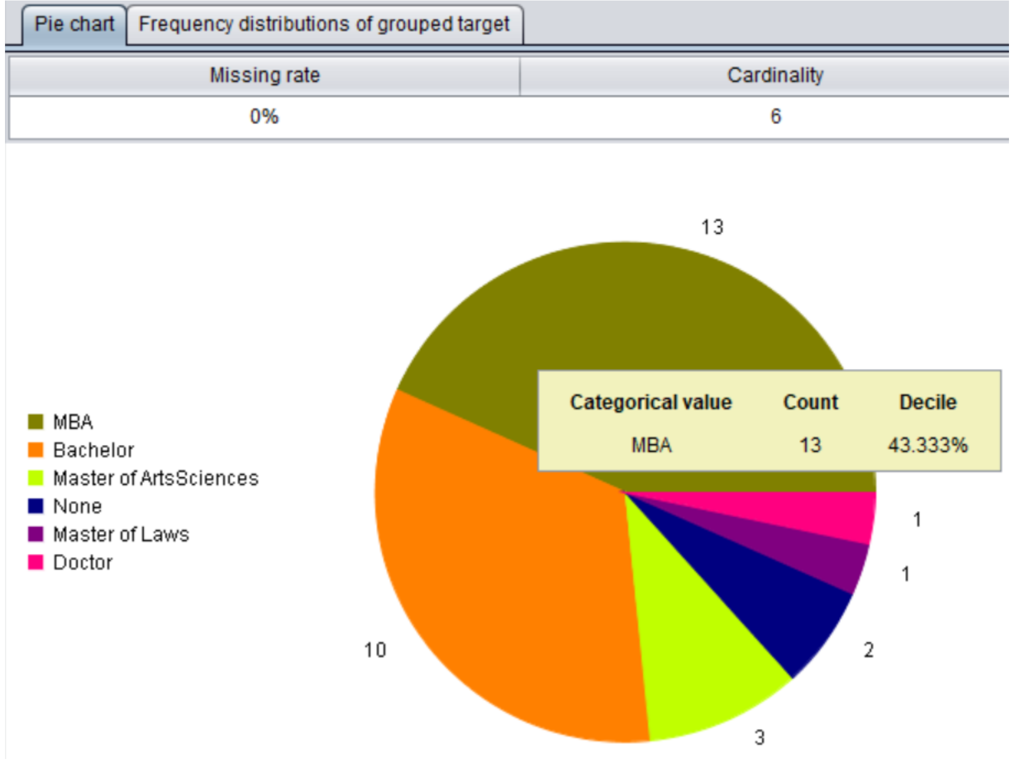


Histogram

Class exercise - Pie chart

On the right is the education background of Forbes ' top 30 CEOs with the highest incomes:

Please use pie chart to represent the statistical index of the variable.



Missing rate	Cardinality
0%	6

Categorical value	Count	Decile
MBA	13	43.333%

ID	CEO education	ID	CEO education
1	Bachelor	16	Master of Arts and Sciences
2	MBA	17	Bachelor
3	Bachelor	18	No university degree
4	Bachelor	19	Bachelor
5	MBA	20	Bachelor
6	No university degree	21	MBA
7	Doctor	22	Bachelor
8	MBA	23	Bachelor
9	MBA	24	MBA
10	MBA	25	MBA
11	Master of Arts and Sciences	26	MBA
12	MBA	27	Master of law
13	MBA	28	Bachelor
14	Master of Arts and Sciences	29	MBA
15	MBA	30	Bachelor

Glossary - for reference

class

class frequency

class relative frequency

class percentage

sample size

pie chart

bar graph

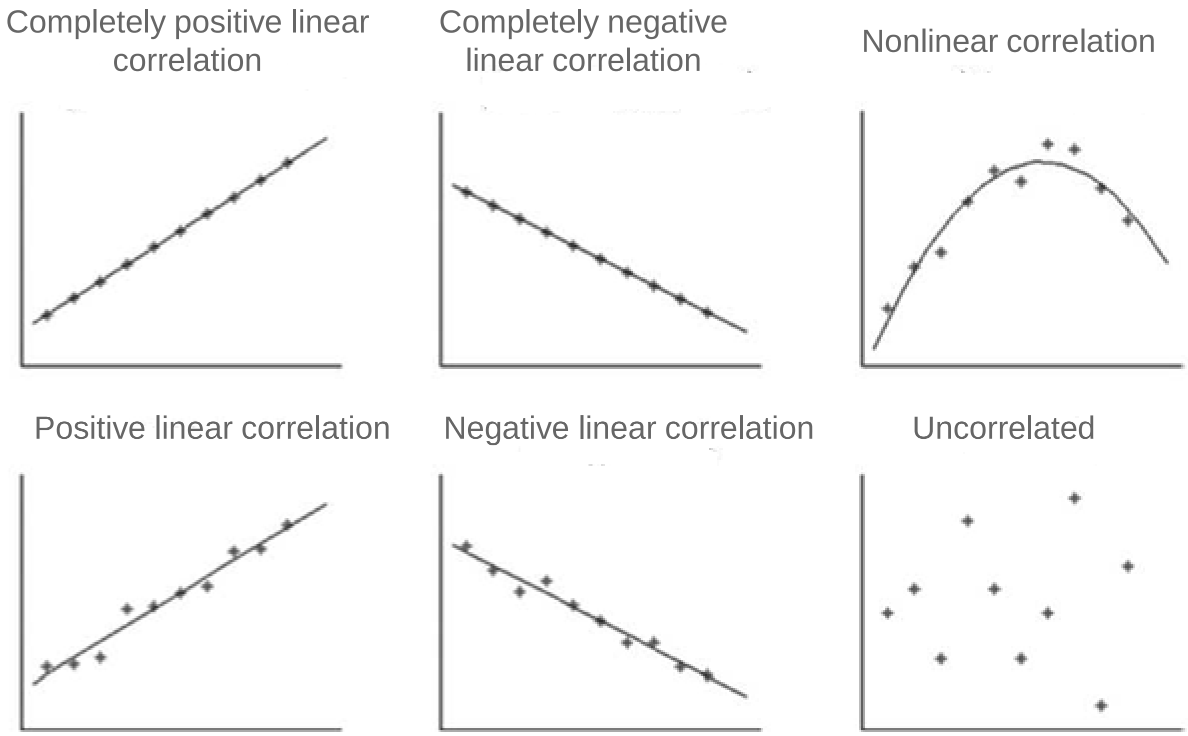
Variable correlation - continuous variable and continuous variable

Pearson correlation coefficient

Spearman correlation coefficient

Kendall correlation coefficient

Hoeffding correlation coefficient



Completely positive linear correlation

Completely negative linear correlation

Nonlinear correlation

Positive linear correlation

Negative linear correlation

Uncorrelated

Pearson and Spearman coefficients are commonly used.

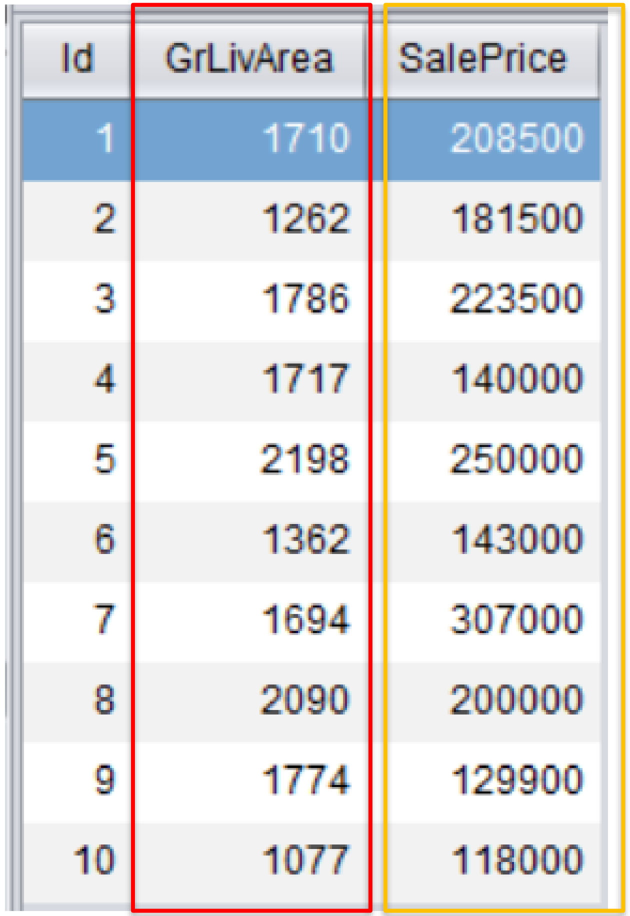
Variable correlation - continuous variable and continuous variable

	Definition	Value	Scope of application
Pearson	Also known as the product moment correlation coefficient, it is a method to calculate the linear correlation between two variables.	[-1,1] The closer the absolute value is to 1, the stronger the correlation is, the closer the absolute value is to 0, and the weaker the correlation is.	Evaluate linear correlation Continuous variables, paired data, general normal distribution
Spearman	Also known as rank correlation coefficient, it is a nonparametric correlation measure based on the order of data values.	[-1,1] The closer the absolute value is to 1, the stronger the correlation is, the closer the absolute value is to 0, and the weaker the correlation is.	Evaluate monotonic relationships (linear or not) Applicable to both continuous and categorical variables There is no requirement for the overall distribution of variables and sample size.
Kendall	Commonly known as Kendall's tau coefficient, statistics used to measure the ordinal correlation between two variables.	[-1,1] The closer the absolute value is to 1, the stronger the correlation is, the closer the absolute value is to 0, and the weaker the correlation is.	It is applicable to the case that both categorical variables are ordered. Nonparametric correlation test for correlated ordered variables.

For more principles and calculation methods, please refer to relevant statistics books.

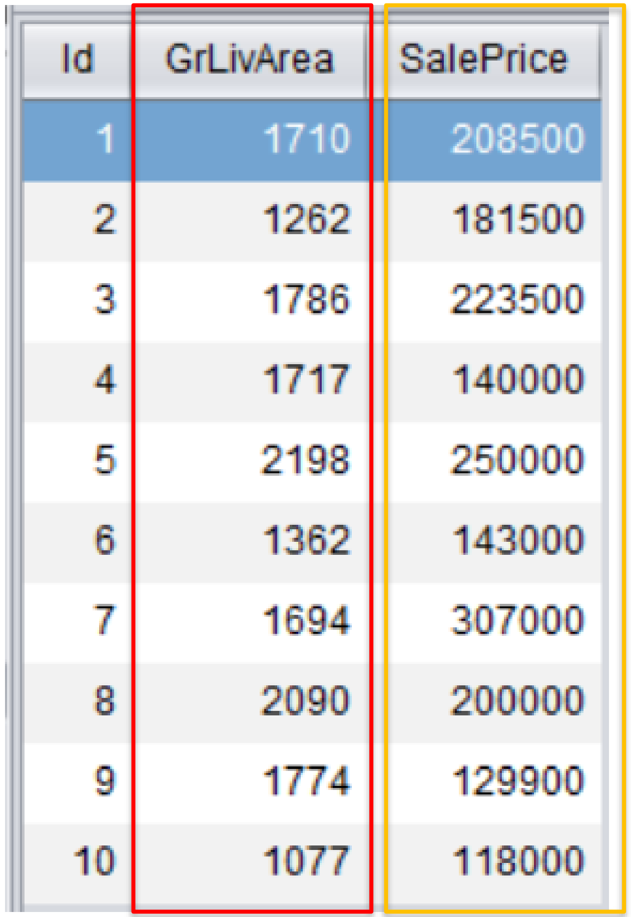
Class exercise - correlation analysis

For example, in the case of housing price prediction on kaggle, practice using tools to analyze Pearson and Spearman correlation coefficients of GrLivArea” and " SalePrice ".



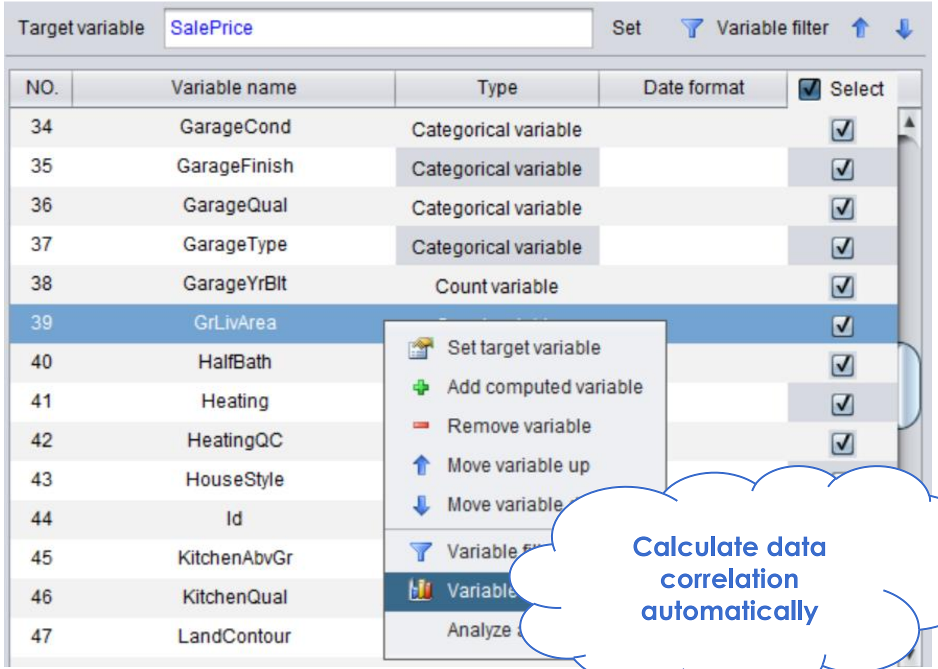
Id	GrLivArea	SalePrice
1	1710	208500
2	1262	181500
3	1786	223500
4	1717	140000
5	2198	250000
6	1362	143000
7	1694	307000
8	2090	200000
9	1774	129900
10	1077	118000

Class exercise - correlation analysis



Id	GrLivArea	SalePrice
1	1710	208500
2	1262	181500
3	1786	223500
4	1717	140000
5	2198	250000
6	1362	143000
7	1694	307000
8	2090	200000
9	1774	129900
10	1077	118000

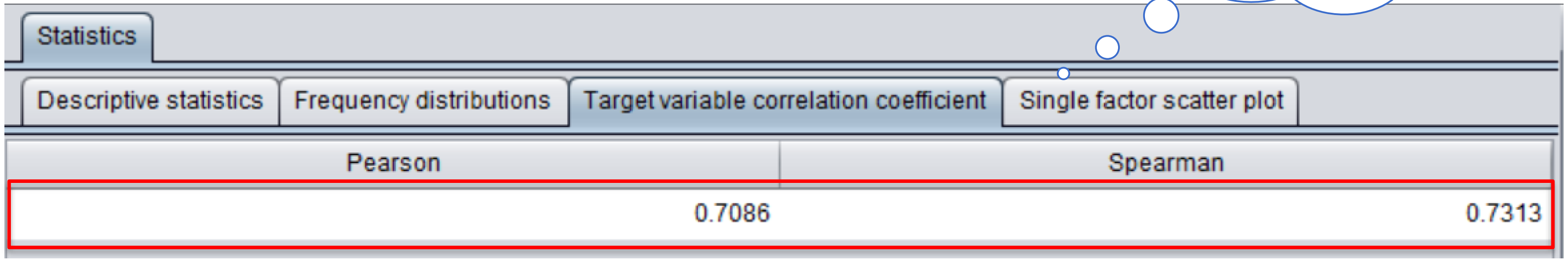
For example, in the case of house price prediction on kaggle, use YModel to view the correlation between " GrLivArea " residential area and "SalePrice".



Target variable: SalePrice

NO.	Variable name	Type	Date format	☑ Select
34	GarageCond	Categorical variable		☑
35	GarageFinish	Categorical variable		☑
36	GarageQual	Categorical variable		☑
37	GarageType	Categorical variable		☑
38	GarageYrBlt	Count variable		☑
39	GrLivArea			☑
40	HalfBath			☑
41	Heating			☑
42	HeatingQC			☑
43	HouseStyle			
44	Id			
45	KitchenAbvGr			
46	KitchenQual			
47	LandContour			

Calculate data correlation automatically


Statistics
Descriptive statistics
Frequency distributions
Target variable correlation coefficient
Single factor scatter plot

Pearson	Spearman
0.7086	0.7313

Both the two correlation coefficients are greater than 0.7, which shows that the linear relationship between them is very strong.

Graph visualization of correlation analysis

Scatter plot: provides the following information about the relationship between two variables

- Correlation strength
- Straight line or curve
- Positive correlation or negative correlation
- Outliers

Graph visualization of correlation analysis

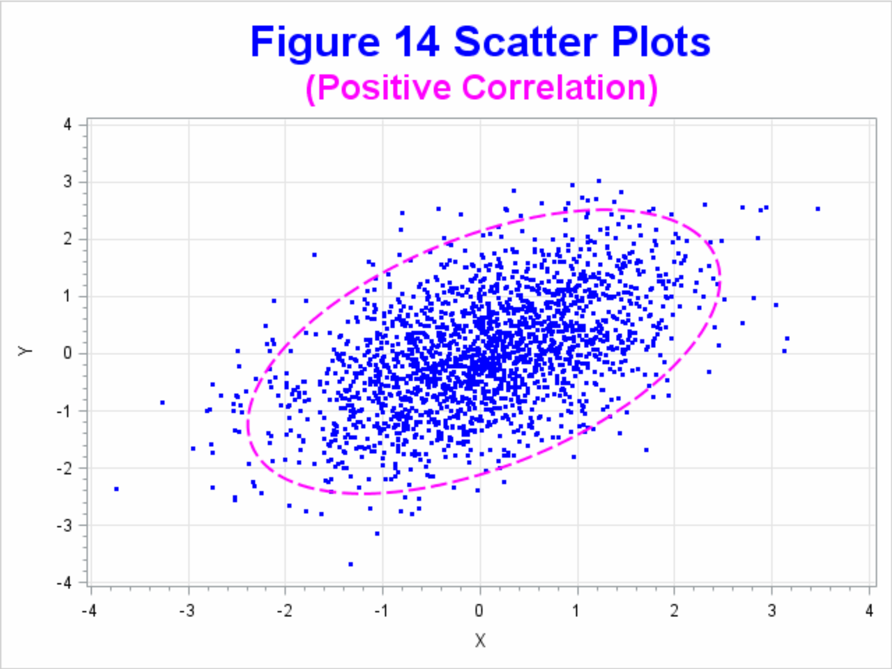


Figure 14 Scatter Plots

(Positive Correlation)

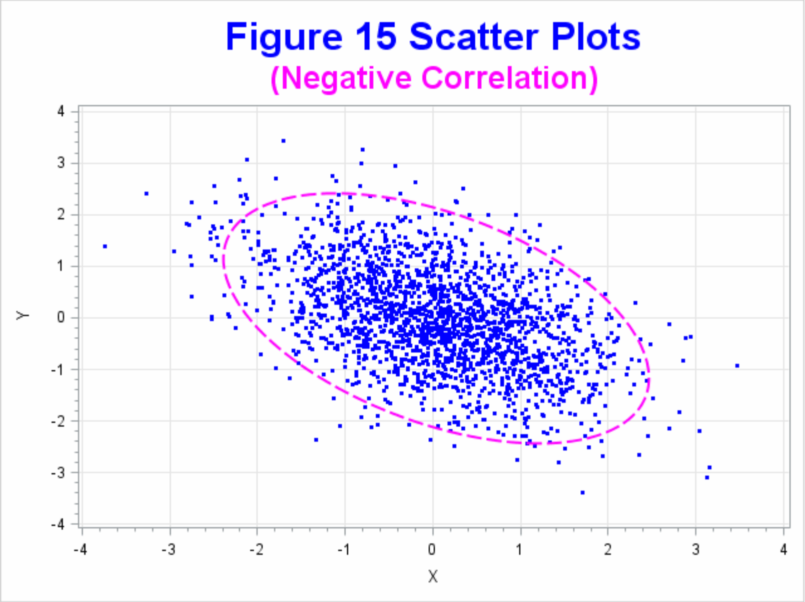


Figure 15 Scatter Plots

(Negative Correlation)

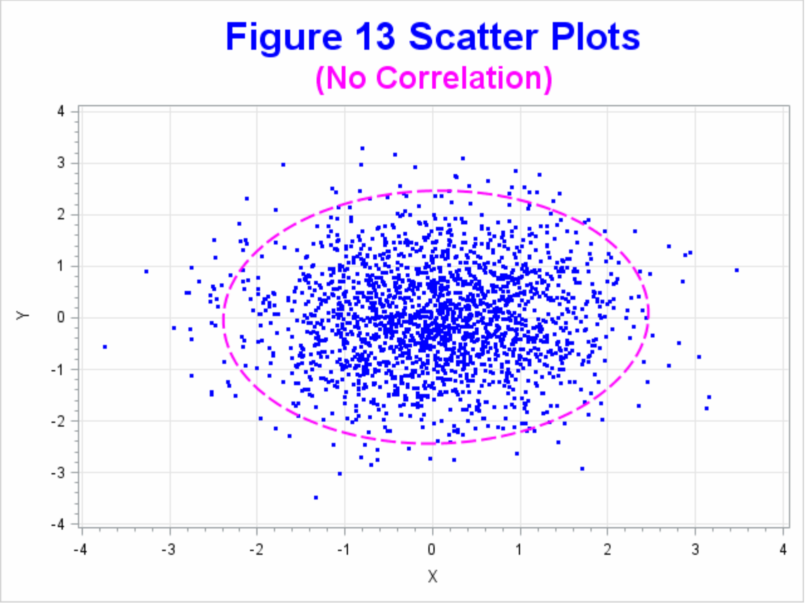
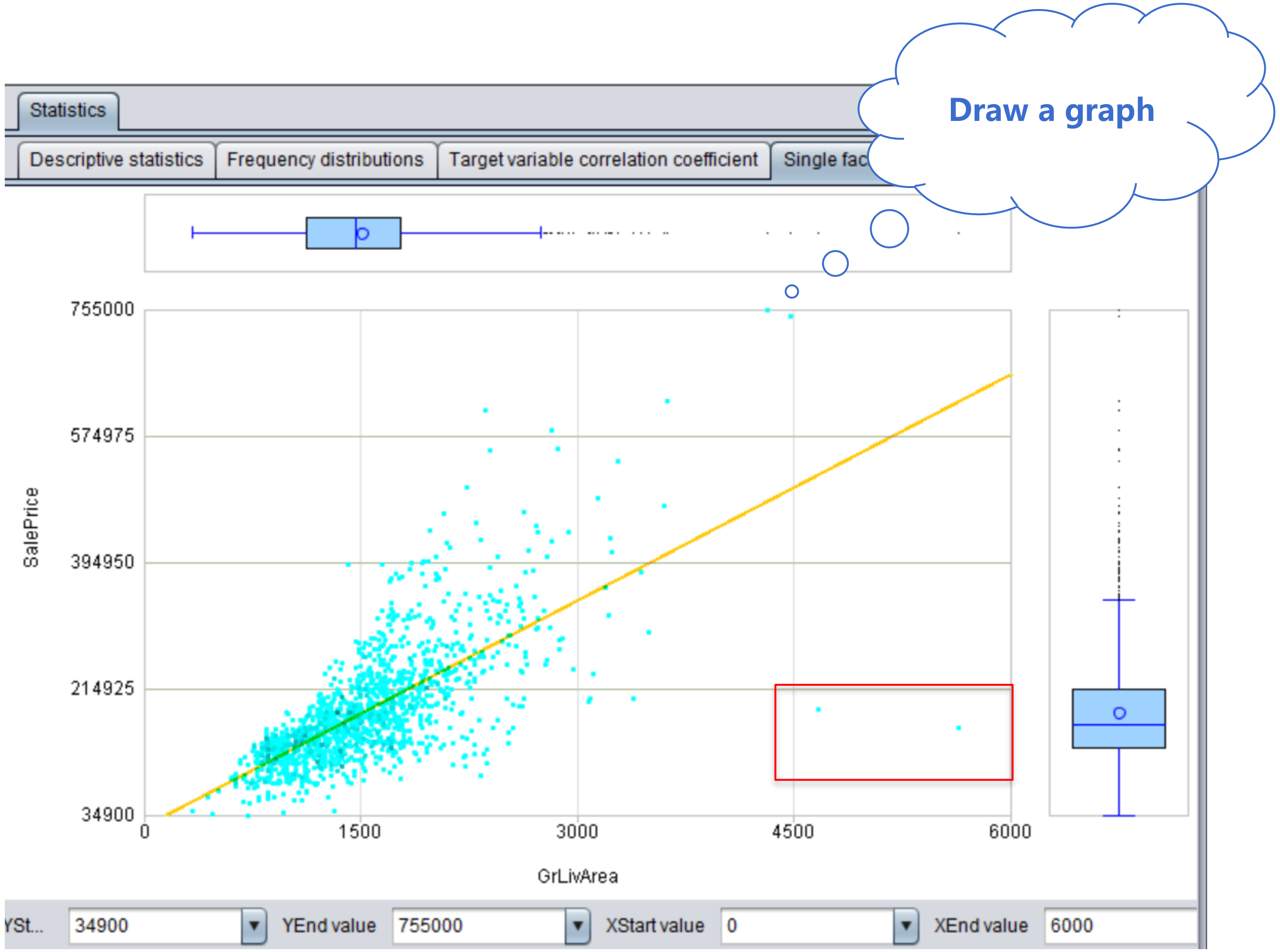


Figure 13 Scatter Plots

(No Correlation)

Class exercise - scatter plot



YSt... 34900 YEnd value 755000 XStart value 0 XEnd value 6000

Take Kaggle house price prediction as an example, use YModel to view the relationship between residential area and house price.

Observing the scatter plot, it shows the trend that the larger the living area is, the higher the house price is, which shows that the correlation between them is very strong.
However, the two points in the lower right corner are very special, with a large living area, but the house price is very low, which affects the overall linear relationship, so they can be deleted as exception values.

Variable correlation - categorical variable and categorical variable

Chi square test, also known as χ2 test, is a statistical hypothesis test to test whether the data distribution is chi square (when the null hypothesis is true).

Chi square test is used to determine whether there is a significant difference between the expected cell count and the observed cell count in one or more categories.

If they are close, χ 2 is relatively small(when two values are completely equal, χ 2 is 0). When χ 2 is relatively large, it means that the difference between them is relatively large.

Here, the expected cell count refers to the expected cell count of a certain type of result, and the observed cell count refers to the actual sample number of that type.

For example:

there are 200 people in total (male: 100, female: 100), including 110 people who make up (male: 15, female: 95)

Then: the expected cell count of male make-up is 100 * (110 / 200) = 55

The observation cell count of male make-up is: 15

The expected cell count and observed cell count are both statistical concepts. Here, they are only simply explained. Please refer to relevant statistical books for specific definitions.

Class exercise - chi square test

non make up

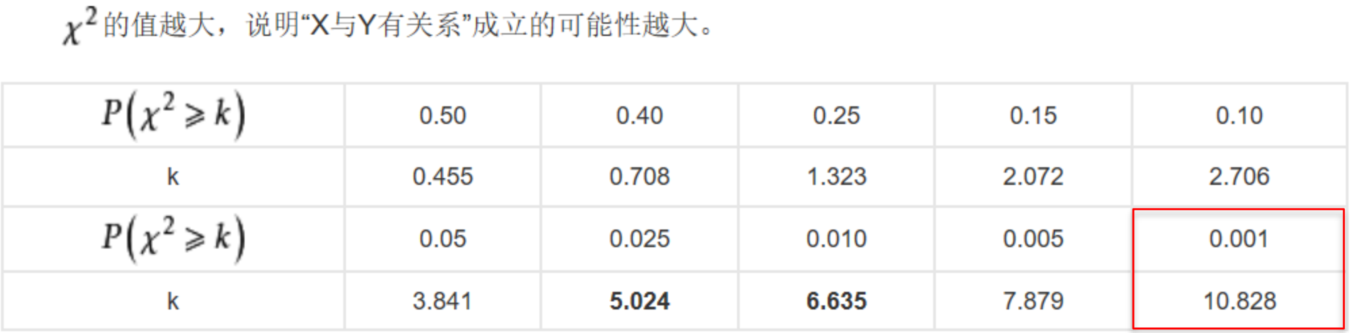
For example: Chi square test to analyze the relationship between gender and make-up.

	male	female	
make up	15（55）	95（55）	110
non make up	85（45）	5（45）	90
	100	100	200

Formula： $$\chi_c^2 = \frac{(95-55)^2}{55} + \frac{(15-55)^2}{55} + \frac{(85-45)^2}{45} + \frac{(5-45)^2}{45} = 129.3 > 10.828$$

The probability of correlation between gender and make-up $p > 0.999$

Significant correlation between the two



χ^2的值越大，说明“X与Y有关系”成立的可能性越大。

$P(\chi^2 \geqslant k)$	0.50	0.40	0.25	0.15	0.10
k	0.455	0.708	1.323	2.072	2.706
$P(\chi^2 \geqslant k)$	0.05	0.025	0.010	0.005	0.001
k	3.841	**5.024**	**6.635**	7.879	10.828

Variable correlation - categorical variable and continuous variable

T-test: also known as student's t test, it uses t-distribution theory and hypothesis test principle to compare the sample mean with the overall mean, as well as to compare the two sample mean.

Analysis of variance (ANOVA): also known as "variance analysis" or "F-test", is a mathematical statistical method used to test whether the mean of two or more groups of samples has significant difference.

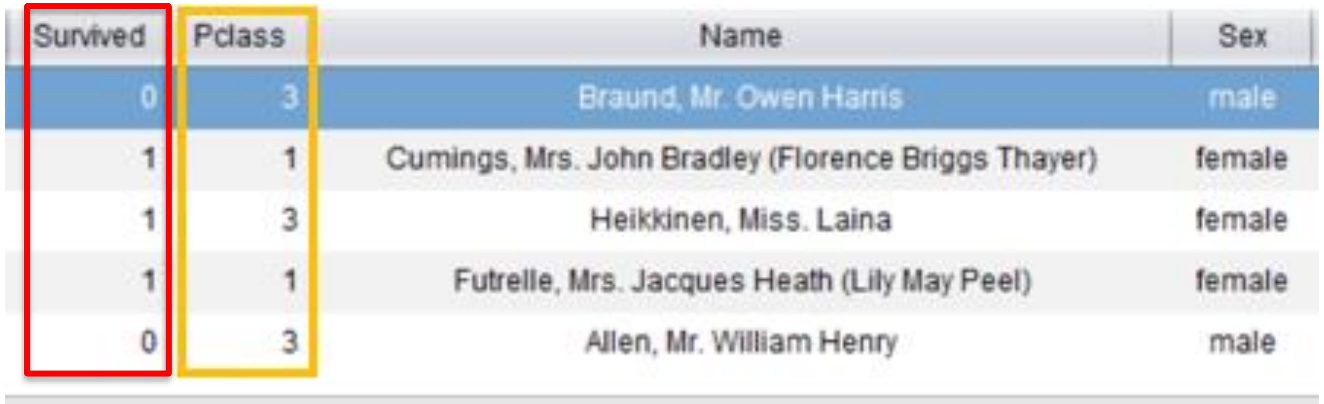
T-test can be used to compare the differences of two variables. ANOVA can be used to compare the differences of multiple variables.

Conditions of use:

samples follow normal distribution, and the total variance of two groups of samples is equal.

These contents are a little abstruse. Beginners don't need to care about the details. For specific principles, please refer to relevant statistical books.

Class exercise - Correlation analysis of categorical variables



Survived	Pclass	Name	Sex
0	3	Braund, Mr. Owen Harris	male
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female
1	3	Heikkinen, Miss. Laina	female
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
0	3	Allen, Mr. William Henry	male

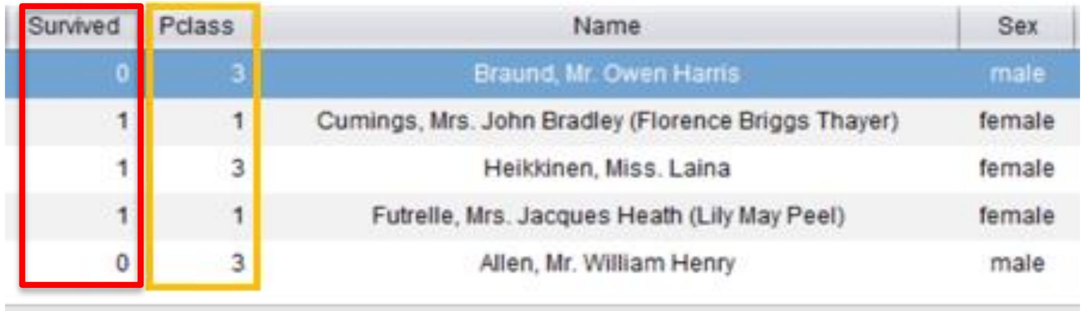
Explore the “Pclass" variable in Titanic data

What are the total categories?

What are the sample size and proportion of each category?

Is there a relationship between the different categories of the variable and the survival of the target variable?

Class exercise - Correlation analysis of categorical variables

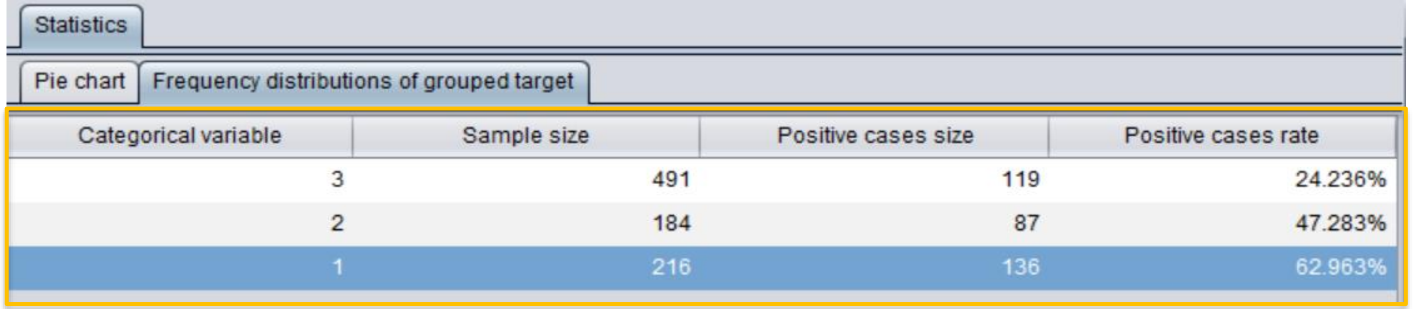


Survived	Pclass	Name	Sex
0	3	Braund, Mr. Owen Harris	male
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female
1	3	Heikkinen, Miss. Laina	female
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
0	3	Allen, Mr. William Henry	male

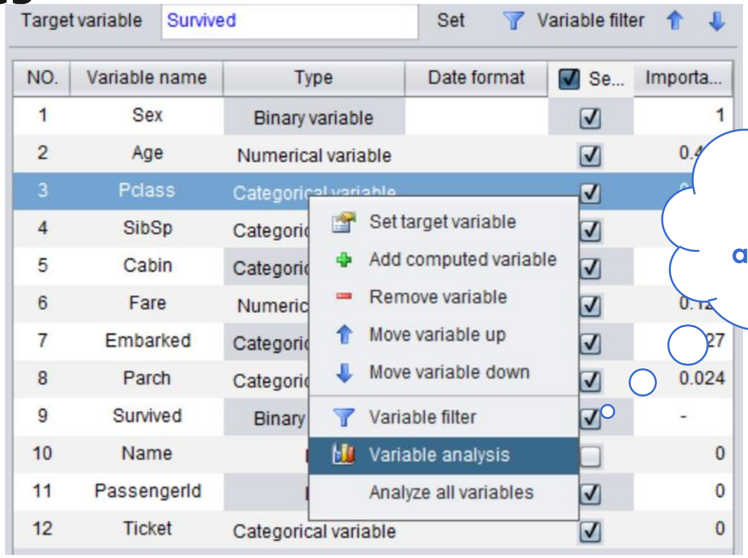
There is a variable "Pclass" in the Titanic data to represent the cabin level, and data exploration is carried out for it.

The statistical values are shown in the figure below, and the graphical distribution is shown in the figure below on the right,

It can be seen that there are three classes in total. The number of people in class 3 accounts for more than half of the total. The higher the cabin class is, the greater the proportion of survival is.



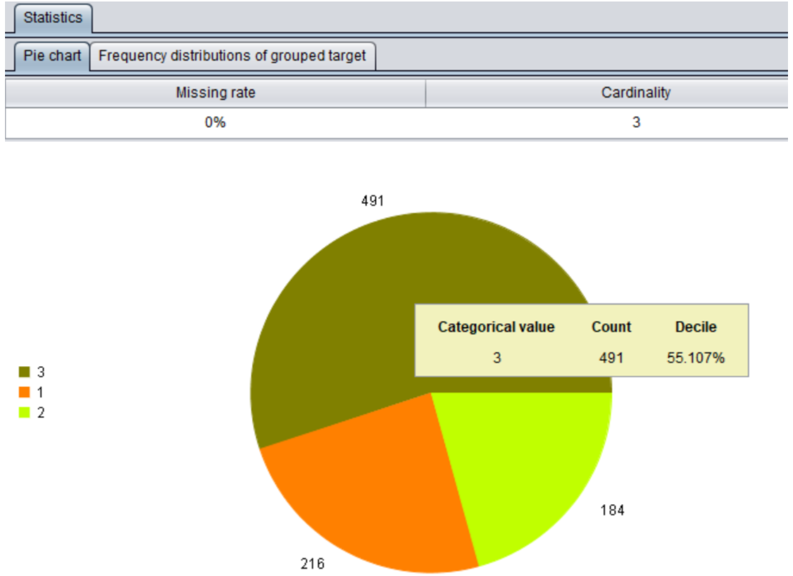
Categorical variable	Sample size	Positive cases size	Positive cases rate
3	491	119	24.236%
2	184	87	47.283%
1	216	136	62.963%



Target variable: Survived

NO.	Variable name	Type	Date format	☑ Se...	Importa...
1	Sex	Binary variable		☑	1
2	Age	Numerical variable		☑	0.4
3	Pclass	Categorical variable		☑	
4	SibSp	Categori		☑	
5	Cabin	Categori		☑	
6	Fare	Numeric		☑	0.1
7	Embarked	Categori		☑	27
8	Parch	Categori		☑	0.024
9	Survived	Binary		☑	-
10	Name			☐	0
11	PassengerId			☑	0
12	Ticket	Categorical variable		☑	0

Explore variant automatically



Missing rate	Cardinality
0%	3

Categorical value	Count	Decile
3	491	55.107%

Glossary - for reference

- correlation
- coefficient of correlation
- scatter plot
- chi-square test
- chi-square distribution
- expectation
- ordinal value
- expected cell count
- observed cell count
- null hypothesis
- statistical hypothesis testing
- student's t test
- t-distribution
- analysis of variance
- normal distribution

Chapter 3 Data pre-processing

Significance of data pre-processing

The purpose of data pre-preprocessing:

- Improve the quality of data;
- Better adapt data to specific mining technologies or tools

Based on the results of data exploration, according to different variable types, the corresponding processing is carried out, such as filling the missing values, rectifying the biased variables, discretizing the numerical data, add derived variables, removing redundant variables, etc.

The model built by directly throwing the original data into the open source algorithm is not effective, often because the preprocessing is not done in place.

Variable preliminary filtering

There are often poor quality or meaningless variables in the data. You can define some rules and delete them directly to reduce the amount of calculation, such as

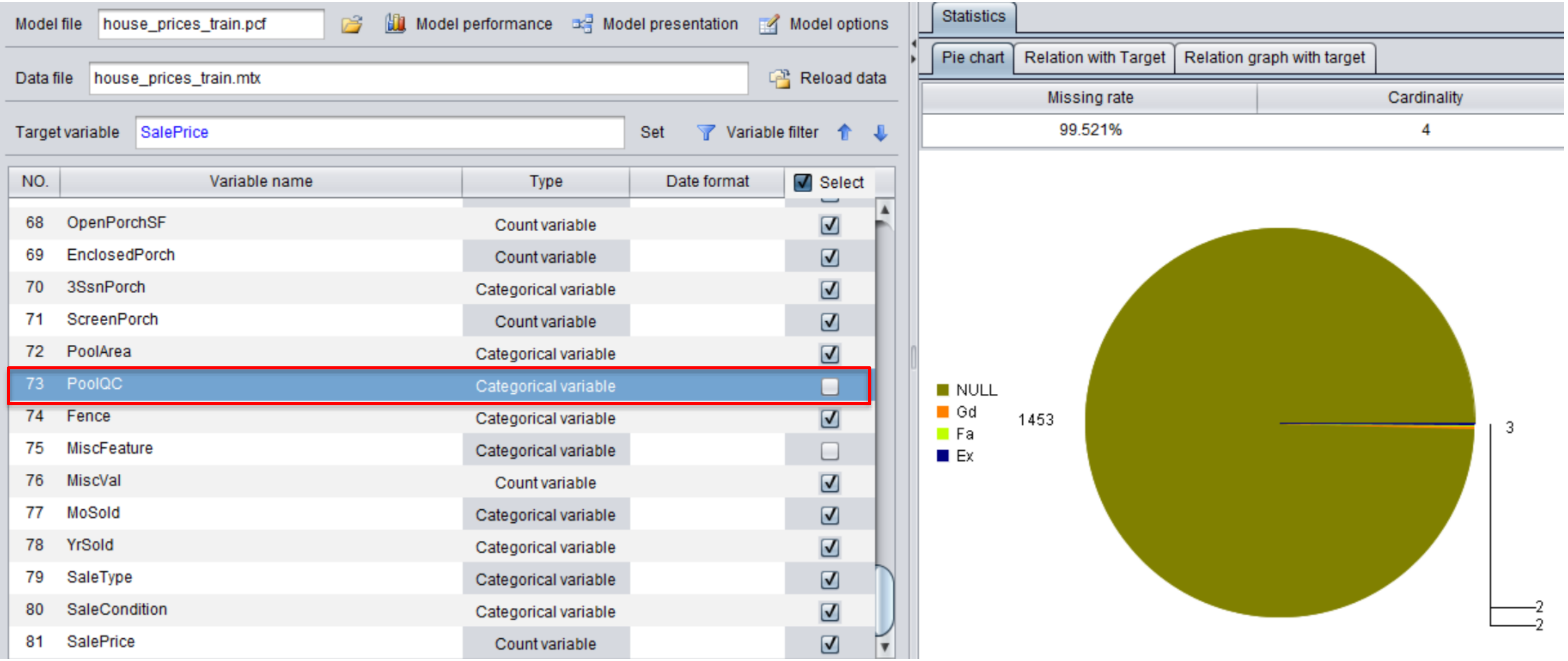
(1) Variables with high missing rate

(2) Unary variable

(3) Variable with too many categories

Variable preliminary filtering

For example, in Houseprice.csv , the missing rate of “PoolIQC” is more than 99%, and Raqsoft YModel will automatically eliminate this variable.



Model file: house_prices_train.pcf

Data file: house_prices_train.mtx

Target variable: SalePrice

NO.	Variable name	Type	Date format	☑ Select
68	OpenPorchSF	Count variable		☑
69	EnclosedPorch	Count variable		☑
70	3SsnPorch	Categorical variable		☑
71	ScreenPorch	Count variable		☑
72	PoolArea	Categorical variable		☑
73	PoolQC	Categorical variable		☐
74	Fence	Categorical variable		☑
75	MiscFeature	Categorical variable		☐
76	MiscVal	Count variable		☑
77	MoSold	Categorical variable		☑
78	YrSold	Categorical variable		☑
79	SaleType	Categorical variable		☑
80	SaleCondition	Categorical variable		☑
81	SalePrice	Count variable		☑

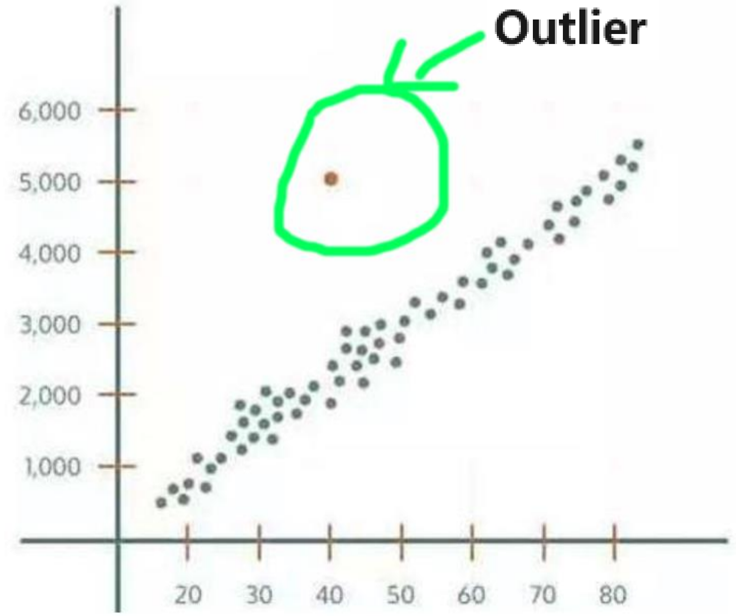
Missing rate	Cardinality
99.521%	4

Outlier analysis

In a data set, observations (or measurements) that are significantly larger or smaller than other data values are called outliers.

The causes of outliers are as follows:

1. The measured value is incorrectly observed, recorded, or entered into computer.
2. The measurements come from different source
3. The measurement is correct, but describes a rare (accidental) event.



Detection method of outliers

Box plot

z-score

Clustering

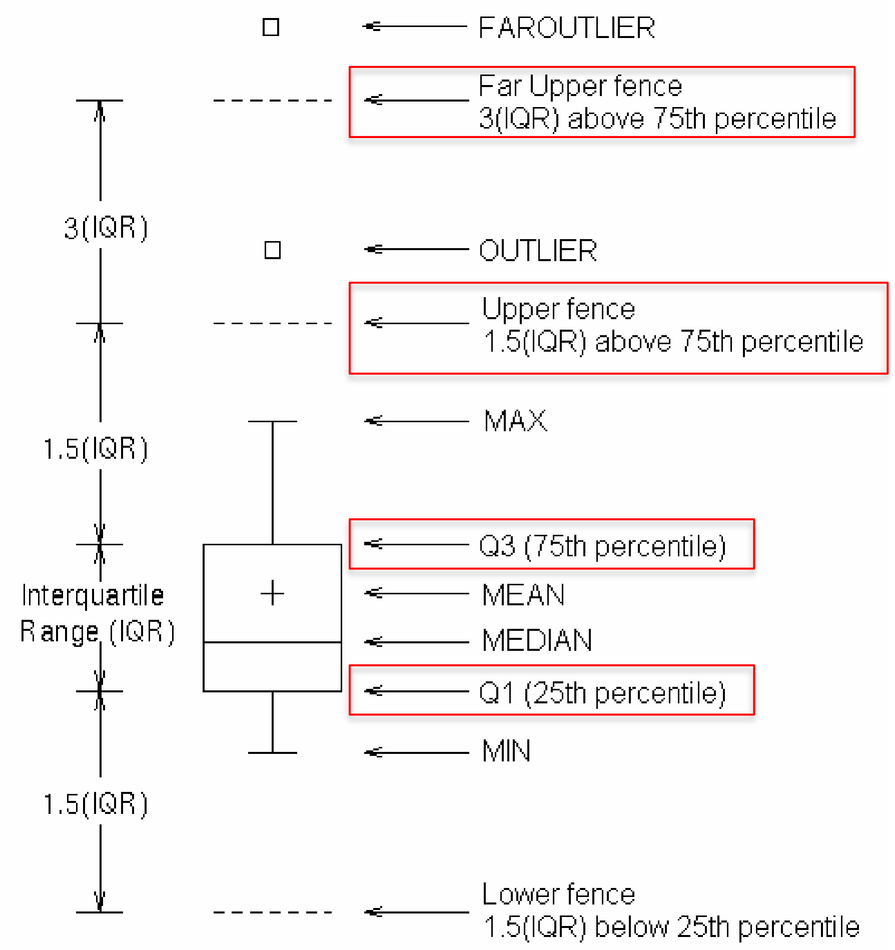
Outlier analysis

Box plot method:

Based on **IQR**, i.e. the distance between the upper and lower quartiles

Observations falling between the inner and outer fences are considered suspicious outliers

Observations falling outside the outer fences are considered highly suspicious outliers



Review:

$IQR=Q_3-Q_1$

Draw two lines at $Q_3+1.5IQR$ and $Q_1 - 1.5IQR$, which are the same as the median line, called the inner fences.

Draw two lines at Q_3+3IQR and $Q_1 - 3IQR$, which are called outer fences.

Outlier analysis

The rule of thumb for Z-score to detect outliers (the same as the rule of N times standard deviation):

Observations with an absolute value of Z-score greater than 3 are considered outliers (for some highly skewed datasets, observations with an absolute value of Z-score greater than 2 may be outliers)

Possible outliers	Highly suspicious outliers
$\lvert z\rvert > 2$	**$\lvert z\rvert > 3$**

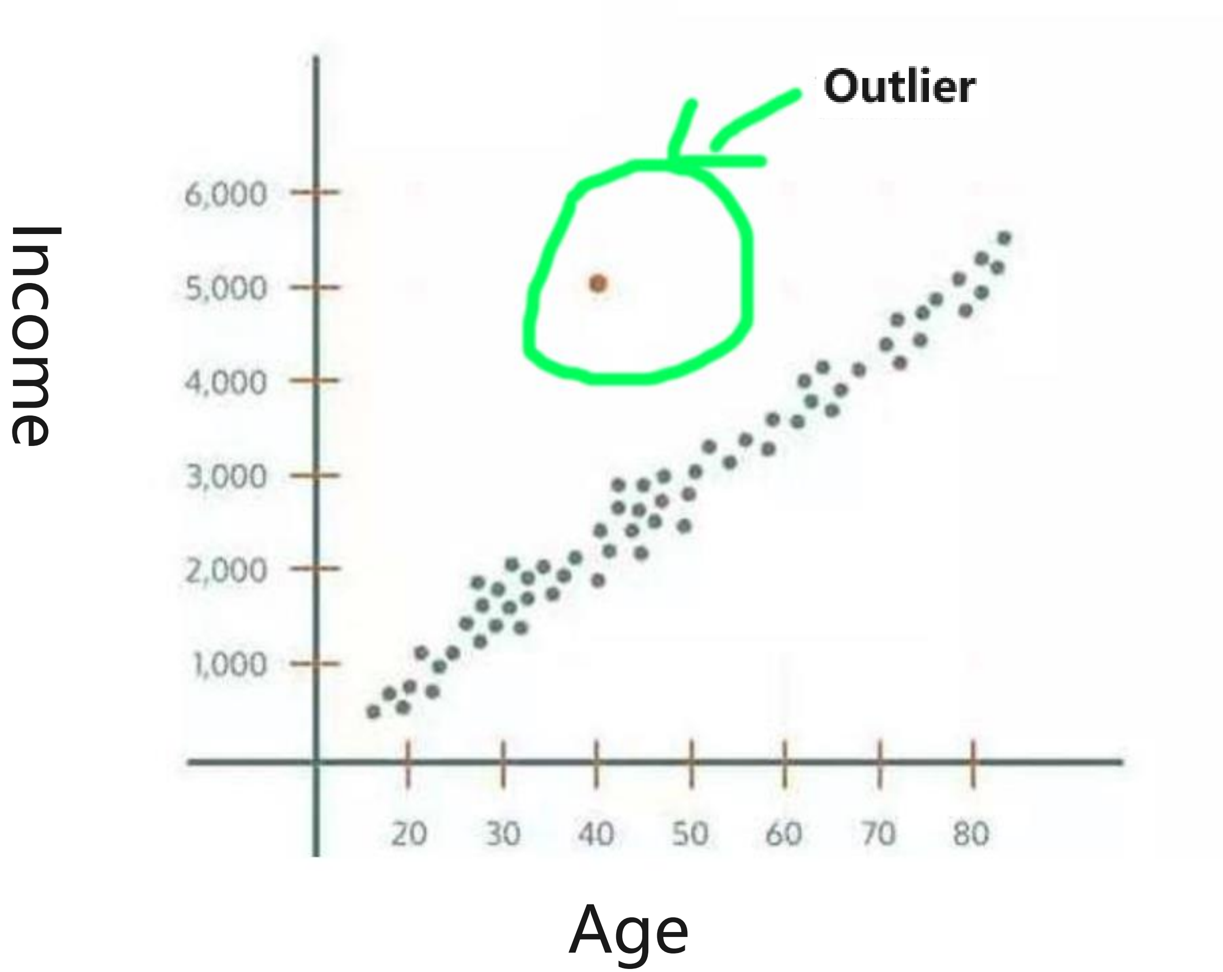
Review：

Z-score describes the distance between a given measurement value x and the average value, which is expressed in standard deviation.

Z-score formula： $Z=\frac{x-\bar{x}}{s}$

Outlier analysis

Clustering: it deals with the outliers of multiple variables. The common algorithm k-means is applicable to the case of large data volume.



As shown in the figure, 40 years old is not an outlier only from the perspective of age, and 5000 is not an outlier only from the perspective of income. However, this point is really different from the behavior mode of other points, so cluster analysis can integrate multiple variable values to judge the existence of outliers.

Outlier handling

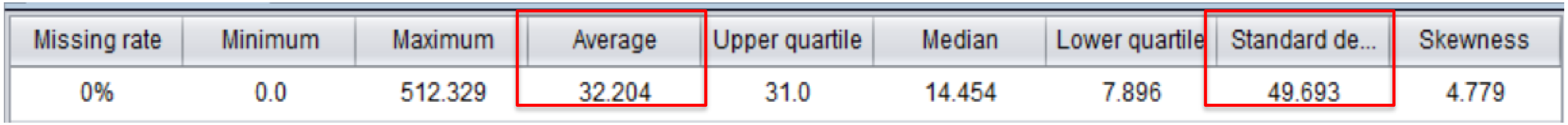
Handling method of outliers:

- **Delete records with outliers**： directly delete records with outliers;
- **Treat as missing value**： treat outlier as missing value, and use the missing value processing method to process；
- **Correction of outliers**： the outliers can be corrected by the endpoint value or the average of the two observed values；
- **No processing**： data mining directly on datasets with outliers；

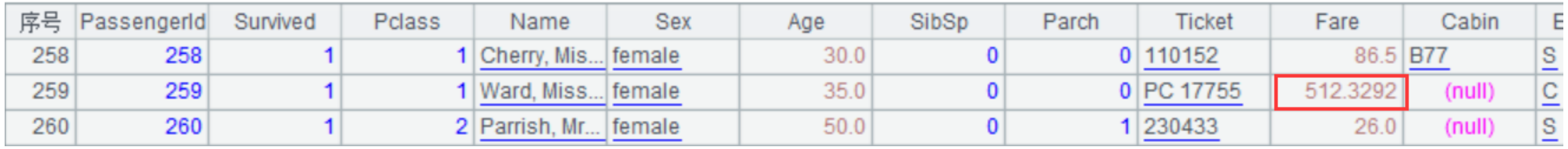
It should be emphasized that how to determine and handle outliers needs to be combined with practice. Because some models are not very sensitive to outliers, even if there are outliers, the model effect will not be affected. However, some models, such as logistic regression LR and AdaBoost, are very sensitive to outliers. If they are not processed, they may have very poor effects such as over fitting.

Class exercise - Outlier handling

For example, for the "fare" variable in Titanic.csv, we use | z| = 5 as the endpoint value to correct the outliers outside this range. According to the formula： $Z=\frac{x-\bar{x}}{s}$, the average value and standard deviation of the variable are brought in, and the maximum value of the endpoint is 280.671, while 512.329 of the variable has exceeded the value, and the correction result is shown in the figure:

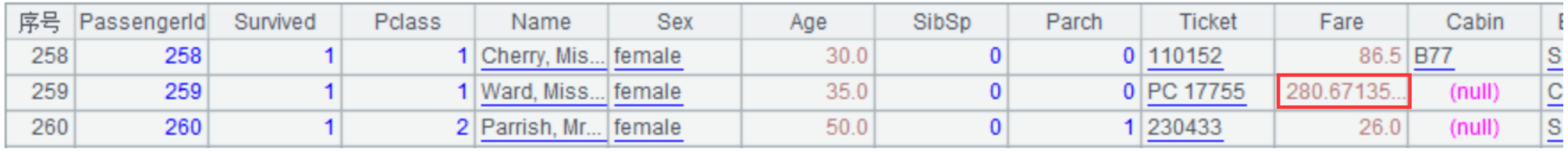


Missing rate	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard de...	Skewness
0%	0.0	512.329	32.204	31.0	14.454	7.896	49.693	4.779



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
258	258	1	1	Cherry, Mis...	female	30.0	0	0	110152	86.5	B77	S
259	259	1	1	Ward, Miss...	female	35.0	0	0	PC 17755	512.3292	(null)	C
260	260	1	2	Parrish, Mr...	female	50.0	0	1	230433	26.0	(null)	S

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
258	258	1	1	Cherry, Mis...	female	30.0	0	0	110152	86.5	B77	S
259	259	1	1	Ward, Miss...	female	35.0	0	0	PC 17755	280.67135...	(null)	C
260	260	1	2	Parrish, Mr...	female	50.0	0	1	230433	26.0	(null)	S

After preprocessing

Class exercise - Outlier handling

SPL Code for outlier correction :

	A	B	C	
1	D:/test/titanic.csv	5	-5	/Define data path and values for Z
2	=file(A1).import@qtc()			/Read data
3	==sqrt(var(A2.(Fare)))	=A2.(Fare).avg()		
4	=B1*A3+B3	=C1*A3+B3		/Calculate extreme value of fare variable
5	=A2.run(Fare=if(Fare>A4,A4,if(Fare<B4,B4,Fare)))			/Correct outlier

Missing value analysis

Type of missing value

1. MCAR (missing completely at random): data loss is completely random, does not depend on any other variables, does not affect the unbiasedness of samples. (for example, home address is missing)

2. MAR (missing at random): data loss is not completely random, that is, the loss of this kind of data depends on other variables. (for example, the lack of financial data is related to the size of the enterprise))

3. MNAR (missing not at random): there are two possible cases. The missing value depends on its hypothesis (for example, high-income people usually do not want to disclose their income in the survey); or, the missing value depends on other variable values (assuming that women usually do not want to disclose their age, the missing value of age variable is affected by gender variable).

Missing value processing method

1

Delete directly

2

Fill in missing values

3

No processing

Missing value handling - delete directly

Advantage	Simple and rough, high efficiency
Disadvantage	At the expense of a large number of data, when the proportion of missing data is large, especially when the missing data is not randomly distributed, direct deletion may lead to data distribution deviation and model deviation.
Applicable scenarios	The dataset is very large and there is not much missing data.

But this kind of situation is rare. It is more the case that there is a lot of missing data. For example, each variable is missing 30, but if the missing is not coincident, as long as one variable is missing, the whole piece of data will be deleted, and then the 10 variables will delete 300 pieces of data, so when there are many missing variables, deleting directly will lead to a large reduction of sample size.

Missing value handling - fill in missing values

Simple filling	Mean, median, mode filling: the numerical variable is filled with mean or median, and the categorical variable is filled with mode.
Hotdecking	It is also called nearest complement. In the complete data, find a record that is most similar to the record with missing value to fill in, but it is difficult to define the similar standard.
Cluster filling	After clustering, the missing values in a class are filled with the sample mean in the class. It can get better effect of complement, but when there is a large amount of data or a large number of missing attribute values, the cost of calculation is very high.

Missing value handling - fill in missing values

Fit missing value

Using other variables as input of the model to predict the missing variable.

Regression prediction: for the object with missing value, bring the known data set into the regression equation to estimate the predicted value, and fill in the predicted value, but when the variable is not linearly correlated, it will lead to the estimation of deviation;

Maximum expectation prediction: in the case of incomplete data to calculate the maximum likelihood estimation or posterior distribution of iterative algorithm, this method may fall into local extreme value, convergence speed is not very fast, and calculation is very complex.

Multiple imputation prediction: a set of possible imputation values is generated for each missing value. These values reflect the uncertainty of the missing value. Then, the imputation set is selected according to the scoring function to generate the final imputation value.

Dummy variable

Derive a variable with a value of 0, 1 to mark whether a variable is missing.

Missing value handling - no processing

Direct modeling	Some algorithms can deal with missing values themselves, and can directly model datasets containing missing values, such as XGB, LGB

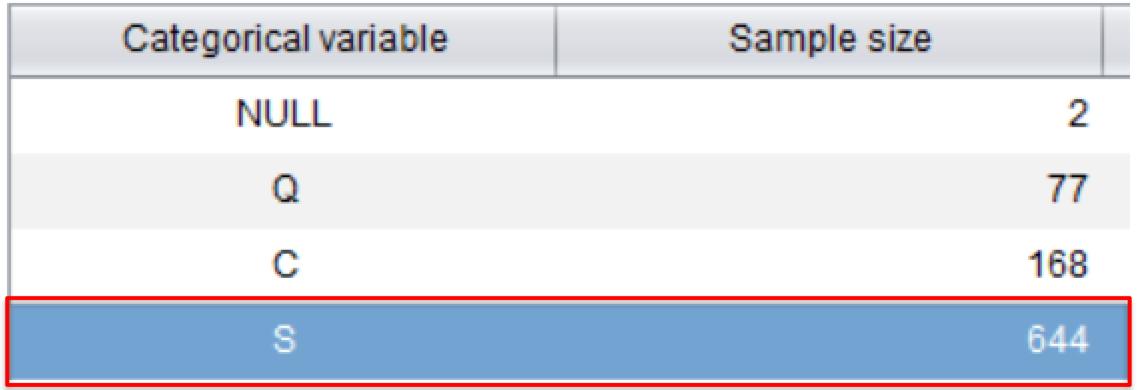
Thinking：

Is there a best way to deal with missing values?

Class exercise - Missing value handling

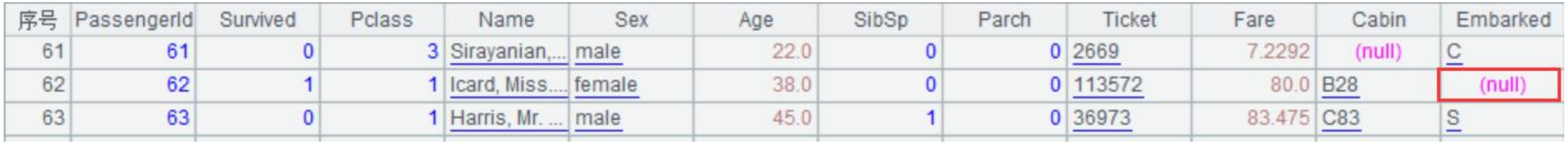
"Embanked" in Titanic.csv is a categorical variable. We use mode to fill in the missing value.

The result is as follows:



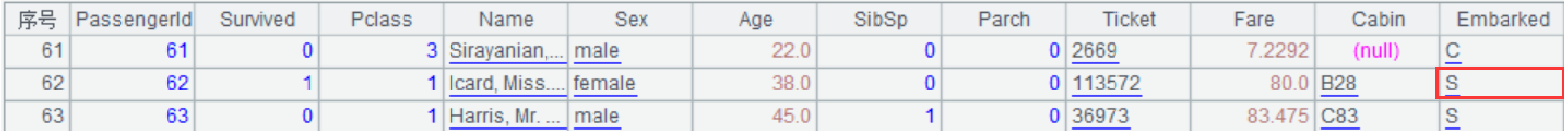
Categorical variable	Sample size
NULL	2
Q	77
C	168
S	644

Mode is "s"



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
61	61	0	3	Sirayanian,...	male	22.0	0	0	2669	7.2292	(null)	C
62	62	1	1	Icard, Miss....	female	38.0	0	0	113572	80.0	B28	(null)
63	63	0	1	Harris, Mr. ...	male	45.0	1	0	36973	83.475	C83	S

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
61	61	0	3	Sirayanian,...	male	22.0	0	0	2669	7.2292	(null)	C
62	62	1	1	Icard, Miss....	female	38.0	0	0	113572	80.0	B28	S
63	63	0	1	Harris, Mr. ...	male	45.0	1	0	36973	83.475	C83	S

After preprocessing

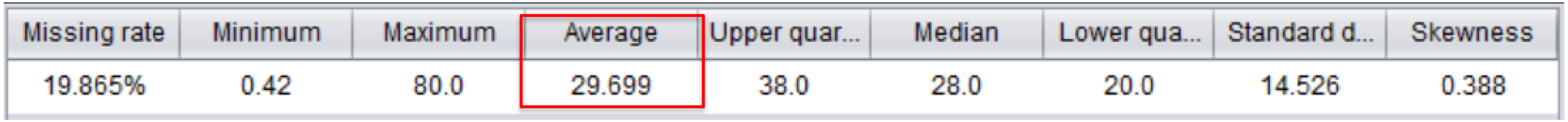
Class exercise - Missing value handling

SPL Code for mode filling :

	A	B
1	D:/test/titanic.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=A2.groups(Embarked;count(~):count)	
4	=A3.maxp(count).Embarked	/Calculate mode of Embarked
5	=A2.run(Embarked=if(Embarked==null,A4,Embarked))	/Fill in missing value with mode

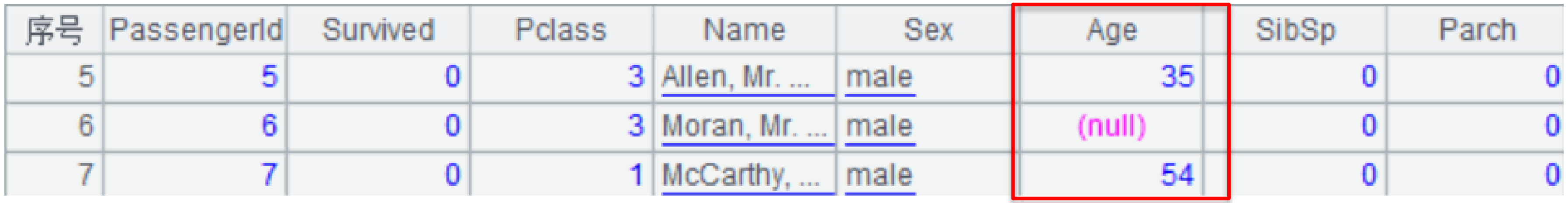
Class exercise - Missing value handling

The "age" in Titanic.csv is a numerical variable. We use the average value to fill in the missing value.
The result is as follows:



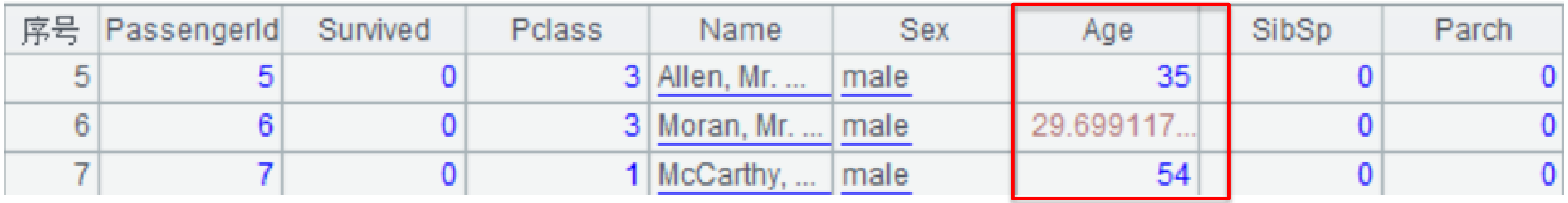
Missing rate	Minimum	Maximum	Average	Upper quar...	Median	Lower qua...	Standard d...	Skewness
19.865%	0.42	80.0	29.699	38.0	28.0	20.0	14.526	0.388

The average value of "age" is 29.699



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
5	5	0	3	Allen, Mr. ...	male	35	0	0
6	6	0	3	Moran, Mr. ...	male	(null)	0	0
7	7	0	1	McCarthy, ...	male	54	0	0

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
5	5	0	3	Allen, Mr. ...	male	35	0	0
6	6	0	3	Moran, Mr. ...	male	29.699117...	0	0
7	7	0	1	McCarthy, ...	male	54	0	0

After preprocessing

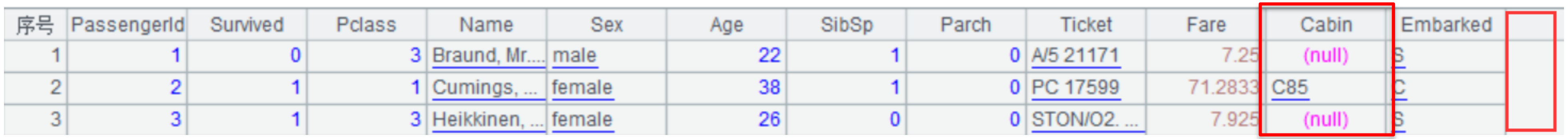
Class exercise - Missing value handling

SPL Code for average filling :

	A	B
1	D:/test/titanic.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=A2.avg(Age)	/Calculate average of age
4	=A2.run(Age=if(Age==null,A3,Age))	/Fill in missing value with average

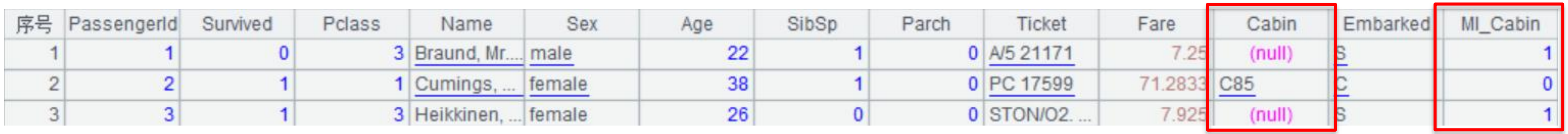
Class exercise - Missing value handling

The "cabin" in Titanic.csv is a variable with missing value. We derive a new dummy variable to mark the missing information of the variable:



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S	
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C	
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O2. ...	7.925	(null)	S	

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	MI_Cabin
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S	1
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C	0
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O2. ...	7.925	(null)	S	1

After preprocessing

Class exercise - Missing value handling

SPL code(dummy variable to mark the missing value):

	A	B
1	D:/test/titanic.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=A2.derive(if(Cabin==null,1,0):MI_Cabin)	/Generate dummy variable

Missing value handling

The filling process is to supplement the unknown value with our subjective estimate, which is not necessarily in line with the objective facts. The above analysis is theoretical analysis. As for the missing value, it is impossible to know its missing type and estimate the effect of filling method because it cannot be observed.

In addition, these methods are universal in various fields, so the filling effect for a field's specialty will not be very ideal. For this reason, many professional data mining personnel through their understanding of the industry, the effect of filling the missing value manually may be better than these methods.

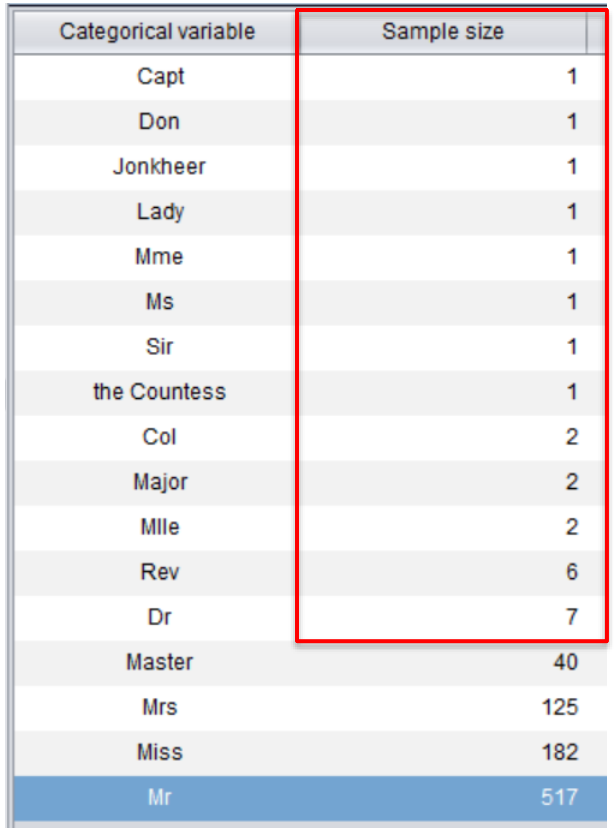
Missing value filling is a kind of human intervention method for missing value in the process of data mining without giving up information. No matter what kind of processing method, it will affect the relationship between variables. When filling incomplete information, we change the original data information system more or less, which has potential impact on later analysis, so we must be careful about missing value handling.

Categorical variable handling

Noise reduction of categorical variables	When the number of categories of categorical variable is large, there may be noise, such as category with very few sample, abnormal category, suspected error category, etc. in this case, the number of category can be reduced by combining low frequency variables.
Numerical the categorical variables	Categorical variables are usually in the form of characters, which can not be directly recognized and calculated by the algorithm, and must be converted into numerical data. Simplest way: enumerate all values, use integer mapping, one-hot-encoder

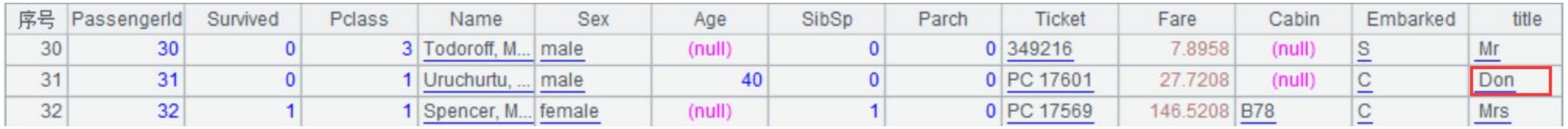
Class exercise - noise reduction of categorical variables

The "title" in Titanic.csv is a categorical variable. There are some low frequency categories, such as "Capt", “Don" ..., we combine them to reduce data noise:



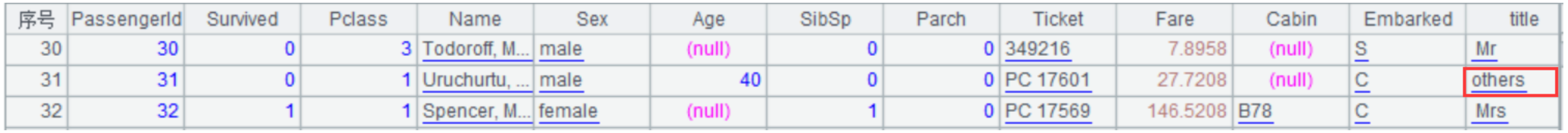
Categorical variable	Sample size
Capt	1
Don	1
Jonkheer	1
Lady	1
Mme	1
Ms	1
Sir	1
the Countess	1
Col	2
Major	2
Mlle	2
Rev	6
Dr	7
Master	40
Mrs	125
Miss	182
Mr	517

Low frequency variables



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title
30	30	0	3	Todoroff, M...	male	(null)	0	0	349216	7.8958	(null)	S	Mr
31	31	0	1	Uruchurtu, ...	male	40	0	0	PC 17601	27.7208	(null)	C	Don
32	32	1	1	Spencer, M...	female	(null)	1	0	PC 17569	146.5208	B78	C	Mrs

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title
30	30	0	3	Todoroff, M...	male	(null)	0	0	349216	7.8958	(null)	S	Mr
31	31	0	1	Uruchurtu, ...	male	40	0	0	PC 17601	27.7208	(null)	C	others
32	32	1	1	Spencer, M...	female	(null)	1	0	PC 17569	146.5208	B78	C	Mrs

After preprocessing

Class exercise - noise reduction of categorical variables

SPL code for noise reduction of categorical variables:

	A	B
1	D:/test/titanic.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=A2.group(title)	
4	=A3.align@a([true,false],~.len()<10)	
5	=A4(1).(~.run(~.title="others"))	/Merge those with classification frequency less than 10 into others
6	=A2	

Numerical categorical variables

Data mapping

Goods	Value
Quilt	1
Pen	2
Table	3
Towel	4
Cup	5
Tea	6
Noodle	7

Can the converted value be calculated directly?

No, the value itself has no mathematical meaning, only represents the text category. If it is directly involved in the calculation, the program will think that the size of the value itself has an impact.

Numerical categorical variables

One-hot-encoder

Goods	Value
Quilt	1
Pen	2
Table	3
Towel	4
Cup	5
Tea	6
Noodle	7

1	2	3	4	5	6	7
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

Numerical categorical variables

Training dataset

Goods		Value
Quilt	→	1
Pen	→	2
Table	→	3
Towel	→	4
Cup	→	5
Tea	→	6
Noodle	→	7

Testing dataset

Goods	Value
Quilt	?
Pen	?
Tea table	?
Towel	?
Cup	?
Milk	?
Beer	?

Thinking:
How should test set be mapped when new classifications appear on test data?

Numerical categorical variables

Training dataset

Goods		Value
Quilt	→	1
Pen	→	2
Table	→	3
Towel	→	4
Cup	→	5
Tea	→	6
Noodle	→	7

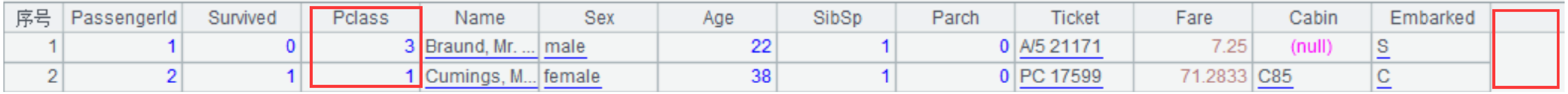
Testing dataset

Goods	✗ Value	✓ Value
Quilt	1	1
Pen	2	2
Tea table	3	0
Towel	4	4
Cup	5	5
Milk	6	0
Beer	7	0

The numerical mapping between the test set and the training set should be consistent. Rules should be determined on how to deal with words that did not appear before.

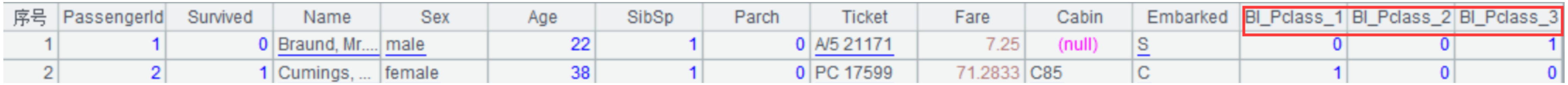
Class exercise - Numerical categorical variables

The “Pclass" in Titanic.csv is a categorical variable, with categories of "1", "2" and "3". We use one hot encoder to make it numerical:



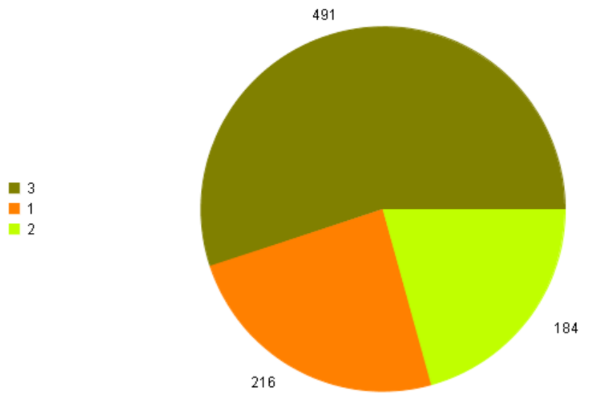
序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. ...	male	22	1	0	A/5 21171	7.25	(null)	S
2	2	1	1	Cumings, M...	female	38	1	0	PC 17599	71.2833	C85	C

Before preprocessing



序号	PassengerId	Survived	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	BI_Pclass_1	BI_Pclass_2	BI_Pclass_3
1	1	0	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(null)	S	0	0	1
2	2	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85	C	1	0	0

After preprocessing



“Pclass” pie chart

Class exercise - Numerical categorical variables

SPL code for one-hot-encoder:

	A	B
1	D:/test/titanic.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=A2.derive(if(Pclass==1,1,0):BI_Pclass_1,if(Pclass==2,1,0):BI_Pclass_2,if(Pclass==3,1,0):BI_Pclass_3)	/One hot transformation of pclass
4	=A3.fname().select(~!="Pclass")	
5	=A4.concat@c()	
6	=A3.new(${A5})	/Delete original Pclass

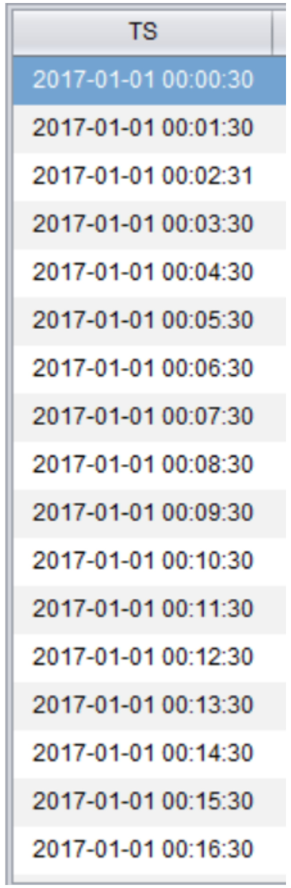
Processing of date time variable

Date time variable is a common variable in business. There are many ways to extract information from the variable.

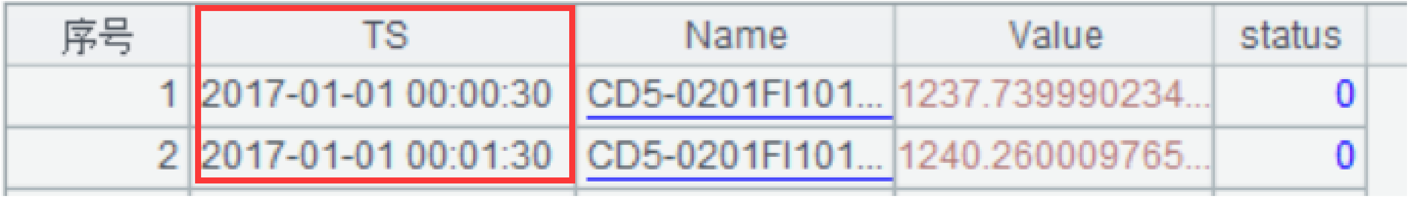
Date feature disassemble	For example, 2020/01/01, it can be disassembled into year (2020), month (01), day (01), season (winter), day (Wednesday), working day / holiday (holiday) ...
Time feature disassemble	For example, 05:10:30, it can be disassembled into hour, minute, second, which time period of the day: early morning, morning, noon, afternoon, evening, late night ...
Date interval calculation	Two date variables are subtracted, such as the number of days since the last purchase, the interval between the loan date and the repayment date, the interval between the due date and the actual repayment date.

Class exercise - date time variable processing

The "TS" in Meter_ data.csv is a date time variable. We process it as follows:

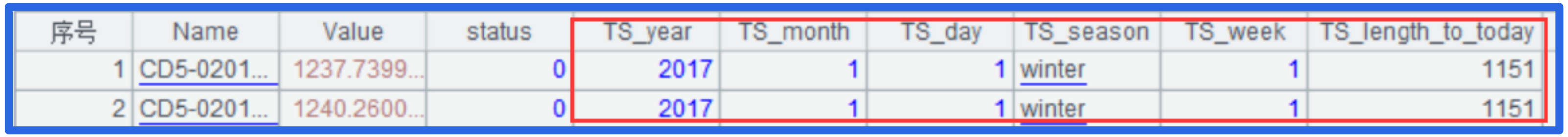


TS
2017-01-01 00:00:30
2017-01-01 00:01:30
2017-01-01 00:02:31
2017-01-01 00:03:30
2017-01-01 00:04:30
2017-01-01 00:05:30
2017-01-01 00:06:30
2017-01-01 00:07:30
2017-01-01 00:08:30
2017-01-01 00:09:30
2017-01-01 00:10:30
2017-01-01 00:11:30
2017-01-01 00:12:30
2017-01-01 00:13:30
2017-01-01 00:14:30
2017-01-01 00:15:30
2017-01-01 00:16:30



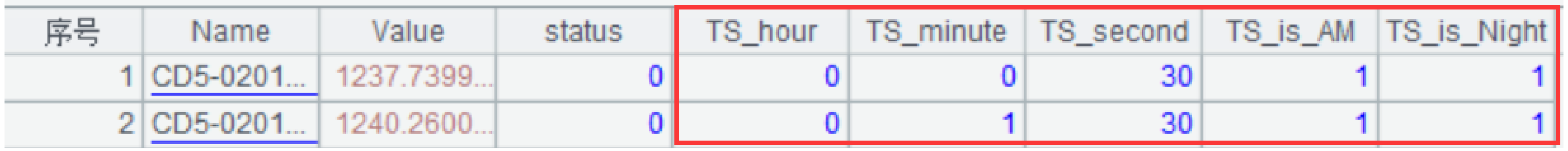
序号	TS	Name	Value	status
1	2017-01-01 00:00:30	CD5-0201FI101...	1237.739990234...	0
2	2017-01-01 00:01:30	CD5-0201FI101...	1240.260009765...	0

Before preprocessing



序号	Name	Value	status	TS_year	TS_month	TS_day	TS_season	TS_week	TS_length_to_today
1	CD5-0201...	1237.7399...	0	2017	1	1	winter	1	1151
2	CD5-0201...	1240.2600...	0	2017	1	1	winter	1	1151

After preprocessing- Extract year, month, day, season, week and days to now



序号	Name	Value	status	TS_hour	TS_minute	TS_second	TS_is_AM	TS_is_Night
1	CD5-0201...	1237.7399...	0	0	0	30	1	1
2	CD5-0201...	1240.2600...	0	0	1	30	1	1

After preprocessing- Extract hour, minute, second and time period from time

Class exercise - date time variable processing

SPL Code for date variable processing :

<table>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th></tr>
<tr><td>1</td><td>D:/test/meter_data.csv</td><td>/Define file path</td><td></td><td></td></tr>
<tr><td>2</td><td>=file(A1).import@qtc()</td><td>=(["winter"]*2)|(["spring"]*3)|(["summer"]*3)|(["autumn"]*3)|["winter"]</td><td colspan="2">/Define season</td></tr>
<tr><td>3</td><td></td><td>=["year","month","day","season","week","length_to_today"].("TS"+"_"+~)</td><td colspan="2">/Define field name</td></tr>
<tr><td>4</td><td></td><td>=A2.(date(TS))</td><td>=B4.(year(~))</td><td>=B4.(month(~))</td></tr>
<tr><td>5</td><td></td><td>=B4.(day(~))</td><td>=D4.((m=~,B2(m)))</td><td>=B4.(day@w(~))</td></tr>
<tr><td>6</td><td></td><td>=B4.(interval(~,now()))</td><td>=A2.fname()\"TS"</td><td>=C6.((f=~,A2.(eval(f))))</td></tr>
<tr><td>7</td><td></td><td>=D6|[C4,D4,B5,C5,D5,B6]</td><td>=A2.len().((ind=#,B7.(~(ind))))</td><td>=C6|B3</td></tr>
<tr><td>8</td><td></td><td>=create(${D7.concat@c()})</td><td>=B8.record(C7.conj())</td><td>/Extract year, month, day, season, week and days to now</td></tr>
</table>

Class exercise - date time variable processing

SPL Code for time variable processing :

	A	B	C	D	
1	D:/test/meter_data.csv	/Define file path			
2	=file(A1).import@qtc()	=["hour","minute","second","is_AM","is_Night"].("TS"+"_"+~)	/Define field name		
3		=A2.(time(TS))	=B3.(hour(~))	=B3.(minute(~))	
4		=B3.(second(~))	=C3.(if(~>=0&&~<=11,1,0))	=C3.(if(~>=6&&~<=17,0,1))	/Extract hour, minute, second and time period
5		=A2.fname()\"TS"	=B5.((f=~,A2.(eval(f))))	=C5|[C3,D3,B4,C4,D4]	
6		=A2.len().((ind=#,D5.(~(ind))))	=create(${(B5|B2).concat@c()})	=C6.record(B6.conj())	/Merge into a new table

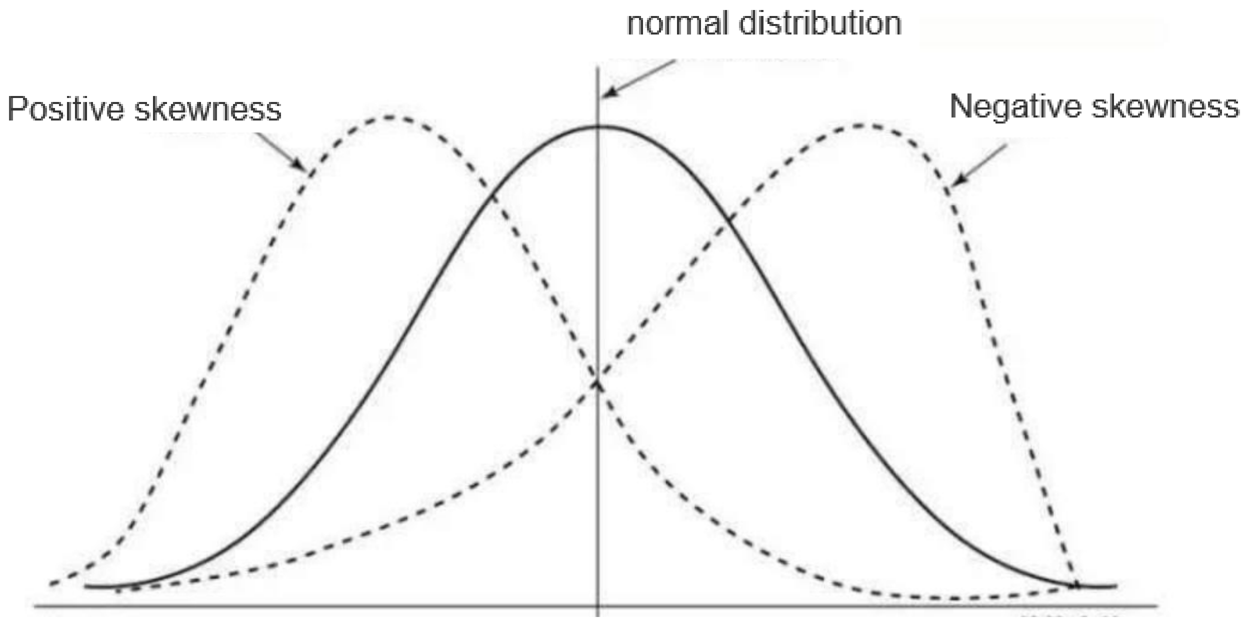
Skewness handling

Review:

Skewness: it is a measure of data symmetry. If its value is 0, it means a distribution of symmetry; if its value is positive, it means the distribution is right biased (positive skewness); if its value is negative, it means the distribution is left biased (negative skewness). The larger the absolute value of the skewness is, the greater the skewness is.

Many data analysis algorithms are based on the distribution of data which is similar to normal distribution, and the data is distributed around the mean.

If the absolute value of skewness is too large, the data can only be used after correction, otherwise the value of the indicator will be greatly reduced.

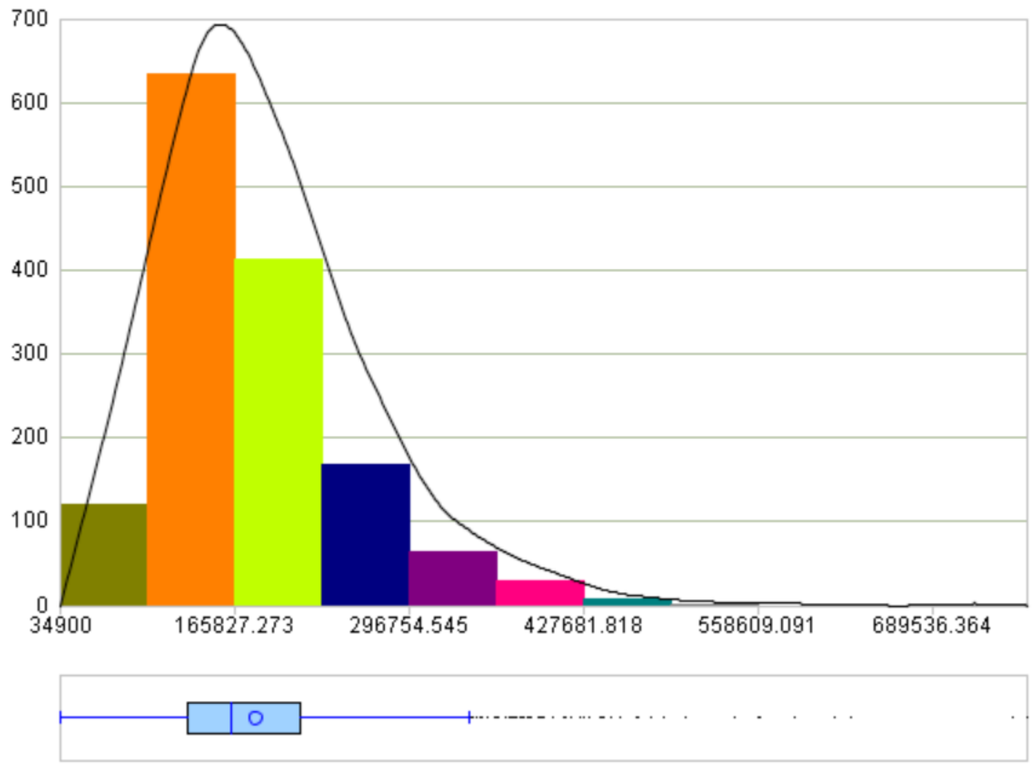


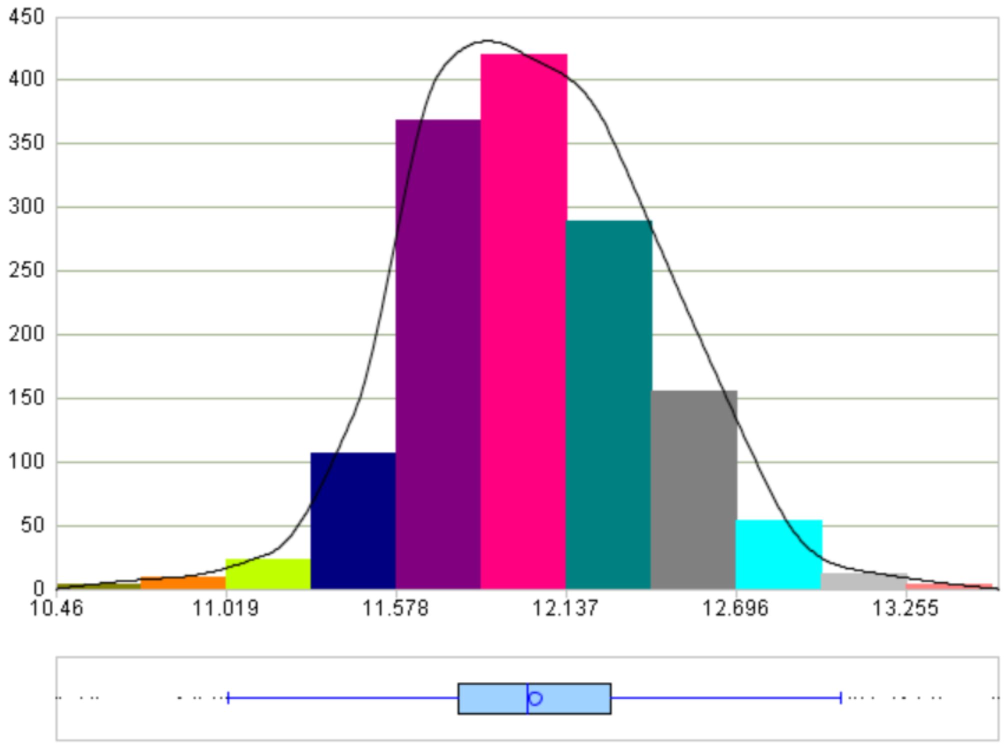
Skewness handling

Skewness handling is to make the distribution of variables present or approximate normal distribution through various mathematical transformations, and the fitting of the model often has obvious improvement.

Commonly used mathematical transformations include: logarithmic transformation, power transformation (such as square root, square, etc.), reciprocal transformation, exponential transformation, etc

Example： There is a variable saleprice in houseprise.csv. Its original skewness is 1.881. We perform log transformation on it. After transformation, the skewness is 0.121.





Class exercise - Skewness handling

SPL Code for skewness calculation :

Before skewness handling, we first calculate the skewness of the variable. Here we write a script to calculate the skewness separately and save it as skew_ calc.dfx , which makes it easy to call.

Seq is the parameter name of the script, the sequence used to calculate the skewness.

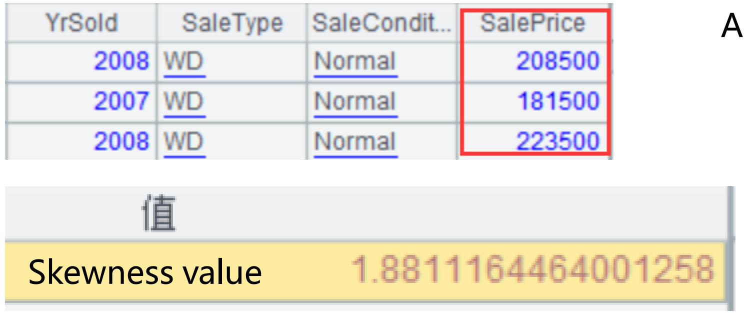
	A	B
1	=seq.avg()	/Average value
2	=seq.count()	/number of samples
3	=seq.sum(power((~-A1),3))/A2	
4	=power(seq.sum(power((~-A1),2))/A2,1.5)	
5	=sqrt(A2*(A2-1))/(A2-2)	
6	return A5*A3/A4	/Calculate skewness value and return

Class exercise - Skewness handling

SPL code for skewness handling (log transformation) :

	A	B
1	D:/test/houseprice_train.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=call("D:/test/skew_calc.dfx",A2.(SalePrice))	/Call the script to calculate the pre transformation skewness
4	=A2.run(SalePrice=ln(SalePrice))	/Carry out logarithmic transformation
5	=A2.rename(SalePrice:log_SalePrice)	
6	=call("D:/test/skew_calc.dfx",A2.(log_SalePrice))	/Skewness after transformation

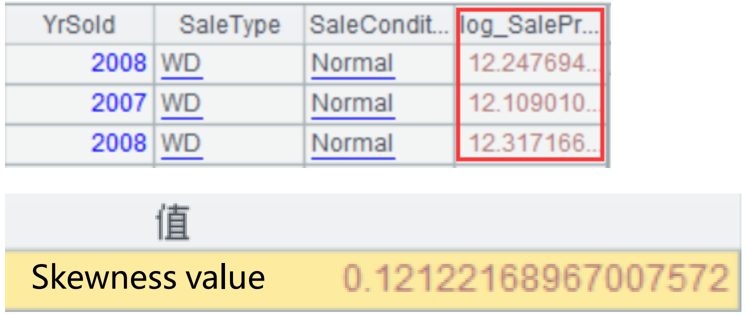
Before preprocessing



YrSold	SaleType	SaleCondit...	SalePrice
2008	WD	Normal	208500
2007	WD	Normal	181500
2008	WD	Normal	223500

	值
Skewness value	1.8811164464001258

After preprocessing

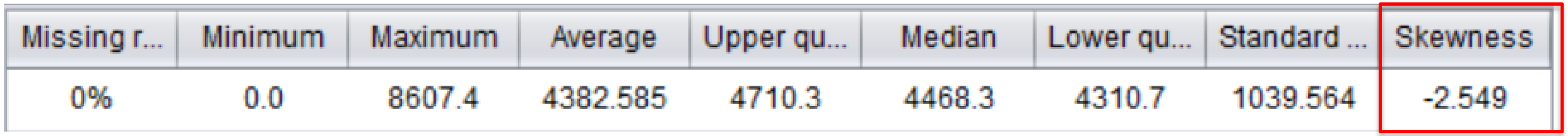


YrSold	SaleType	SaleCondit...	log_SalePr...
2008	WD	Normal	12.247694...
2007	WD	Normal	12.109010...
2008	WD	Normal	12.317166...

	值
Skewness value	0.12122168967007572

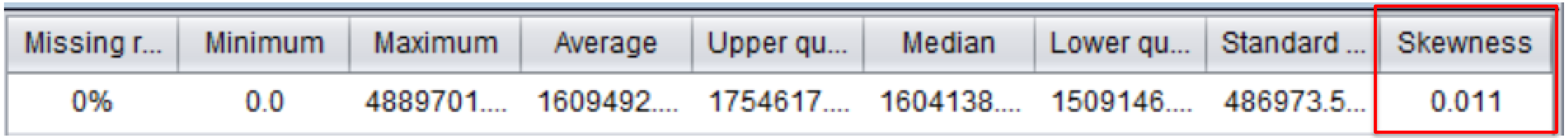
Class exercise - Skewness handling

The "sales" in Catering_ sale.csv is a variable with an original skewness of - 2.549. We perform power transformation on it, and the skewness after transformation is 0.011



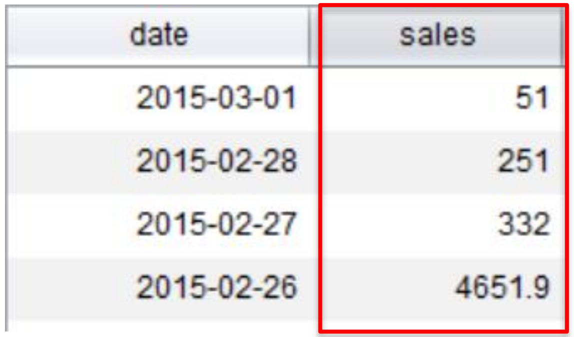
Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	0.0	8607.4	4382.585	4710.3	4468.3	4310.7	1039.564	-2.549

Skewness before transformation



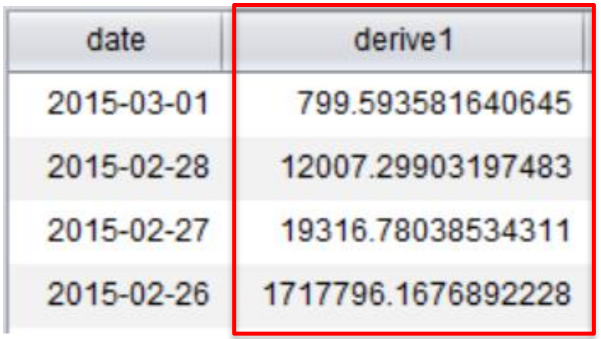
Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	0.0	4889701....	1609492....	1754617....	1604138....	1509146....	486973.5...	0.011

Skewness after transformation



date	sales
2015-03-01	51
2015-02-28	251
2015-02-27	332
2015-02-26	4651.9

Before preprocessing



date	derive1
2015-03-01	799.593581640645
2015-02-28	12007.29903197483
2015-02-27	19316.78038534311
2015-02-26	1717796.1676892228

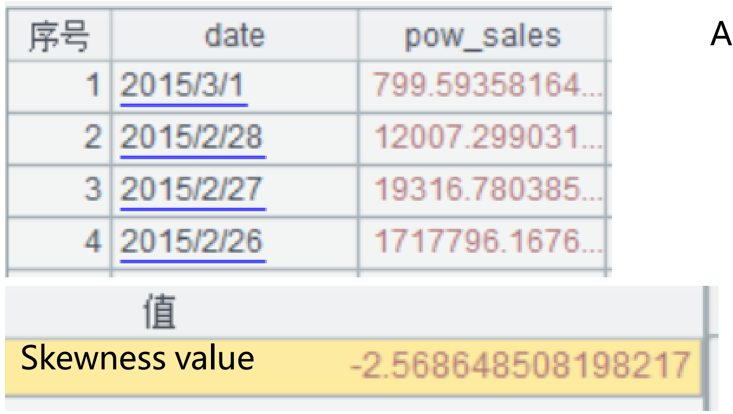
After preprocessing

Class exercise - Skewness handling

SPL code for skewness handling (power transformation) :

	A	B
1	D:/test/catering_sale.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=call("D:/test/skew_calc.dfx",A2.(sales))	/skewness before transformation
4	=A2.run(sales=power(sales,1.7))	/Carry out power transformation
5	=A2.rename(sales:pow_sales)	
6	=call("D:/test/skew_calc.dfx",A2.(pow_sales))	/skewness after transformation

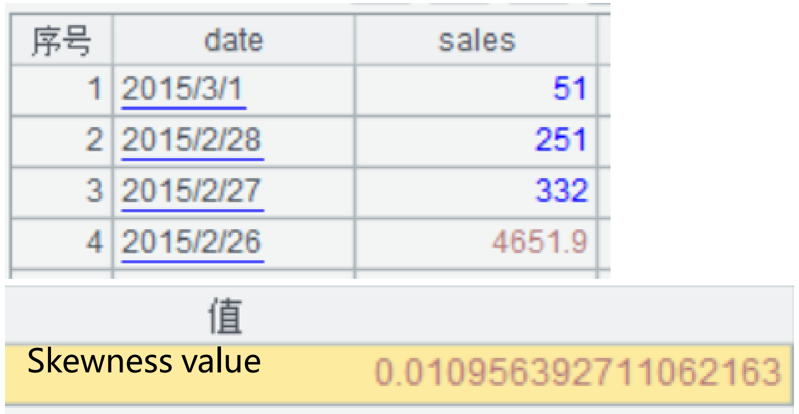
Before preprocessing



序号	date	pow_sales
1	2015/3/1	799.59358164...
2	2015/2/28	12007.299031...
3	2015/2/27	19316.780385...
4	2015/2/26	1717796.1676...

值	
Skewness value	-2.568648508198217

After preprocessing



序号	date	sales
1	2015/3/1	51
2	2015/2/28	251
3	2015/2/27	332
4	2015/2/26	4651.9

值	
Skewness value	0.010956392711062163

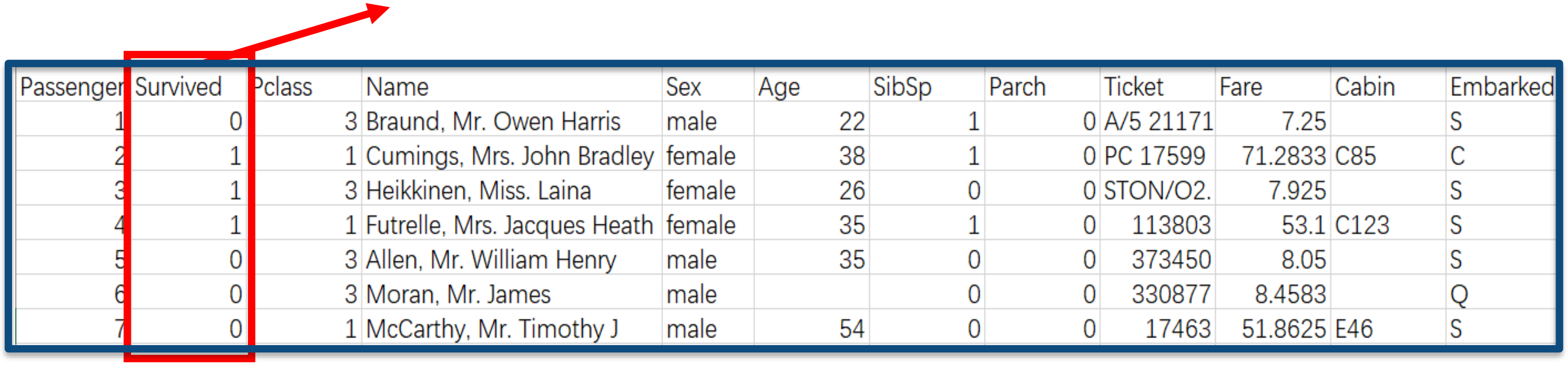
Balanced sampling

Concept understanding:

What are positive samples and negative samples?

For example, in Titanic data, samples with a target variable "survived" value of 1 (for survival) are positive samples, and samples with a target variable value of 0 (for death) are negative samples.

In the two classification problem, the positive rate (i.e. the proportion of positive samples) is usually an indicator that needs attention.



Target variable

Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S

Balanced sampling

What is an unbalanced sample like?

(1) Of 10000 samples, 5000 are positive and 5000 are negative

(2) Of the 10000 samples, only 500 are positive, and the rest are all negative

Can a good model be directly trained on unbalanced samples?

In the classical hypothesis of machine learning, it is often assumed that all kinds of training samples are equal, that is, the number of all kinds of samples is balanced, but the actual problems encountered in the real scene often do not meet this assumption. For example, credit default scenario, customer churn scenario...

Generally speaking, unbalanced samples will lead the training model to focus on the categories with a large number of samples, while "despise" the categories with a small number of samples, so the generalization ability of the model in the test data will be affected.

For example, there are 99 positive samples and 1 negative sample in the training set. In many cases without considering the imbalance of samples, the learning algorithm will make the classifier give up the negative case prediction, because it can get up to 99% training classification accuracy by dividing all samples into positive ones.

Balanced sampling

Over sampling	Over sampling is to achieve sample balance by increasing the data amount of small sized samples. Among them, the simpler way is to copy small samples directly to form a quantitative equilibrium.
Under sampling	Under sampling is to achieve sample balance by reducing the number of samples of most classes. The simple and direct method is to remove some data randomly to reduce the size of most class samples.

Balanced sampling

At first glance, over sampling and under sampling technologies seem to be equivalent in function. Both of them can change the sample size of the original dataset and achieve a balance of the same proportion. However, this common ground is only a superficial phenomenon. The two methods will have different negative effects of reducing the algorithm effect.

Under sampling

Deleting most class samples may lose important information about most classes.

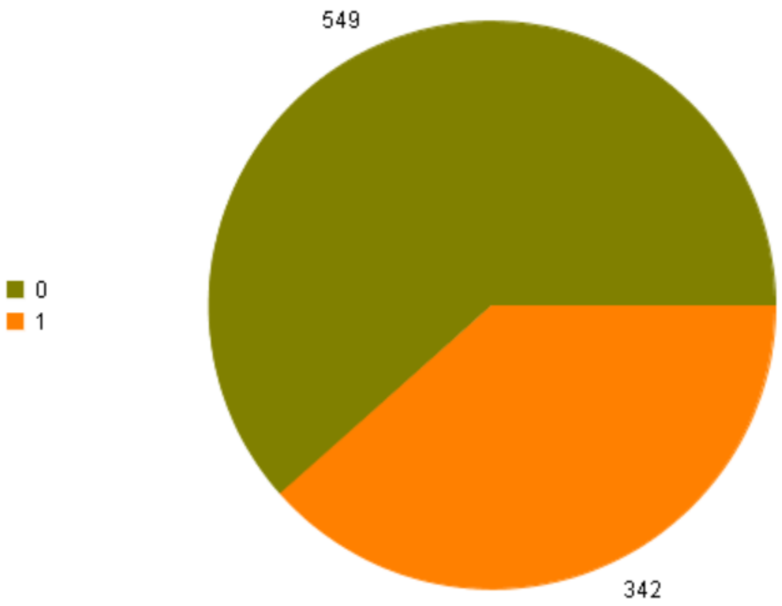
Over sampling

Multiple instances of some samples are "juxtaposed" and may be over fitted.

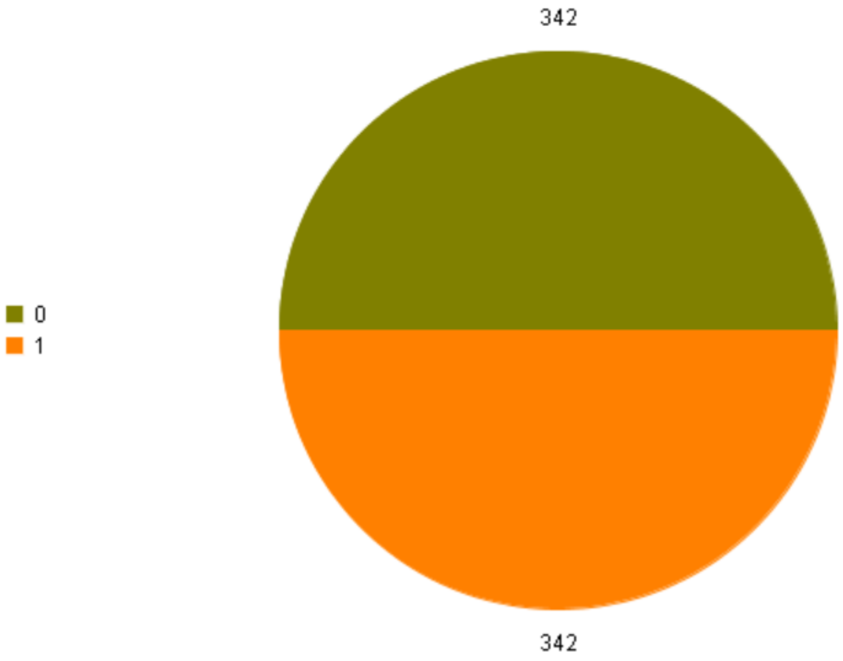
Class exercise - under sampling

The target variable in Titanic.csv is "survived", which is a binary variable. The number of positive samples (represented by 1) is 342, and the number of negative samples (represented by 0) is 549, totaling 891 samples.

After under sampling according to the ratio of 1:1, the positive sample is still 342, and the negative sample is reduced to 342.



Before preprocessing



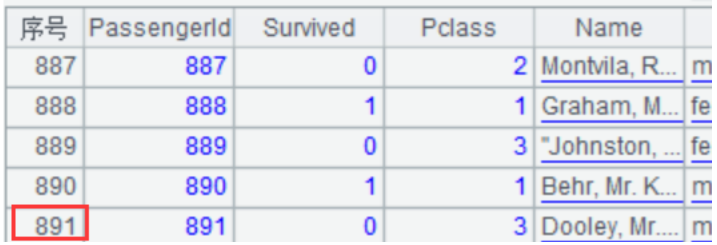
After preprocessing

Class exercise - under sampling

SPL code for under sampling:

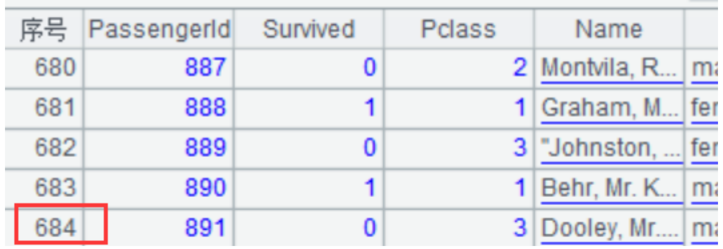
	A	B	C	D
1	D:/KDD/titanic.csv			/File directory
2	=file(A1).import@qtc()		1	/Data, minority to majority ratio
3	=A2.group@p(Survived)	=A3.sort(~.len())	=ceil(min(B3(2).len(),B3(1).len()*C2))	
4	=to(B3(2).len()).sort(rand())	=A4(to(C3)).sort()	=(B3(2)(B4)|B3(1)).sort()	/Ensure the order of samples is constant and carry out balanced sampling
5	=A2(C4)			

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	
887	887	0	2	Montvila, R...	m
888	888	1	1	Graham, M...	fe
889	889	0	3	"Johnston, ...	fe
890	890	1	1	Behr, Mr. K...	m
891	891	0	3	Dooley, Mr....	m

After preprocessing

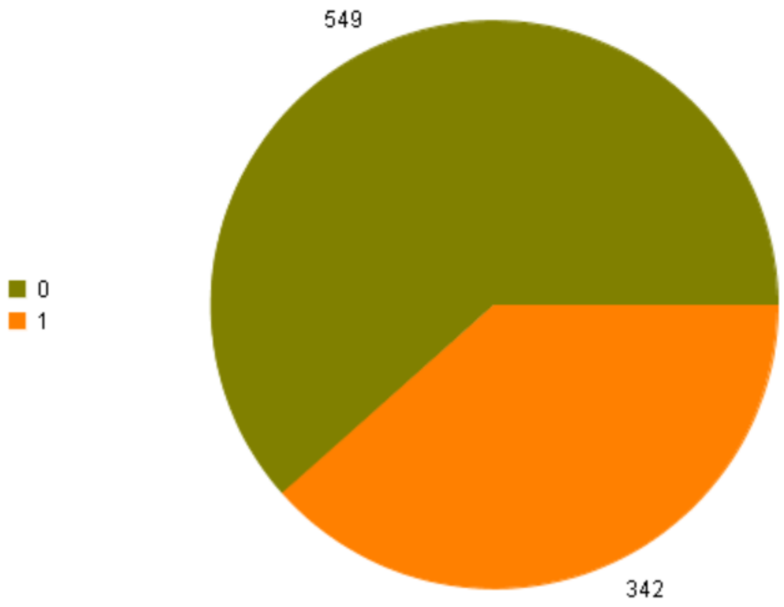


序号	PassengerId	Survived	Pclass	Name	
680	887	0	2	Montvila, R...	ma
681	888	1	1	Graham, M...	fer
682	889	0	3	"Johnston, ...	fer
683	890	1	1	Behr, Mr. K...	ma
684	891	0	3	Dooley, Mr....	ma

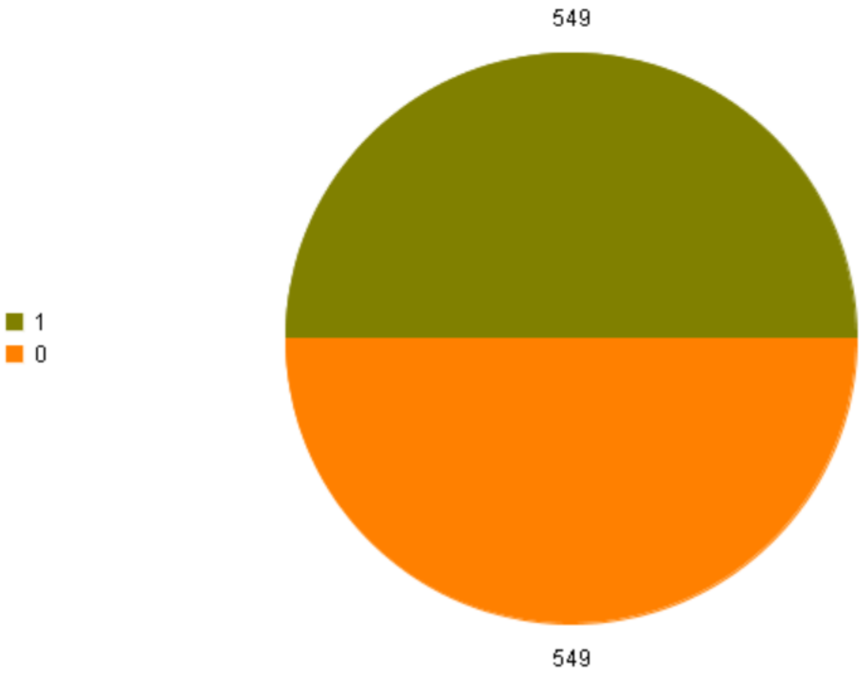
Class exercise - over sampling

The target variable in Titanic.csv is "survived", which is a binary variable. The number of positive samples (represented by 1) is 342, and the number of negative samples (represented by 0) is 549, totaling 891 samples.

After over sampling according to the ratio of 1:1, the positive sample is 549, and the negative sample is 549.



Before preprocessing



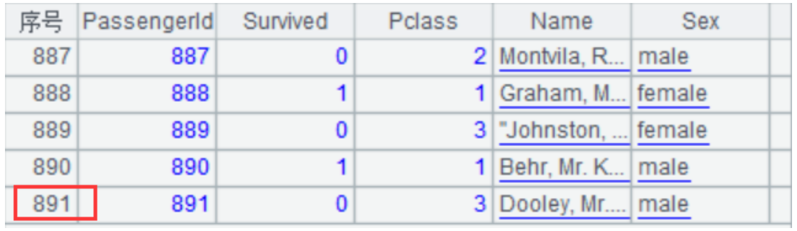
After preprocessing

Class exercise - over sampling

SPL code for over sampling:

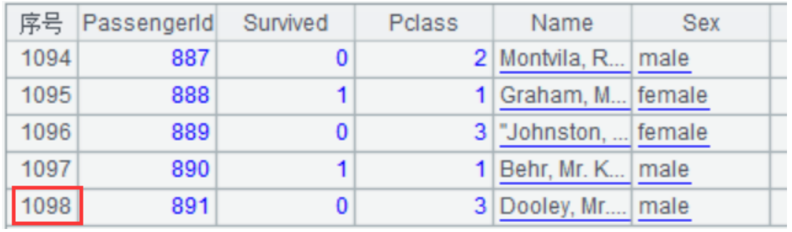
	A	B	C	D
1	D:/KDD/titanic.csv			/File directory
2	=file(A1).import@qtc()		1	/Data, minority to majority ratio
3	=A2.group@p(Survived)	=A3.sort(~.len())	=B3(2).len()/C2-B3(1).len()	/Ensure the order of samples is constant and carry out balanced sampling
4	=if(C3>0,C3,0)	=A4.(B3(1)(rand(B3(1).len())+1))	=(to(A2.len())|B4).sort()	
5	=A2(C4)			

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex
887	887	0	2	Montvila, R...	male
888	888	1	1	Graham, M...	female
889	889	0	3	"Johnston, ...	female
890	890	1	1	Behr, Mr. K...	male
891	891	0	3	Dooley, Mr....	male

After preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex
1094	887	0	2	Montvila, R...	male
1095	888	1	1	Graham, M...	female
1096	889	0	3	"Johnston, ...	female
1097	890	1	1	Behr, Mr. K...	male
1098	891	0	3	Dooley, Mr....	male

Data standardization

Why standardization

In some practical problems, the sample data we get are multi-dimensional, that is, a sample is characterized by multiple features.

For example, in the prediction of house price, the factors (features) that affect house price include house area, bedroom quantity, etc. obviously, the dimensions and magnitude of these features are different. When predicting house price, if the original data value is used directly, their influence on house price will be different. Through standardized processing, different features can be made have the same scale.

In short, when the scales (units) of features in different dimensions of the original data are inconsistent, standardized steps are needed to preprocess the data.

Data standardization - standardization (normalization) method

min-max normalization

(Min-Max Normalization)

It is a linear transformation of the original data to map the result value to [0 - 1].

Conversion function： $x^*=\frac{x-min}{max-min}$， max:max value of sample， min:min value of sample

One drawback of this method is that when new data is added, it may lead to changes in max and min, which need to be redefined.

Z-score normalization

(0-1 Normalization)

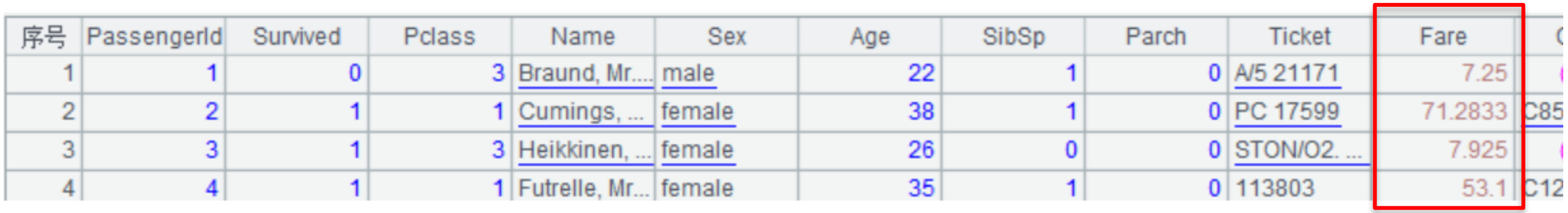
This method standardizes the mean and standard deviation of the original data. The processed data are in accordance with the standard normal distribution, i.e. the mean value is 0 and the standard deviation is 1.

Conversion function： $Z=\frac{x-\bar{x}}{s}$

Class exercise - Data Standardization

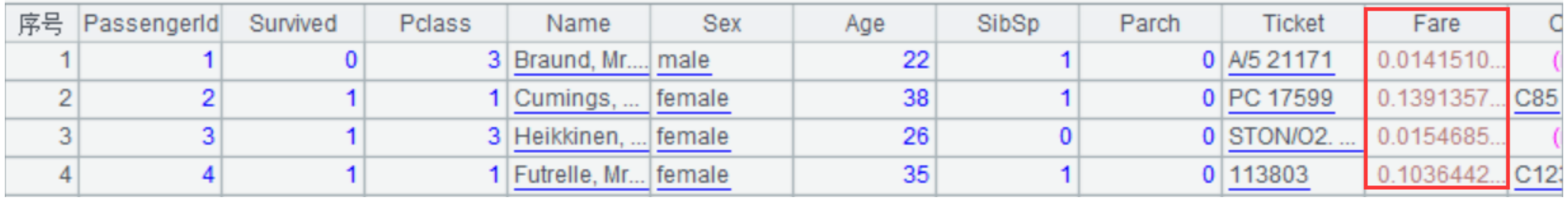
"Fare" in Titanic.csv is a numerical variable, and we standardize it with min max normalization.

As shown in the figure below, it can be seen that the result of standardization falls between [0,1]



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	7.25	(
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	71.2833	C85
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O2. ...	7.925	(
4	4	1	1	Futrelle, Mr...	female	35	1	0	113803	53.1	C12

Before preprocessing



序号	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
1	1	0	3	Braund, Mr....	male	22	1	0	A/5 21171	0.0141510...	(
2	2	1	1	Cumings, ...	female	38	1	0	PC 17599	0.1391357...	C85
3	3	1	3	Heikkinen, ...	female	26	0	0	STON/O2. ...	0.0154685...	(
4	4	1	1	Futrelle, Mr...	female	35	1	0	113803	0.1036442...	C12

After preprocessing

Class exercise - Data Standardization

SPL code for min-max normalization:

	A	B
1	D:/test/titanic.csv	/Define file path
2	=file(A1).import@qtc()	/Read data
3	=A2.max(Fare)	
4	=A2.min(Fare)	
5	=A2.run(Fare=(Fare-A4)/(A3-A4))	/Standardize fare variable

Dataset split

Machine learning requires datasets：

Train: used to train models

Test: used to test model effect

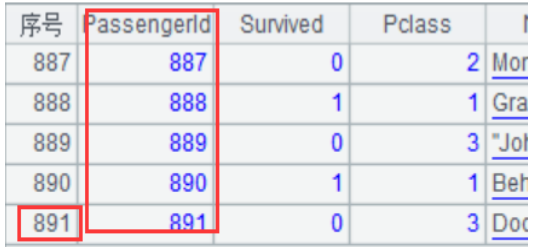
Generally, 70% of the data is divided into training set and 30% into test set (the division proportion is not fixed and can be set by yourself).

Class exercise – Divide dataset at random

Randomly divide training set and test set according to 7:3 ratio.

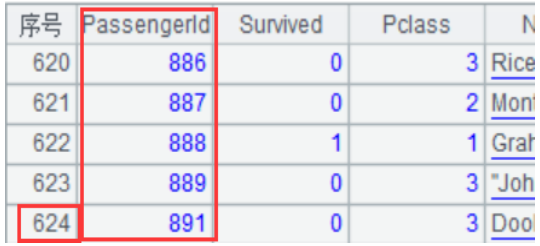
	A	B
1	D:/test/titanic.csv	/Define data path
2	=file(A1).import@qtc()	/Read data
3	=A2.group(rand()<=0.3)	/Randomly divided into two groups in proportion
4	=train=A3(1)	/Train set
5	=test=A3(2)	/Test set

Before division

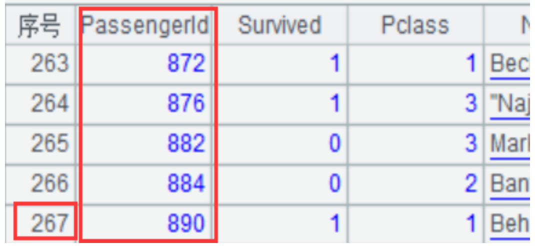


序号	PassengerId	Survived	Pclass	
887	887	0	2	Mor
888	888	1	1	Gra
889	889	0	3	"Jol
890	890	1	1	Beh
891	891	0	3	Doc

After division



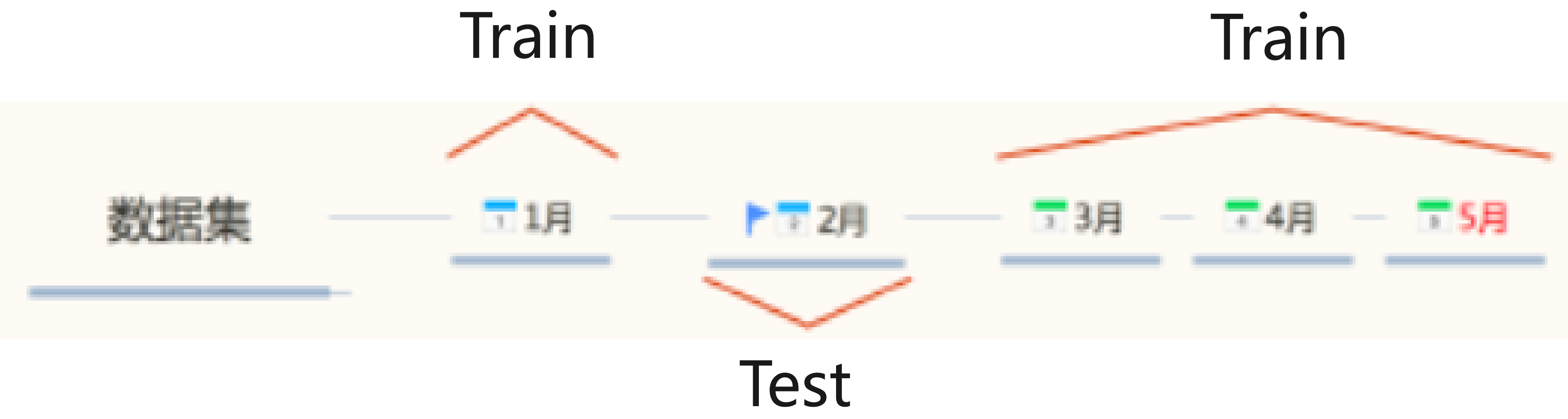
序号	PassengerId	Survived	Pclass	N
620	886	0	3	Rice
621	887	0	2	Mon
622	888	1	1	Grah
623	889	0	3	"Joh
624	891	0	3	Dool



序号	PassengerId	Survived	Pclass	N
263	872	1	1	Bec
264	876	1	3	"Naj
265	882	0	3	Marl
266	884	0	2	Ban
267	890	1	1	Beh

Dataset split

But randomly dividing dataset can sometimes cause errors.



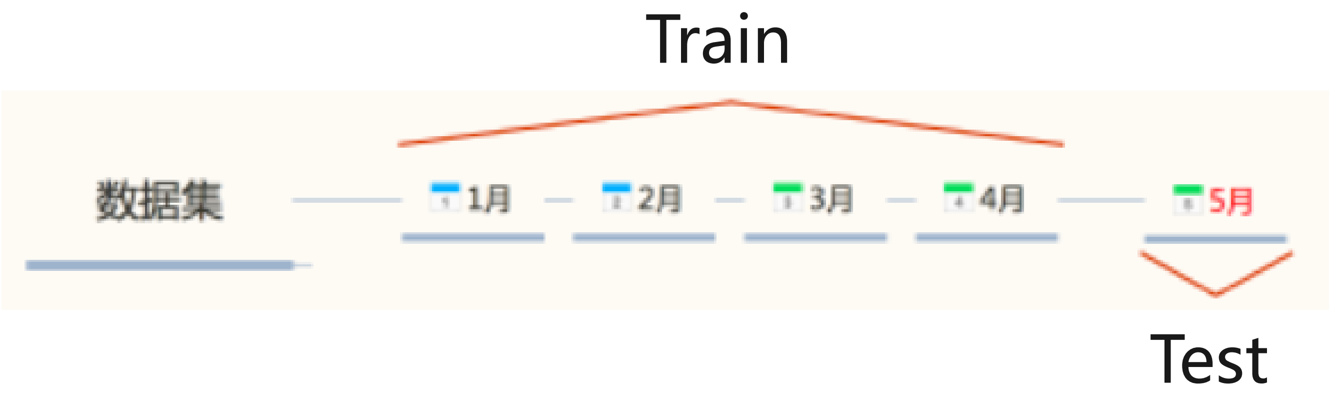
As shown in the figure, the data characteristics of March, April and May may include the data of February.

This is to use the training data to test.

Not in line with the concept of machine learning, we hope to make predictions on unknown dataset.

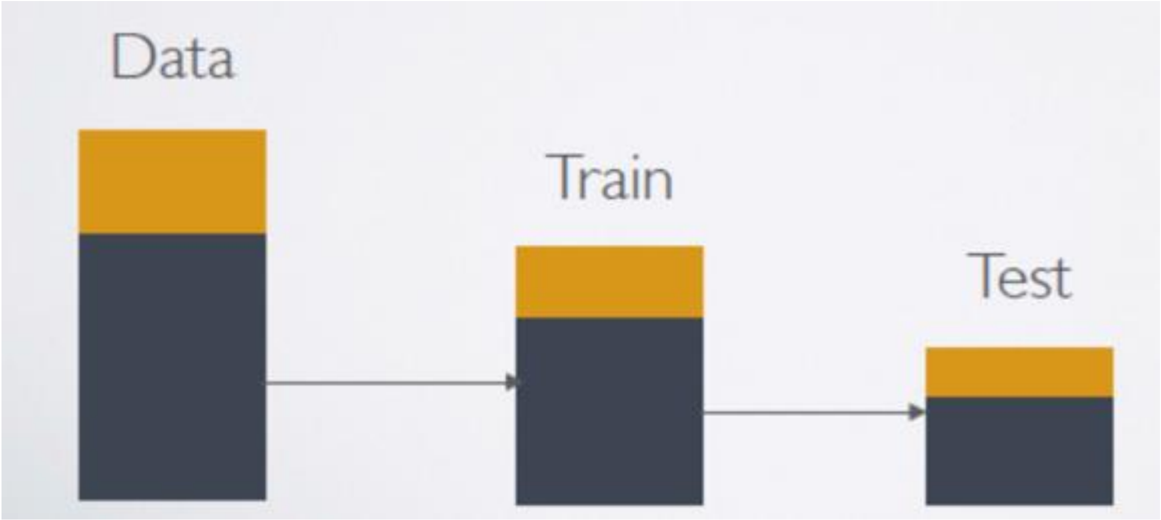
Dataset split

Divided by time dimension:



Training with historical known data and testing with future unknown data.

stratified split：



The proportion of Class 0 / 1 (target variable) in training set and test set is nearly the same.

Glossary - for reference

data preparation
missing value
discretization
derived variables
redundant variables
missing rate
outlier
logistic regression
AdaBoost, adaptive boosting
MCAR, Missing Completely at Random
MAR，Missing at Random
MNAR，Missing not at Random
unbiasedness
hotdecking
regression
linear correlation
expectation
MLE, maximum likelihood estimation
Posterior distribution
XGBoost(eXtreme Gradient Boosting)

Glossary - for reference

LGB(light gradient boosting machine)
low frequency variable
enumerate
one-hot-encoder
mapping
training set
testing set
normal distribution
positive sample
negative sample
positive rate

unbalanced sample
generalization ability
over sampling
under sampling
standardization/ normalization
scale
min-max normalization
z-score normalization

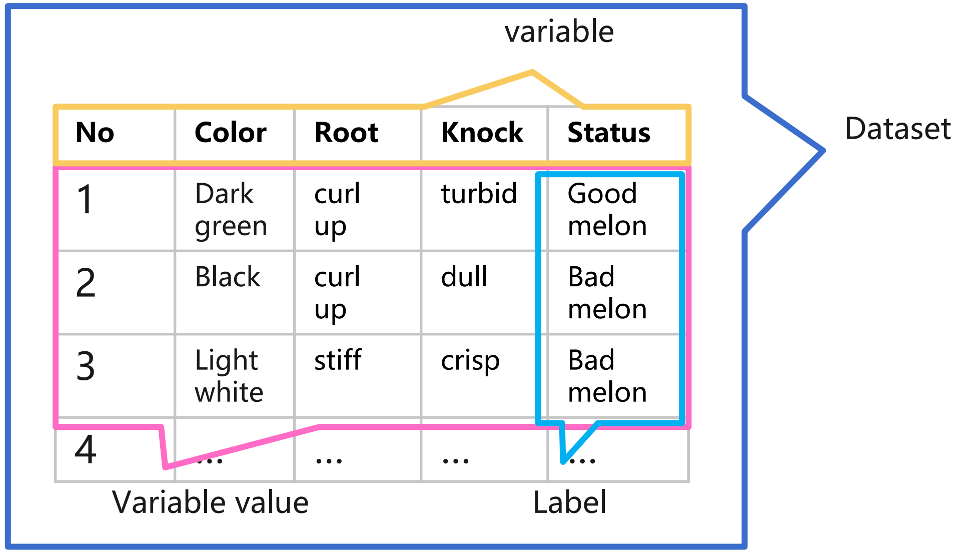
Chapter 4 Modeling

Supervised learning

Supervised learning: in supervised learning, the training data has both features and label (target variable). Through training, the machine can find the relationship between features and label by itself. When facing the data with only features but no label, it can judge the label.

For example, the accuracy of the exercises with standard answers and then to the exam is higher than that of the exercises without answers and then to the exam

Another example: when we were young, we didn't know whether cattle and birds belonged to the same category, but as we grew up, we kept inputting all kinds of knowledge, and the models in our brains became more and more accurate, and the judgment of animals became more and more accurate.

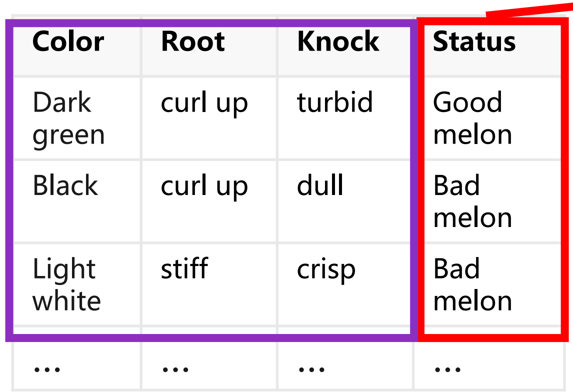


No	Color	Root	Knock	Status
1	Dark green	curl up	turbid	Good melon
2	Black	curl up	dull	Bad melon
3	Light white	stiff	crisp	Bad melon
4	…	…	…	…

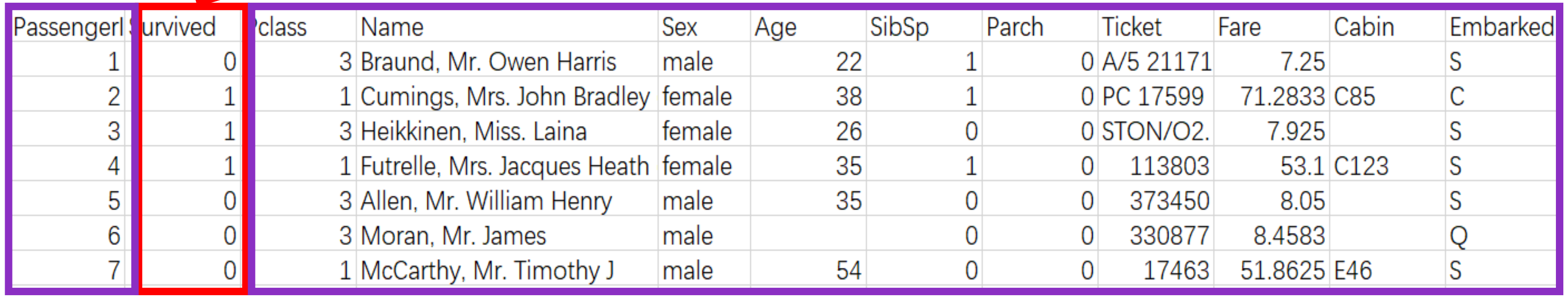
Supervised learning

Classification and regression are two major problems that can be solved by supervised learning. From the perspective of the type of prediction value, the quantitative output of continuous variable prediction is called regression; the qualitative output of discrete variable prediction is called classification. For example, predicting the degree of tomorrow is a regression task; predicting whether tomorrow is cloudy, sunny or rainy is a classification task.

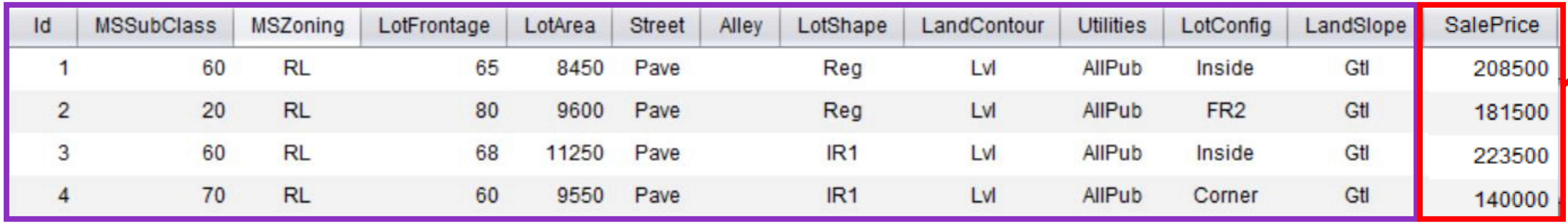
Classification task



Color	Root	Knock	Status
Dark green	curl up	turbid	Good melon
Black	curl up	dull	Bad melon
Light white	stiff	crisp	Bad melon
…	…	…	…



PassengerI	urvived	class	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S



Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	SalePrice
1	60	RL	65	8450	Pave		Reg	Lvl	AllPub	Inside	Gtl	208500
2	20	RL	80	9600	Pave		Reg	Lvl	AllPub	FR2	Gtl	181500
3	60	RL	68	11250	Pave		IR1	Lvl	AllPub	Inside	Gtl	223500
4	70	RL	60	9550	Pave		IR1	Lvl	AllPub	Corner	Gtl	140000

Regression task

Class exercise

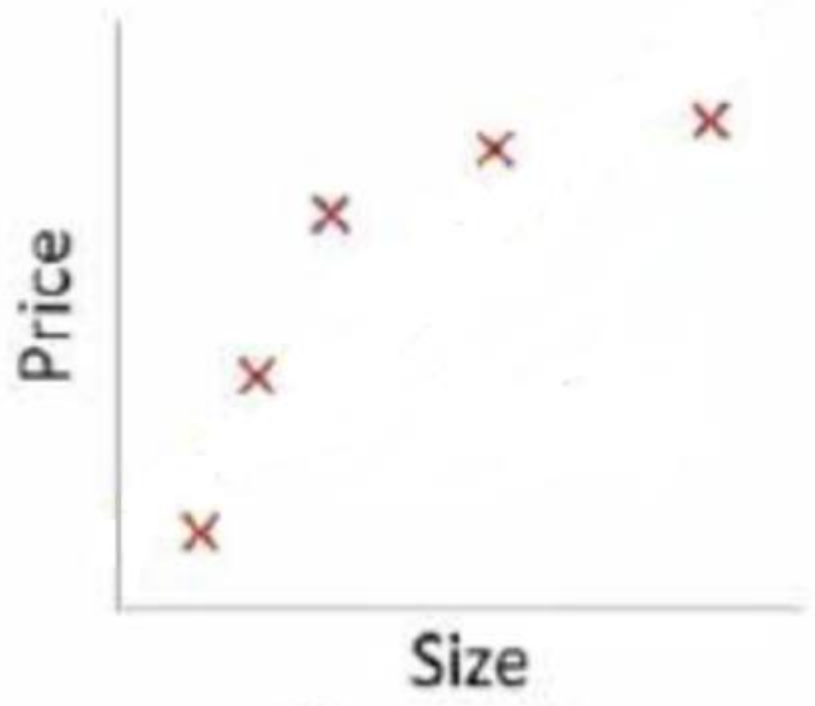
1. **Predict sales volume**： the dataset has different product properties, usages, main user characteristics and historical sales volume. Establish a model to predict its future sales volume.
2. **Estimate the nature of the tumor**： the dataset includes the patient's gender, age , tumor size, location, ... label of benign or malignant tumor. Use it to establish a model to judge the nature of the tumor of the new patient.
3. **Purchase prediction**： the dataset contains product information, user information and historical purchase users. Build a model to predict which customers will purchase products.
4. **Default prediction**： the dataset contains basic information of loan users, historical loan situation, credit information, etc. Establish a model to determine which users will default .

Thinking： Which are classification problems and which are regression problems?

Common concepts

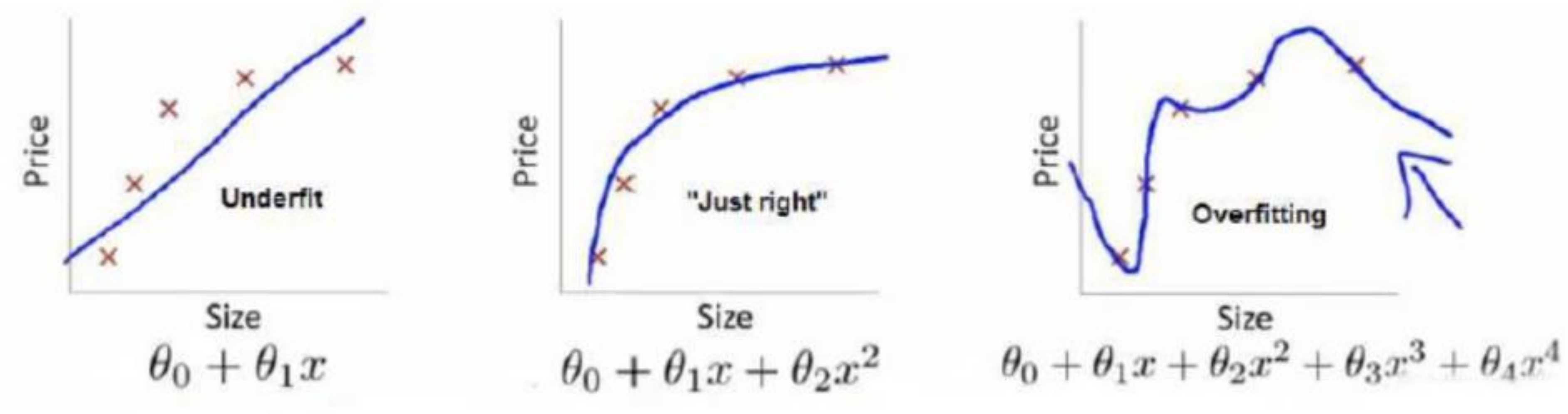
About machine learning, we often hear some concepts, such as objective function, loss function, over fitting. What do they mean?   Let's use an example to explain:

There are two sets of data. The abscissa size represents the size of the house, and the ordinate price represents the sales price of the house. A model needs to be built to represent the relationship between the two.



Picture from Andrew Ng machine learning open class

Loss function

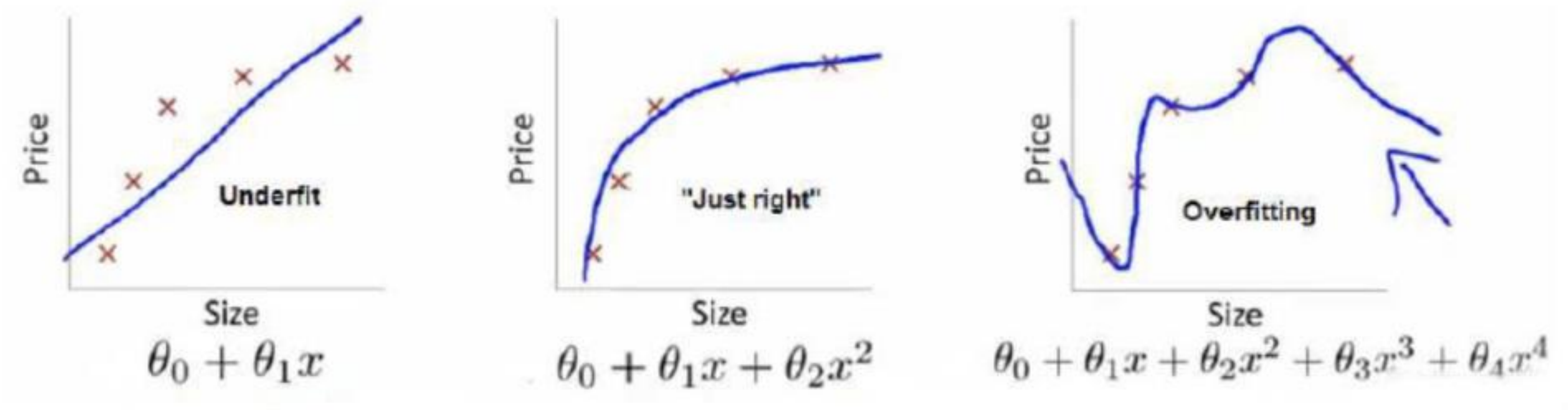


The functions of the above three graphs are in turn $f_1\ (X)$, $f_2\ (X)$, $f_3\ (X)$. We use these three functions to fit price respectively. The real value of price is recorded as Y.

Given X, all three functions will output a $f(X)$, the $f(X)$ of this output may be the same as or different from the real value Y. In order to express the quality of our fitting, we use a function to measure the degree of fitting.

This function is called loss function, also known as cost function. The smaller the loss function, the better the model fitting.

Over fitting and under fitting



Let's see the picture above, on the far right $f_3\ (X)$ is the best fit for historical data, so the loss function is the smallest.

But let's see from the picture, $f_3\ (X)$ is certainly not the best, because it over learns historical data, leading to its poor effect in real prediction, this situation is called over fitting. On the contrary, $f_1\ (X)$ the poor fitting of historical data is called under fitting.

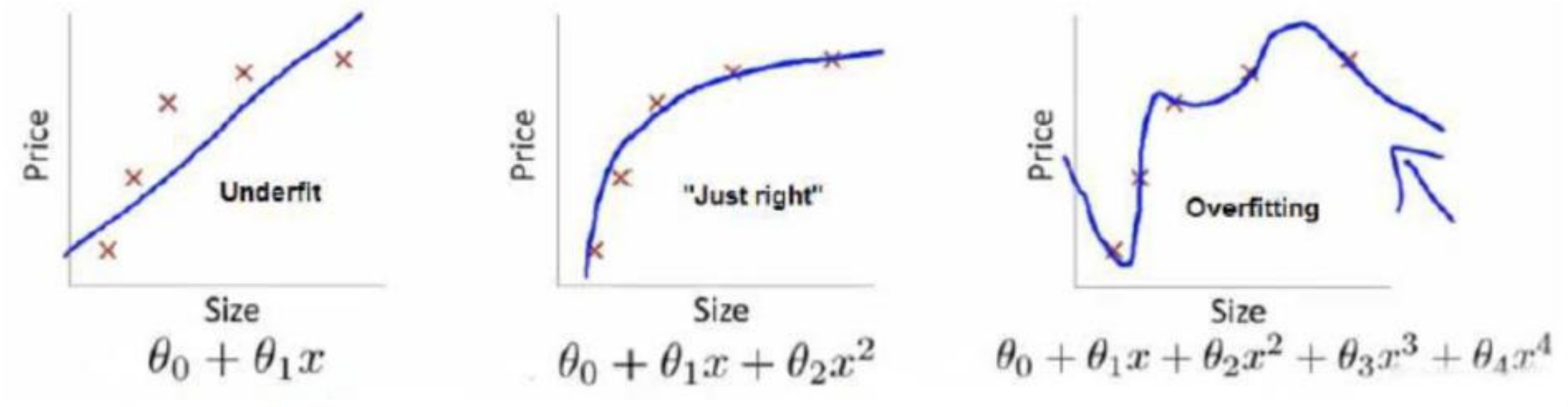
Regularization

Why over fitting?

In short, its function is too complex. It even has the fourth power, which leads to the following concept:
A function is defined to measure the complexity of the model. It is also called regularization in machine learning. The commonly used regularization functions are L_1 and L_2 .

At this stage, we can say that our final optimization function is:
Loss function + regularization function
This function is called the **objective function**.

Objective function



Combine the above examples to analyze:

On the far left $f_1(X)$, the model structure is the simplest, but the fitting of historical data is the worst;

On the far right $f_3(X)$, the best fit for historical data, but model structure is the most complex;

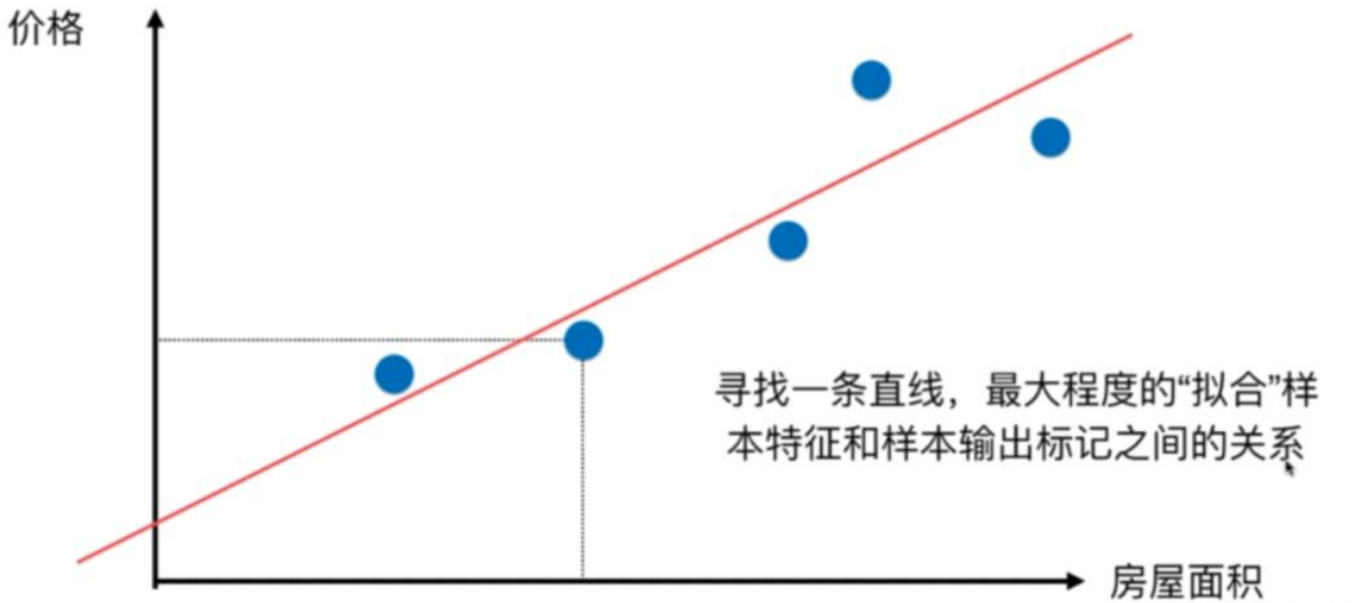
$f_2(X)$ achieves a good balance between the two and is most suitable for predicting unknown dataset.

Linear regression

A simple example is the area of the house and the price of the house. By learning the sample points, we can find a straight line y=ax+b that best fits all the sample points, so that each time we get a new sample into the model, we can give the predicted y value.

If there is only one eigenvalue like this, it is called simple linear regression.

Multiple eigenvalues are called multiple linear regression.



Linear regression

Advantage	Solve the problem of regression; The idea is simple and easy to realize; The basis of many powerful nonlinear models; The results are well explainable; Contains many important ideas in machine learning.

Lasso, ridge and elastic net

Lasso regression is to add L1 regularization on the basis of standard linear regression.

Ridge regression is based on standard linear regression with L2 regularization.

Elastic net is a combination of L1 and L2 regularization.

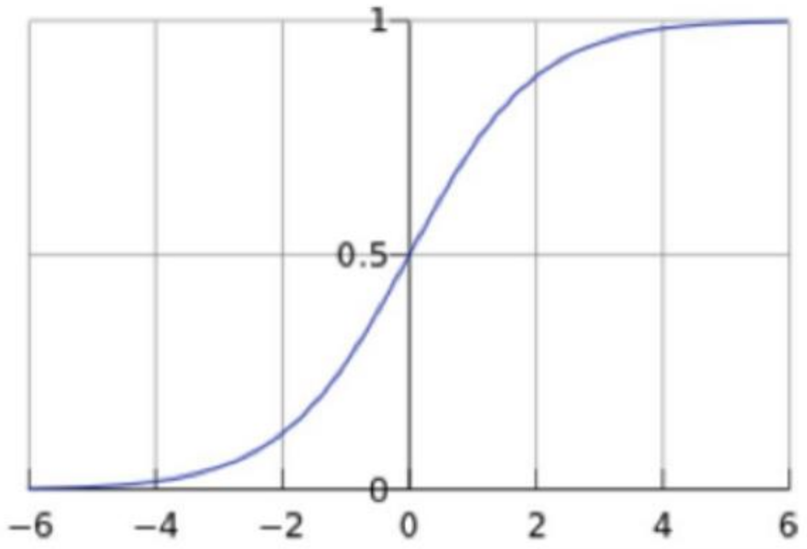
Use scenarios of Lasso, ridge and elastic net

Similarity	Can be used to solve the over fitting problem of standard linear regression.
Difference	• As long as the data is linearly correlated, it is not well fitted with linear regression and needs to be regularized, so ridge regression (L2) can be considered. • For the data with high dimension (many feature variables), especially the linear relationship is sparse (most of the data are missing or zero), L1 regularization (lasso regression) is used, or to find out the main features in a pile of features, so L1 regularization (lasso regression) is preferred. • When we find that lasso regression is too much (too many features are sparse to 0), and ridge regression is not enough (regression coefficient decay is too slow), we can consider using elastic net to get better result.

Logistic Regression

Logistic regression, LR for short. Although regression appears in the name of the algorithm, it is actually a classification algorithm, and the output value is always between 0-1.

Logistic regression and linear regression are both linear models. The biggest difference between them is the data type of Y (target variable). Y of linear regression belongs to quantitative data, while Y of logistic regression belongs to classified data.



Logistic Regression

Advantage	Low computational cost, easy to understand and implement. LR directly models the possibility of classification without assuming the distribution of data in advance, which avoids the problems caused by inaccurate distribution of data. LR can output in the form of probability instead of knowledge 0,1 decision, which is very useful for many tasks using probability to assist decision making.
Disadvantage	Cannot use logistic regression to solve nonlinear problems. It is easy to under fit. In most cases, it is necessary to carry out feature engineering manually to build composite features, and the classification accuracy is not high.

Logistic Regression

Applicable scenarios

LR is the basic component of many classification algorithms. Its advantage is that the output value naturally falls between 0 and 1, and has probability significance. It is essentially a linear classifier, which can not deal with the correlation between features.

Although the effect is general, the model is clear, and the probability behind it can stand the scrutiny.

The fitted parameters represent the influence of each feature on the result, and it is also a good tool to understand the data.

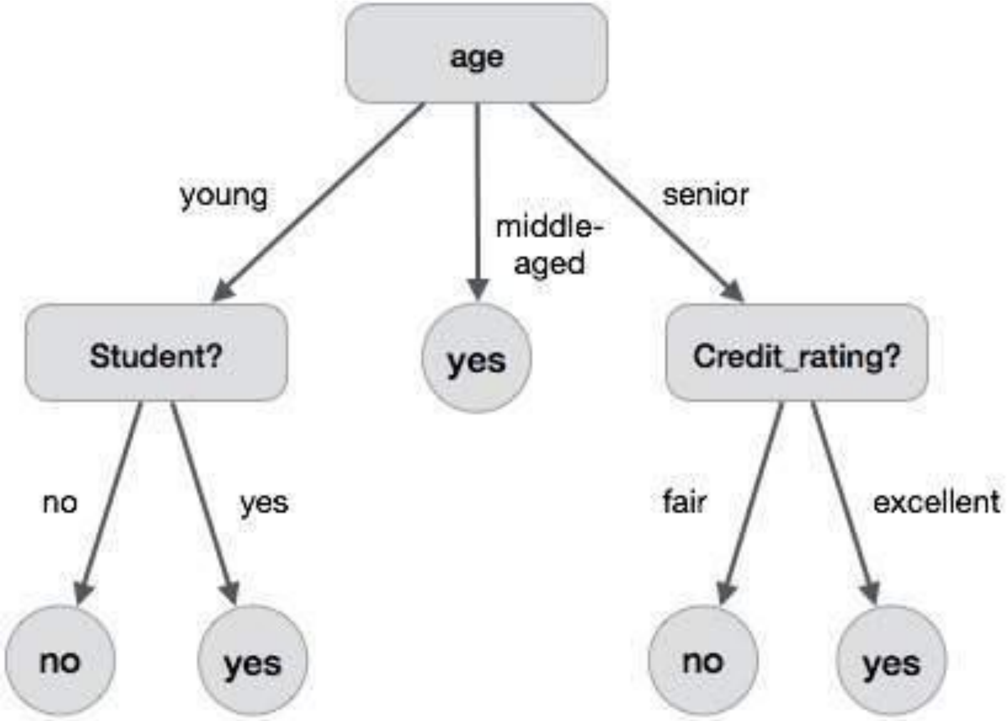
Decision tree

The decision tree can be seen as a set of if then rules, that is, a rule is constructed for each path from the root node of the decision tree to the leaf node. The characteristics of the internal nodes on the path correspond to the conditions of the rule, and the classes of the leaf nodes correspond to the conclusions of the rule.

Decision tree is a kind of nonparametric learning supervision method for classification or regression.

Decision tree learning algorithm includes feature selection, decision tree generation and decision tree pruning process.

At present, the most popular algorithm of decision tree is CART, which can generate both classification tree and regression tree.



Decision tree

Advantage

- The decision tree is simple and intuitive.
- Basically, there is no need for preprocessing, normalization in advance, and missing value processing.
- It can handle both discrete and continuous values. Many algorithms only focus on discrete or continuous values.
- It can handle the classification of multi-dimensional output.
- Compared with the black box classification model such as neural network, the decision tree can be logically well explained.
- Cross validation pruning can be used to select models to improve generalization ability.
- It has good fault tolerance and robustness for abnormal points.

Decision tree

Disadvantage

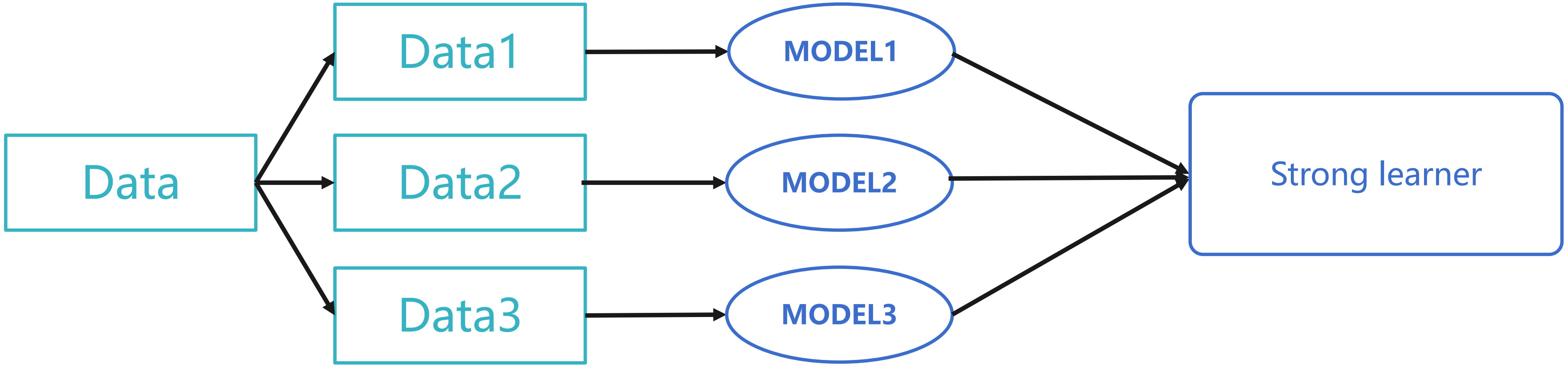
- The decision tree algorithm is very easy to over fit, resulting in poor generalization ability. It can be improved by setting the minimum number of samples and limiting the depth of decision tree.
- A little change of the sample will lead to the drastic change of the tree structure. It can be solved through ensemble learning and other methods.
- It is difficult to find the optimal decision tree. Generally, it is easy to fall into local optimum by heuristic method. It can be improved by ensemble learning and other methods.
- In some complex relationships, decision trees are difficult to learn, such as XOR.
- If the sample proportion of some features is too large, the decision tree is easy to lean to these features. It can be improved by adjusting the weight of samples.

Ensemble learning

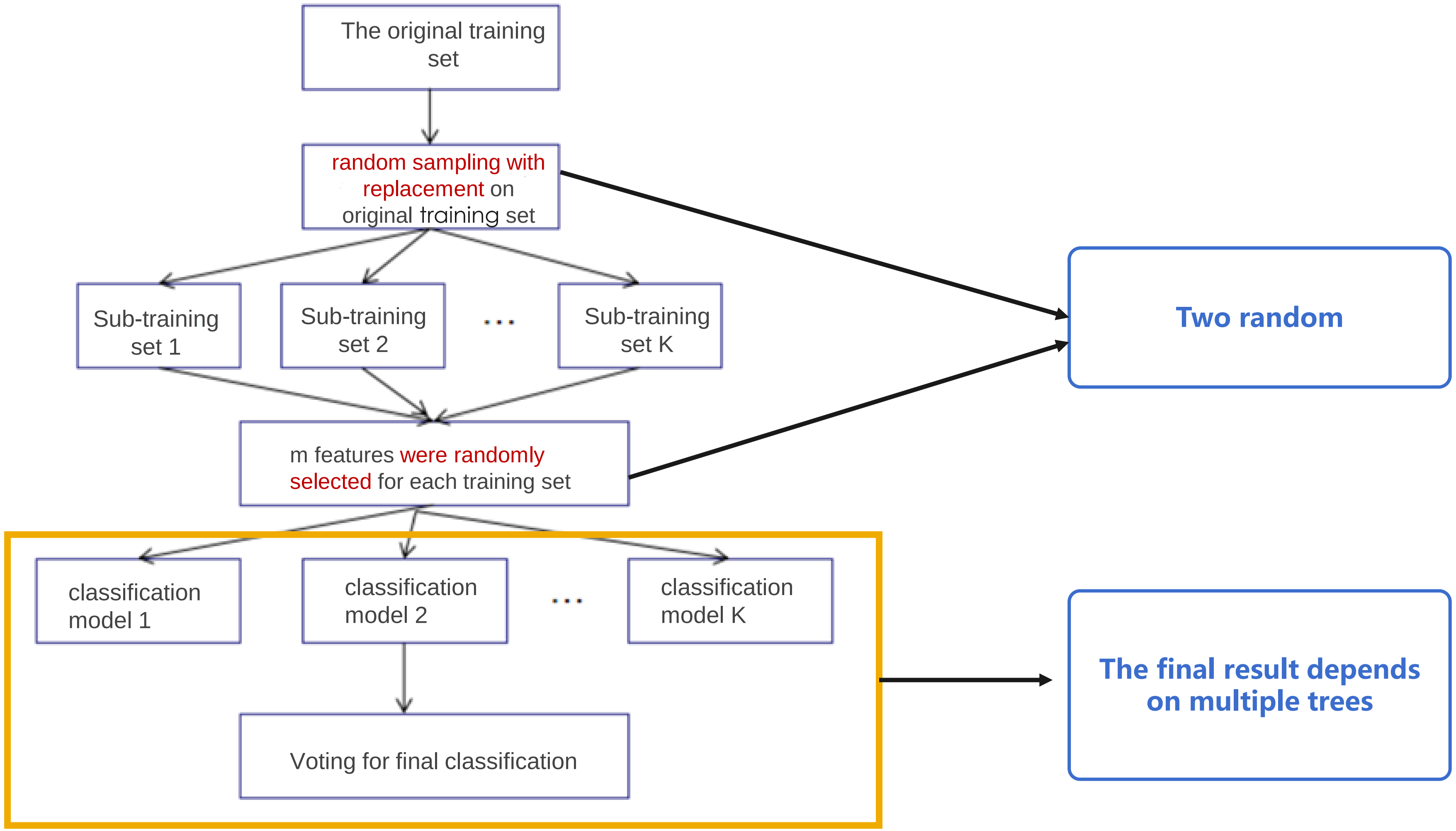
The purpose of ensemble learning is to combine the prediction values of multiple basic models according to an algorithm, so as to improve the generalization ability of the model.

There are two main methods:

- Averaging (bagging) method: build multiple models independently and average their prediction results, for example, random forest.
- Boosting method: establish multiple base models with sequence and dependency, and the latter model is used to modify the previous model's bias, such as GBDT, XGBoost , AdaBoost



Random forest



The original training set

random sampling with replacement on original training set

Sub-training set 1

Sub-training set 2

…

Sub-training set K

m features were randomly selected for each training set

classification model 1

classification model 2

…

classification model K

Voting for final classification

Two random

The final result depends on multiple trees

Random forest training process

Random forest

Random forest is divided into "random" and "forest".

"Forest" is composed of many trees, so the result of random forest depends on the results of many decision trees, which is an ensemble learning idea. The CART decision tree is used as a weak learner for random forest.

Classification algorithm: the maximum number of votes of K weak classifiers is sample category.
Regression algorithm: arithmetic mean of prediction results of K weak classifiers as final output.

There are two meanings of "**random**", one is to randomly select samples, the other is to randomly select features. For each tree, there are randomly selected training samples that are put back, and then there are randomly selected M features that are put back as the basis for branching this tree.

The introduction of two randomness is very important to the classification performance of random forest. Because of their introduction, random forest is not easy to fall into over fitting.

Random forest

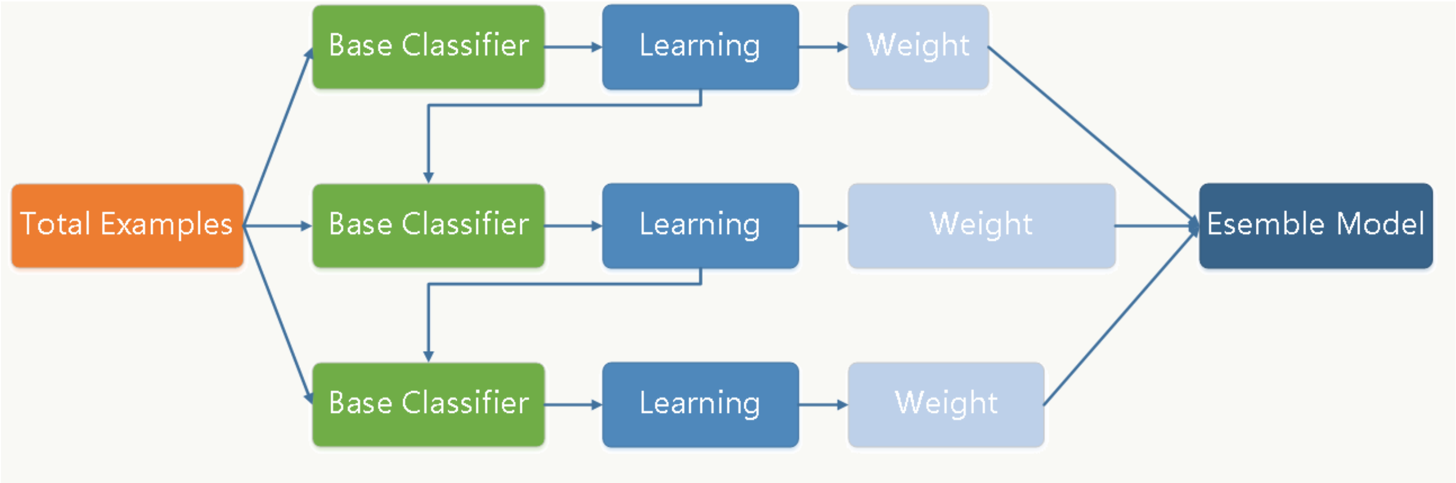
Advantage	It can process high-dimensional data without feature selection (feature subset is randomly selected). The generalization ability of the model is strong. The training speed of the model is fast and parallel, that is, the trees are independent of each other. The model can deal with unbalanced data and balance errors. The final training result can rank the special amount and select the more important features. Random forest has out of bag data (OOB), so there is no need to separate cross validation set. The accuracy of the model training result is high.
Disadvantage	When the data noise is large, there will be over fitting. For the data with different values, the attributes with more values will have a greater impact on the random forest.

GBDT

GBDT (Gradient Boosting Decision Tree), also known as GBM, can be used for classification or regression. It is a kind of boosting algorithm. It uses the error of the previous weak learner to update the sample weight value, and then iterations. GBDT requires that the weak learner must be CART model.

GBDT calculates every time to reduce the last residual. The next model mainly establishes the model in the gradient direction of the residual reduction, so that the residual decreases in the gradient direction. More vividly, it can be understood from the literal meaning, just like climbing a mountain. Every step goes in the ascending direction of the hillside, and always reaches or approaches the top of the mountain.

Training process of GBDT



GBDT

The core of GBDT is that every tree learns the residual of all previous tree conclusions, and the residual is the accumulation that can get the real value after adding the predicted value.

For example, the real age of A is 18 years old, but the predicted age of the first tree is 12 years old, the difference is 6 years, that is, the residual is 6 years;
Then in the second tree, we set the age of A as 6 years old to learn. If the second tree can really divide A into 6-year-old leaf nodes, the result of accumulating two trees is the real age of A;
If the conclusion of the second tree is 5 years old, then A still has a 1-year residual, and the age of A in the third tree becomes 1 year old, so continue to learn.

GBDT

Advantage

1) Further improvement based on RF
2) It can handle all kinds of data flexibly, including continuous value and discrete value.
3) In the case of a relatively small parameter adjustment time, the prediction accuracy can also be relatively high.

Disadvantage

Because it is boosting, it is difficult to train data in parallel because there is a serial relationship between the basic learners.

XGBoost

The full name of XGBoost is eXtreme Gradient Boosting, which was proposed by Dr. Chen Tianqi of Washington University. Because of its outstanding efficiency and high prediction accuracy, XGBoost has attracted wide attention.

Only in 2015, 17 of the 29 algorithms that won in the kaggle [2] competition used XGBoost library. As a comparison, the number of deep neural network methods that have been popular in recent years is 11. In KDDCUP 2015 [3] competition, the top ten teams all use XGBoost library.

XGBoost is an improvement of boosting algorithm based on GBDT, such as:
Traditional GBDT uses CART as base classifier, XGBoost also supports linear classifier
Traditional GBDT only uses first derivative information in optimization, XGBoost uses second-order Taylor expansion for cost function, and the definition of loss function is more accurate

……

Deep learning

Deep learning, or DL for short, is a branch of machine learning, and also one of the most popular machine learning at present. The concept of deep learning originates from the research of artificial neural network, and artificial neural network (ANN) abstracts the neural network of human brain from the perspective of information processing, establishes a simple model, and forms different networks according to different connection ways, which is called neural network or neural network like for short. Therefore, deep learning is also called deep neural networks(DNN).

Like machine learning, deep machine learning can be divided into supervised learning and unsupervised learning.
For example, convolutional neural networks (CNN) is a kind of machine learning model under deep supervised learning, while deep belief networks (DBNs) is a kind of machine learning model under unsupervised learning.

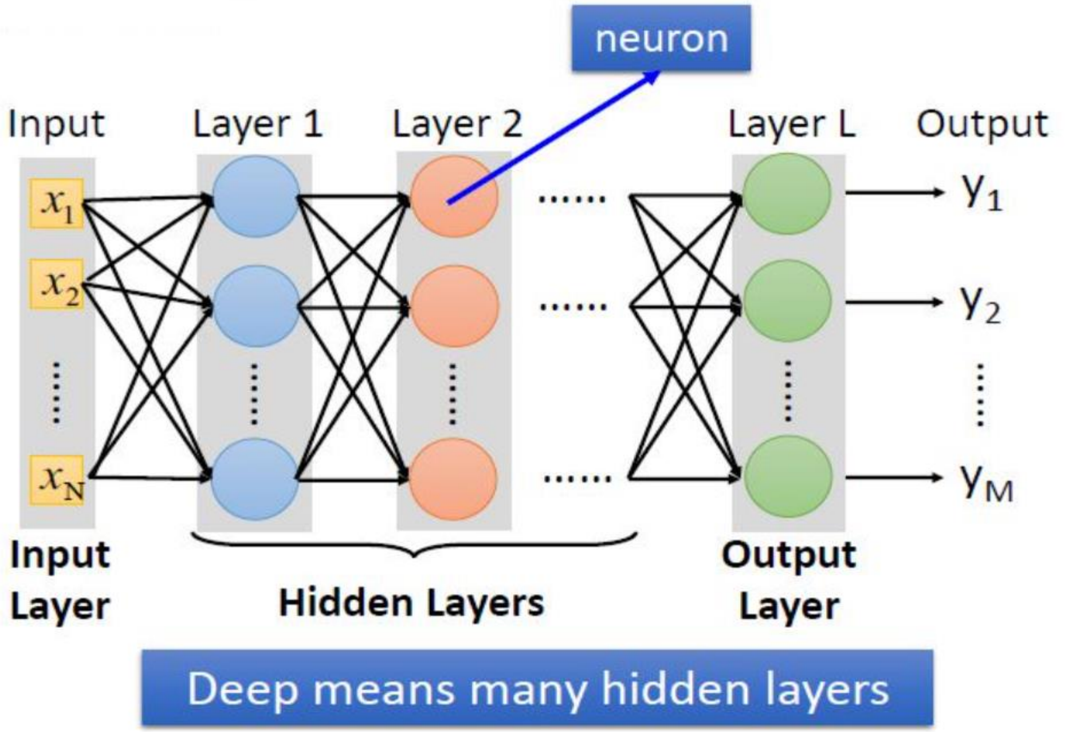
Deep learning

"Depth" of deep learning refers to the number of layers experienced from "input layer" to "output layer", that is, the number of layers of "hidden layer". The more layers, the deeper the depth.

So the more complex the selection problem, the more depth and layers are needed. In addition to the number of layers, there are more "neurons" - small circles in each layer. For example, Alphago's network is 13 layers, each layer has 192 neurons.

The essence of deep learning is to learn more useful features by building machine learning models with many hidden layers and massive training data, so as to ultimately improve the accuracy of classification or prediction.

Therefore, "deep model" is the means and "feature learning" is the purpose.



Deep means many hidden layers

Deep learning

Advantage	Deep learning puts forward a method to let computer learn pattern features automatically, and integrates feature learning into the process of building model, so as to reduce the imperfection caused by artificial design features.
Disadvantage	In the case of limited data, deep learning algorithm can not estimate the law of data without deviation. In order to achieve good accuracy, big data support is needed. In order to ensure the real-time performance of the algorithm, we need more parallel programming skills and better hardware support.
Application fields	Computer vision, speech recognition, memory network, natural language processing and other fields.

Automatic modeling

What we mentioned above is only a small part of the common data mining algorithms, and there are more and more complex algorithms in the industry. The principles of these algorithms are different, parameters are different, application scenarios are different, and the requirements for data preprocessing are different. The actual use process is not simple, and the requirements for modelers are very high. Generally, professional statistical background is required to complete these algorithms.

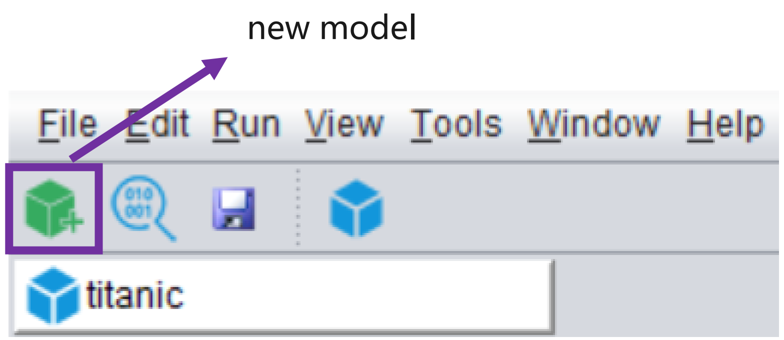
Therefore, it is difficult for us to arrange direct exercises of these algorithms in the course.

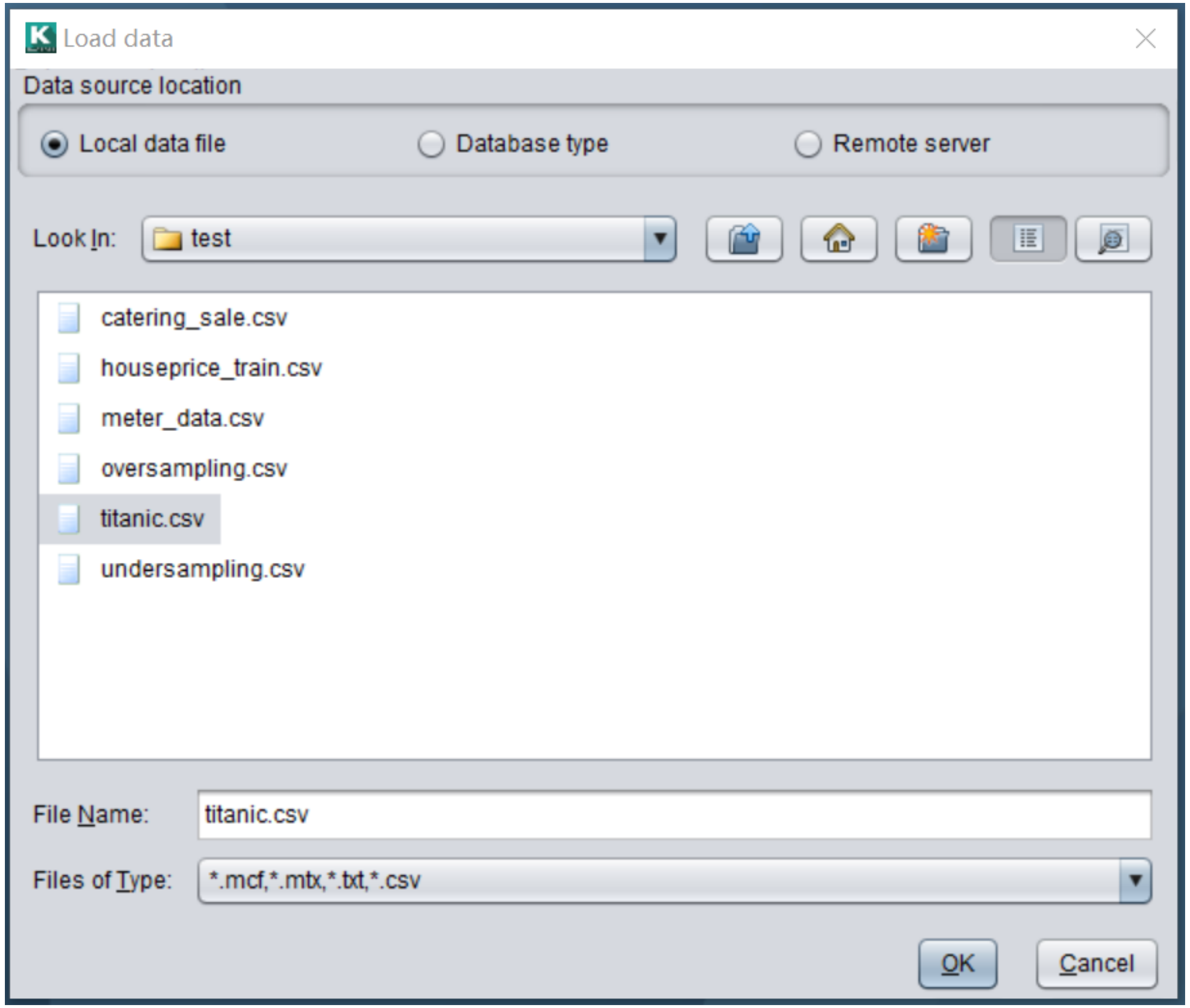
Fortunately, in recent years, the concept of automatic modeling has emerged, that is, integrating the rich experience of the experts into the software, so that it can automatically complete the actions of data preprocessing, model algorithm and parameter selection.

Now let's use " Titanic.csv ” (classification) and “ Houseprice.csv ” (regression) data as an example, using Raqsoft YModel tool to try the process of automatic modeling.

Class exercise - Classification Model

Click "new model", select Titanic data, click "OK" to import data.





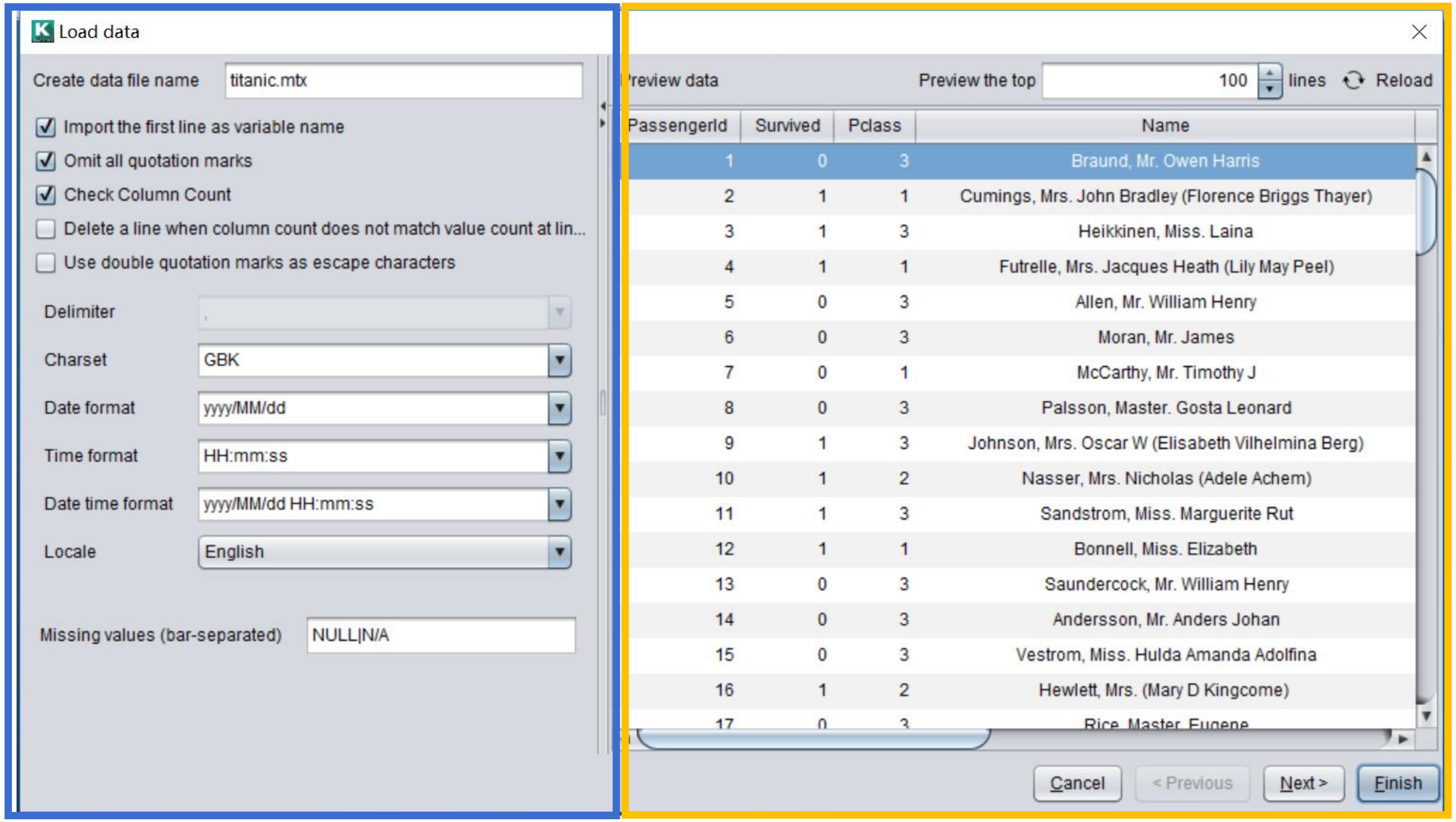
File Name: titanic.csv

Files of Type: *.mcf,*.mtx,*.txt,*.csv

Class exercise - Classification Model

Data and variables can be previewed on the right side of the page

The left side of the page is configured with character set format, date time format and missing value format for automatic recognition by software.



Create data file name: titanic.mtx

☑ Import the first line as variable name
☑ Omit all quotation marks
☑ Check Column Count
☐ Delete a line when column count does not match value count at lin...
☐ Use double quotation marks as escape characters

Delimiter: ,
Charset: GBK
Date format: yyyy/MM/dd
Time format: HH:mm:ss
Date time format: yyyy/MM/dd HH:mm:ss
Locale: English

Missing values (bar-separated): NULL|N/A

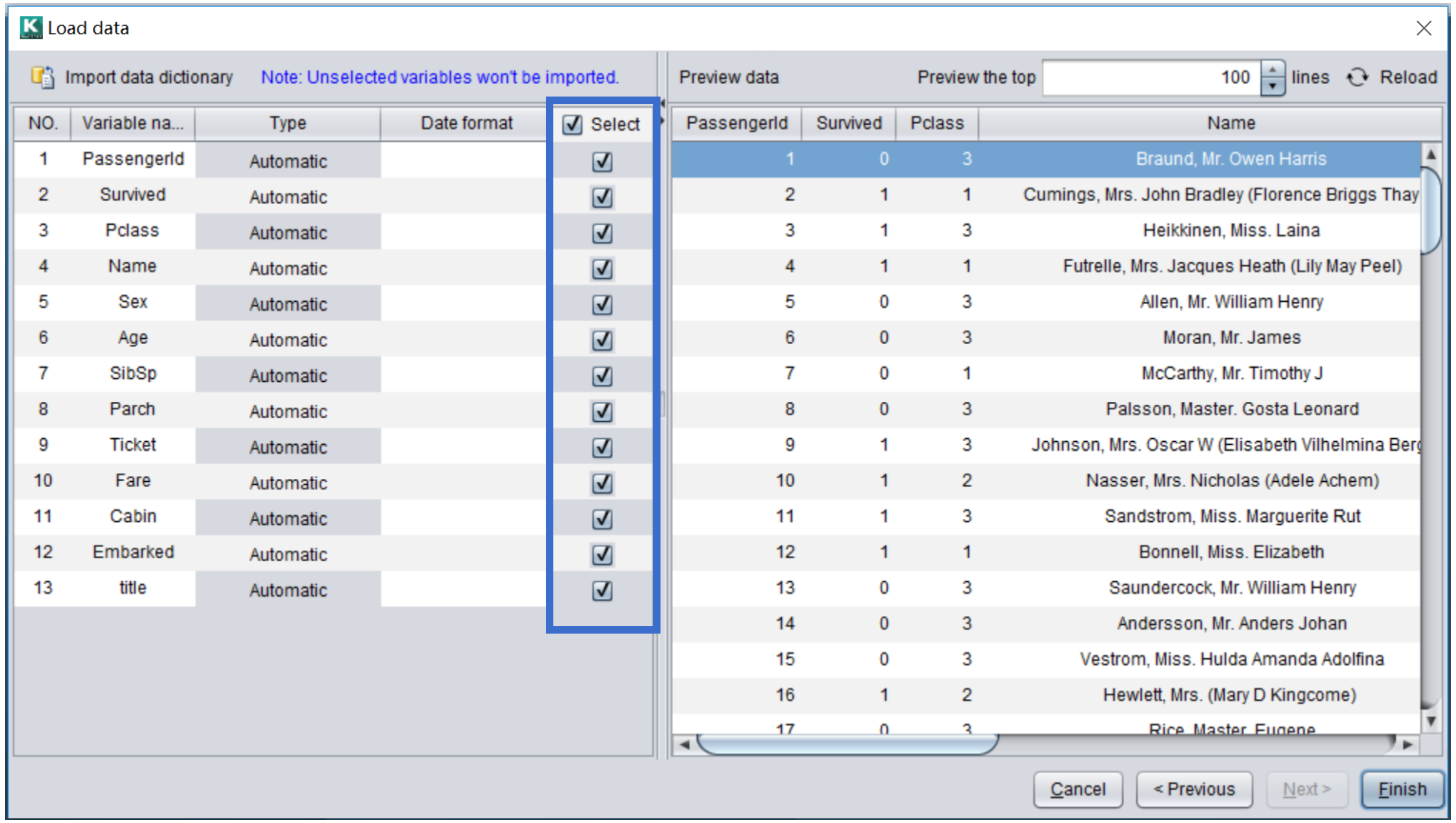
Preview the top 100 lines

PassengerId	Survived	Pclass	Name
1	0	3	Braund, Mr. Owen Harris
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)
3	1	3	Heikkinen, Miss. Laina
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)
5	0	3	Allen, Mr. William Henry
6	0	3	Moran, Mr. James
7	0	1	McCarthy, Mr. Timothy J
8	0	3	Palsson, Master. Gosta Leonard
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)
11	1	3	Sandstrom, Miss. Marguerite Rut
12	1	1	Bonnell, Miss. Elizabeth
13	0	3	Saundercock, Mr. William Henry
14	0	3	Andersson, Mr. Anders Johan
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina
16	1	2	Hewlett, Mrs. (Mary D Kingcome)
17	0	3	Rice, Master. Eugene

Class exercise - Classification Model

Select the variables involved in the modeling and click Finish.

Here we choose all variables.



Note: Unselected variables won't be imported.

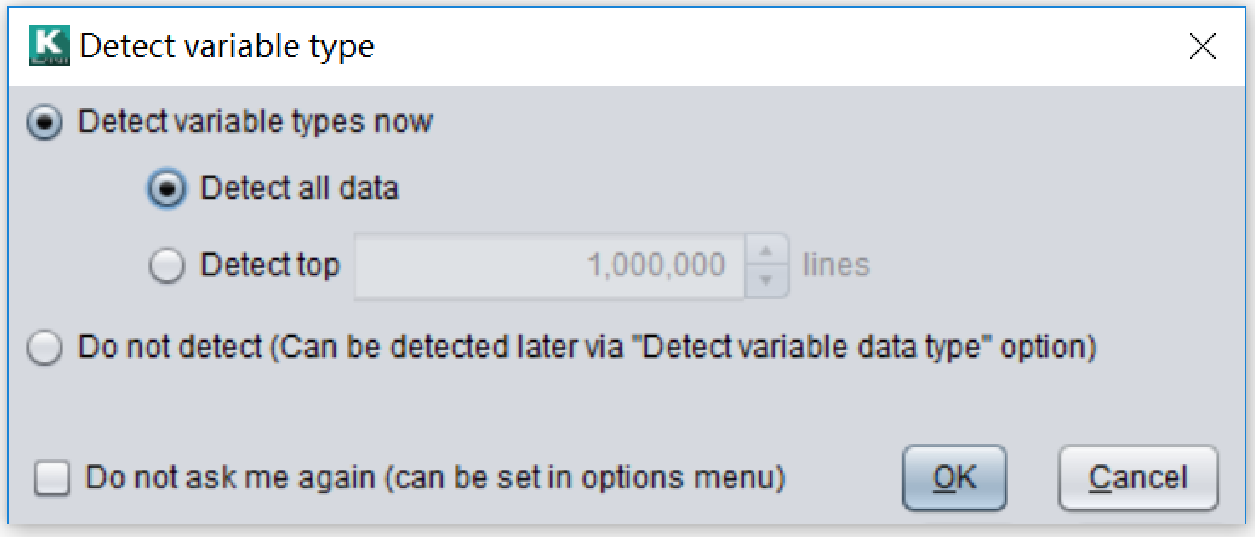
NO.	Variable na...	Type	Date format	☑ Select
1	PassengerId	Automatic		☑
2	Survived	Automatic		☑
3	Pclass	Automatic		☑
4	Name	Automatic		☑
5	Sex	Automatic		☑
6	Age	Automatic		☑
7	SibSp	Automatic		☑
8	Parch	Automatic		☑
9	Ticket	Automatic		☑
10	Fare	Automatic		☑
11	Cabin	Automatic		☑
12	Embarked	Automatic		☑
13	title	Automatic		☑

Preview data   Preview the top 100 lines   Reload

PassengerId	Survived	Pclass	Name
1	0	3	Braund, Mr. Owen Harris
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thay
3	1	3	Heikkinen, Miss. Laina
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)
5	0	3	Allen, Mr. William Henry
6	0	3	Moran, Mr. James
7	0	1	McCarthy, Mr. Timothy J
8	0	3	Palsson, Master. Gosta Leonard
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)
11	1	3	Sandstrom, Miss. Marguerite Rut
12	1	1	Bonnell, Miss. Elizabeth
13	0	3	Saundercock, Mr. William Henry
14	0	3	Andersson, Mr. Anders Johan
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina
16	1	2	Hewlett, Mrs. (Mary D Kingcome)
17	0	3	Rice, Master. Eugene

Class exercise - Classification Model

Select the amount of data to be detected. When the amount of data is small, all can be detected. When the amount of data is large, some can be detected, such as 50000 pieces, to improve efficiency.
Here we check all.

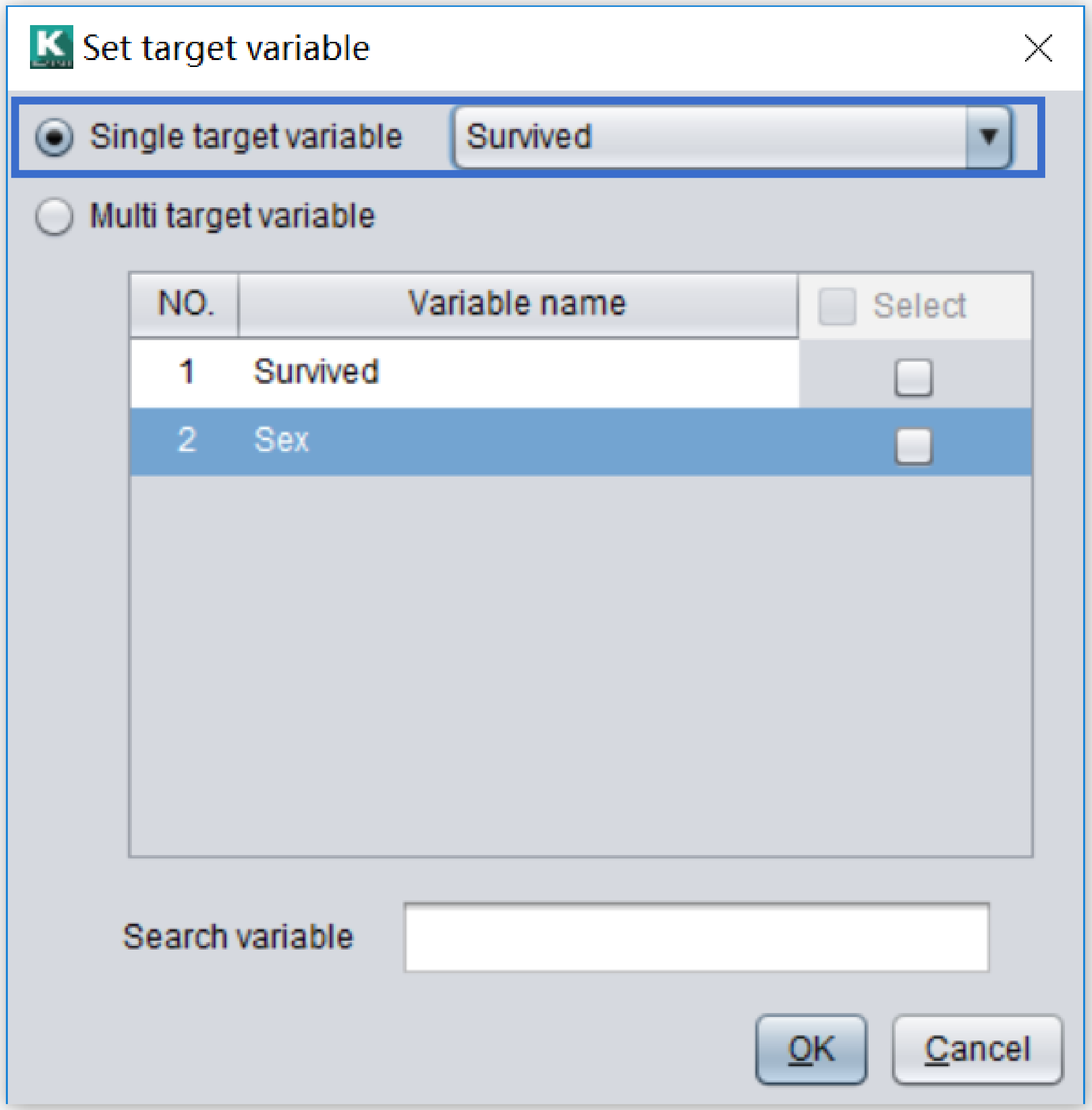


Detect variable type
● Detect variable types now
● Detect all data
○ Detect top 1,000,000 lines
○ Do not detect (Can be detected later via "Detect variable data type" option)
☐ Do not ask me again (can be set in options menu)
OK
Cancel

Class exercise - Classification Model

Set target variable

In this case, we only have one target variable "survived", so we choose a single target variable. "Survived" is a binary variable, so we need to build a classification model.



Set target variable

● Single target variable: Survived

○ Multi target variable

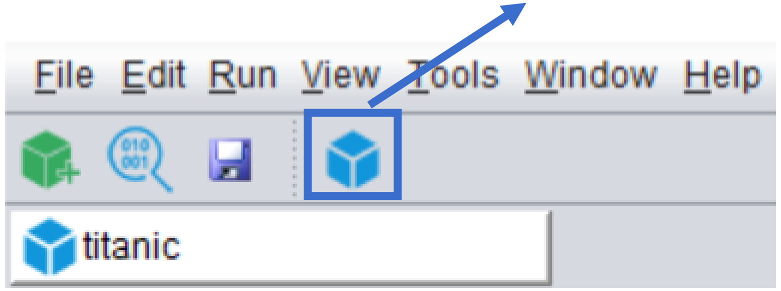
NO.	Variable name	☐ Select
1	Survived	☐
2	Sex	☐

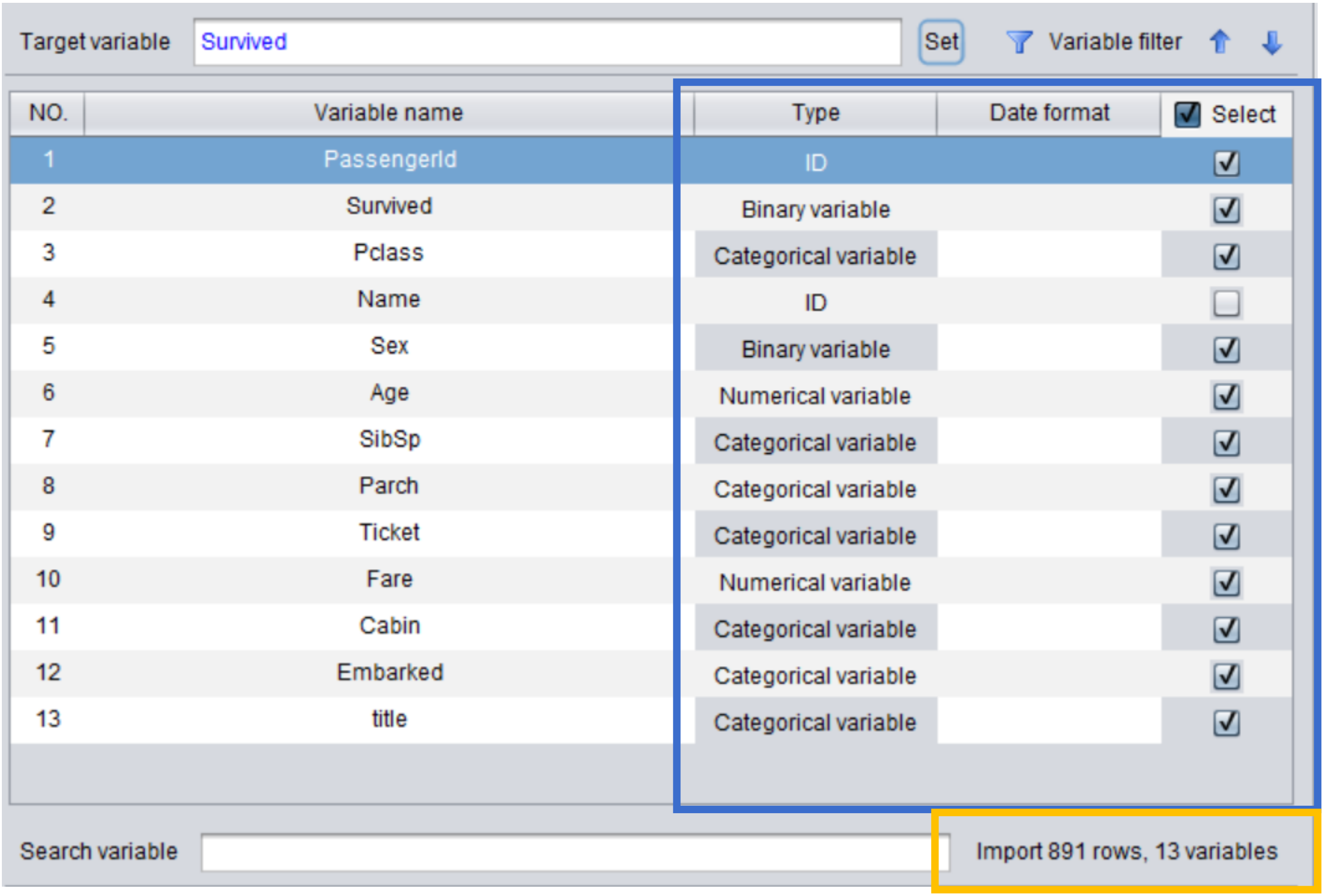
Search variable

Class exercise - Classification Model

The software automatically counts 891 samples and 13 variables, and automatically identifies the data type of each variable, and eliminates useless variables.

Click the modeling button to start modeling.





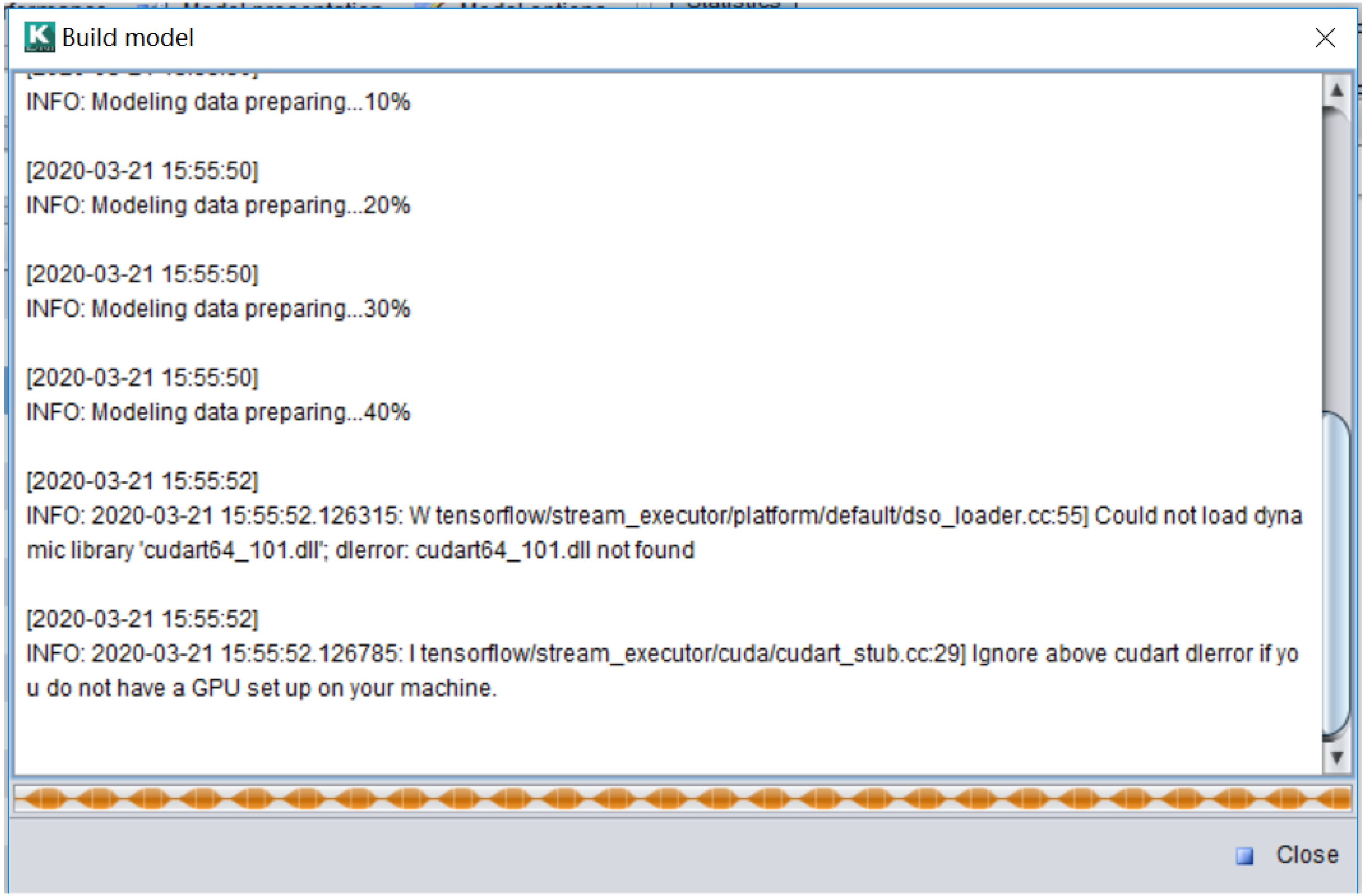
Target variable: Survived

NO.	Variable name	Type	Date format	☑ Select
1	PassengerId	ID		☑
2	Survived	Binary variable		☑
3	Pclass	Categorical variable		☑
4	Name	ID		☐
5	Sex	Binary variable		☑
6	Age	Numerical variable		☑
7	SibSp	Categorical variable		☑
8	Parch	Categorical variable		☑
9	Ticket	Categorical variable		☑
10	Fare	Numerical variable		☑
11	Cabin	Categorical variable		☑
12	Embarked	Categorical variable		☑
13	title	Categorical variable		☑

Import 891 rows, 13 variables

Class exercise - Classification Model

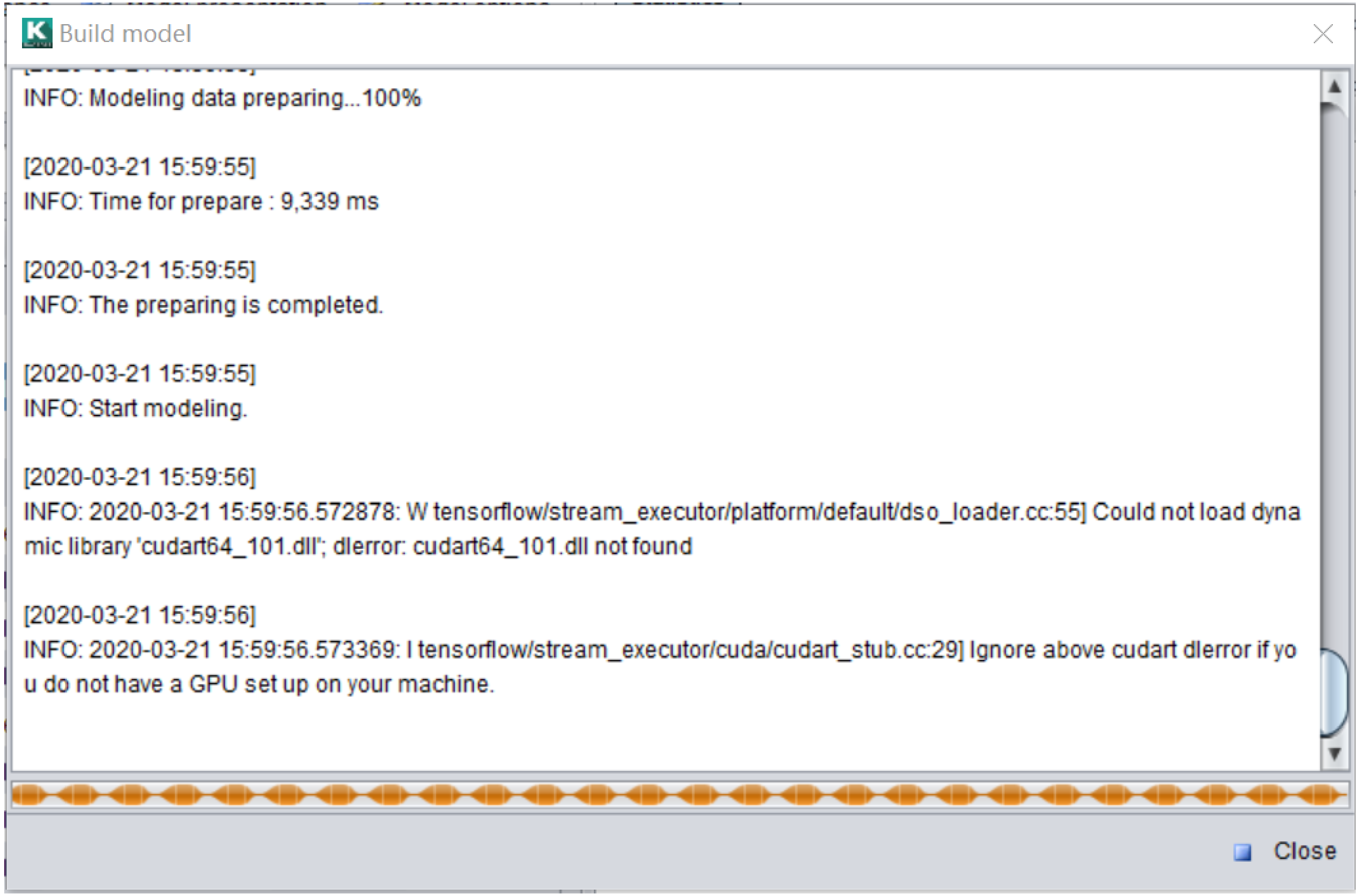
Automatic data preparation and display the preparing progress.



```
INFO: Modeling data preparing...10%

[2020-03-21 15:55:50]
INFO: Modeling data preparing...20%

[2020-03-21 15:55:50]
INFO: Modeling data preparing...30%

[2020-03-21 15:55:50]
INFO: Modeling data preparing...40%

[2020-03-21 15:55:52]
INFO: 2020-03-21 15:55:52.126315: W tensorflow/stream_executor/platform/default/dso_loader.cc:55] Could not load dyna
mic library 'cudart64_101.dll'; dlerror: cudart64_101.dll not found

[2020-03-21 15:55:52]
INFO: 2020-03-21 15:55:52.126785: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if yo
u do not have a GPU set up on your machine.
```

Class exercise - Classification Model

Preparation complete, start modeling.

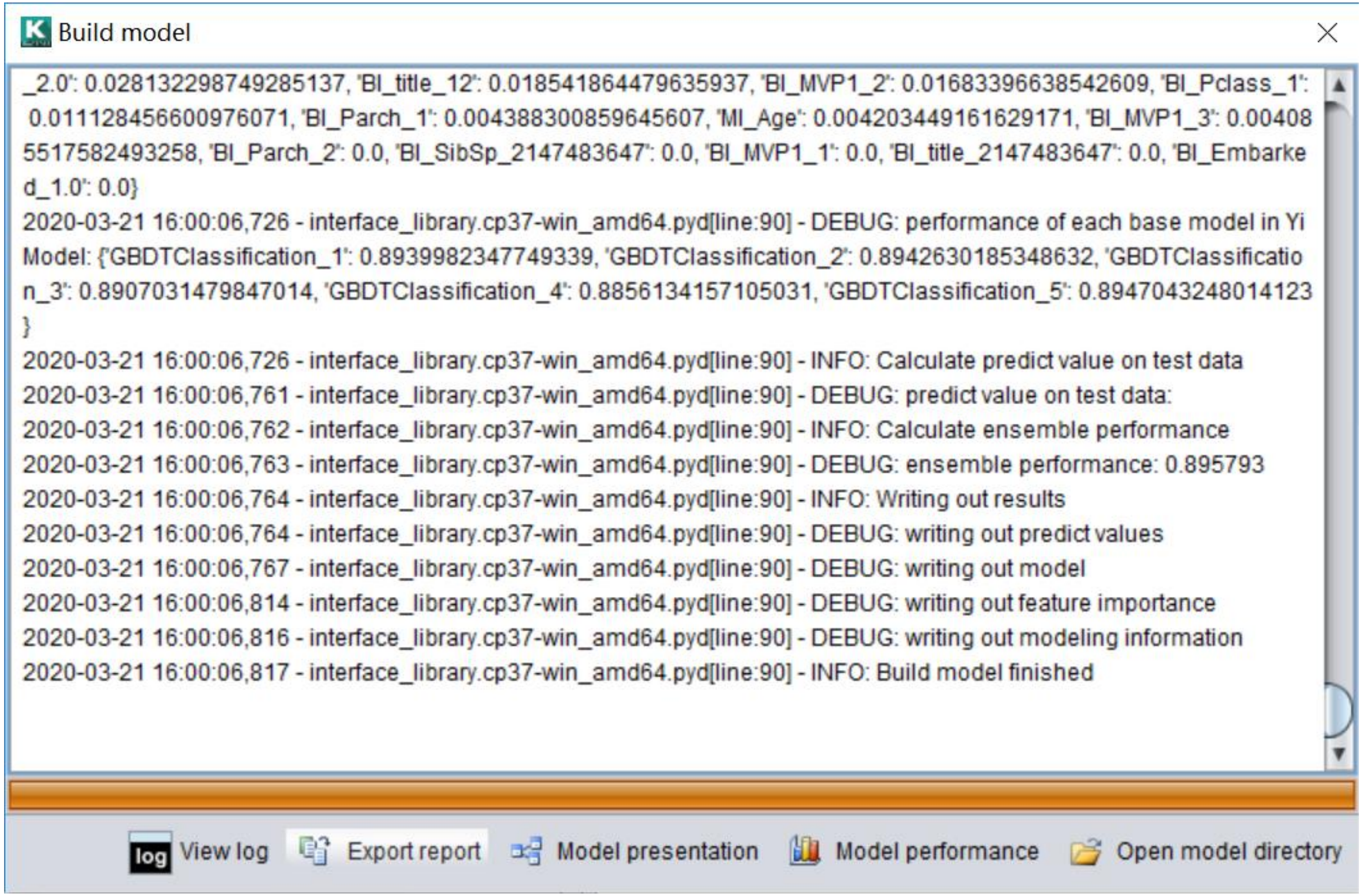


```
INFO: Modeling data preparing...100%

[2020-03-21 15:59:55]
INFO: Time for prepare : 9,339 ms

[2020-03-21 15:59:55]
INFO: The preparing is completed.

[2020-03-21 15:59:55]
INFO: Start modeling.

[2020-03-21 15:59:56]
INFO: 2020-03-21 15:59:56.572878: W tensorflow/stream_executor/platform/default/dso_loader.cc:55] Could not load dyna
mic library 'cudart64_101.dll'; dlerror: cudart64_101.dll not found

[2020-03-21 15:59:56]
INFO: 2020-03-21 15:59:56.573369: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if yo
u do not have a GPU set up on your machine.
```

Class exercise - Classification Model

Automatic modeling completed,

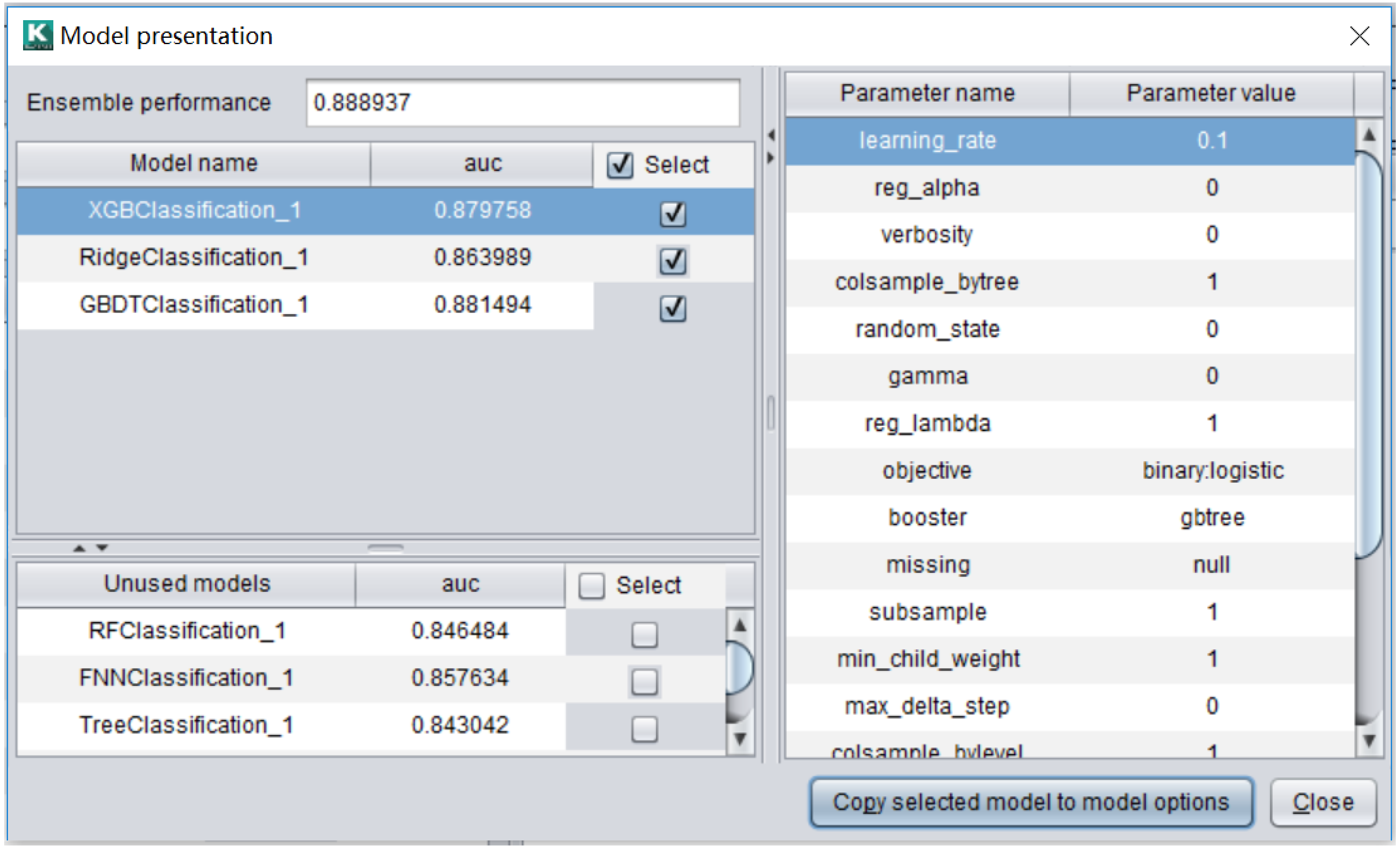
time consumed 10s



```
_2.0': 0.028132298749285137, 'BI_title_12': 0.018541864479635937, 'BI_MVP1_2': 0.01683396638542609, 'BI_Pclass_1':
 0.011128456600976071, 'BI_Parch_1': 0.004388300859645607, 'MI_Age': 0.004203449161629171, 'BI_MVP1_3': 0.00408
5517582493258, 'BI_Parch_2': 0.0, 'BI_SibSp_2147483647': 0.0, 'BI_MVP1_1': 0.0, 'BI_title_2147483647': 0.0, 'BI_Embarke
d_1.0': 0.0}
2020-03-21 16:00:06,726 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: performance of each base model in Yi
Model: {'GBDTClassification_1': 0.8939982347749339, 'GBDTClassification_2': 0.8942630185348632, 'GBDTClassificatio
n_3': 0.8907031479847014, 'GBDTClassification_4': 0.8856134157105031, 'GBDTClassification_5': 0.8947043248014123
}
2020-03-21 16:00:06,726 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Calculate predict value on test data
2020-03-21 16:00:06,761 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: predict value on test data:
2020-03-21 16:00:06,762 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Calculate ensemble performance
2020-03-21 16:00:06,763 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: ensemble performance: 0.895793
2020-03-21 16:00:06,764 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Writing out results
2020-03-21 16:00:06,764 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out predict values
2020-03-21 16:00:06,767 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out model
2020-03-21 16:00:06,814 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out feature importance
2020-03-21 16:00:06,816 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out modeling information
2020-03-21 16:00:06,817 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Build model finished
```

Class exercise - Classification Model

The models and parameters with better effect are selected automatically by using various algorithms to model respectively.



Ensemble performance: 0.888937

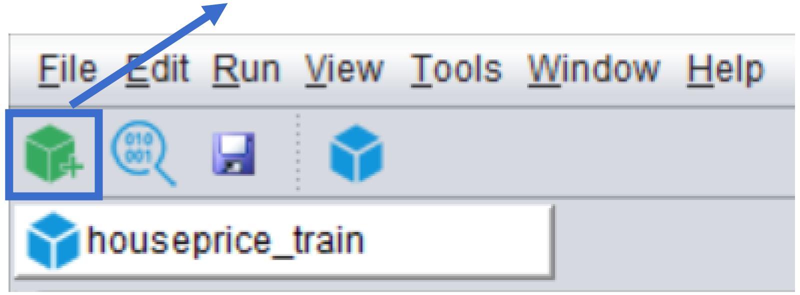
Model name	auc	☑ Select
XGBClassification_1	0.879758	☑
RidgeClassification_1	0.863989	☑
GBDTClassification_1	0.881494	☑

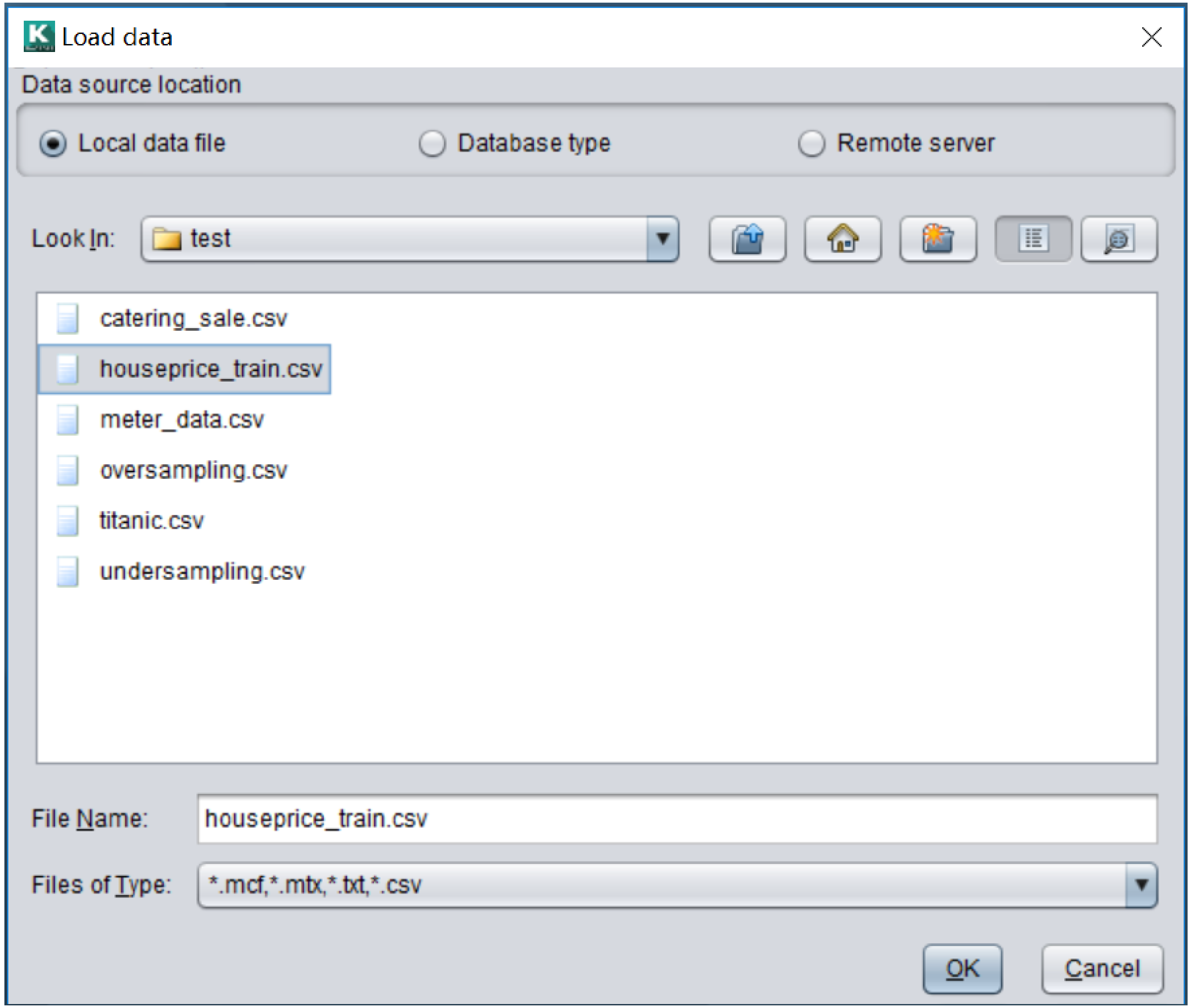
Unused models	auc	☐ Select
RFClassification_1	0.846484	☐
FNNClassification_1	0.857634	☐
TreeClassification_1	0.843042	☐

Parameter name	Parameter value
learning_rate	0.1
reg_alpha	0
verbosity	0
colsample_bytree	1
random_state	0
gamma	0
reg_lambda	1
objective	binary:logistic
booster	gbtree
missing	null
subsample	1
min_child_weight	1
max_delta_step	0
colsample_bylevel	1

Class exercise - Regression Model

Click new model, select the data of house price prediction, and click OK to import the data.


new model


Load data

Data source location

● Local data file ○ Database type ○ Remote server

Look In: test

catering_sale.csv
houseprice_train.csv
meter_data.csv
oversampling.csv
titanic.csv
undersampling.csv

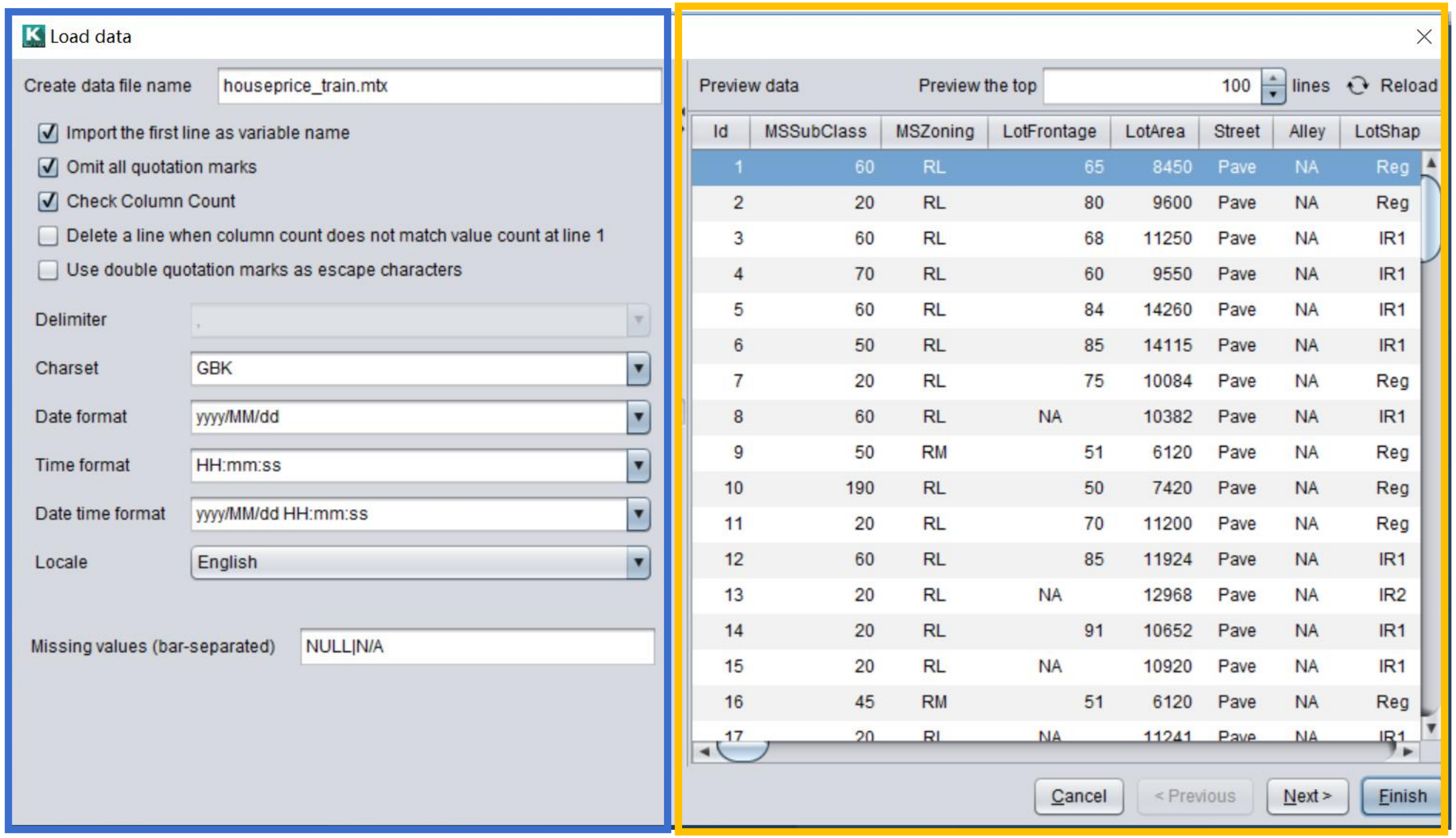
File Name: houseprice_train.csv

Files of Type: *.mcf,*.mtx,*.txt,*.csv

Class exercise - Regression Model

Data and variables can be previewed on the right side of the page.

The left side of the page is configured with character set format, date time format and missing value format for automatic recognition by software.



Load data

Create data file name: houseprice_train.mtx

☑ Import the first line as variable name
☑ Omit all quotation marks
☑ Check Column Count
☐ Delete a line when column count does not match value count at line 1
☐ Use double quotation marks as escape characters

Delimiter: ,
Charset: GBK
Date format: yyyy/MM/dd
Time format: HH:mm:ss
Date time format: yyyy/MM/dd HH:mm:ss
Locale: English

Missing values (bar-separated): NULL|N/A

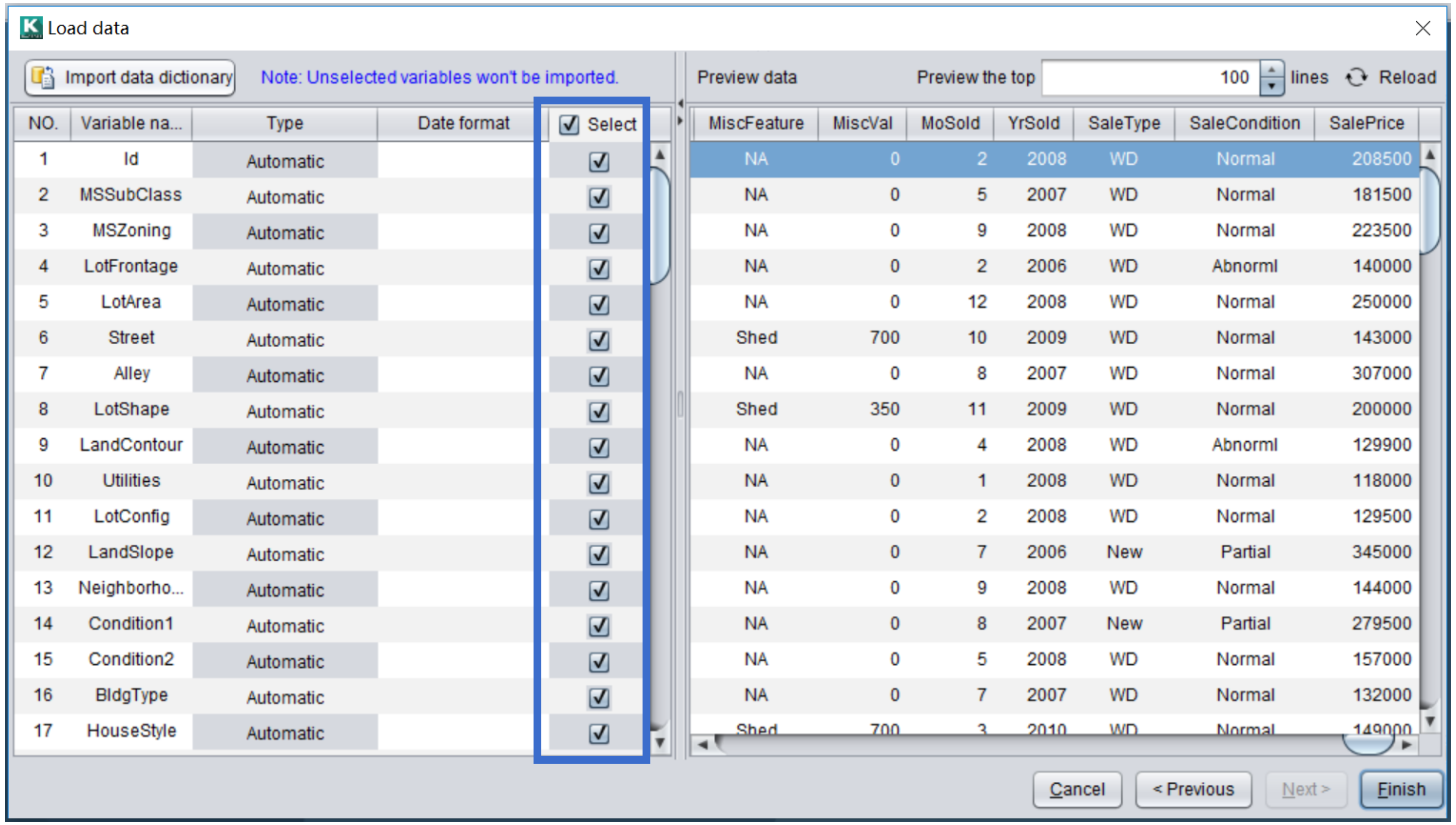
Preview data — Preview the top 100 lines

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShap
1	60	RL	65	8450	Pave	NA	Reg
2	20	RL	80	9600	Pave	NA	Reg
3	60	RL	68	11250	Pave	NA	IR1
4	70	RL	60	9550	Pave	NA	IR1
5	60	RL	84	14260	Pave	NA	IR1
6	50	RL	85	14115	Pave	NA	IR1
7	20	RL	75	10084	Pave	NA	Reg
8	60	RL	NA	10382	Pave	NA	IR1
9	50	RM	51	6120	Pave	NA	Reg
10	190	RL	50	7420	Pave	NA	Reg
11	20	RL	70	11200	Pave	NA	Reg
12	60	RL	85	11924	Pave	NA	IR1
13	20	RL	NA	12968	Pave	NA	IR2
14	20	RL	91	10652	Pave	NA	IR1
15	20	RL	NA	10920	Pave	NA	IR1
16	45	RM	51	6120	Pave	NA	Reg
17	20	RL	NA	11241	Pave	NA	IR1

Class exercise - Regression Model

Select the variables involved in the modeling and click Finish.

Here we choose all variables.



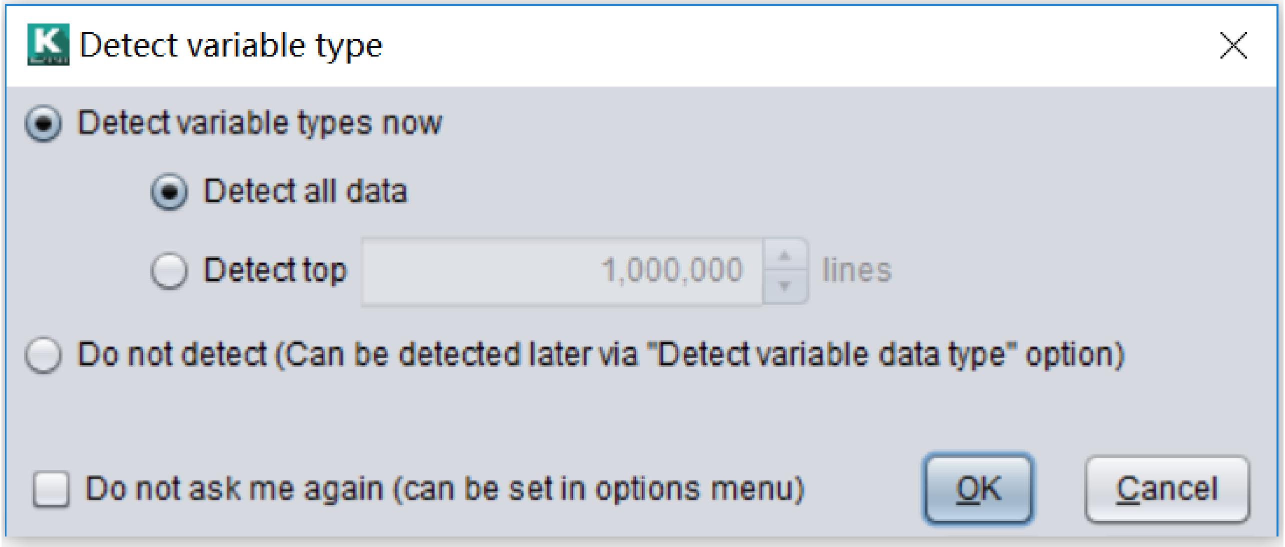
NO.	Variable na...	Type	Date format	☑ Select
1	Id	Automatic		☑
2	MSSubClass	Automatic		☑
3	MSZoning	Automatic		☑
4	LotFrontage	Automatic		☑
5	LotArea	Automatic		☑
6	Street	Automatic		☑
7	Alley	Automatic		☑
8	LotShape	Automatic		☑
9	LandContour	Automatic		☑
10	Utilities	Automatic		☑
11	LotConfig	Automatic		☑
12	LandSlope	Automatic		☑
13	Neighborho...	Automatic		☑
14	Condition1	Automatic		☑
15	Condition2	Automatic		☑
16	BldgType	Automatic		☑
17	HouseStyle	Automatic		☑

MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
NA	0	2	2008	WD	Normal	208500
NA	0	5	2007	WD	Normal	181500
NA	0	9	2008	WD	Normal	223500
NA	0	2	2006	WD	Abnorml	140000
NA	0	12	2008	WD	Normal	250000
Shed	700	10	2009	WD	Normal	143000
NA	0	8	2007	WD	Normal	307000
Shed	350	11	2009	WD	Normal	200000
NA	0	4	2008	WD	Abnorml	129900
NA	0	1	2008	WD	Normal	118000
NA	0	2	2008	WD	Normal	129500
NA	0	7	2006	New	Partial	345000
NA	0	9	2008	WD	Normal	144000
NA	0	8	2007	New	Partial	279500
NA	0	5	2008	WD	Normal	157000
NA	0	7	2007	WD	Normal	132000
Shed	700	3	2010	WD	Normal	149000

Class exercise - Regression Model

Select the amount of data to be detected. When the amount of data is small, all can be detected. When the amount of data is large, some can be detected, such as 50000 pieces, to improve efficiency.

Here we check all.



Detect variable type

● Detect variable types now

● Detect all data

○ Detect top 1,000,000 lines

○ Do not detect (Can be detected later via "Detect variable data type" option)

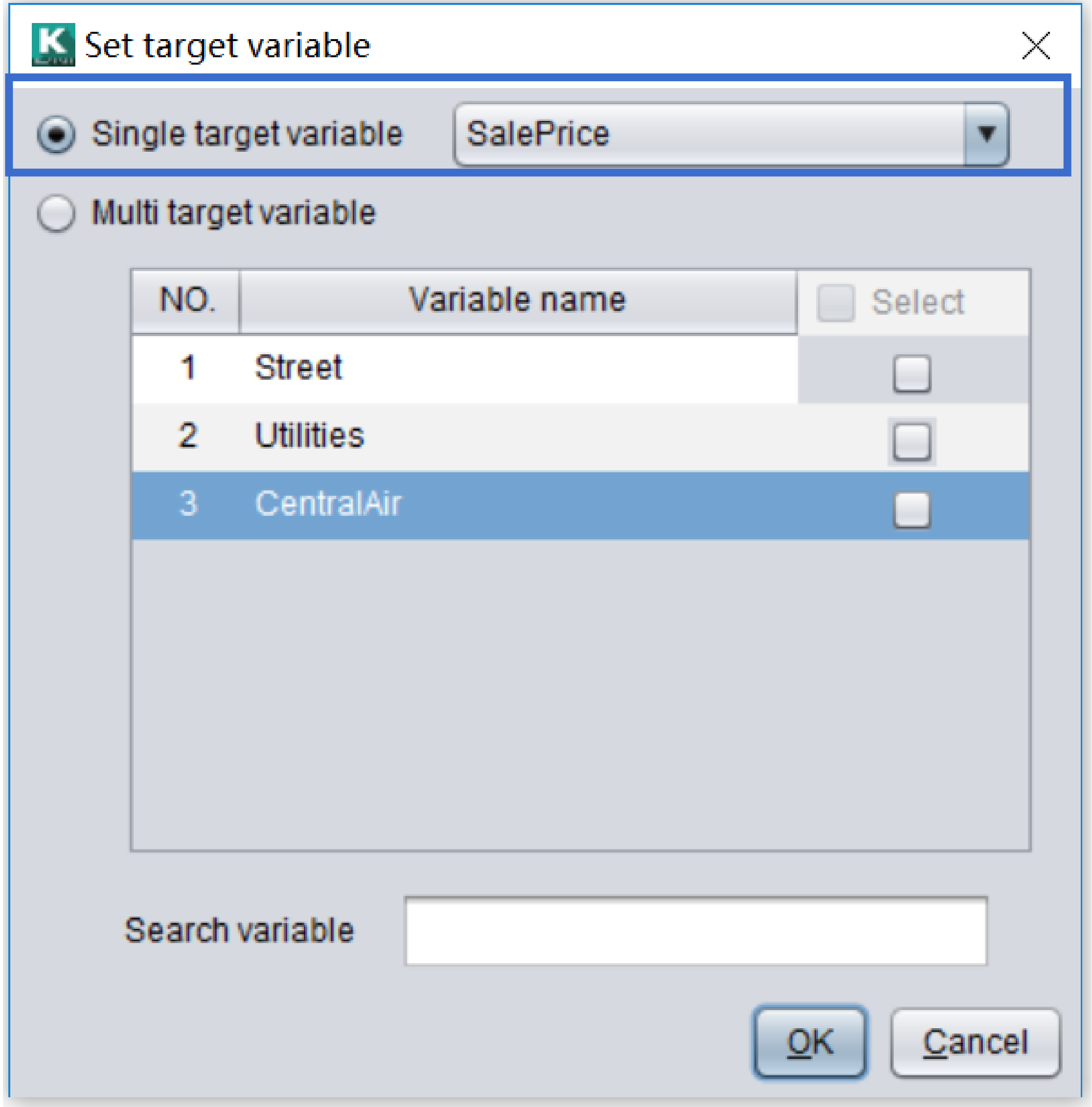
☐ Do not ask me again (can be set in options menu)

Class exercise - Regression Model

Set target variable

In this case, we only have one target variable "SalePrice", so we choose a single target variable.

"SalePrice" is a numerical variable, so a regression model is needed.



Set target variable

● Single target variable: SalePrice

○ Multi target variable

NO.	Variable name	☐ Select
1	Street	☐
2	Utilities	☐
3	CentralAir	☐

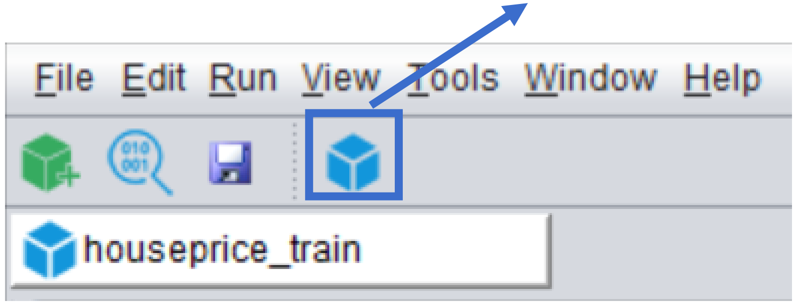
Search variable:

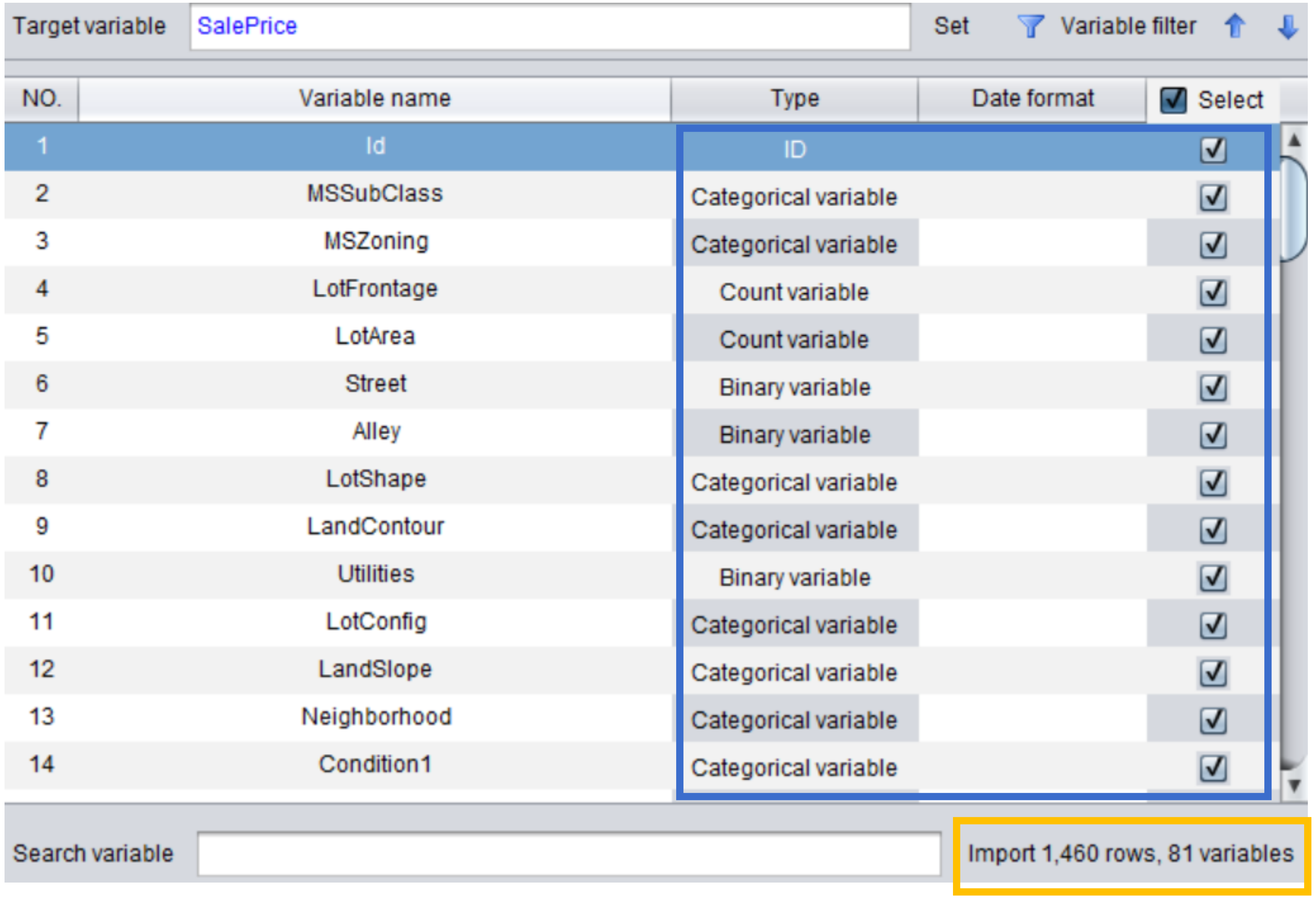
Class exercise - Regression Model

The software automatically counts 1460 samples and 81 variables, and automatically identifies the data type of each variable, and eliminates useless variables.

Click the modeling button to start modeling.

modeling





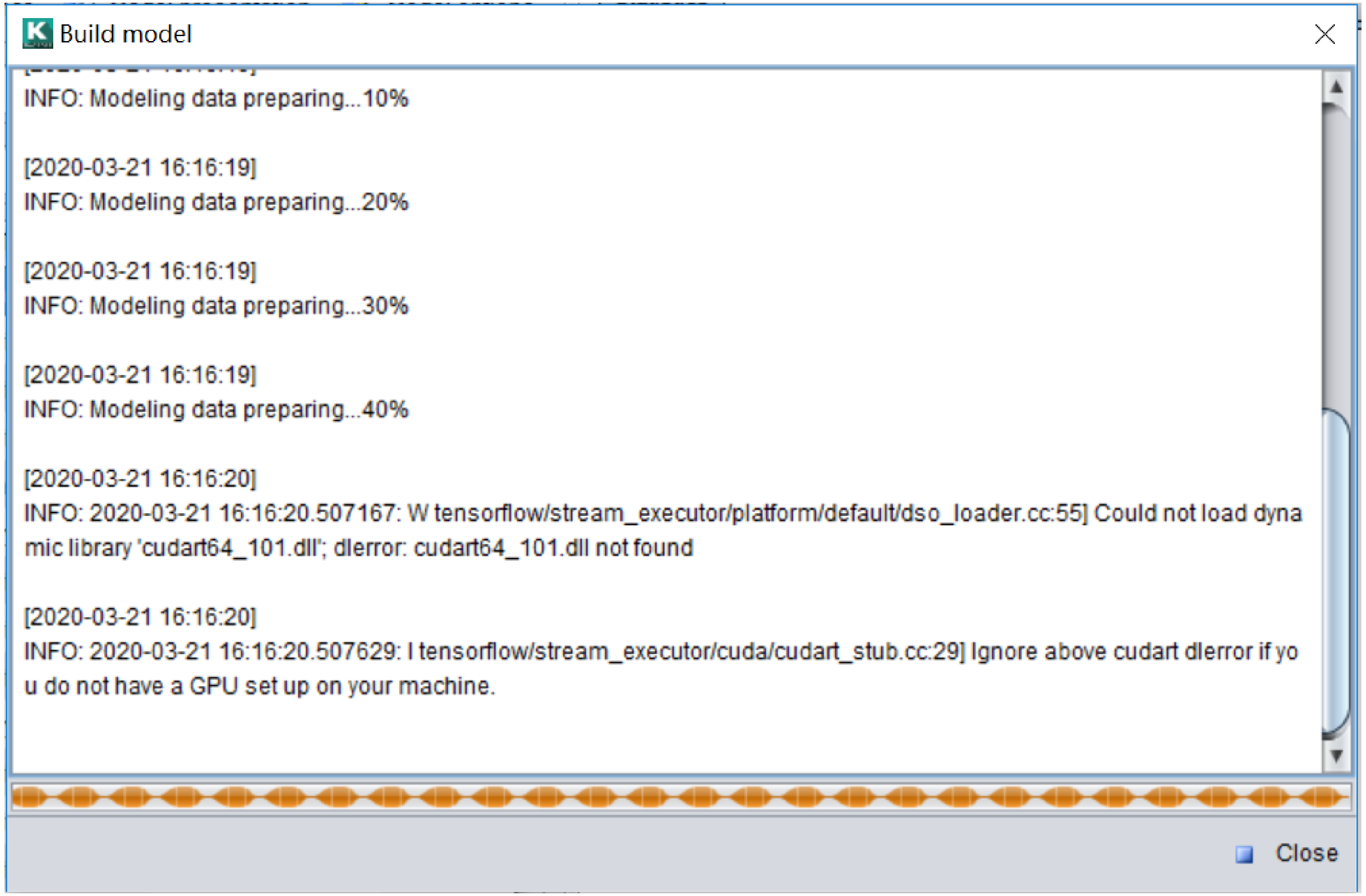
Target variable: SalePrice

NO.	Variable name	Type	Date format	☑ Select
1	Id	ID		☑
2	MSSubClass	Categorical variable		☑
3	MSZoning	Categorical variable		☑
4	LotFrontage	Count variable		☑
5	LotArea	Count variable		☑
6	Street	Binary variable		☑
7	Alley	Binary variable		☑
8	LotShape	Categorical variable		☑
9	LandContour	Categorical variable		☑
10	Utilities	Binary variable		☑
11	LotConfig	Categorical variable		☑
12	LandSlope	Categorical variable		☑
13	Neighborhood	Categorical variable		☑
14	Condition1	Categorical variable		☑

Import 1,460 rows, 81 variables

Class exercise - Regression Model

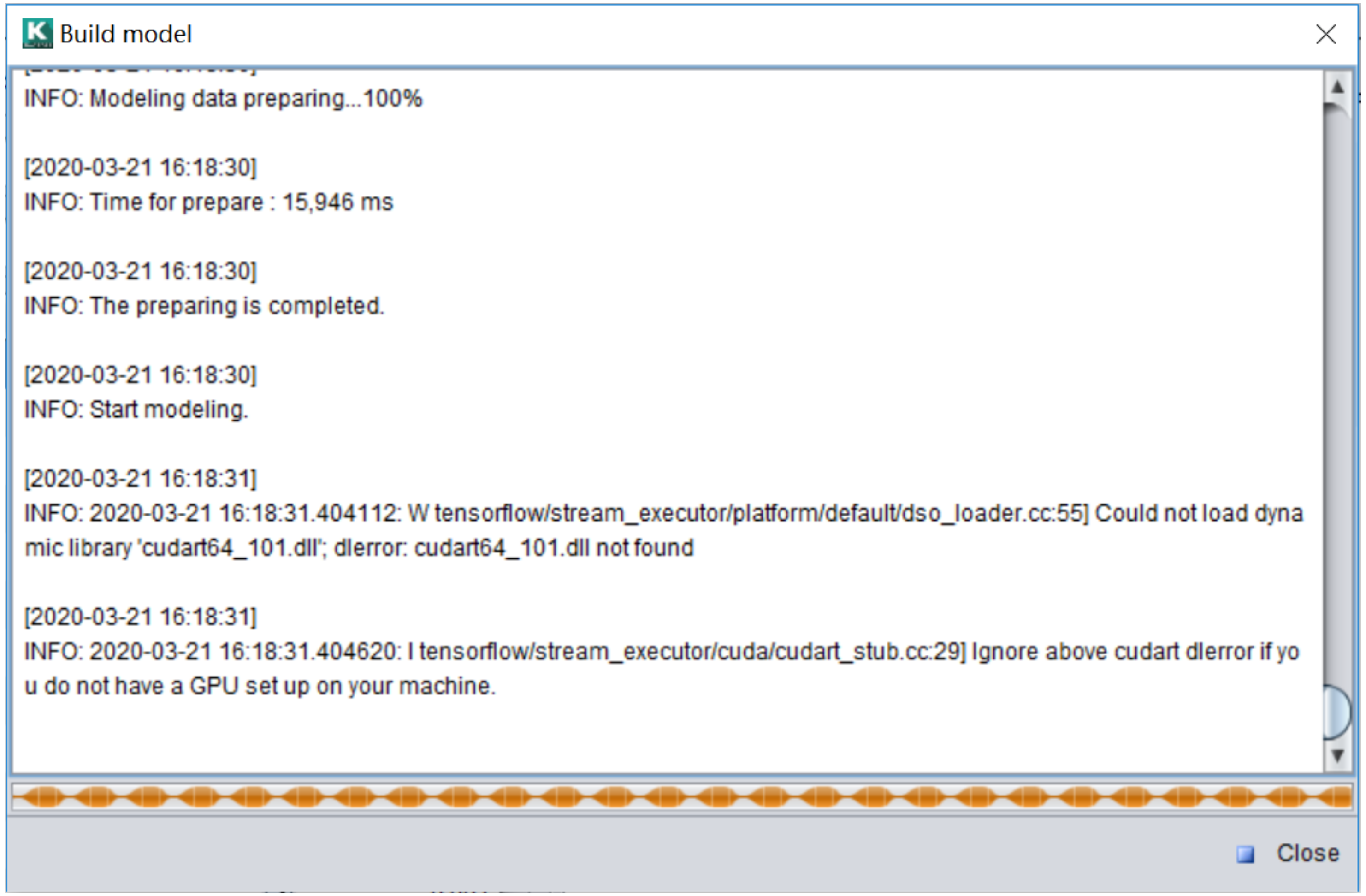
Automatic data preparation and display the preparing progress.



```
INFO: Modeling data preparing...10%

[2020-03-21 16:16:19]
INFO: Modeling data preparing...20%

[2020-03-21 16:16:19]
INFO: Modeling data preparing...30%

[2020-03-21 16:16:19]
INFO: Modeling data preparing...40%

[2020-03-21 16:16:20]
INFO: 2020-03-21 16:16:20.507167: W tensorflow/stream_executor/platform/default/dso_loader.cc:55] Could not load dyna
mic library 'cudart64_101.dll'; dlerror: cudart64_101.dll not found

[2020-03-21 16:16:20]
INFO: 2020-03-21 16:16:20.507629: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if yo
u do not have a GPU set up on your machine.
```

Class exercise - Regression Model

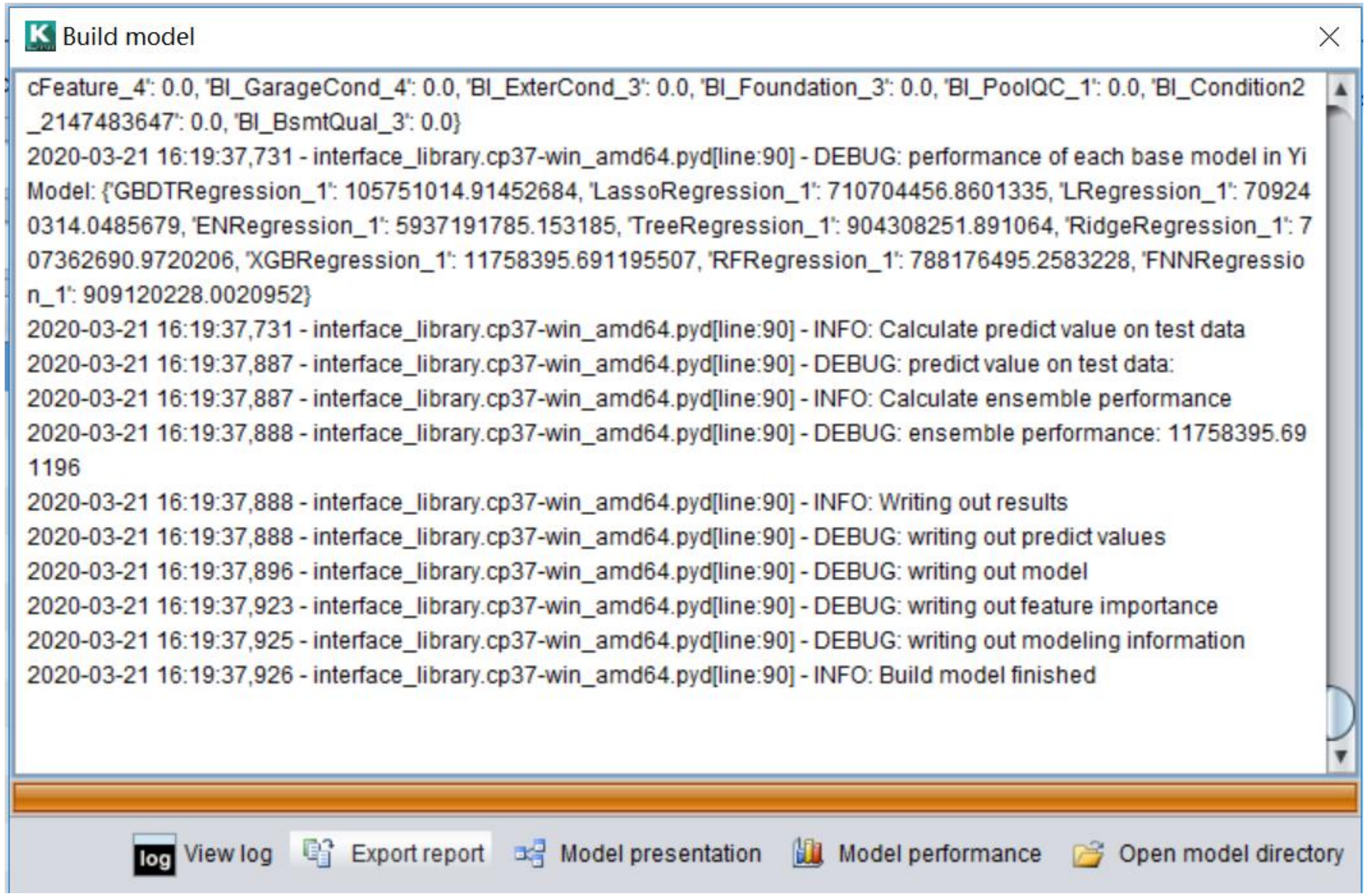
Preparation complete, start modeling.



```
INFO: Modeling data preparing...100%

[2020-03-21 16:18:30]
INFO: Time for prepare : 15,946 ms

[2020-03-21 16:18:30]
INFO: The preparing is completed.

[2020-03-21 16:18:30]
INFO: Start modeling.

[2020-03-21 16:18:31]
INFO: 2020-03-21 16:18:31.404112: W tensorflow/stream_executor/platform/default/dso_loader.cc:55] Could not load dyna
mic library 'cudart64_101.dll'; dlerror: cudart64_101.dll not found

[2020-03-21 16:18:31]
INFO: 2020-03-21 16:18:31.404620: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if yo
u do not have a GPU set up on your machine.
```

Class exercise - Regression Model

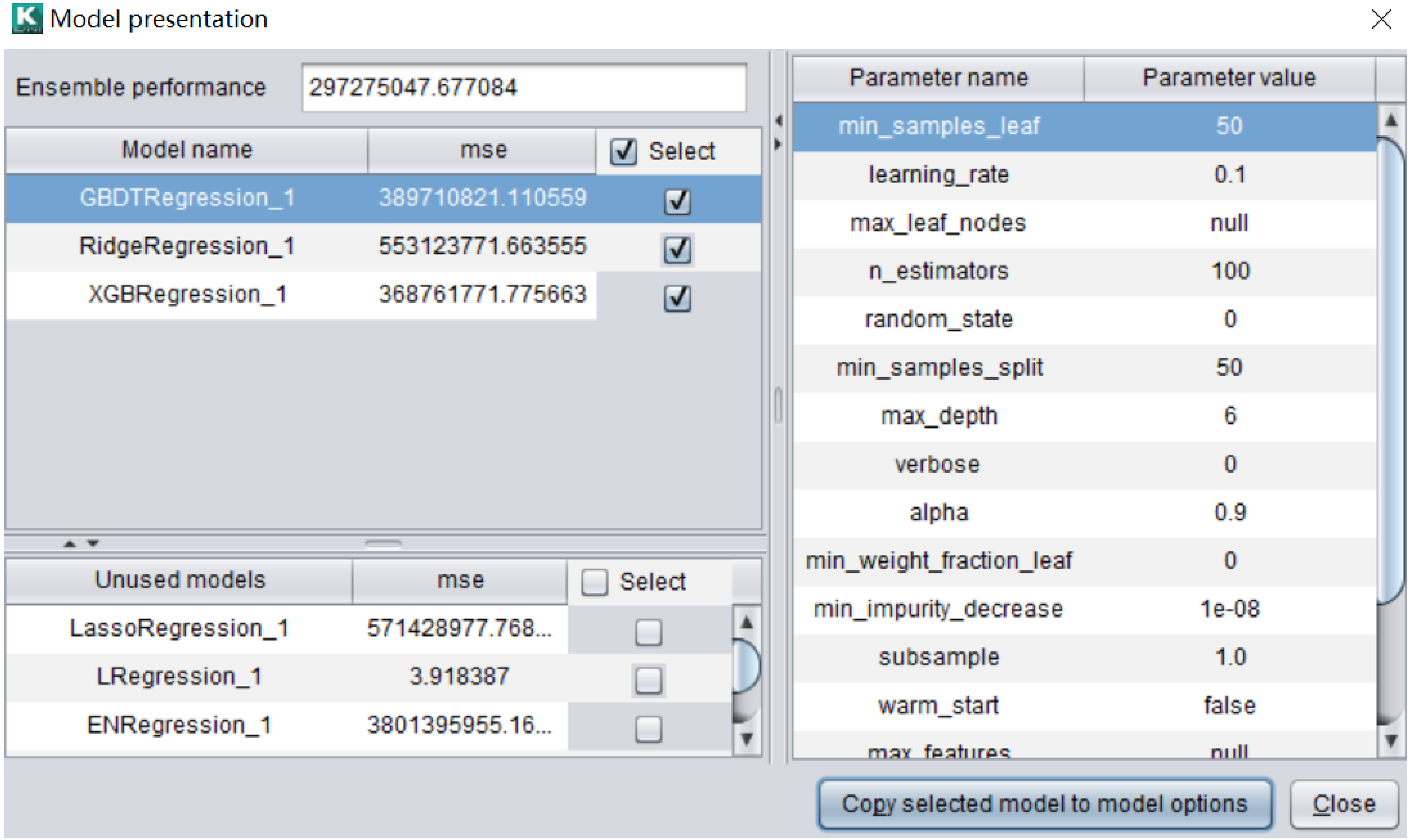
Automatic modeling completed, time consumed 31s



```
cFeature_4': 0.0, 'BI_GarageCond_4': 0.0, 'BI_ExterCond_3': 0.0, 'BI_Foundation_3': 0.0, 'BI_PoolQC_1': 0.0, 'BI_Condition2
_2147483647': 0.0, 'BI_BsmtQual_3': 0.0}
2020-03-21 16:19:37,731 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: performance of each base model in Yi
Model: {'GBDTRegression_1': 105751014.91452684, 'LassoRegression_1': 710704456.8601335, 'LRegression_1': 70924
0314.0485679, 'ENRegression_1': 5937191785.153185, 'TreeRegression_1': 904308251.891064, 'RidgeRegression_1': 7
07362690.9720206, 'XGBRegression_1': 11758395.691195507, 'RFRegression_1': 788176495.2583228, 'FNNRegressio
n_1': 909120228.0020952}
2020-03-21 16:19:37,731 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Calculate predict value on test data
2020-03-21 16:19:37,887 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: predict value on test data:
2020-03-21 16:19:37,887 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Calculate ensemble performance
2020-03-21 16:19:37,888 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: ensemble performance: 11758395.69
1196
2020-03-21 16:19:37,888 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Writing out results
2020-03-21 16:19:37,888 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out predict values
2020-03-21 16:19:37,896 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out model
2020-03-21 16:19:37,923 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out feature importance
2020-03-21 16:19:37,925 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out modeling information
2020-03-21 16:19:37,926 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Build model finished
```

Class exercise - Regression Model

The models and parameters with better effect are selected automatically by using various algorithms to model respectively.



Ensemble performance: 297275047.677084

Model name	mse	☑ Select
GBDTRegression_1	389710821.110559	☑
RidgeRegression_1	553123771.663555	☑
XGBRegression_1	368761771.775663	☑

Unused models	mse	☐ Select
LassoRegression_1	571428977.768...	☐
LRegression_1	3.918387	☐
ENRegression_1	3801395955.16...	☐

Parameter name	Parameter value
min_samples_leaf	50
learning_rate	0.1
max_leaf_nodes	null
n_estimators	100
random_state	0
min_samples_split	50
max_depth	6
verbose	0
alpha	0.9
min_weight_fraction_leaf	0
min_impurity_decrease	1e-08
subsample	1.0
warm_start	false
max_features	null

Conclusion

All kinds of algorithms have complex principles and many parameters. Users need to have theoretical knowledge and rich experience to build high-quality models.

The operation of automatic modeling is simple, and users can use it without knowing the algorithm.

Automatic modeling can automatically complete data preparation and modeling, and select high-quality models, with high modeling efficiency.

Data mining personnel can use tools to improve work efficiency and reduce workload.

Glossary - for reference

- supervised learning
- variable/feature
- label/target
- classification
- regression
- objective function
- loss function/cost function
- under-fitting
- over-fitting
- regularization
- L1 regularization
- L2 regularization
- linear regression
- linear model
- non-linear model
- lasso regression
- ridge regression
- Elastic Net
- high-dimensional
- sparse data
- logistic regression
- Decision tree
- Parametric learning

Glossary - for reference

neural network

cross validation

sample weight

ensemble learning

averaging / bagging

boosting

random forest

weak learner

out-of-bag/OOB

gradient boosting decision tree/GBDT

residual

gradient

XGBoost(eXtreme Gradient Boosting)

deep learning

artificial neural network/ANN

neuron

input layer

output layer

hidden layer

unbiased estimate

time complexity

auto machine learning/AutoML

Chapter 5 Model evaluation

5.1 classification model evaluation

5.2 Regression model evaluation

Significance of model evaluation

For a machine learning project, we can choose many models. For the previous chapters, for example, we can choose: linear regression, logistic regression, decision tree, ensemble learning, etc.

For a model, we also have many model indexes and graphs to evaluate the model.
Many analysts are not even willing to check the robustness of their models. Once the model is built, they rush to apply predictions to invisible data. This method is very dangerous.

Our goal is not to simply build a prediction model, but to create and select a model that can achieve high precision for data other than samples. Therefore, it is very important to check the accuracy of the model before calculating the predicted value.

This chapter will introduce the evaluation method of the model.

Classification model index

How to evaluate the model?

Classification model evaluation

Common indexes:

Accuracy, precision, recall, AUC, GINI, KS, etc.

Graphs:

ROC curve, lift chart, recall chart

Confusion matrix

Confusion matrix is a visual tool in supervised learning, which is mainly used to compare the real information of classification results and instances. Each row in the matrix represents the real category of the instance, and each column represents the predicted category of the instance.

True Positive , TP： positive samples predicted as positive by the model.

False Positive , FP： negative samples predicted as positive by the model.

False Negative , FN： positive samples predicted as negative by the model.

True Negative , TN： negative samples predicted as negative by the model.

		Predicted value	
		Positive	Negative
Real value	Positive	TP	FN
	Negative	FP	TN

Confusion matrix

Accuracy, ACC=(TP+TN)/N,

How many of the predictions are correct.

Precision, PPV=TP/(TP+FP),

How many of the results predicted to be positive samples are really positive samples.

Sensitivity/Recall , TPR=TP/(TP+FN),

How many of the actual positive samples are correctly predicted.

Specificity , TNR=TN/(FP+TN),

How many of the actual negative samples are correctly predicted.

		Predicted value		
		P	N	
Real value	P	**TP**	FN	TPR=TP/(TP+FN)
	N	FP	**TN**	TNR=TN/(FP+TN)
		PPV=TP/(TP+FP)	ACC=(TP+TN)/N	N=TP+FN+FP+TN

Precision is easily confused with accuracy. In fact, precision is only for the correct positive samples, not all the correct samples. The meaning is: How many of the results predicted to be positive samples are really positive samples.

Confusion matrix

The confusion matrix is represented by the example of judging good and bad melons as follows (there are 10 known good and bad melons in the test set)

True Positive （TP） =3, actually good melon, predicted as good melon
False Positive （FP） =1, actually bad melon, predicted as good melon
False Negative （FN） =2, actually good melon, predicted as bad melon
True Negative （TN） =4, actually bad melon, predicted as bad melon

		Predicted value	
		Good melon	Bad melon
Real value	Good melon	3， TP	2， FN
	Bad melon	1， FP	4， TN

Confusion matrix

Accuracy , ACC=$\frac{7}{10}$=0.7, 7 of all 10 prediction results are correct (that is, good melon is predicted to be good, bad melon is predicted to be bad), with an accuracy of 0.7

Precision ,PPV= $\frac{3}{4}$=0.75, 3 of the predicted 4 good melons are really good melons, with a precision of 0.75

Sensitivity/Recall ,TPR= $\frac{3}{5} = 0.6$, 3 of the actual 5 good melons have been correctly predicted, and the recall rate is 0.6

Specificity ,TNR= $\frac{4}{5}$, 4 of 5 bad melons were predicted correctly, the specificity is 0,8

		Predicted value		
		Good melon	Bad melon	
Real value	Good melon	3，TP	2，FN	TPR=3/(3+2)
	Bad melon	1，FP	4，TN	TNR=4/(1+4)
		PPV=3/(3+1)	ACC=(3+4)/10	(3+2+1+4)，N

Index application

Why do we have so many indexes? Isn't it enough to have an accuracy index? Well, not enough.

Different scene objectives need different evaluation indexes. For example, an enterprise wants to sell 50 products, it has established two models to select the customers to be promoted. The confusion matrix is as follows:

Model A		**Predicted value**		
		Buy	Do not buy	Total
Real value	Buy	100，TP	20，FN	120
	Do not buy	50，FP	70，TN	120
	Total	150	90	240，N
Precision=100/150=0.667			**Accuracy=(100+70)/240=0.708**	

Model B		**Predicted value**		
		Buy	Do not buy	Total
Real value	Buy	50，TP	70，FN	120
	Do not buy	10，FP	110，TN	120
	Total	60	180	240，N
Precision=50/60=0.833			**Accuracy=(50+110)/240=0.667**	

Thinking： Which model should I choose?

Index application

Model A		Predicted value		
		Buy	Do not buy	Total
Real value	Buy	100，TP	20，FN	120
	Do not buy	50，FP	70，TN	120
	Total	150	90	240，N
Precision=100/150=0.667		**Accuracy=(100+70)/240=0.708**		

Model B		Predicted value		
		Buy	Do not buy	Total
Real value	Buy	50，TP	70，FN	120
	Do not buy	10，FP	110，TN	120
	Total	60	180	240，N
Precision=50/60=0.833		**Accuracy=(50+110)/240=0.667**		

Only considering the accuracy index, it seems that we should choose model A, but at this time, we need to sell 50 products to 75 (= 50 / 0.667, 66.7% of the predicted purchasers will actually buy, that is, the precision index) customers; and model B, as long as we sell 50 products to 60 (= 50 / 0.833) customers, the selling cost will be reduced.

Because, we only care about those customers who can be successfully promoted. For those who can't be successfully promoted and are correctly predicted to be unsuccessful, although it helps to improve the accuracy of the model, it doesn't mean much to us. In this scenario, the precision index should be used to evaluate the model.

Index application

Then look at the sensitivity / recall as an evaluation index, which is often used when the data distribution is unbalanced.

For example, recognize terrorists at the airport, because terrorists are very few, if the accuracy is used to evaluate the model, the accuracy can reach 99.999% or even higher as long as all people are identified as normal people, but obviously this model is useless. At this time, a model with high sensitivity needs to be established. For example, two confusion matrices are as follows:

Model A		**Predicted value**		
		Terrorists	Normal people	Total
Real value	Terrorists	0	5	5
	Normal people	0	999995	999995
	Total	0	1000000	1000000
Recall=0/5=0			Accuracy=99.9995%	

Model B		**Predicted value**		
		Terrorists	Normal people	Total
Real value	Terrorists	5	0	5
	Normal people	95	999900	999995
	Total	100	999900	1000000
Recall=5/5=1			Accuracy=99.9905%	

Considering only the accuracy, we will choose model A, but it can't identify terrorists at all. And model B, although its accuracy is low, can identify all terrorists. Although it may be wronged by several good people, it is better than being exploited by terrorists.

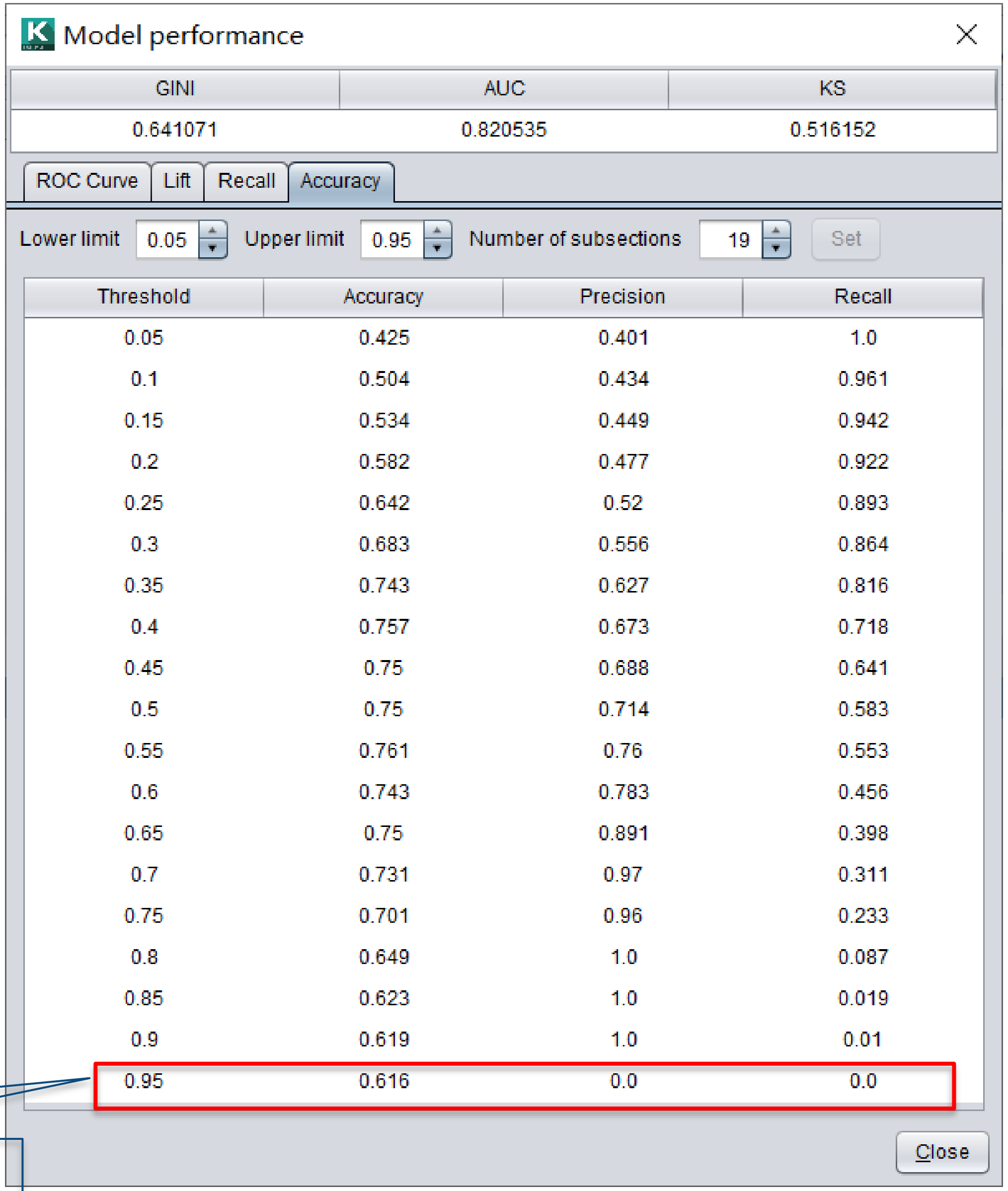
Accuracy table

Review:

The output of the binary classification model is the probability value, which means the probability that the target variable is a positive sample, between 0-1. The critical probability value used to distinguish positive and negative samples is called threshold.

Different thresholds correspond to different accuracy, precision and recall . Users need to choose reasonable thresholds to make decisions according to business objectives, instead of simply taking 0.5 as the judgment basis.

The right figure is the accuracy table of intelligent modeling software, which shows the corresponding accuracy, precision and recall values under different thresholds, which can help users make decisions conveniently.



GINI	AUC	KS
0.641071	0.820535	0.516152

Lower limit 0.05 Upper limit 0.95 Number of subsections 19

Threshold	Accuracy	Precision	Recall
0.05	0.425	0.401	1.0
0.1	0.504	0.434	0.961
0.15	0.534	0.449	0.942
0.2	0.582	0.477	0.922
0.25	0.642	0.52	0.893
0.3	0.683	0.556	0.864
0.35	0.743	0.627	0.816
0.4	0.757	0.673	0.718
0.45	0.75	0.688	0.641
0.5	0.75	0.714	0.583
0.55	0.761	0.76	0.553
0.6	0.743	0.783	0.456
0.65	0.75	0.891	0.398
0.7	0.731	0.97	0.311
0.75	0.701	0.96	0.233
0.8	0.649	1.0	0.087
0.85	0.623	1.0	0.019
0.9	0.619	1.0	0.01
0.95	0.616	0.0	0.0

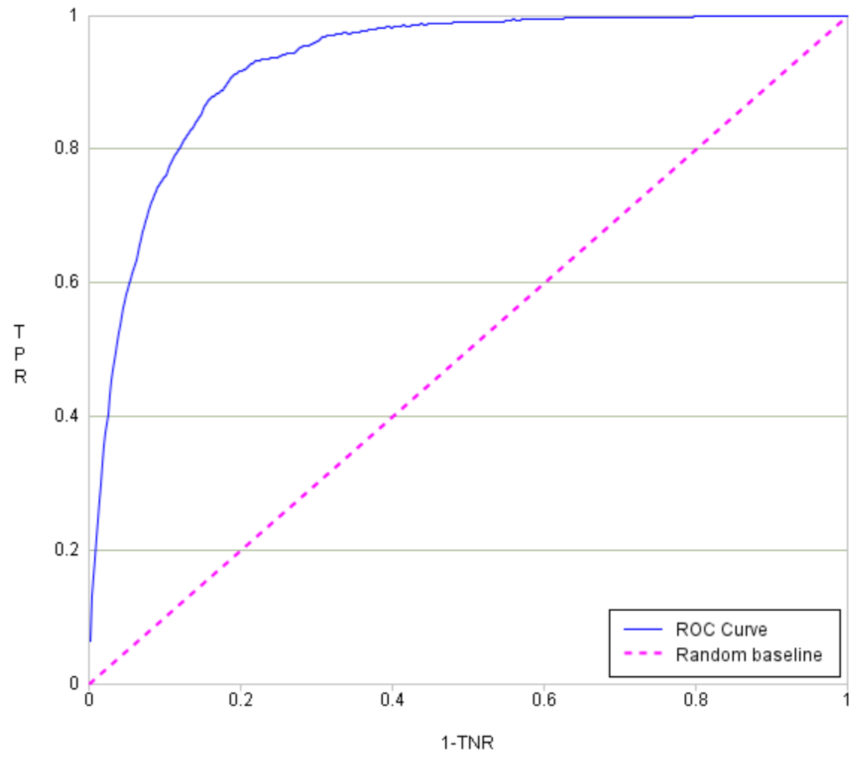
Thinking:
When the threshold value is 0.95, why does the precision and recall rate suddenly change to 0? Why is there no sudden change in accuracy?

ROC curve

In the binary classification model (also known as classifier), for the definition of positive and negative examples, a threshold value is usually set. If the value is greater than the threshold value, it is a positive class, and if the value is less than the threshold value, it is a negative class. If we reduce the threshold value, more samples will be identified as positive class, and the recognition rate of positive class will be improved, but at the same time, more negative classes will be identified as positive class by mistake.

In order to express this phenomenon intuitively, ROC is introduced. According to the classification results, the corresponding points in ROC space are calculated, and the ROC curve is formed by connecting these points.

The horizontal coordinate of ROC curve is "1-TNR" and the vertical coordinate is "sensitivity" (TPR).

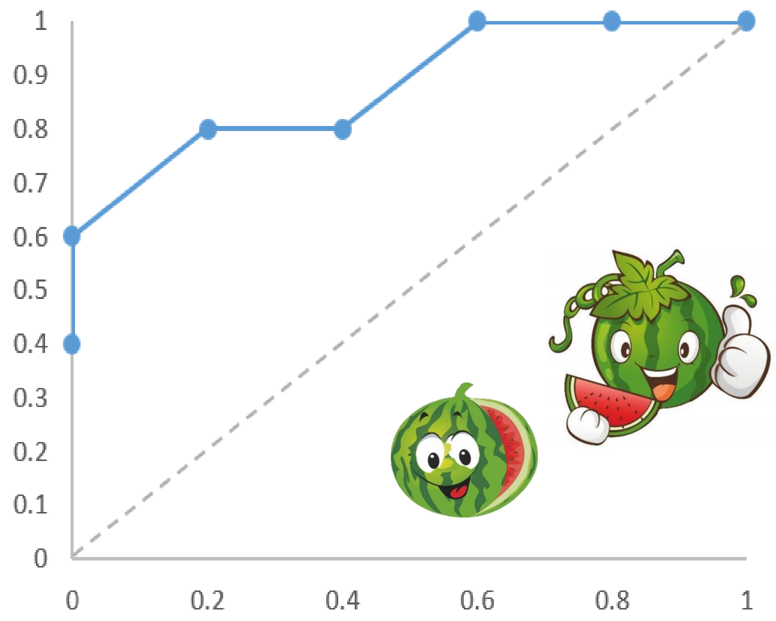


ROC curve

For example, the prediction probability and real situation of good and bad melons are shown in the left table. If 0.9 is selected as the threshold value, only No.1 and No.2 melons will be judged as good melons, and No3-10 melons will be judged as bad melons. According to the confusion matrix, the values of vertical coordinate (sensitivity) and horizontal coordinate (1-TNR) can be calculated as 0.4 and 0 respectively, and so on, the corresponding coordinate values under different thresholds can be calculated as the right table. Connecting these points together is the ROC curve of the model.

No.	**Prediction probability**	**Real value**
1	0.9	1
2	0.9	1
3	0.8	1
4	0.7	0
5	0.7	1
6	0.6	0
7	0.5	1
8	0.5	0
9	0.3	0
10	0.2	0

Threshold	**Sensitivity**	**1-TNR**
0.9	0.4	0
0.8	0.6	0
0.7	0.8	0.2
0.6	0.8	0.4
0.5	1	0.6
0.3	1	0.8
0.2	1	1



ROC curve

ROC curve

Four points and a line in the ROC curve:

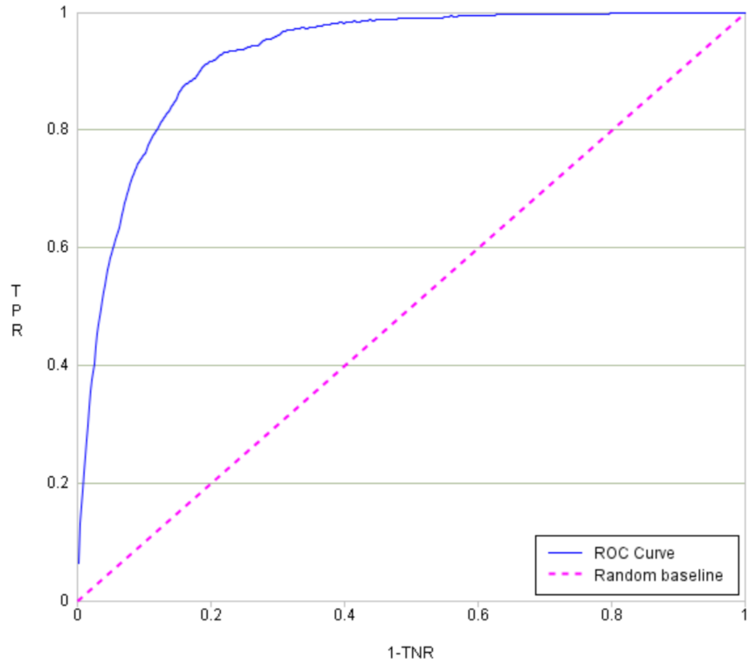
Point (0,1): that is, 1-TNR = 0, TPR = 1, which means FN = 0 and FP = 0. All samples are correctly classified.

Point (1,0): that is, 1-TNR = 1, TPR = 0, worst classifier, avoiding all correct answers.

Point (0,0): that is, 1-TNR = TPR = 0, FP = TP = 0. The classifier predicts every instance as a negative class.

Point (1,1): the classifier predicts every instance as a positive class.

A 45 ° straight line: random model



The closer the ROC curve is to the upper left corner, the better the effect of the classifier is.

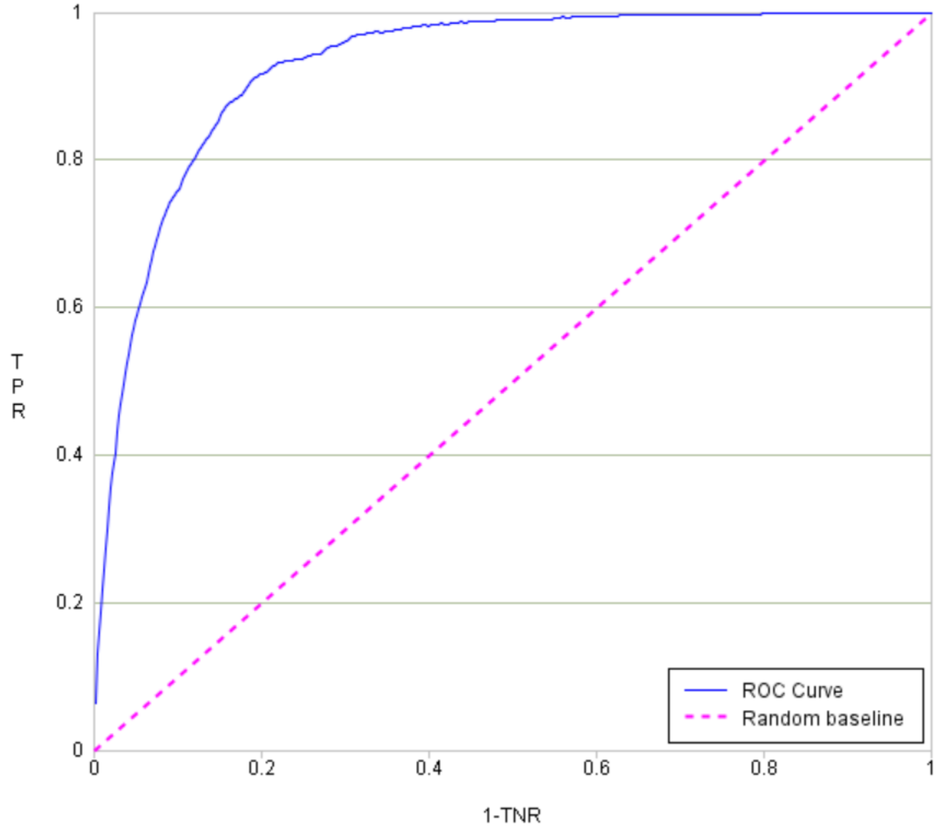
ROC curve

ROC curve is a non decreasing curve.

The ROC curve of the perfect model will follow two axes (from the origin (0,0) to (0,1) and then from (0,1) to (1,1)).

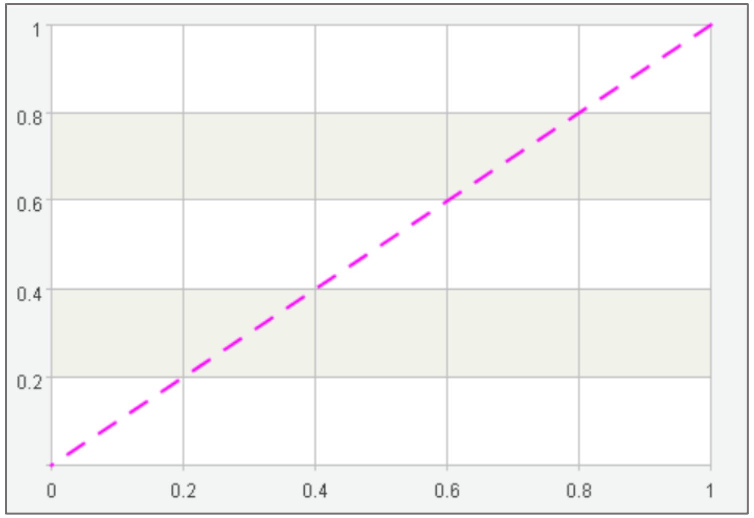
ROC curve of random model is a straight line of 45°

ROC curve better than random model should be between the two cases.

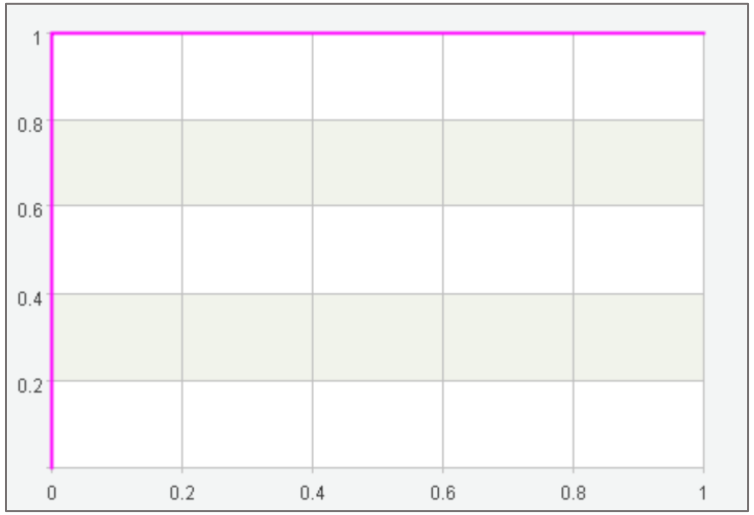


AUC

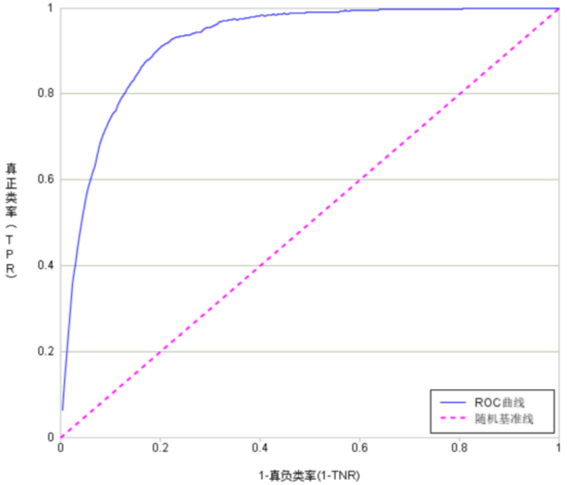
AUC (area under curve) is defined as the area under ROC curve (the integral of ROC), which is usually greater than 0.5 but less than 1.
The classifier with larger AUC value (area) has better effect, as shown in the figure below:



Random model
AUC=0.5
Random guess results
AUC < 0.5 indicates that the model is not as good as coin tossing



Perfect model
AUC=1
All predictions are correct
But if AUC = 1, it's probably over fitting



Normal ROC curve 0.5<AUC<1
Find some data laws, and not over fitting
This is the model that can be used basically

AUC

AUC = 0.5: the prediction model is the same as the random model, i.e. the discrimination between positive and negative samples is not better than the random model.

$0.50 < AUC \leq 0.65$: poor

$0.65 < AUC \leq 0.80$: medium

$0.80 < AUC \leq 0.90$: good

$0.90 < AUC \leq 1.00$: excellent

Index application

AUC can effectively measure the quality of the model, which is the most commonly used model index. However, we must pay attention to the problem of over fitting when using AUC.

For example, the following table has two models, model A and model B. AUC values are shown in the table. Which model should be selected?

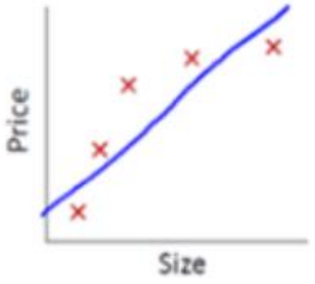
Model performance		
	Training set AUC	Test set AUC
Model A	0.961	0.924
Model B	0.918	0.911

Index application

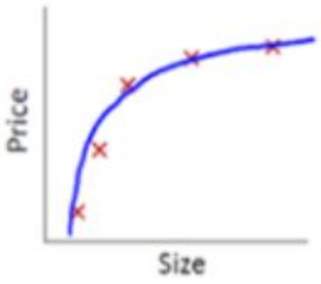
If only judging from AUC value, it is better to use model A, but further analyzing AUC of model in training set and test set, it is found that model A decays very fast, indicating that it is an over fitting model. While although AUC of model B is lower, its generalization ability in unknown data is better.
Therefore, from the perspective of application, model B is more stable.

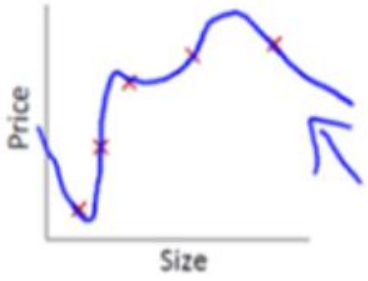
Model performance			
	Training set AUC	Test set AUC	Model attenuation
Model A	0.961	0.924	0.037
Model B	0.918	0.911	0.007

Model attenuation refers to the difference between the indexes of the model in the training set and the test set, which can indicate the degree of decline of the model indexes in the prediction set (i.e. unknown data).



Under fitting





Over fitting

Index selection

As for the importance of selecting evaluation indexes, sometimes it's not that we didn't make a good model, but that we didn't choose the right evaluation indexes.

Take precision and recall as examples. If your model focuses on the problem of “Of all the thieves, how many are caught", then you will care about recall. If your model focuses on "how many of the suspects are real thieves", then you should use precision.

For different scenarios and different problems, the evaluation indexes are different.

Thinking

What is the difference and connection between AUC and accuracy?

Thinking

What is the difference and connection between AUC and accuracy?

First of all, AUC does not correspond to an accuracy, but a series of accuracy. AUC is the "offline area" of ROC, while ROC is a line with FPR-TPR as the coordinate, which is actually a polyline connecting a series of scattered points. Each point on the polyline corresponds to a threshold, and the predicted value determined by the threshold and its measurement of accuracy, precision, recall, etc.

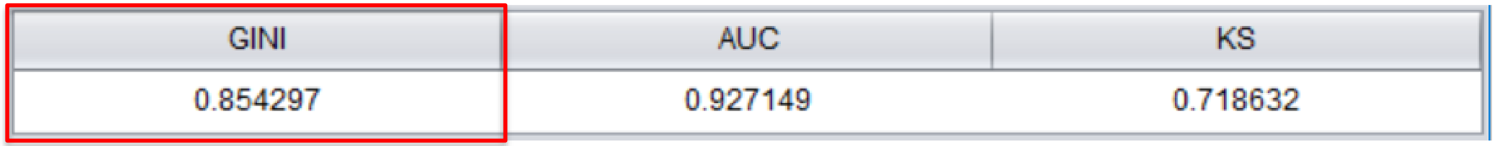
Therefore, AUC measures the quality of a model, which is how reasonable it is to rank all samples (whether it correctly ranks the negative examples ahead of the positive ones); while accuracy measures the prediction accuracy (whether the negative cases are correctly ranked before the threshold and the positive cases after the threshold) of a model at a specific threshold (for example, logistic regression model under threshold of ½).

Gini index

Gini index is usually used in insurance rate making and credit risk management system.

$$Gini\ Index = 2 \times (AUC - 0.5)$$

Using the same data to model, the higher the Gini index, the better the model is in the sense of separating data.

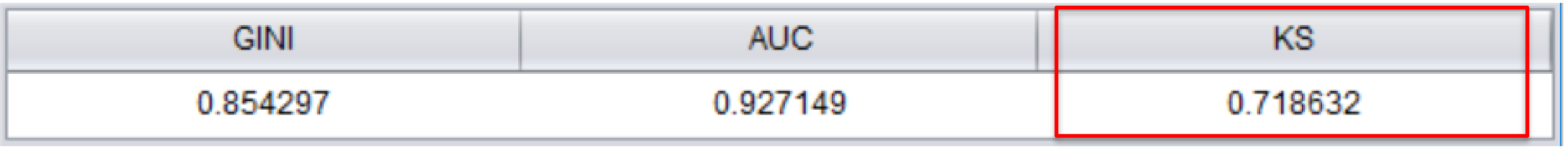


GINI	AUC	KS
0.854297	0.927149	0.718632

$Gini \geq 0.8$: the model is excellent. However, you need to check whether the model is over fitted.
$0.8 > Gini \geq 0.6$: very good model
$0.6 > Gini \geq 0.3$: reasonable model
$0.3 > Gini \geq 0$: there is no difference between the model and the random model, and it needs to be completed.

KS

KS(Kolmogorov-Smirnov): KS value can be used to evaluate the prediction model. Used to measure the ability of the model to distinguish positive and negative samples. The larger the KS value is, the stronger the ability of the model to distinguish positive and negative samples is.



GINI	AUC	KS
0.854297	0.927149	0.718632

$ks \geq 0.3$: ***the model has good predictability.***
$0.3 > ks \geq 0.2$: ***the model is usable.***
0.2> ks ≥0: poor prediction ability of the model
$ks < 0$: ***the model is incorrect***

Generally, if the negative samples have a great impact on the business, then the differentiation must be very important. At this time, K-S is more suitable for model evaluation than AUC. If there is no special impact, then AUC is good.

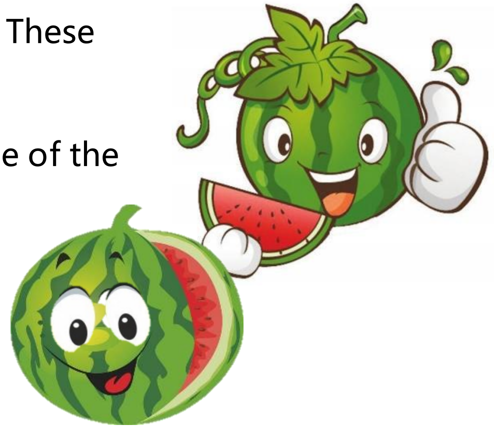
Lift graph

Lift is a measure to evaluate the effectiveness of a prediction model. Its value is the ratio between the results obtained with and without the prediction model.

Suppose there are 100 watermelons, of which 50 are good melons, 50 are bad melons, and the rate of good melons is 50%. These watermelons were predicted by using the model, and arranged in descending order according to the predicted probability,

8 of the top 10 melons are really good melons, and the proportion of correct prediction is 0.8, then the improvement degree of the model in the top 10% melons is 0.8 / 0.5 = 1.6

That is to say, for the top 10% of melons, using the model will be 1.6 times better than random grasping.



Good melon rate	Top X%	Number of melons	Accumulated samples	Number of good melons	Good melon rate	Accumulated good melons	Accumulated good melon rate	Lift	Accumulated lift
0.5	10%	10	10	8	0.8	8	0.8	1.6	1.6
	20%	10	20	7	0.7	15	0.75	1.4	1.5
	30%	10	30	6	0.6	21	0.7	1.2	1.4
	40%	10	40	6	0.6	27	0.675	1.2	1.35
	50%	10	50	5	0.5	32	0.64	1	1.28
	……								

Index application

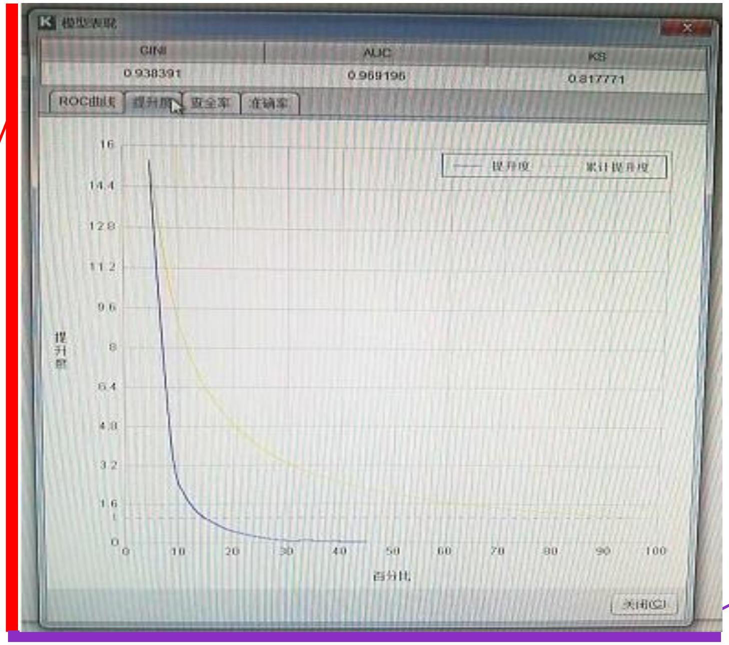
Lift is particularly suitable for targeted marketing scenarios

For example, in a product telemarketing scenario, there are 1 million potential customers, and the purchase rate of customers is 1.5%, that is to say, an average of **1.5** randomly selected **100** customers will buy the product.

After using the model, the lift of the top 5% of the predicted probability is14.4, that is to say, **21.6** (1.5 * 14.4) people in the top 5% of the customers will buy the product, far higher than the randomly selected 1.5 people, greatly improving the marketing efficiency and reducing the ineffective marketing actions.

	Accumulated lift
Top 5%	14.4
Top 10%	9.4
Top 15%	6.3
Top 20%	4.8
Current product purchase rate is 1.5%	

Vertical axis: lift



GINI	AUC	KS
0.938391	0.969196	0.817771

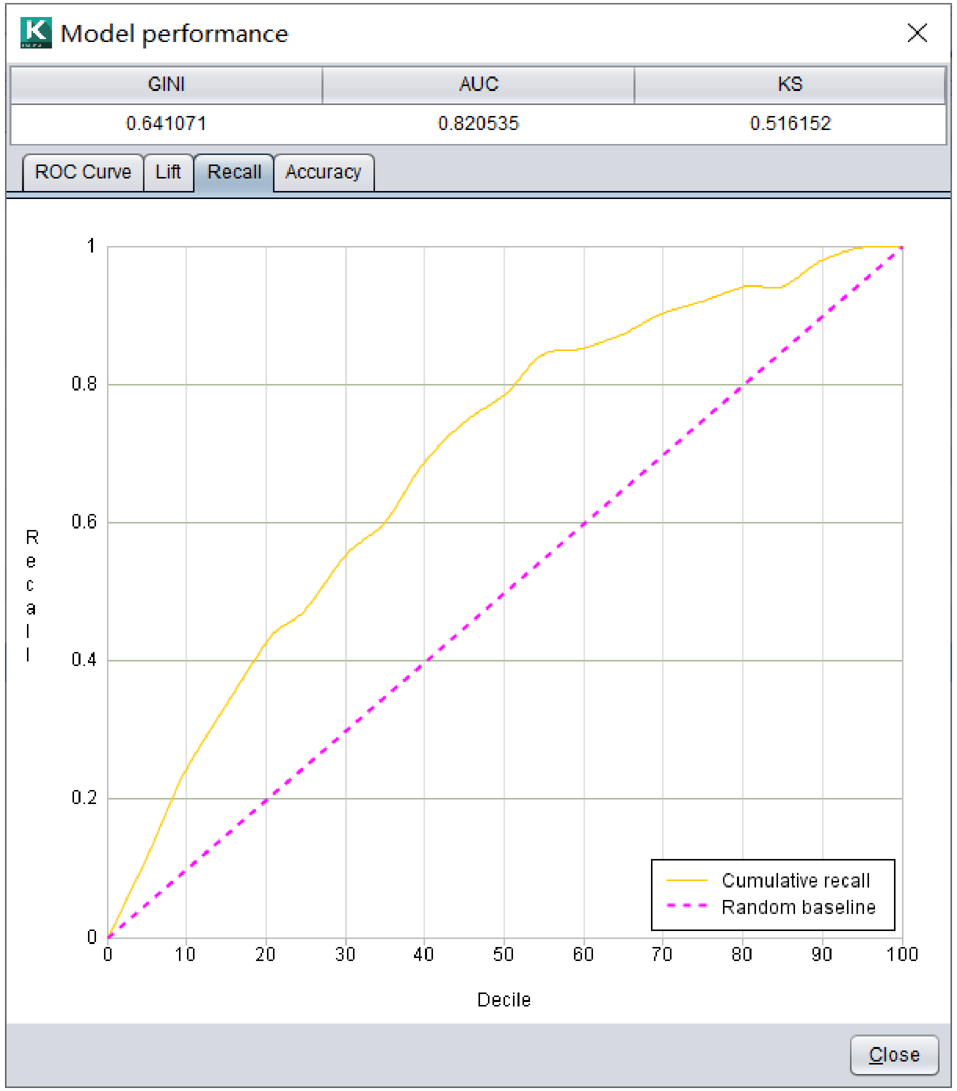
Horizontal axis: grouping

Lift diagram

Recall chart

When the number of positive and negative samples is relatively balanced, ROC curve performs very well. However, the data of many business applications are unbalanced. For example, only less than 1% of automobile insurance policyholders will make a claim at a certain time.

In the case of unbalanced datasets, ROC curves may be misleading. In this case, recall chart can be used.

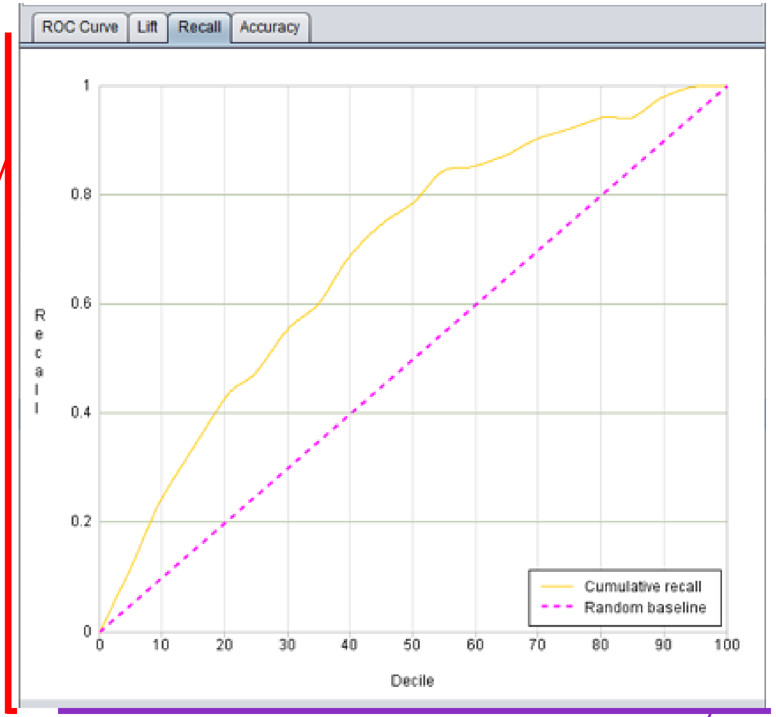


Model performance

GINI	AUC	KS
0.641071	0.820535	0.516152

Recall chart

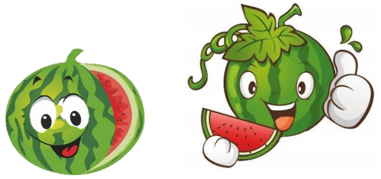
Recall chart shows the model' s ability to find positive samples (good melon), which is mainly used in the scene of data imbalance. For example, only 5 out of 100 watermelons are good ones and 95 are bad ones. Even if all of them are predicted as bad ones, the accuracy of this model is very good (up to 95%), but it is obviously meaningless. At this time, we can use recall chart to evaluate the model. The larger the slope is, the faster the search speed is, the better the model is.



Vertical axis: cumulative recall

Recall chart

Horizontal axis: sorting and grouping by output probability



The cumulative recall is the ratio of the cumulative positive samples of each group to the total positive samples. The calculation process is as follows :

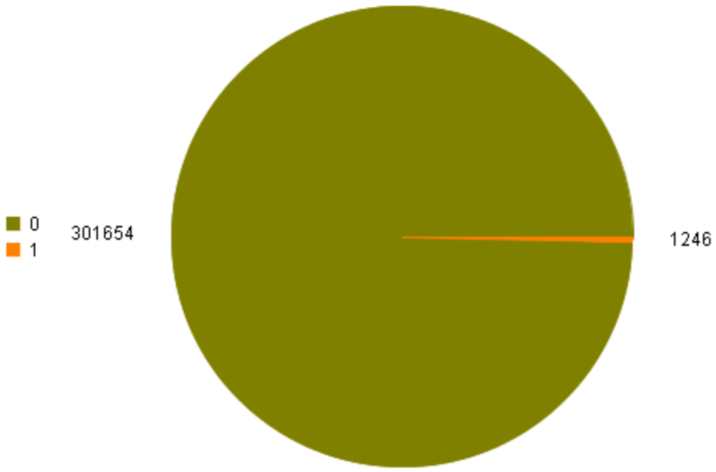
Number of good melons	Top x%	Number of melons	Accumulated samples	Number of good melons	Accumulated number of good melons	Cumulative recall
5	10%	10	10	2	2	0.4
	20%	10	20	1	3	0.6
	30%	10	30	0	3	0.6
	……					

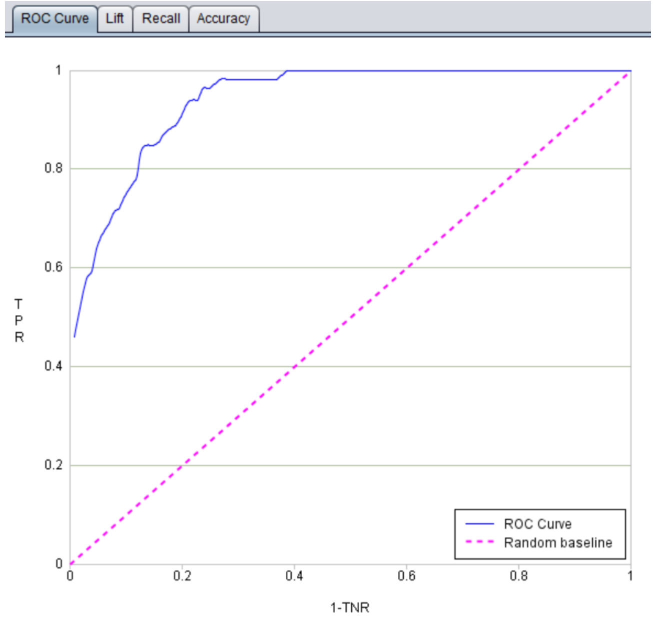
Index application

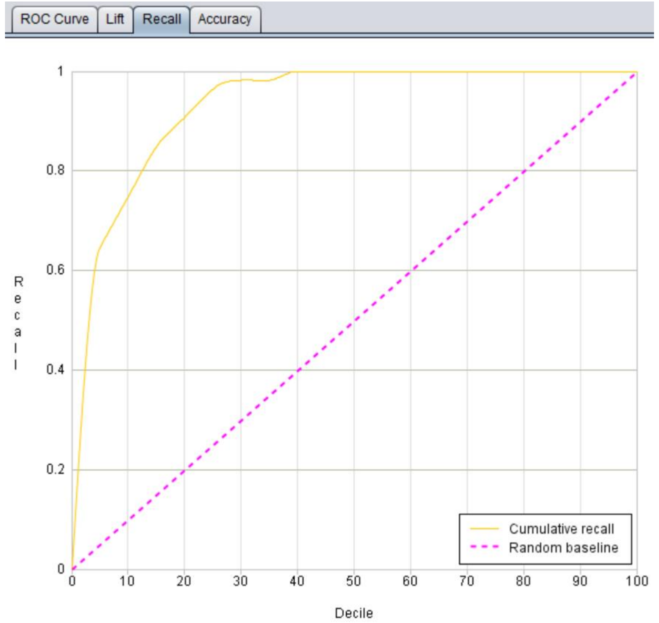
ROC curve reflects the comprehensive performance of the model, but in some unbalanced samples, it is often concerned about how to find the few samples.

For example, in the insurance claim risk scenario shown in the figure, there are only 1246 claims in more than 300000 insurance policies, and the positive sample ratio is only 0.4%. For insurance companies, what they care about is how to quickly find these high-risk customers, so as to take measures to reduce the claim risk loss.

In this case, the ROC curve can not directly meet the target demand, you need to rely on the recall chart, as shown in the figure, you can see that the top 10% of the data can capture about 75% of high-risk customers.

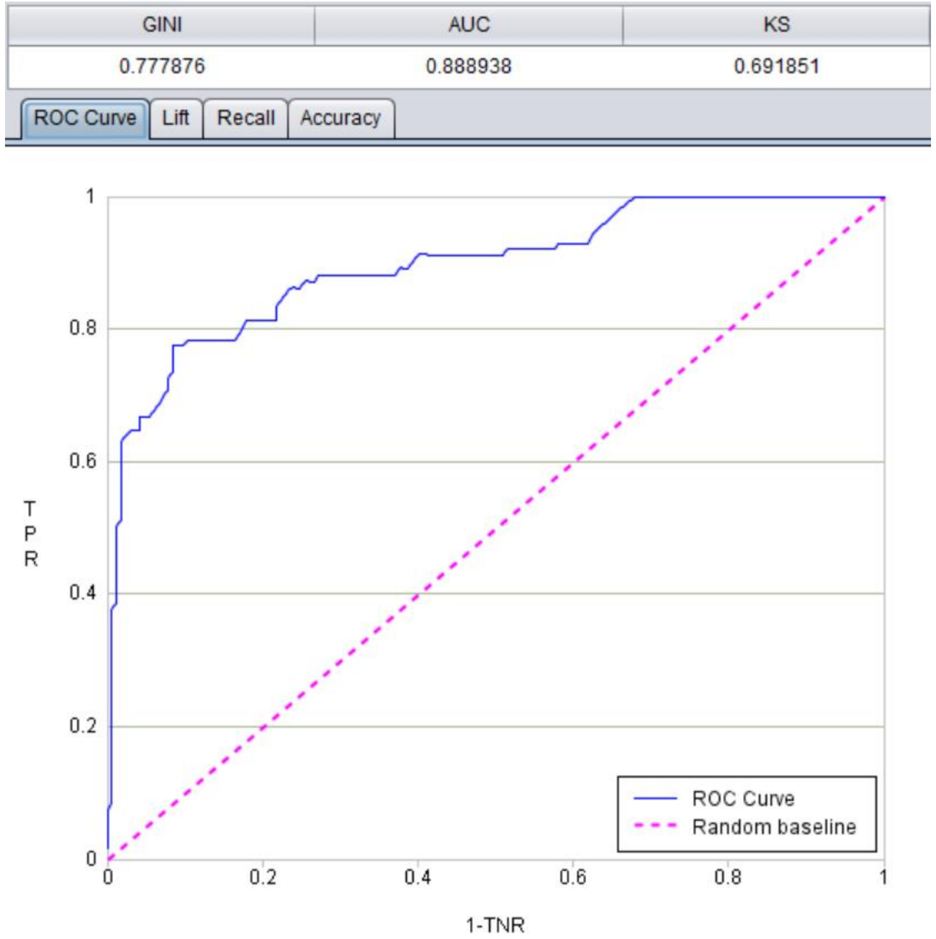




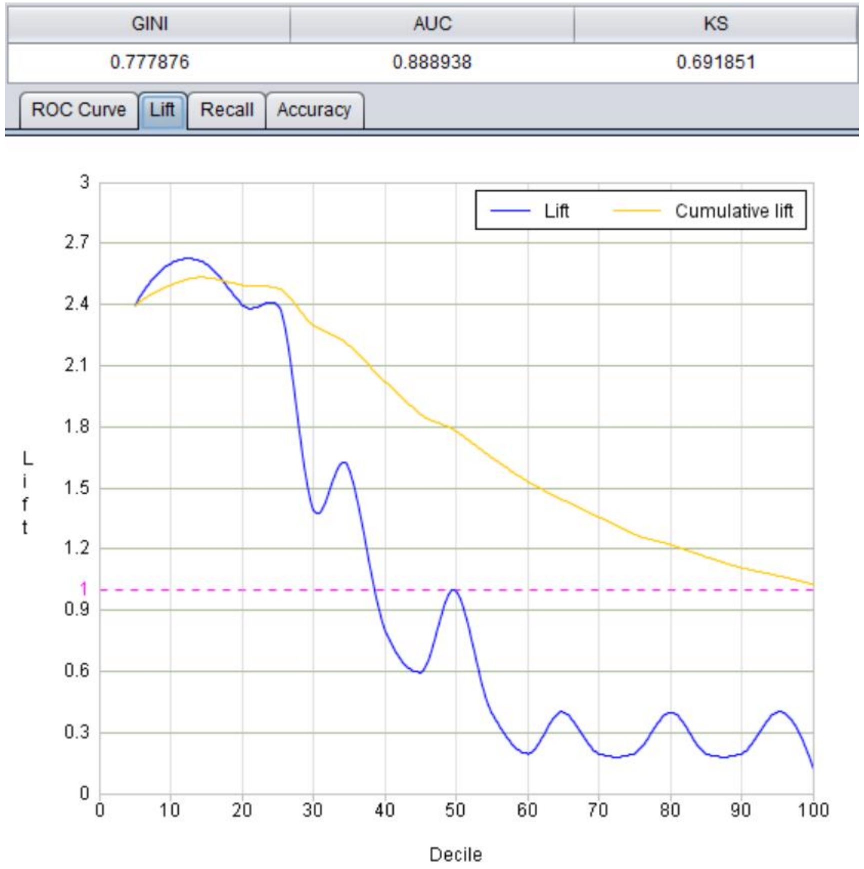


Class exercise

Use YModel to model Titanic's data, and then view various model indexes.



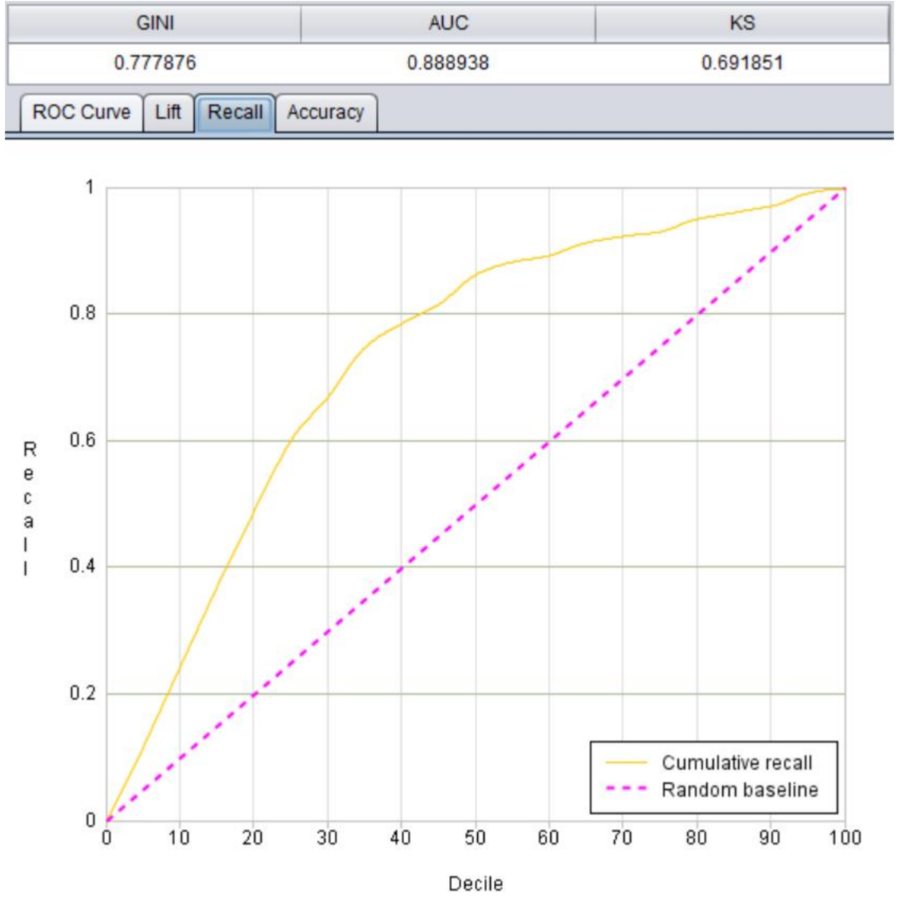
GINI	AUC	KS
0.777876	0.888938	0.691851



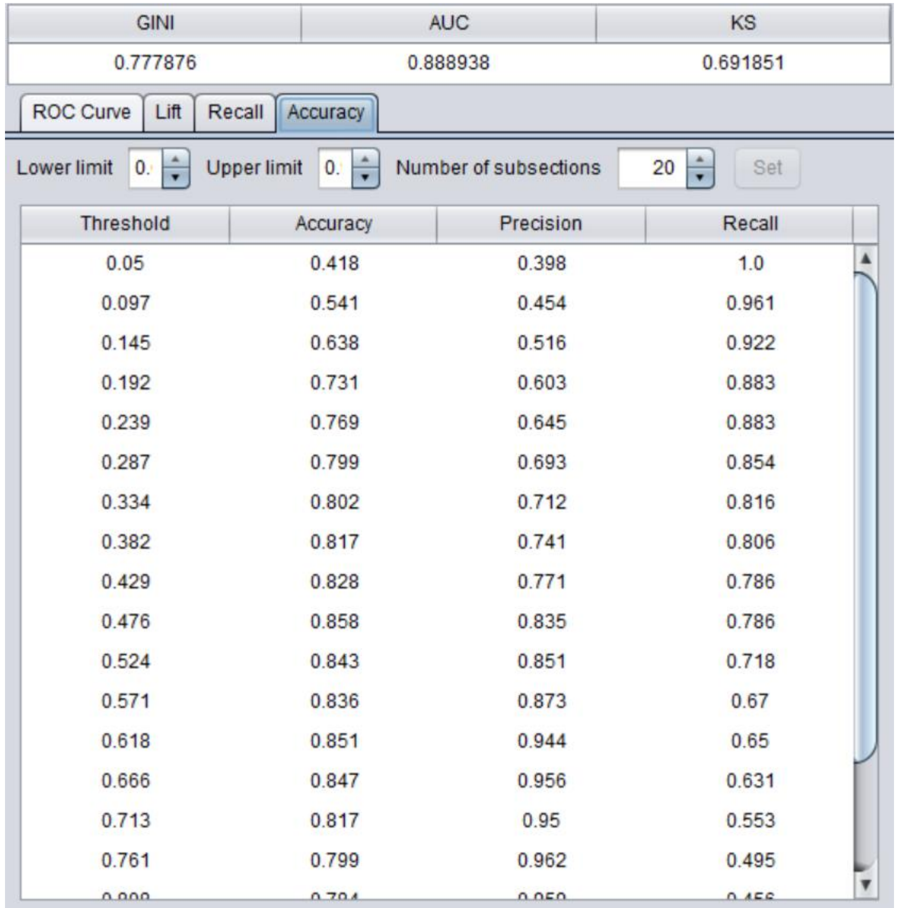
GINI	AUC	KS
0.777876	0.888938	0.691851

Class exercise

Use YModel to model Titanic's data, and then view various model indexes.



GINI	AUC	KS
0.777876	0.888938	0.691851



GINI	AUC	KS
0.777876	0.888938	0.691851

Lower limit 0. Upper limit 0. Number of subsections 20

Threshold	Accuracy	Precision	Recall
0.05	0.418	0.398	1.0
0.097	0.541	0.454	0.961
0.145	0.638	0.516	0.922
0.192	0.731	0.603	0.883
0.239	0.769	0.645	0.883
0.287	0.799	0.693	0.854
0.334	0.802	0.712	0.816
0.382	0.817	0.741	0.806
0.429	0.828	0.771	0.786
0.476	0.858	0.835	0.786
0.524	0.843	0.851	0.718
0.571	0.836	0.873	0.67
0.618	0.851	0.944	0.65
0.666	0.847	0.956	0.631
0.713	0.817	0.95	0.553
0.761	0.799	0.962	0.495

Regression model index

How to evaluate the model?

Regression model evaluation

Common indexes：
R^2，MSE，RMSE，GINI，MAE，MAPE, etc.

Graphs：
Residual diagram, result comparison diagram

Regression model index

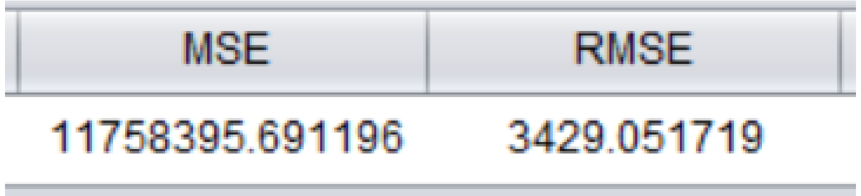
Evaluation index	Description
MSE	Mean of the sum of squares of the deviation between the predicted value and the true value
RMSE	The square root of MSE. The order of magnitude is the same as the real value, for example RMSE = 10. It can be considered that the average difference between the regression effect and the real value is 10.
MAE	The average of the absolute value of the deviation between the predicted value and the true value.
MAPE	The average of the ratio of the deviation between the predicted value and the true value to the absolute value of true value . MAPE not only considers the error between the predicted value and the real value, but also the proportion between the error and the real value. In some scenarios, such as the house price is between 50W and 1000W, there is a big gap between the prediction of 50W for 100W and 950W for 1000W. MAPE can evaluate the model well.

The range of these four standards is [0, + ∞), when the predicted value and the real value are completely consistent, it is equal to 0, that is, the perfect model; the larger the error is, the larger these values are, the worse the model is.

Regression model index

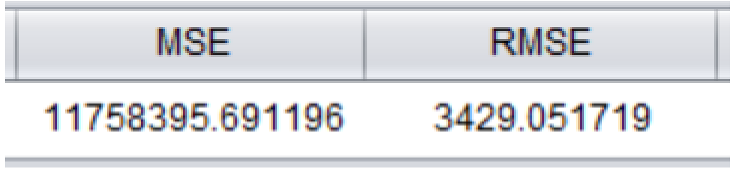
Thinking：

The target variable of the house price prediction data is "saleprice". The MSE and RMSE indexes after modeling are shown in the figure, and their values are very large, indicating that the model quality is poor?

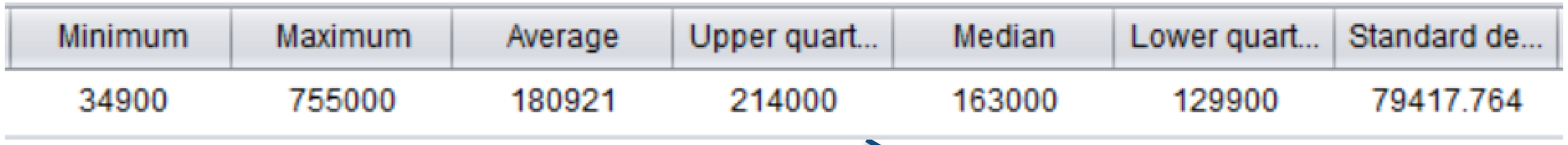


MSE	RMSE
11758395.691196	3429.051719

Regression model index



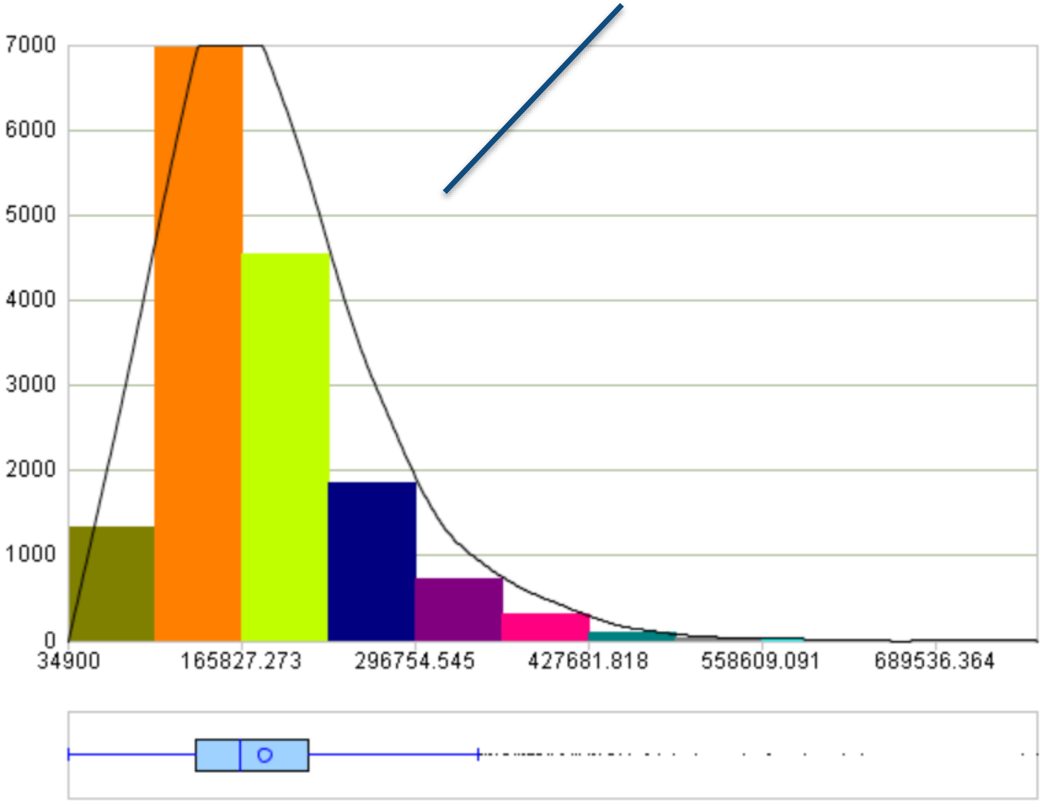
MSE	RMSE
11758395.691196	3429.051719



Minimum	Maximum	Average	Upper quart...	Median	Lower quart...	Standard de...
34900	755000	180921	214000	163000	129900	79417.764

Statistics of target variable

We use the YModel tool to observe the data statistics of the target variable "saleprice". The average value is 180000, the median value is 160000, and the data size level is basically a dozen to several hundred thousand. At such a large data level, RMSE is more than 3000, that is, the difference between predicted value and real value is more than 3000, which can be ignored.



Regression model index

Therefore, there is no unified standard to judge the quality of model indexes, so we must combine the business objectives to make a comprehensive judgment.

In addition, these indexes are calculated on the test set. These indexes can be compared between different models established by the same dataset. When the dataset is different, but these indexes have no comparability. For example, dataset A predicts house prices (between 500,000 and 1,000,000), and dataset B predicts clothes prices (between 50 and 1000). Obviously, the prediction error of data set A must be higher than that of data set B, but it cannot be said that the model built by data set A is not as good as that of data set B.

Regression model index

R ² (r squared, coefficient of determination) reflects the interpretation degree of predicted value to actual value.

R^2 has a range of 0 to 1.

The closer it is to 1, the stronger the ability of the variables of the equation to explain y, and the better the model fits the data.

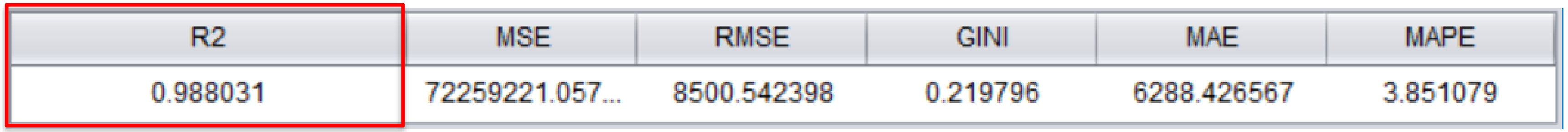
The closer to 0, the worse the model fitting.

R^2 =0.9 indicates that the model explains 90% of the uncertainty, and the model is very good.

Empirical value： >0.4， good fitting effect

R^2 does not mean the square of R, and it may be a negative number: such a model is equal to blind guessing, not as good as directly calculating the average value of the target variable.

In 2015, Chai Jing released "under the sky", in which she wanted to explain the correlation between PM2.5 and mortality, but the video inadvertently showed r^2=0.19. This is a joke(the model is too bad).



R2	MSE	RMSE	GINI	MAE	MAPE
0.988031	72259221.057...	8500.542398	0.219796	6288.426567	3.851079

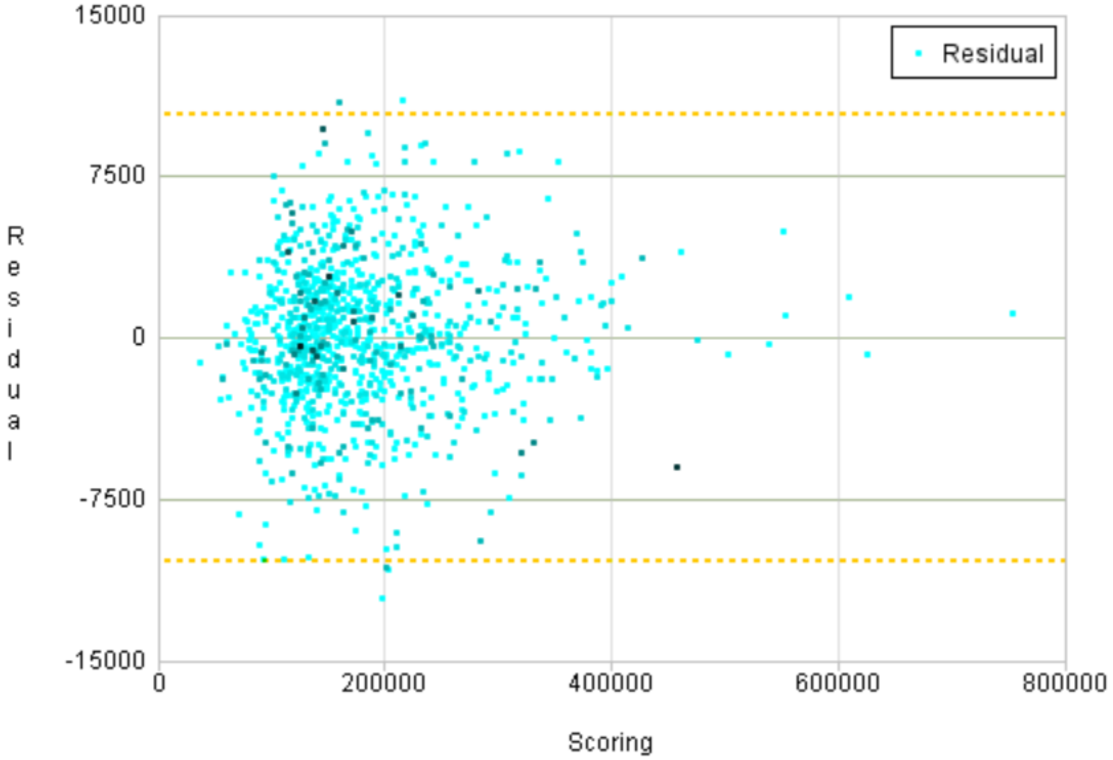
Using the same data to model, the closer test set R^2 is to 1, the stronger the ability to interpret uncertainty, and the better the model.

There is no comparability between R^2 modeled by different datasets.

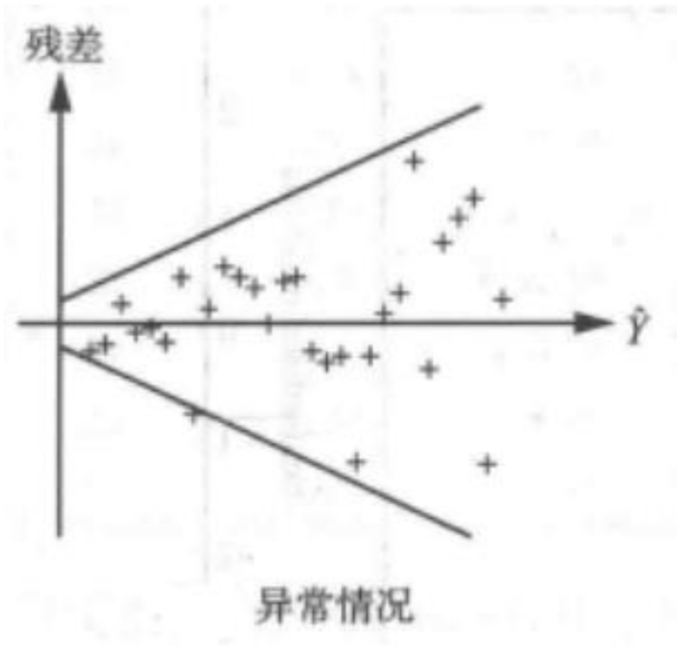
Residual plot

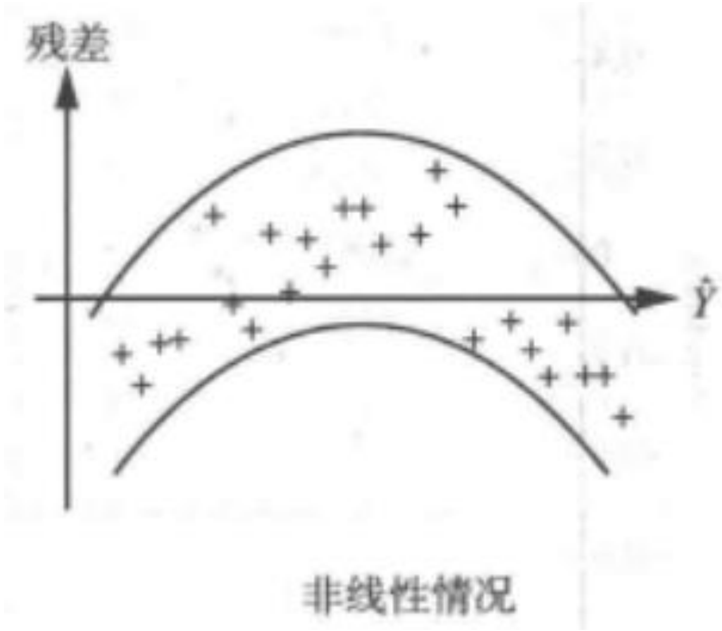
The residual is the difference between the real y value and the y value predicted by the model.
The vertical axis of the residual plot represents the residual and the horizontal axis represents the original value.

It is generally believed that if a regression model satisfies the basic assumption given, all residuals should change randomly near 0 and in a band with small change range.
That is to say, if the residuals are all in a band with a small variation, the model is good.



Residual plot



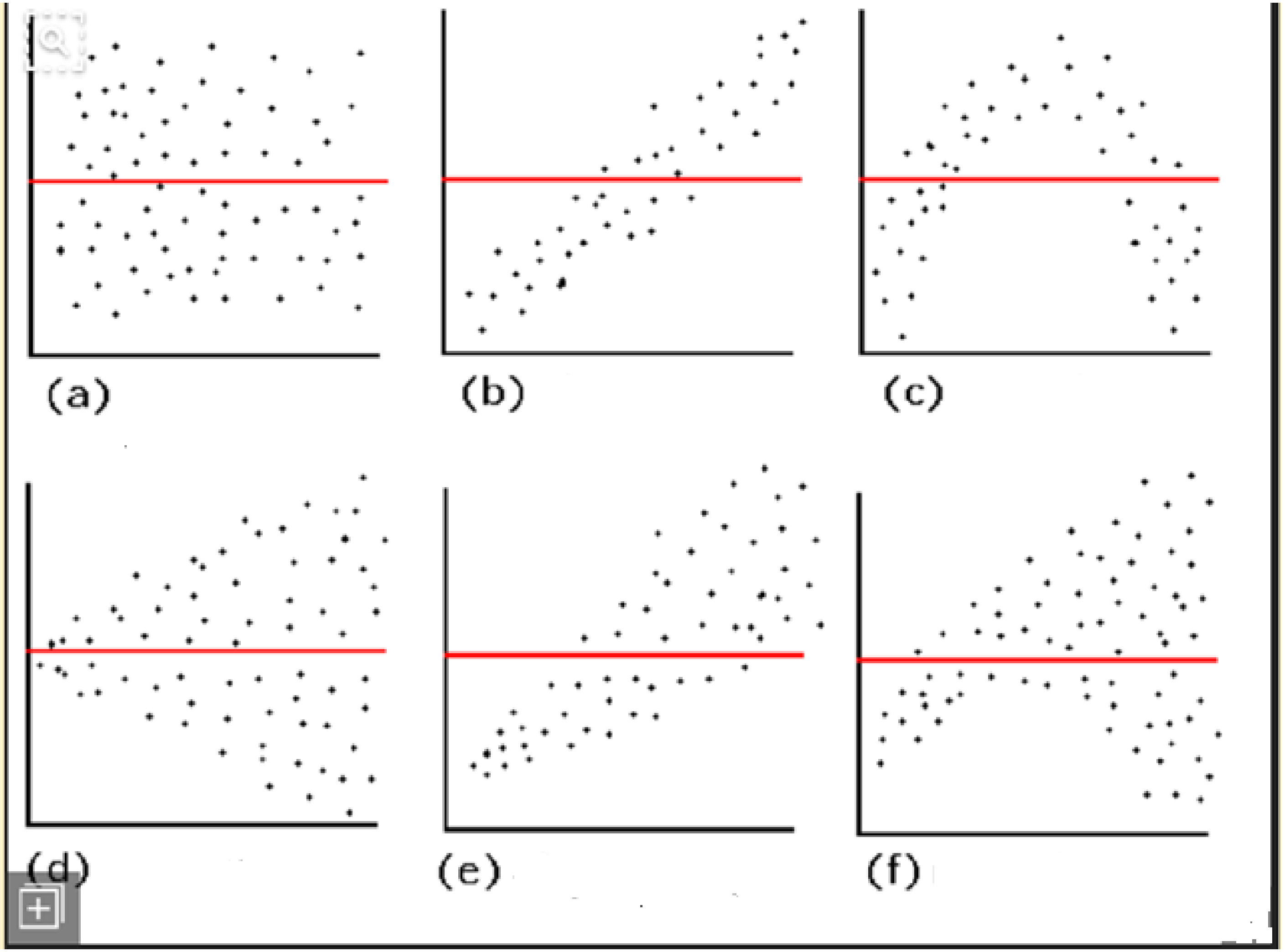


The variance of the residual increases with the increase of the predicted value.
Need to change the form of X.

In the figure above, the non-zero value of the residual can be predicted according to the fitting value, indicating that the model is under fitted, which may be caused by incorrect model structure or the need for additional variables.

Class exercise

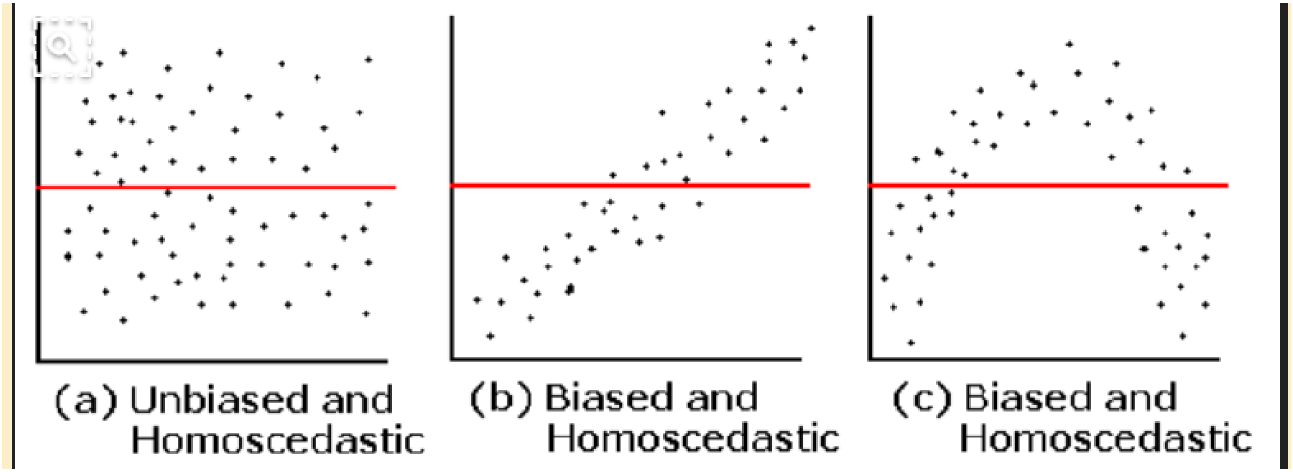
Analyze the residual in the right figure.



Class exercise

1. Normal residual plot
2. The bias of the model indicates that some important variables have not been saved.
3. The model is under fit. The structure of the model is incorrect. The model should contain quadratic terms.

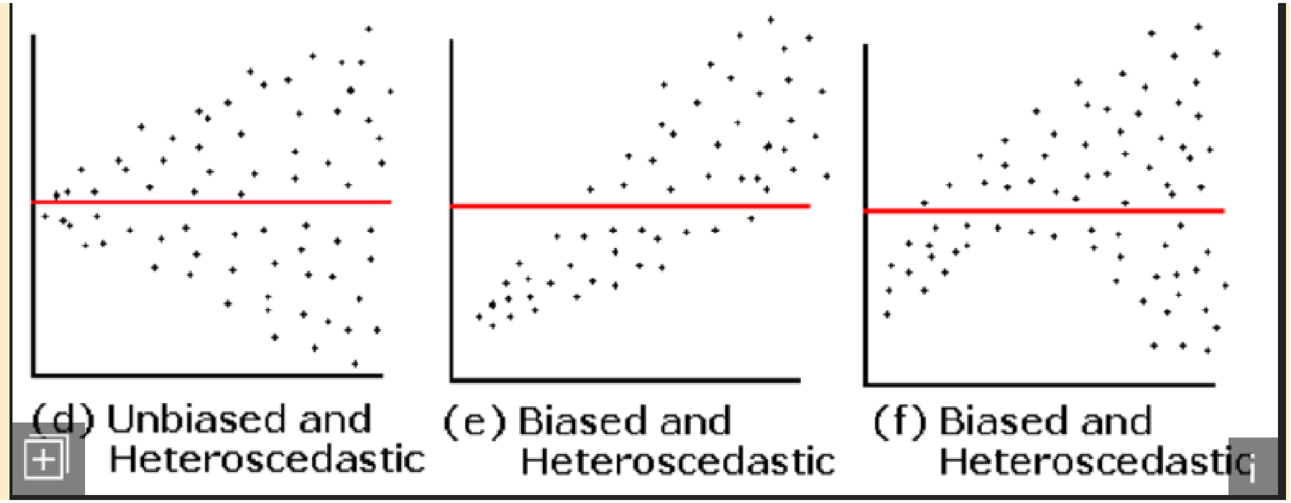
The variance of these three models is constant.



Class exercise

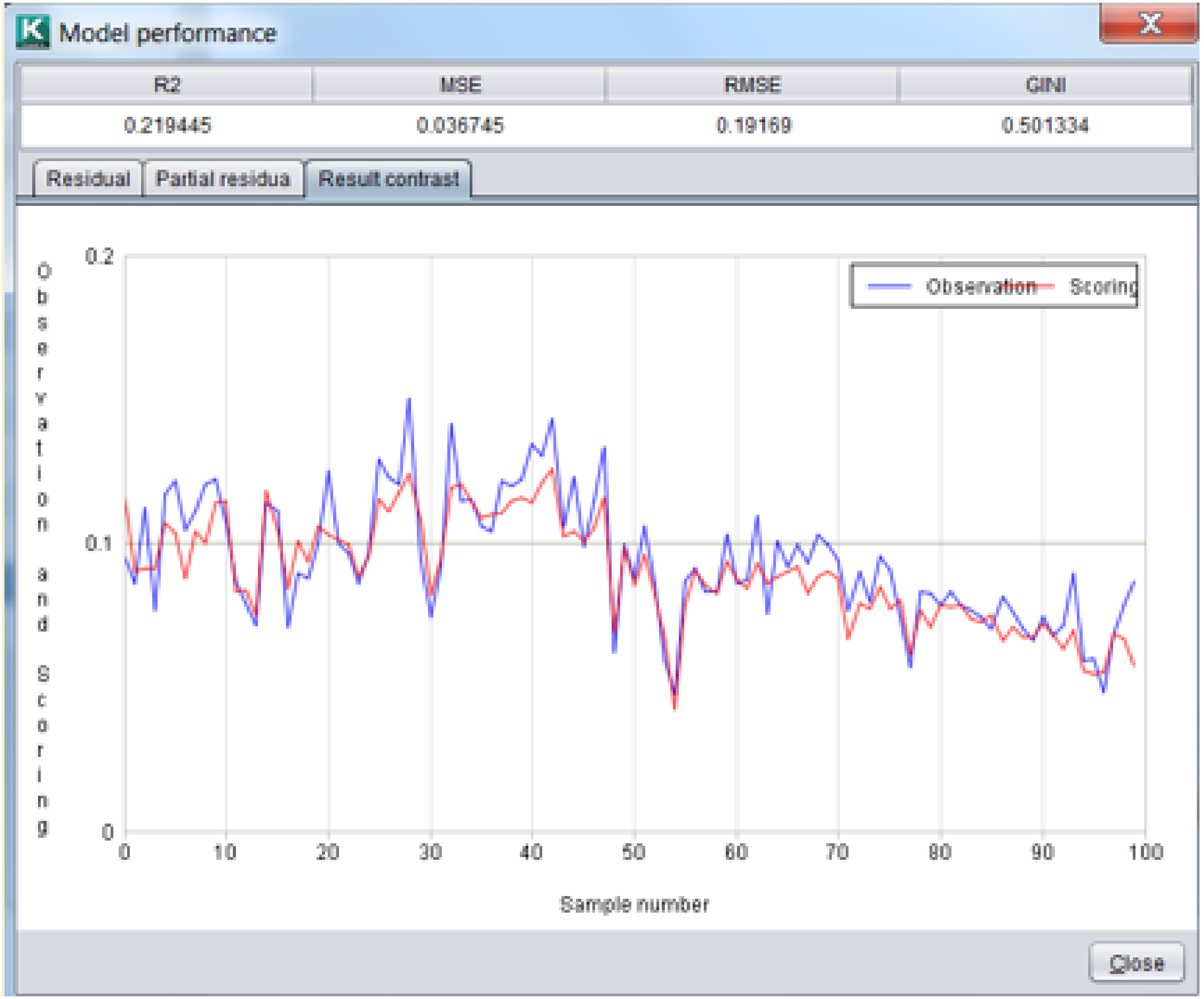
4. The model indicates that the larger the predicted value is, the greater the variance is, namely "heteroscedasticity". However, the model is unbiased.
5. The model has biased variance, indicating the lack of some important variables.
6. The model is not fit. In addition, the model has biased variance. The model is not structured correctly and some important variables are missing.

The variance of these three models is not constant.



Result comparison graph

The fitting of the model can also be observed by comparing the real value with the predicted value.



Model performance

R2	MSE	RMSE	GINI
0.219445	0.036745	0.19169	0.501334

Index application

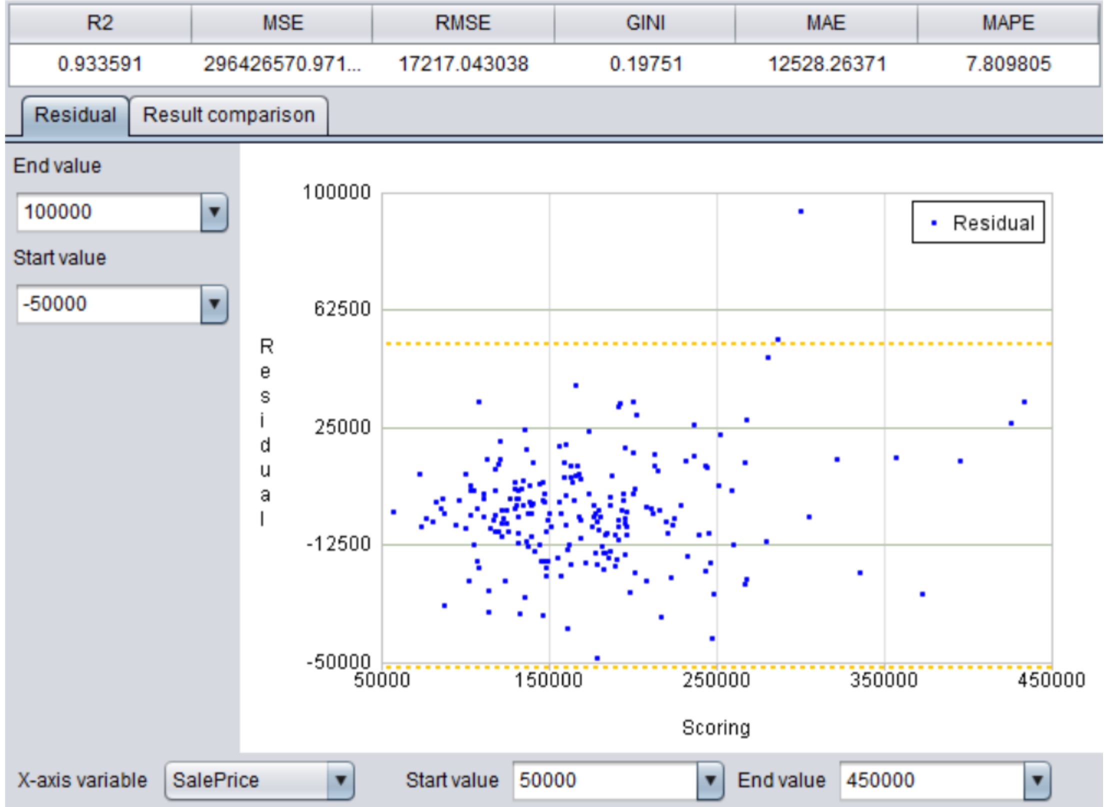
Similarly, the regression model should also consider whether the model is over fitted, that is, whether the model is stable on the unknown data. For example, RMSE changes in training and test sets can be used to represent the attenuation of the model.

Model performance			
	Training set RMSE	Test set RMSE	Model attenuation
Model A	3389	3452	63

The RMSE of the model in training set and test set is almost unchanged, which shows that the stability of the model is good.

Class exercise

Use YModel to model the data of house price prediction, and then view various model indexes and graphs.



R2	MSE	RMSE	GINI	MAE	MAPE
0.933591	296426570.971...	17217.043038	0.19751	12528.26371	7.809805

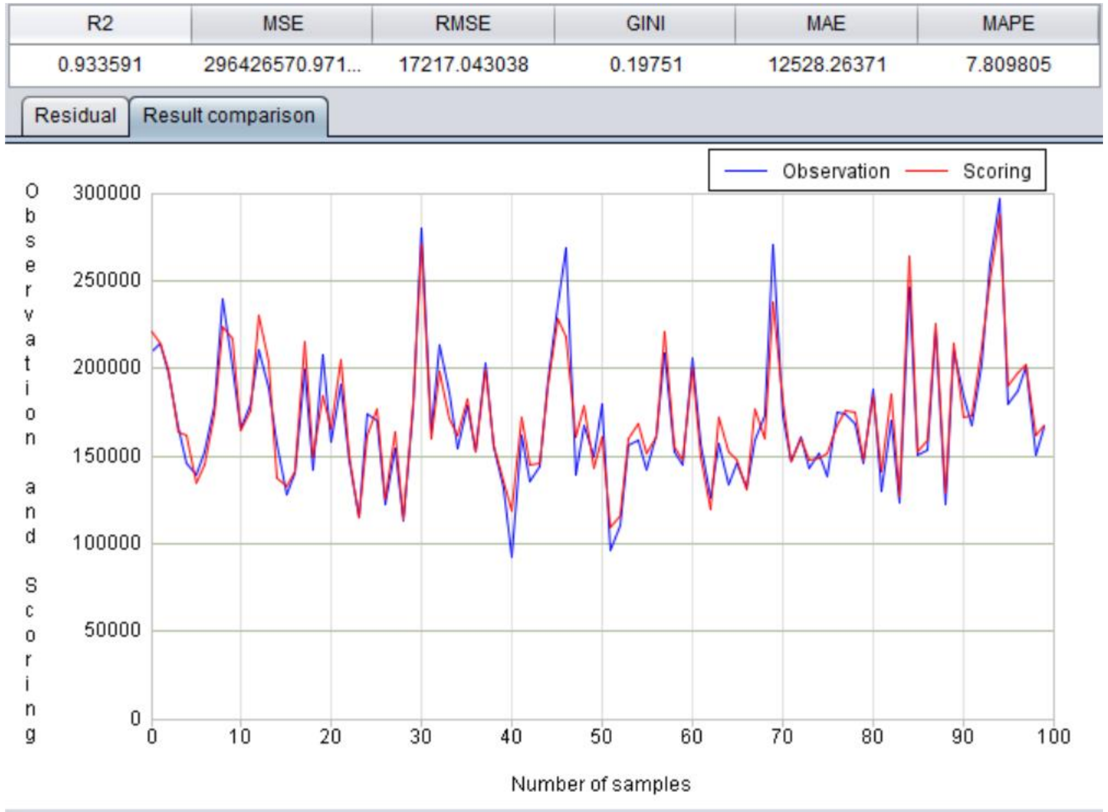
End value: 100000

Start value: -50000

X-axis variable: SalePrice

Start value: 50000

End value: 450000



R2	MSE	RMSE	GINI	MAE	MAPE
0.933591	296426570.971...	17217.043038	0.19751	12528.26371	7.809805

Glossary – for reference

confusion matrix
accuracy
precision
sensitivity/recall
specificity
threshold
ROC curve(receiver operating characteristic)
AUC(area under curve)
attenuation
Gini index
KS(Kolmogorov-Smirnov)

lift
MSE(mean squared error)
RMSE(root mean squared error)
MAE(mean absolute error)
MAPE(mean absolute percentage error)
R squared/ coefficient of determination)
residual

Chapter 6 Model tuning

Model optimization method

The modeling process is not accomplished overnight. Good models often need to be polished repeatedly.

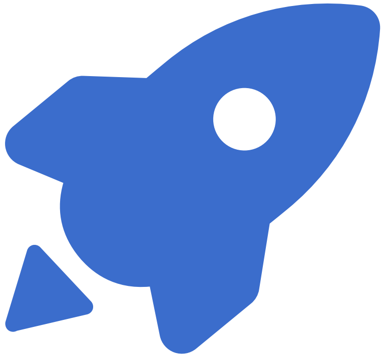
This chapter will introduce several commonly used methods of optimizing models:

- Add derived variable
- Choose different algorithms
- Optimization of algorithm parameters

Derived variable

Derived variable refers to the variable generated by the change of original variable. Good derived variable can effectively improve the model effect, but how to generate good derived variable has no fixed rules and needs repeated attempts. This course only introduces some common methods for readers to take part in.

The extraction of derived variables depends on business domain knowledge, which can be said to be a mathematical representation of business logic. Therefore, business data and business logic should be understood first. Feature extraction can also be regarded as the process of describing business logic by features. The goal of feature extraction is to describe business accurately and comprehensively.



Binning

Data binning is to classify data according to some rules, which is usually used for discretization of continuous data.

The commonly used methods of binning are as follows:

Equi-width binning

Equi-frequency binning

Custom binning rules

Binning- Binning methods

Equi-width binning	The range of variable value is divided into k equal width intervals, each of which is regarded as a bin. Here, only boundary is considered, and the sample size in each bin may be different..
Equi-frequency binning	The variable values are arranged in the order of small to large. According to the number of samples in the dataset, they are equally divided into k parts. Each part is treated as a bin. For example, if the number of bins is 10, each bin contains about 10% of the samples.
Custom binning	Data binning according to business understanding

Binning

For example, the table shows the income data of a group of people, which can be divided into three levels: high, medium and low according to the three ways of binning.

(1) **Equi-width binning**: there are three bins with equal distance between 36 and 4023. There are two samples in the bin with high income, the rest are all low income, and there are no samples in the bin with medium income.

(2) **Equi-frequency binning**: a total of 15 samples, sorted according to the income level, each bin has 5 samples.

income	Equi-width binning	Equi-frequency binning	Custom binning
4023	High	High	High
3274	High	High	High
227	Low	High	Medium
129	Low	High	Medium
154	Low	High	Medium
269	Low	Medium	Medium
463	Low	Medium	Medium
196	Low	Medium	Medium
190	Low	Medium	Medium
232	Low	Medium	Medium
90	Low	Low	Low
84	Low	Low	Low
65	Low	Low	Low
90	Low	Low	Low
36	Low	Low	Low

Binning

（3）**Custom binning rules**:

income>=1000，high-income population

100<income<1000，medium-income population

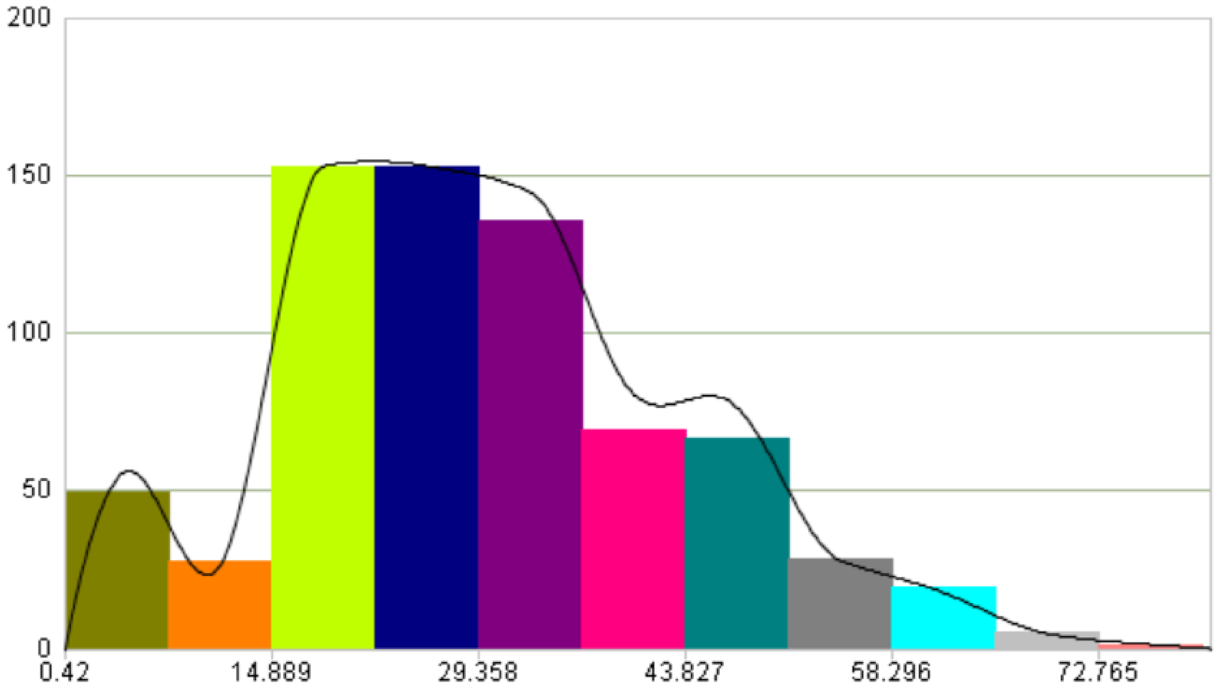
income<=100，low-income population

After binning, as shown in the table

income	Equi-width binning	Equi-frequency binning	Custom binning
4023	High	High	High
3274	High	High	High
227	Low	High	Medium
129	Low	High	Medium
154	Low	High	Medium
269	Low	Medium	Medium
463	Low	Medium	Medium
196	Low	Medium	Medium
190	Low	Medium	Medium
232	Low	Medium	Medium
90	Low	Low	Low
84	Low	Low	Low
65	Low	Low	Low
90	Low	Low	Low
36	Low	Low	Low

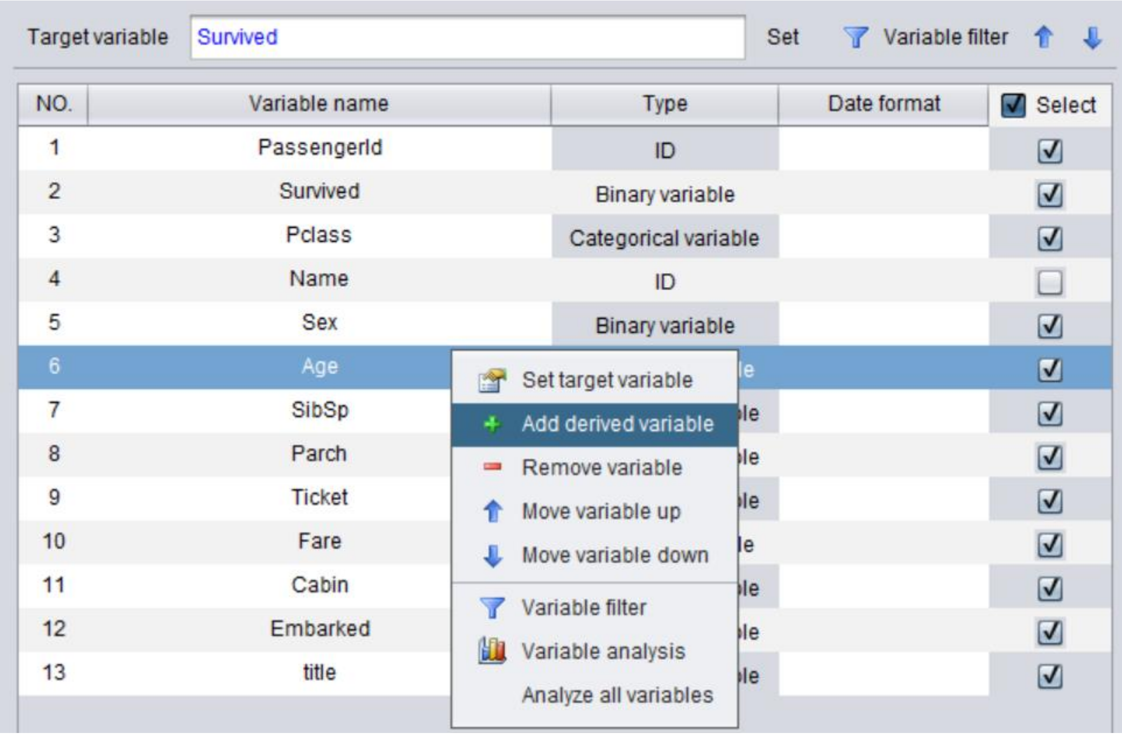
Class exercise - binning

There is a "age" variable in the Titanic data, as shown in the figure, which represents the age of passengers. We hope to group their ages into fewer intervals. Please use YModel tool, use Equi-width binning, Equi-frequency binning, and custom binning rules to practice binning, and view and compare the results.



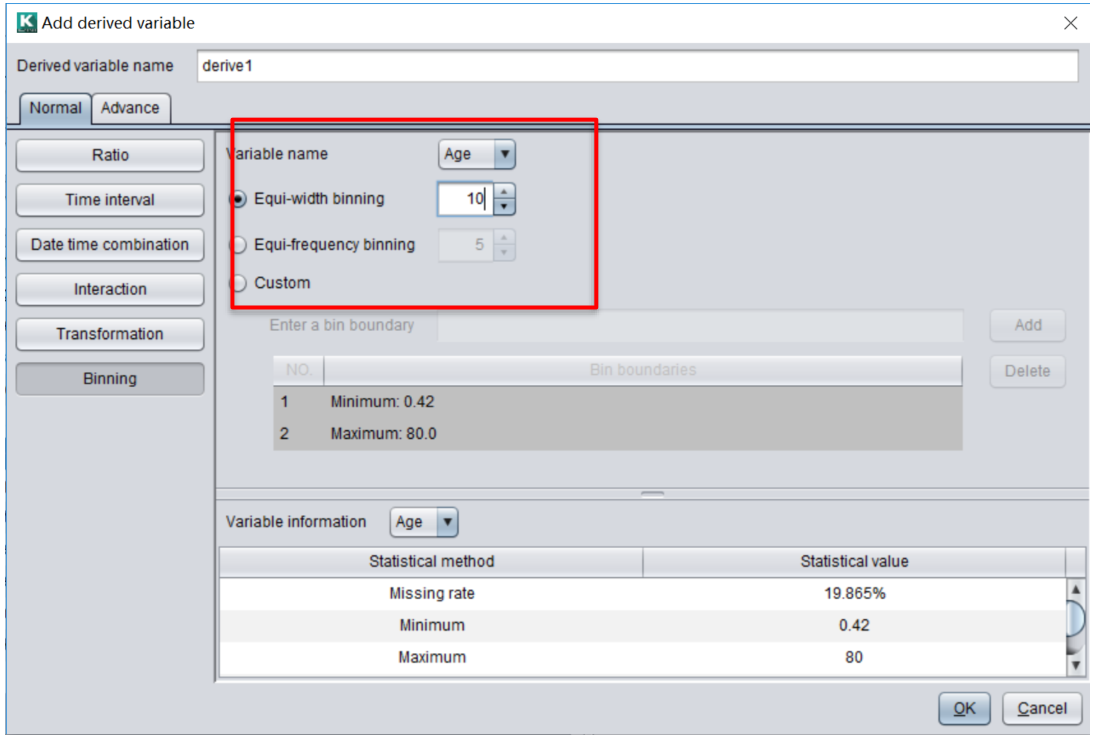
Class exercise - binning

1. Open Raqsoft YModel and import " Titanic.csv " data.
2. Select "age" variable, right-click to select "add derived variable "
3. Select "binning", you can see that there are Equi-width binning, Equi-frequency binning, and custom binning



Target variable: Survived

NO.	Variable name	Type	Date format	Select
1	PassengerId	ID		☑
2	Survived	Binary variable		☑
3	Pclass	Categorical variable		☑
4	Name	ID		☐
5	Sex	Binary variable		☑
6	Age			☑
7	SibSp			☑
8	Parch			☑
9	Ticket			☑
10	Fare			☑
11	Cabin			☑
12	Embarked			☑
13	title			☑



Add derived variable

Derived variable name: derive1

Variable name: Age

● Equi-width binning 10

○ Equi-frequency binning 5

○ Custom

Enter a bin boundary

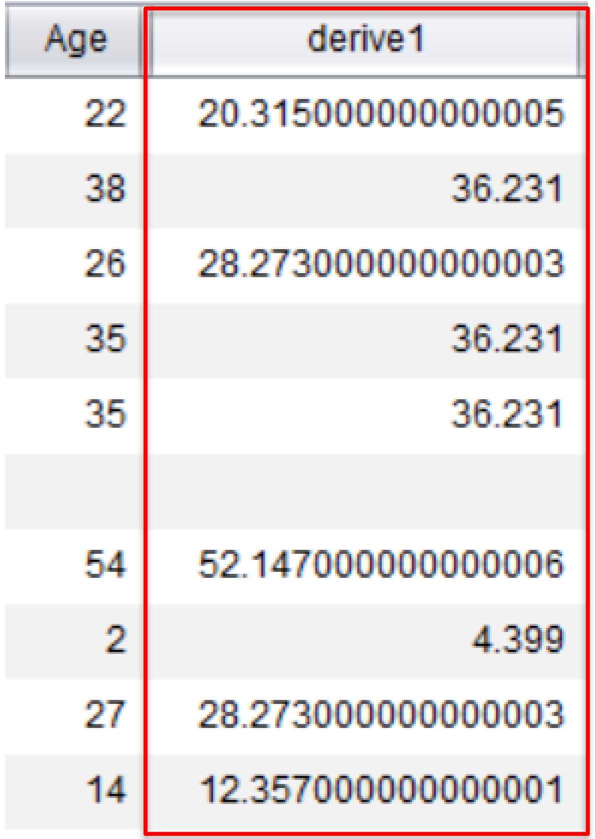
NO.	Bin boundaries
1	Minimum: 0.42
2	Maximum: 80.0

Variable information: Age

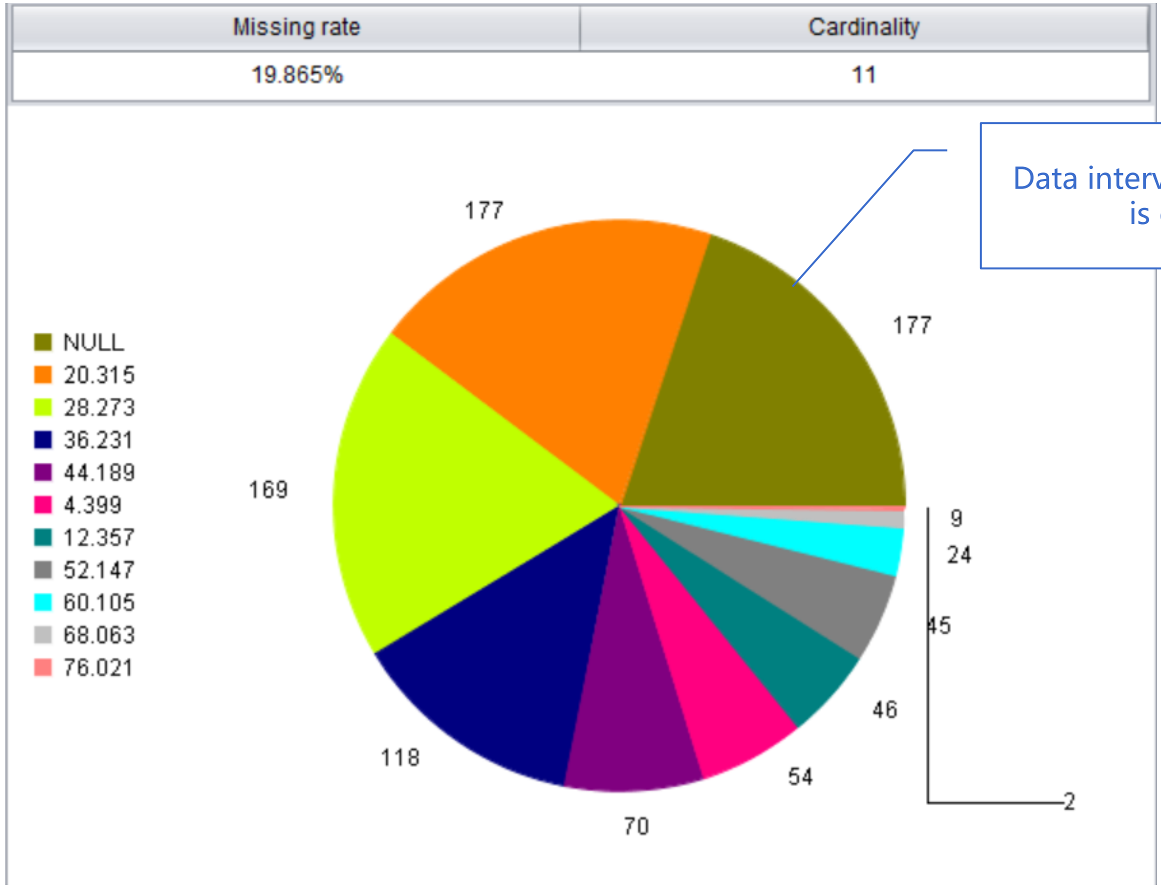
Statistical method	Statistical value
Missing rate	19.865%
Minimum	0.42
Maximum	80

Class exercise - binning

(1) Equi-width binning， interval is 10, generate "derive1"



Age	derive1
22	20.315000000000005
38	36.231
26	28.273000000000003
35	36.231
35	36.231
54	52.147000000000006
2	4.399
27	28.273000000000003
14	12.357000000000001

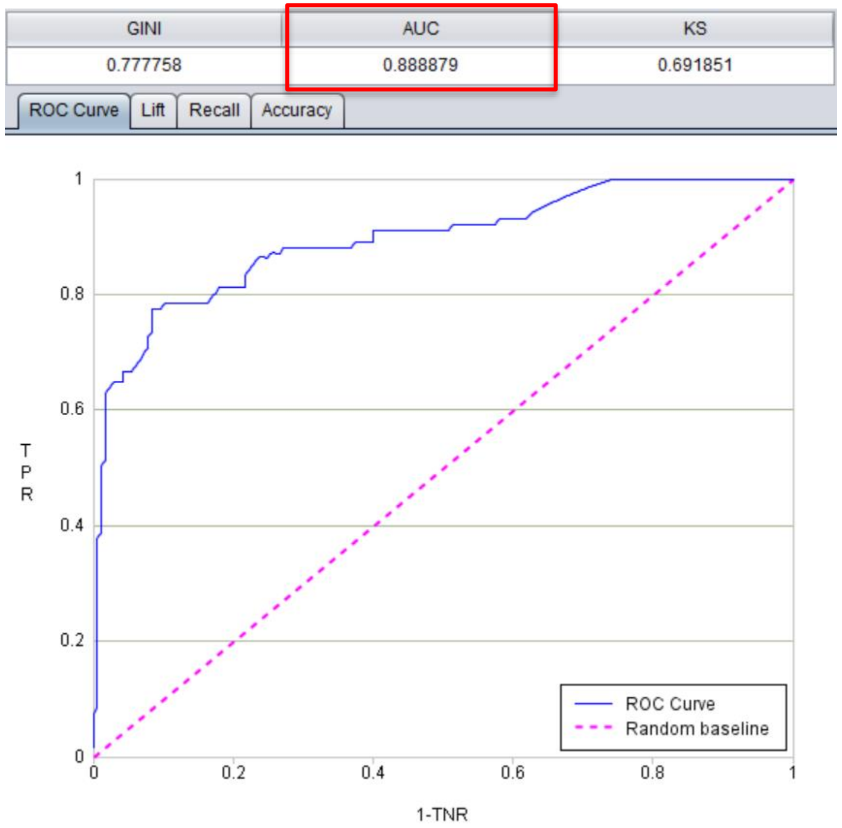


Missing rate	Cardinality
19.865%	11

Data interval in each bin is equal

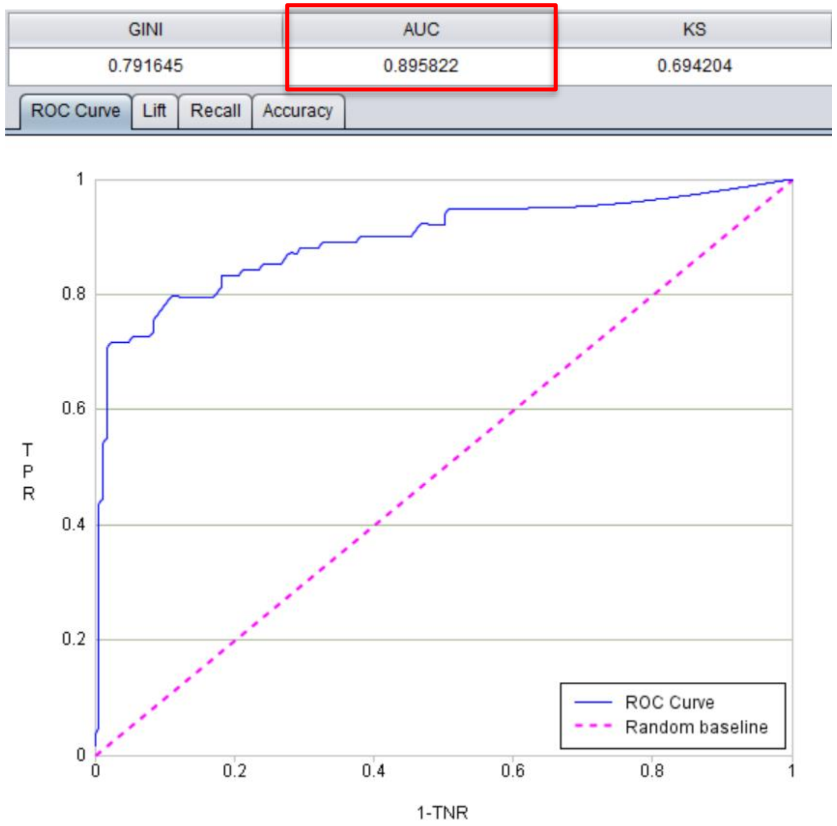
Note： After binning, the data in each bin is represented by the average value of the original data in the bin.

Class exercise - binning



GINI	AUC	KS
0.777758	0.888879	0.691851

Before binning



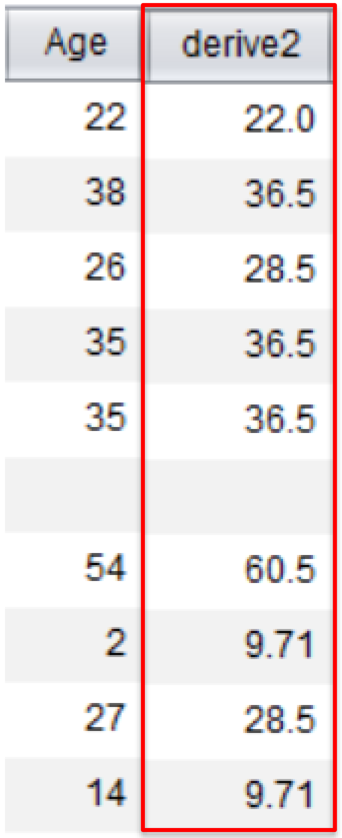
GINI	AUC	KS
0.791645	0.895822	0.694204

After binning

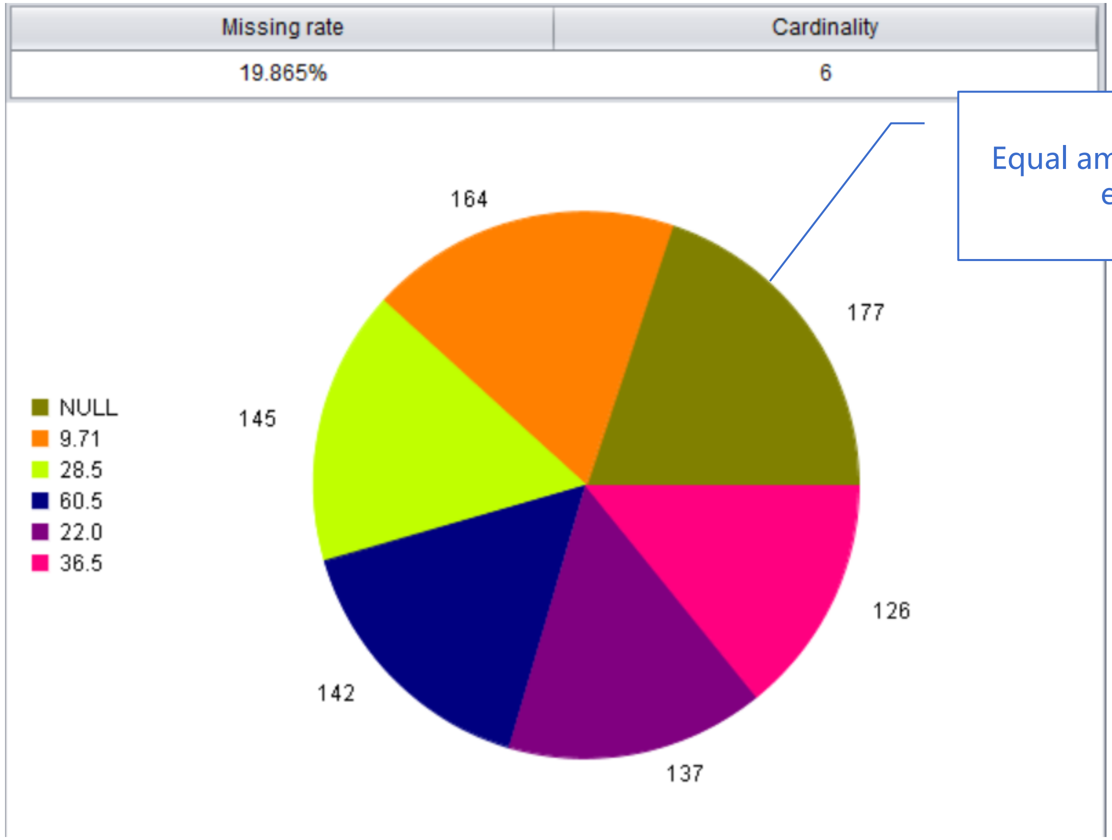
The AUC value has been increased after Equi-width binning of the passenger's age.

Class exercise - binning

(2）Equi-frequency binning，number of bin is 5， generate “derive2”



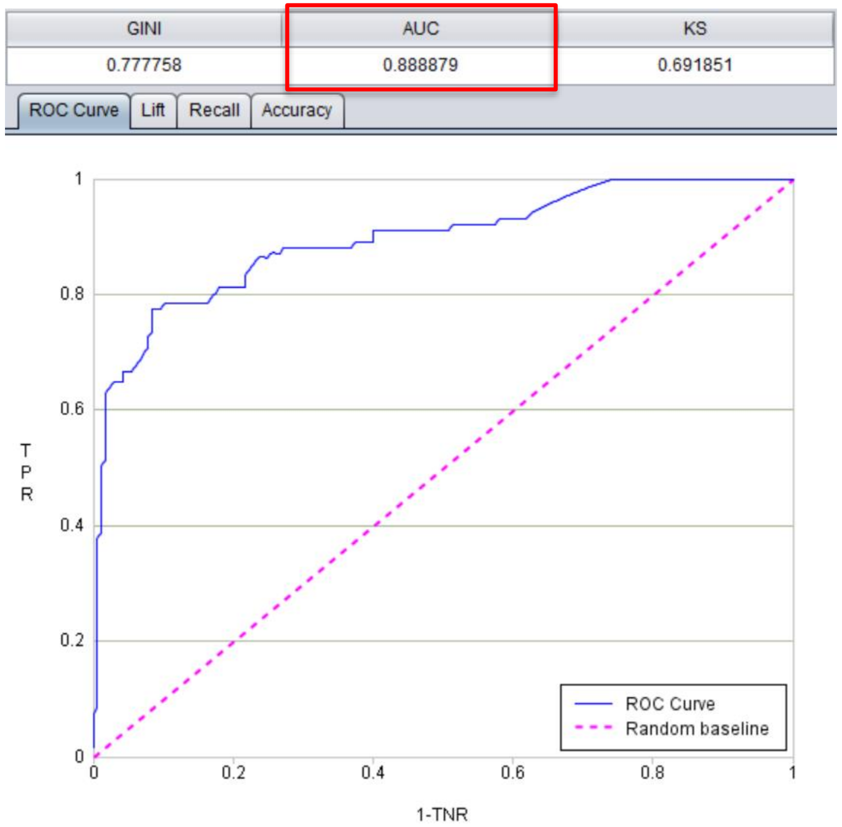
Age	derive2
22	22.0
38	36.5
26	28.5
35	36.5
35	36.5
54	60.5
2	9.71
27	28.5
14	9.71



Missing rate	Cardinality
19.865%	6

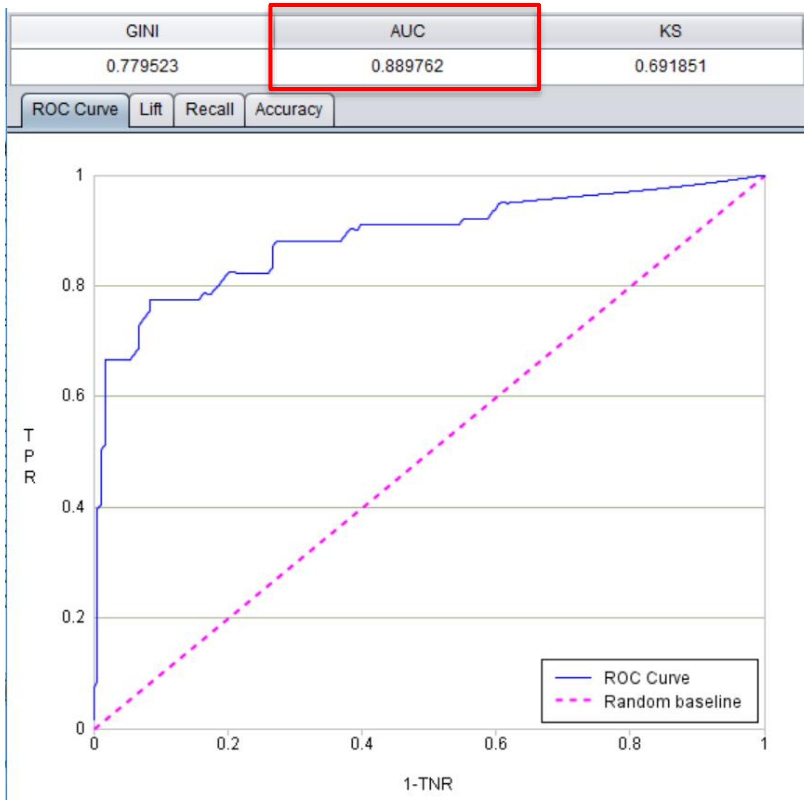
Equal amount of data in each bin

Class exercise - binning



GINI	AUC	KS
0.777758	0.888879	0.691851

Before binning



GINI	AUC	KS
0.779523	0.889762	0.691851

After binning

The AUC value has been slightly increased after Equi-frequency binning of the passenger's age.

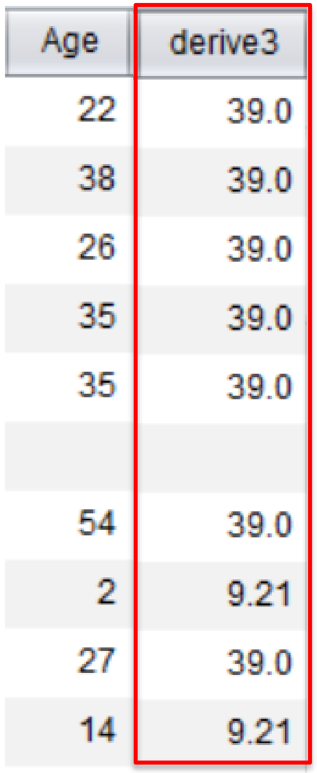
Class exercise - binning

(3) Custom binning rules， generate “derive3”

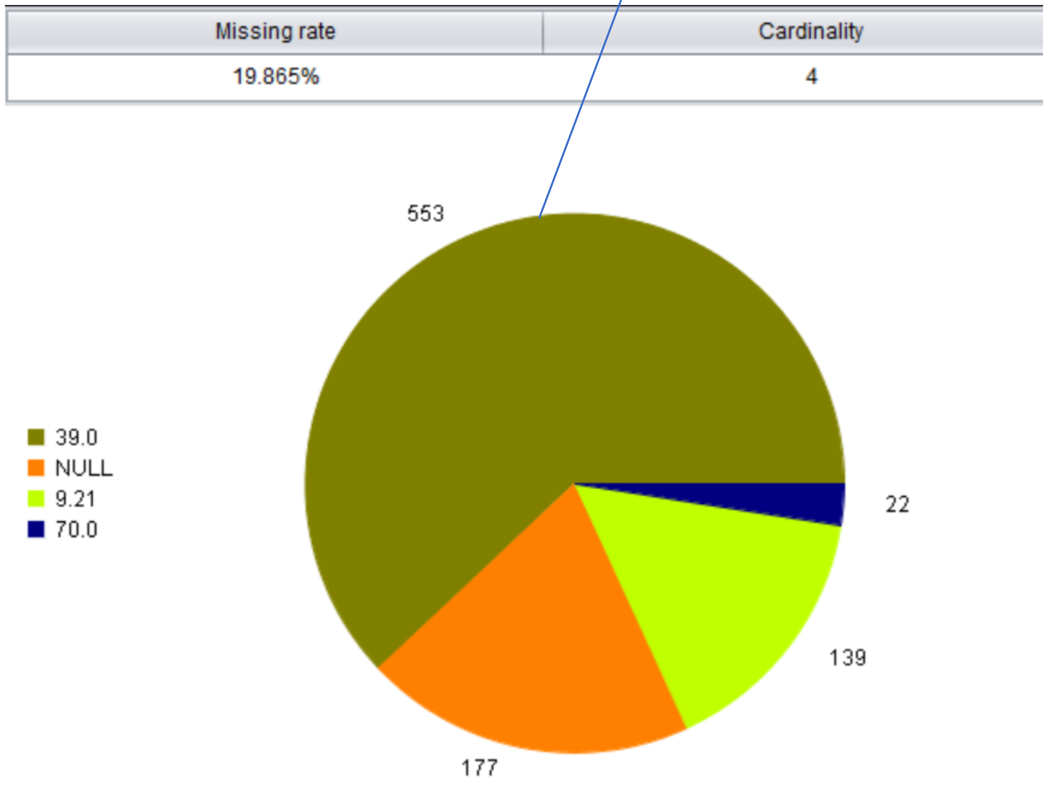
Under18, as a group

18-60, as a group

Greater than 60, as a group



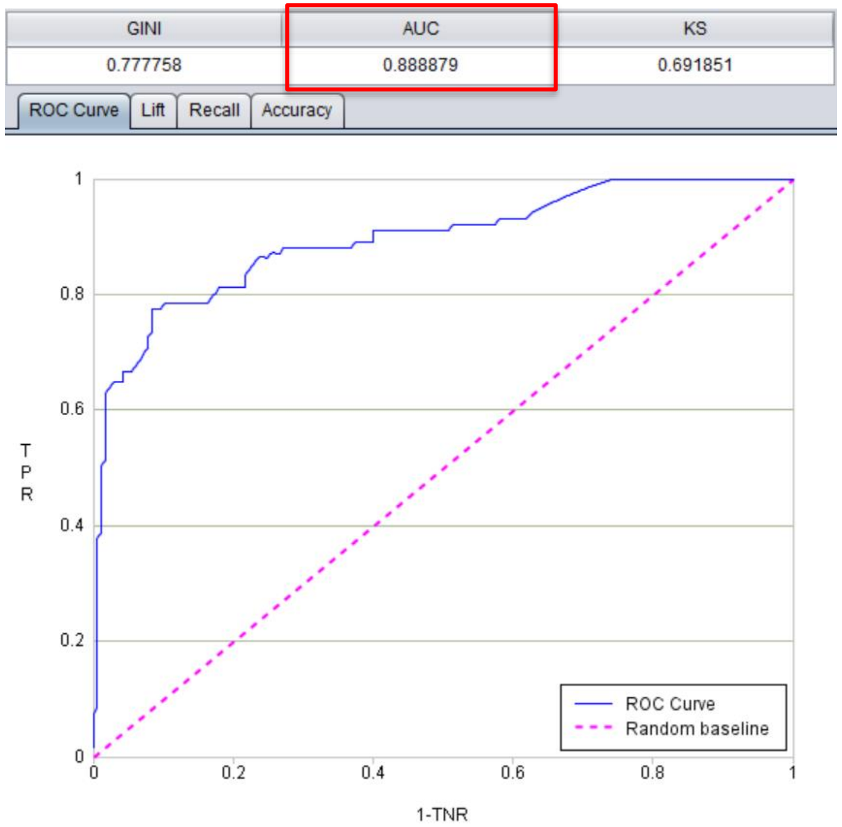
Age	derive3
22	39.0
38	39.0
26	39.0
35	39.0
35	39.0
54	39.0
2	9.21
27	39.0
14	9.21



Missing rate	Cardinality
19.865%	4

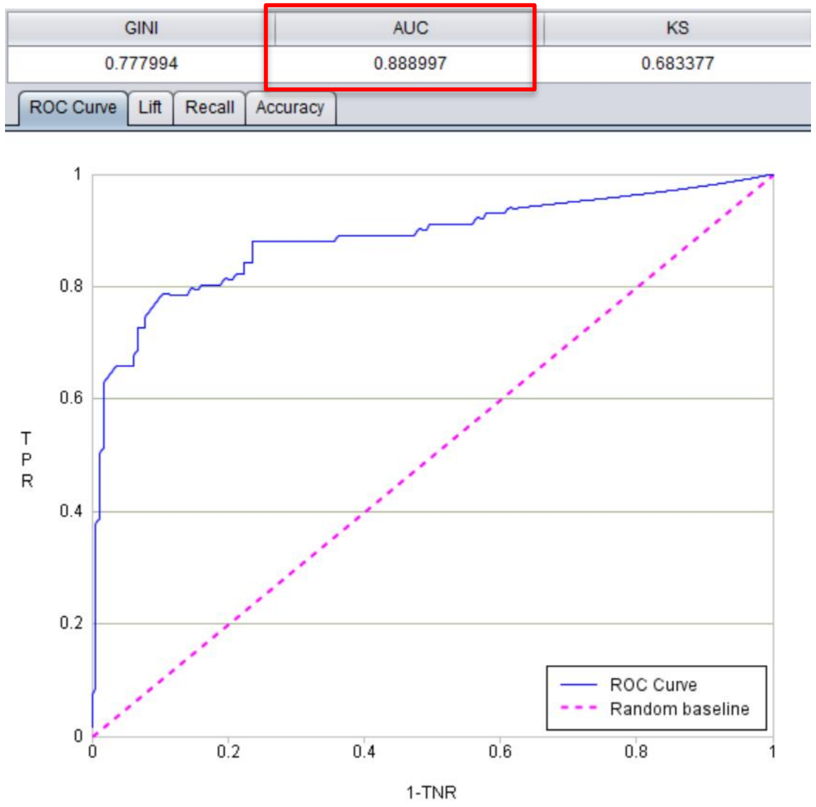
Usually define the rules of binning according to the business understanding

Class exercise - binning



GINI	AUC	KS
0.777758	0.888879	0.691851

Before binning



GINI	AUC	KS
0.777994	0.888997	0.683377

After binning

The AUC value has been slightly increased after custom binning of the passenger's age.

Variable transformation

In order to improve the effect of the model, some mathematical transformations are sometimes made to the variables, such as logarithmic transformation, Box-Cox transformation, tangent, arctangent, hyperbolic tangent, etc.

Logarithm transformation is generally used for financial data, which can transform exponential growth data into linear growth data.

For example, if you subtract yesterday's share price from today's share price, it doesn't make sense to get the difference, because the stock price is high or low, and the caliber is different. But after making logarithmic transformation and subtraction, you get the growth rate, and the caliber is the same. This is to transform the exponential growth data into linear growth data.

Box-Cox transformation is a kind of generalized power transformation method proposed by Box and Cox in 1964. It is a kind of data transformation commonly used in statistical modeling, which is used in the case that continuous variable does not meet normal distribution.

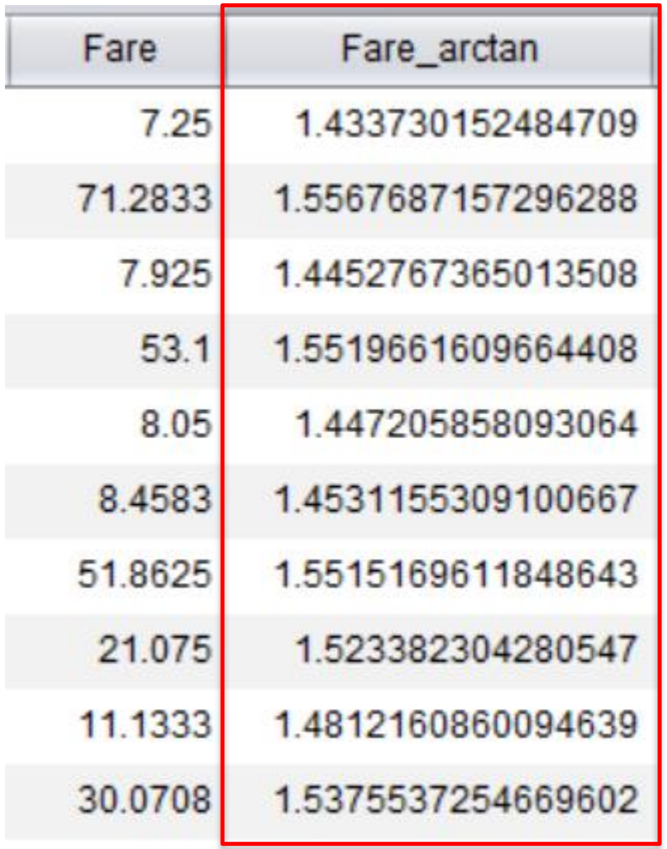
Variable transformation

Tangent transformation and **arctangent** transformation are value domain transformations, which transform data into each other in finite domain and infinite domain.

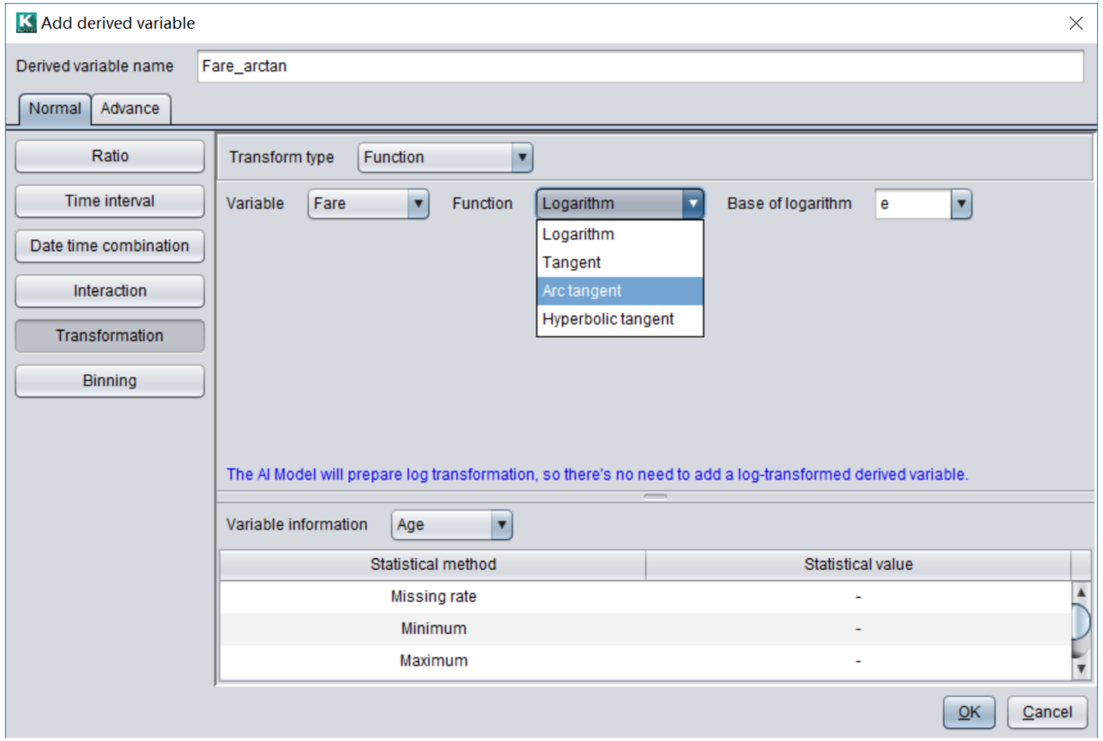
For example, the claim amount, whose value can be between 0 and tens of millions, has no upper limit, and can be limited between 0 and π/2 by arctangent, and does not change the order, which is very helpful for the stability of the model.

Class exercise - Variable transformation

Exercise： Please use YModel tool to perform arctangent transformation on the “Fare" variable in Titanic data to generate the derived variable " Fare_arctan "and run to view the derived results.



Fare	Fare_arctan
7.25	1.433730152484709
71.2833	1.5567687157296288
7.925	1.4452767365013508
53.1	1.5519661609664408
8.05	1.447205858093064
8.4583	1.4531155309100667
51.8625	1.5515169611848643
21.075	1.523382304280547
11.1333	1.4812160860094639
30.0708	1.5375537254669602



Add derived variable

Derived variable name: Fare_arctan

Transform type: Function

Variable: Fare

Function: Logarithm

Base of logarithm: e

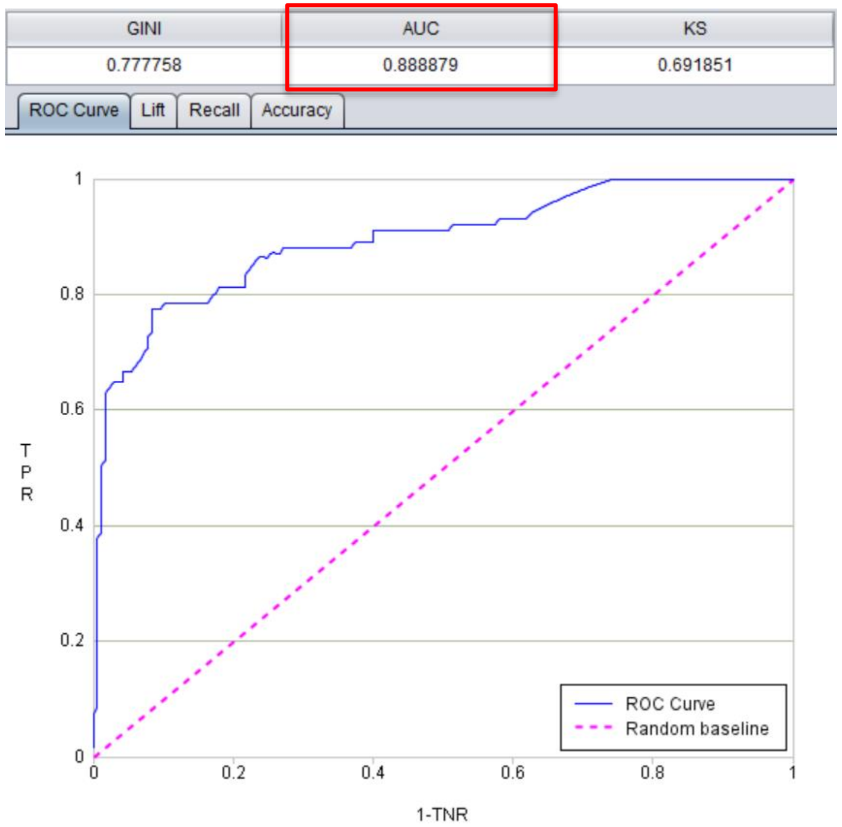
The AI Model will prepare log transformation, so there's no need to add a log-transformed derived variable.

Variable information: Age

Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-

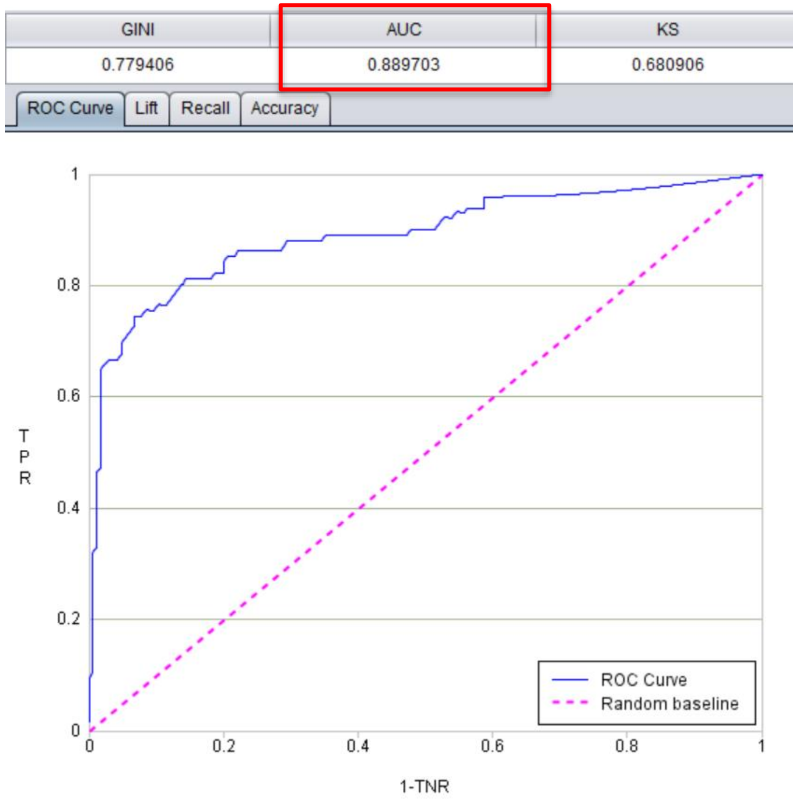
Note： In the process of preprocessing, YModel will automatically do logarithmic transformation, and users generally do not need to do logarithmic derivation .

Class exercise - Variable transformation



GINI	AUC	KS
0.777758	0.888879	0.691851

Before arctangent transformation



GINI	AUC	KS
0.779406	0.889703	0.680906

After arctangent transformation

After arctangent transformation for "fare" , AUC slightly increased.

Variable transformation

In addition to making some mathematical transformations for the variables themselves, some variables reflecting the relationship with the target variable can also be derived, such as the proportion of target positive samples, odds encoding, log odds encoding, and mean encoding.

The target positive sample proportion is used to categorical variables in the classification model, reflecting the positive rate of each category of variables. For example, 10 boys and 10 girls take the university entrance examination, and the admitted results are 7 boys and 3 girls. Then the male positive sample proportion is 7 / 10 = 0.7, and the female positive sample proportion is 3 / 10 = 0.3

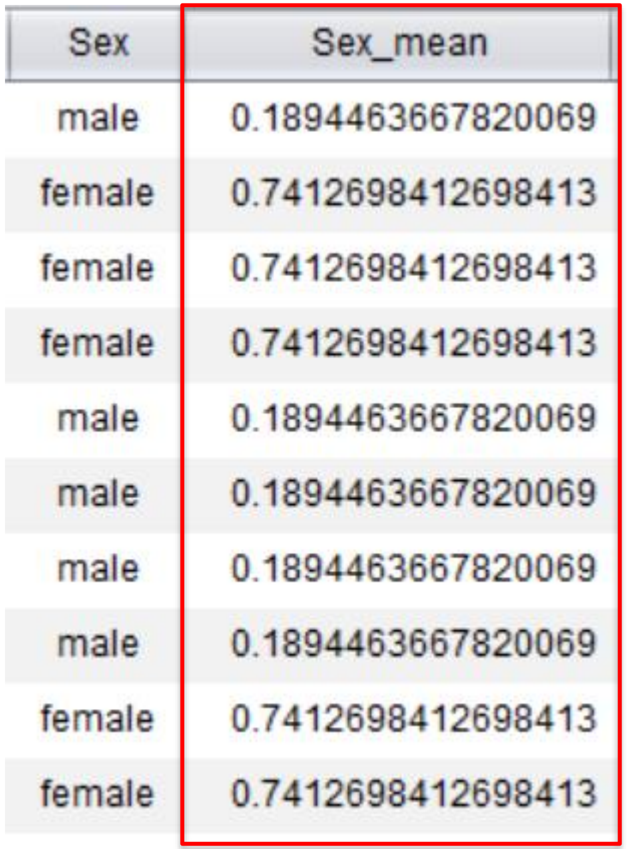
In the above example, odds (boy) = 0.7 / 0.3 = 2.33, odds (girl) = 0.3 / 0.7 = 0.428

The log odds encoding is calculated on the basis of odds.

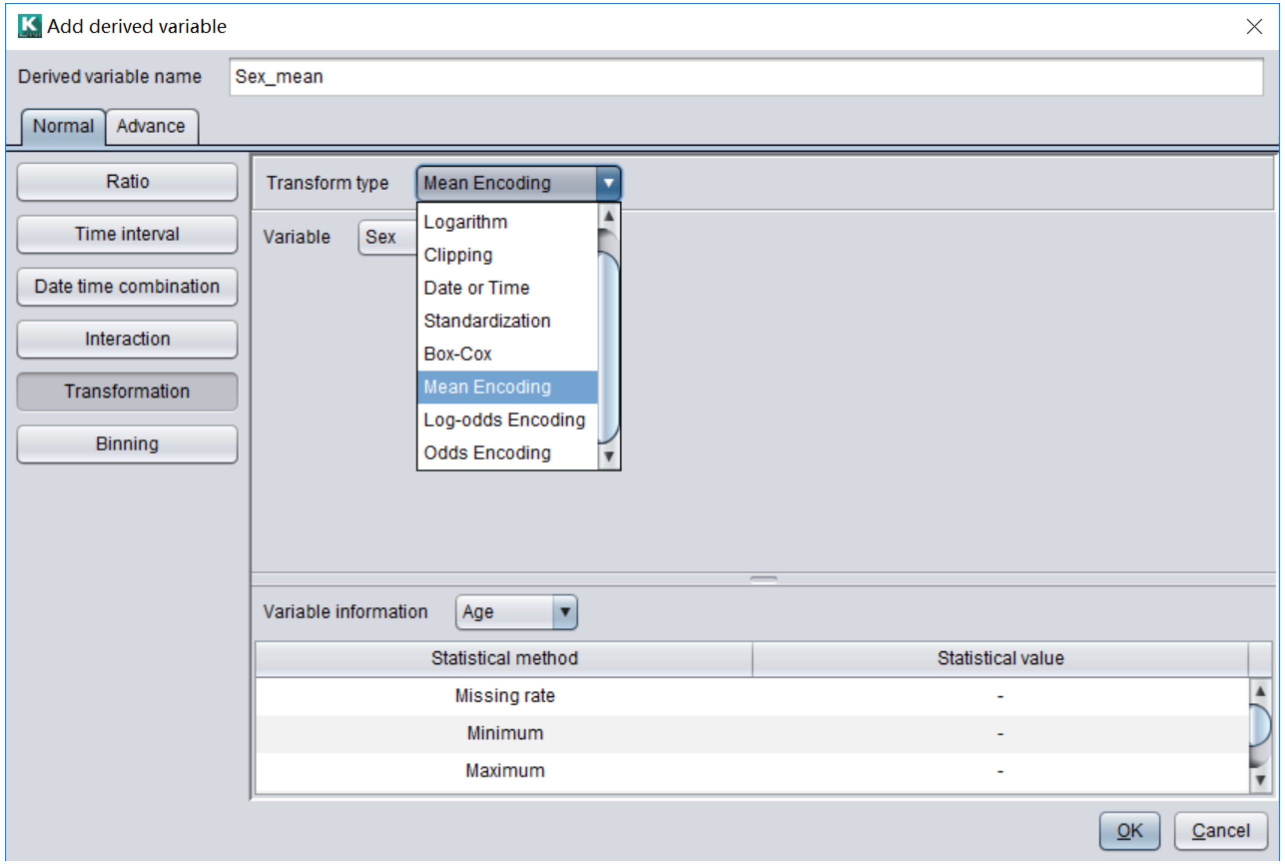
Mean encoding refers to the mean value of each category of categorical variables in the regression model.

Class exercise - Variable transformation

Exercise： Please use YModel to transform the "sex" variable in Titanic data into the target positive sample proportion to generate the derived variable " Sex_mean “ and run to view the derived results.



Sex	Sex_mean
male	0.1894463667820069
female	0.7412698412698413
female	0.7412698412698413
female	0.7412698412698413
male	0.1894463667820069
male	0.1894463667820069
male	0.1894463667820069
male	0.1894463667820069
female	0.7412698412698413
female	0.7412698412698413



Add derived variable

Derived variable name: Sex_mean

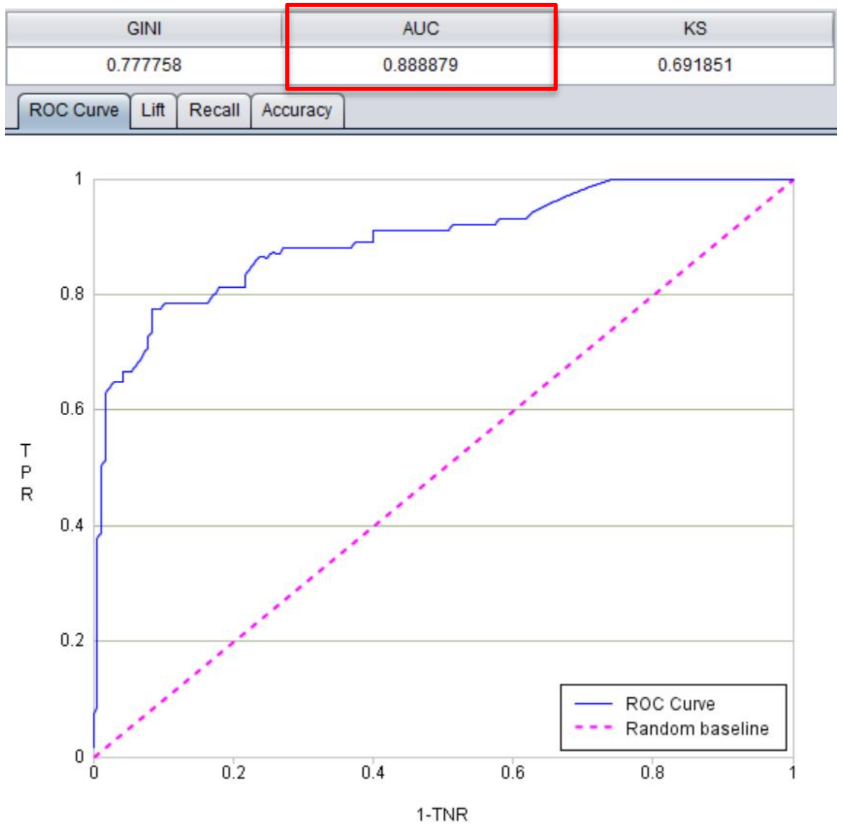
Transform type: Mean Encoding

Variable: Sex

Variable information: Age

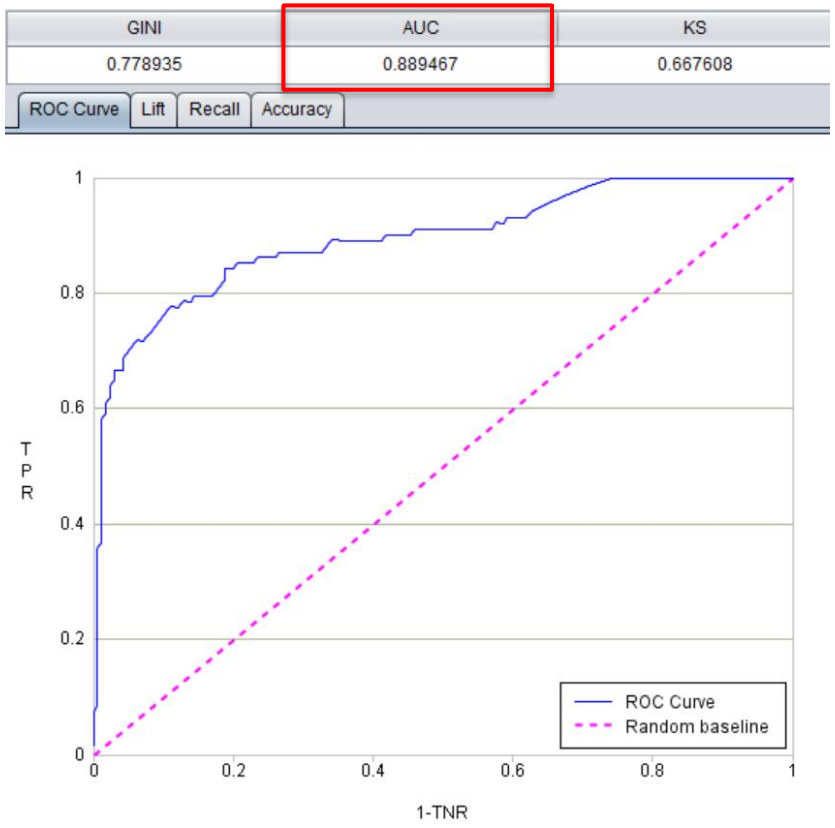
Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-

Class exercise - Variable transformation



GINI	AUC	KS
0.777758	0.888879	0.691851

Before transformation



GINI	AUC	KS
0.778935	0.889467	0.667608

After transformation

AUC slightly increased after doing the target positive samples proportion transformation.

Variable interaction

Variable interaction refers to the multiplication of two variables, which can be numerical variable * numerical variable, or categorical variable * categorical variable. Generally, we can multiply two variables with high importance to improve the model effect.

Combined feature is one of the most important methods in feature engineering. It combines two or more class attributes into one. This is a very useful technique when a combined feature is better than a single feature.

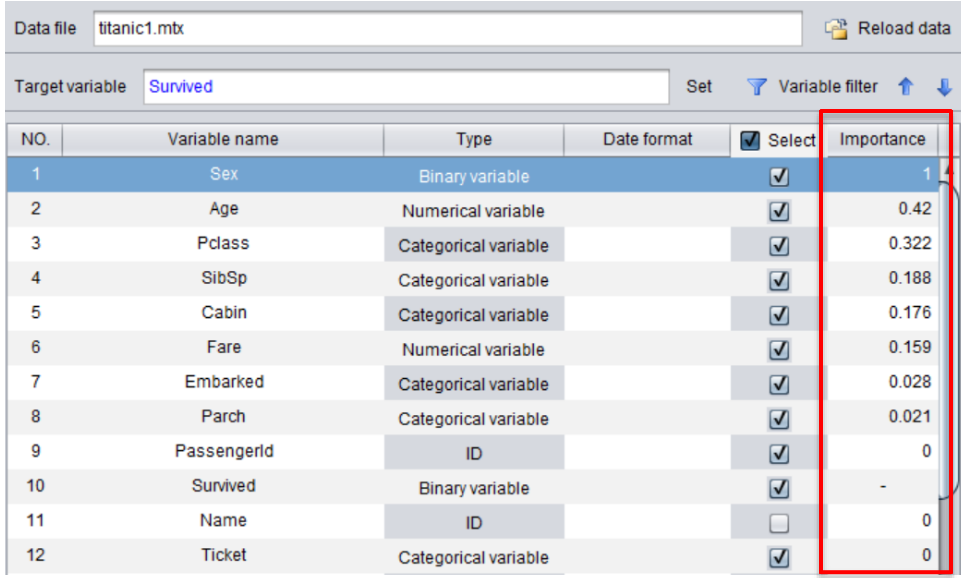
If there is a characteristic A, A has two possible values {A1, A2}. There is a characteristic B, B has two possible values of {B1, B2}. Then, the cross characteristics between A * B are as follows: {(A1, B1), (A1, B2), (A2, B1), (A2, B2)}

Class exercise - Variable interaction

Exercise: Use YModel to interactively derive two important categorical variables of Titanic data, and model to see the comparison results.

(1) Import the data for automatic modeling, and return the importance of variables as shown in the figure below on the left.

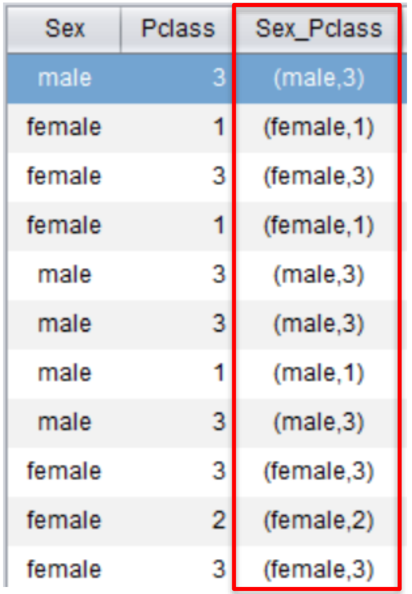
(2) Make variable interaction between "sex" and "pcalss" to generate "sex_ Pclass” .



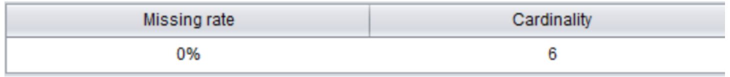
Data file: titanic1.mtx

Target variable: Survived

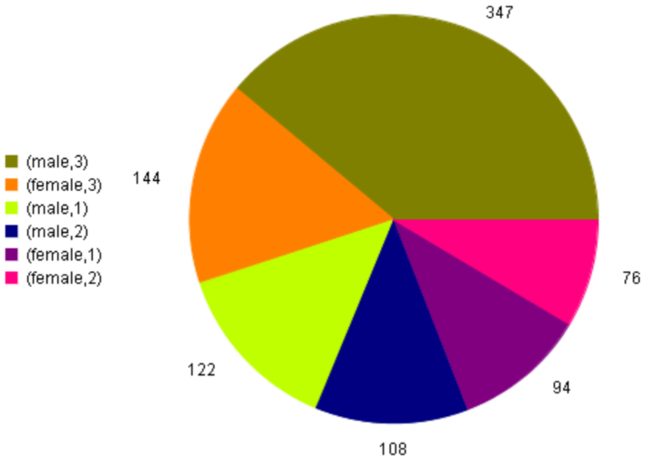
NO.	Variable name	Type	Date format	☑ Select	Importance
1	Sex	Binary variable		☑	1
2	Age	Numerical variable		☑	0.42
3	Pclass	Categorical variable		☑	0.322
4	SibSp	Categorical variable		☑	0.188
5	Cabin	Categorical variable		☑	0.176
6	Fare	Numerical variable		☑	0.159
7	Embarked	Categorical variable		☑	0.028
8	Parch	Categorical variable		☑	0.021
9	PassengerId	ID		☑	0
10	Survived	Binary variable		☑	-
11	Name	ID		☐	0
12	Ticket	Categorical variable		☑	0



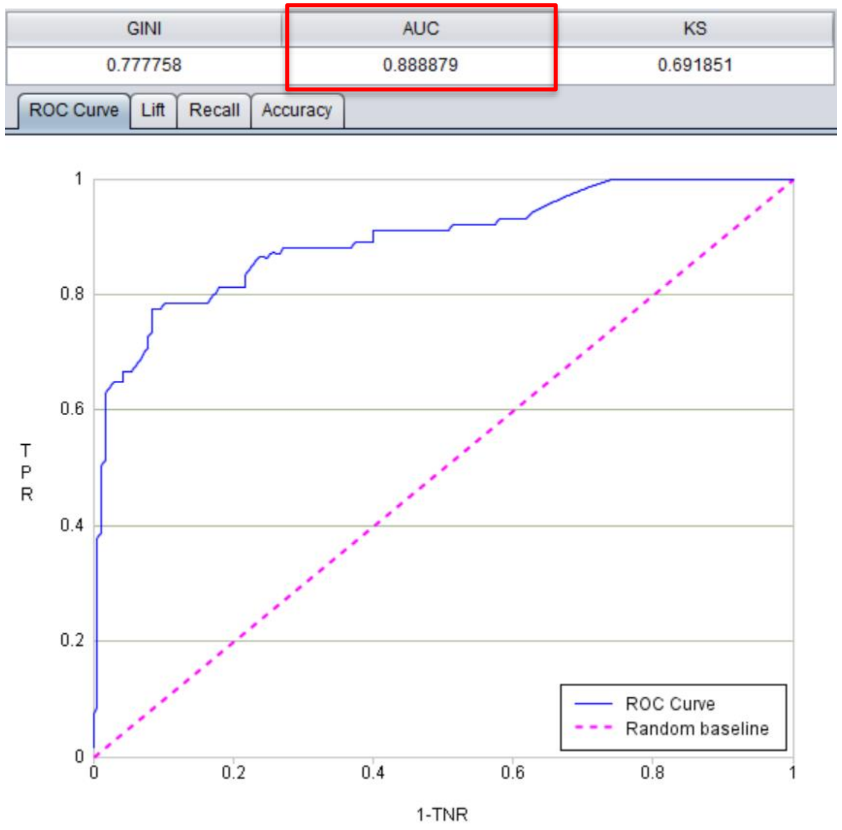
Sex	Pclass	Sex_Pclass
male	3	(male,3)
female	1	(female,1)
female	3	(female,3)
female	1	(female,1)
male	3	(male,3)
male	3	(male,3)
male	1	(male,1)
male	3	(male,3)
female	3	(female,3)
female	2	(female,2)
female	3	(female,3)



Missing rate	Cardinality
0%	6

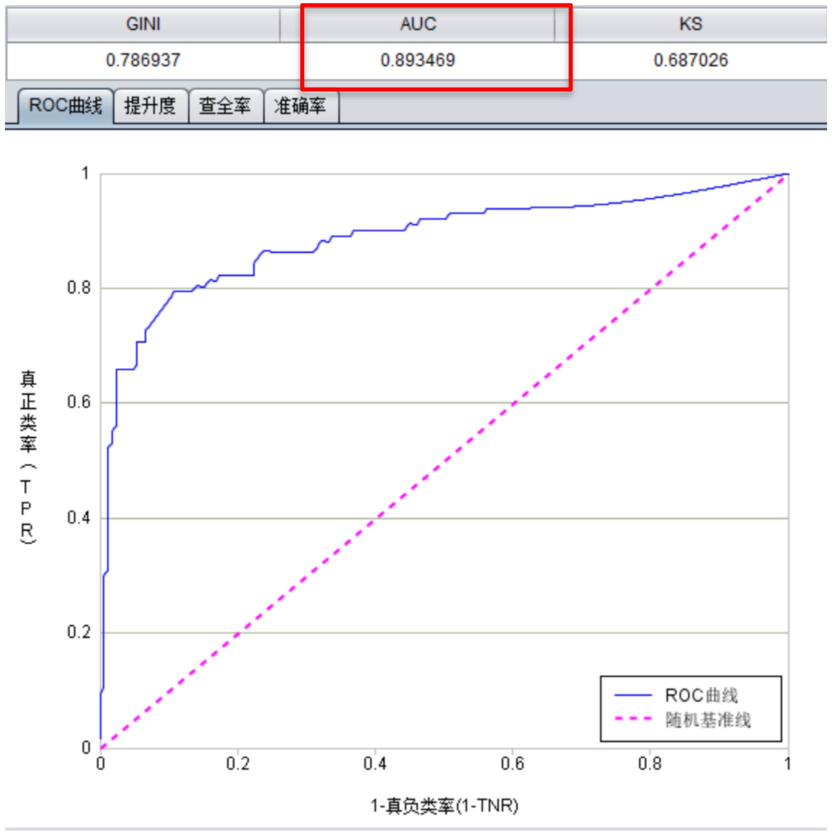


Class exercise - Variable interaction



GINI	AUC	KS
0.777758	0.888879	0.691851

Before interaction



GINI	AUC	KS
0.786937	0.893469	0.687026

After interaction

After multiplying the two important variables, AUC slightly improved.

Ratio

Ratio is the division of two variables, used for continuous variables.

Ratio is also a method of feature combination. Feature combination is a supplement to make up for the lack of linear model which can not express nonlinear properties, which is helpful to improve the expression ability of linear model.

Note: in the ratio calculation, the denominator variable cannot have 0.

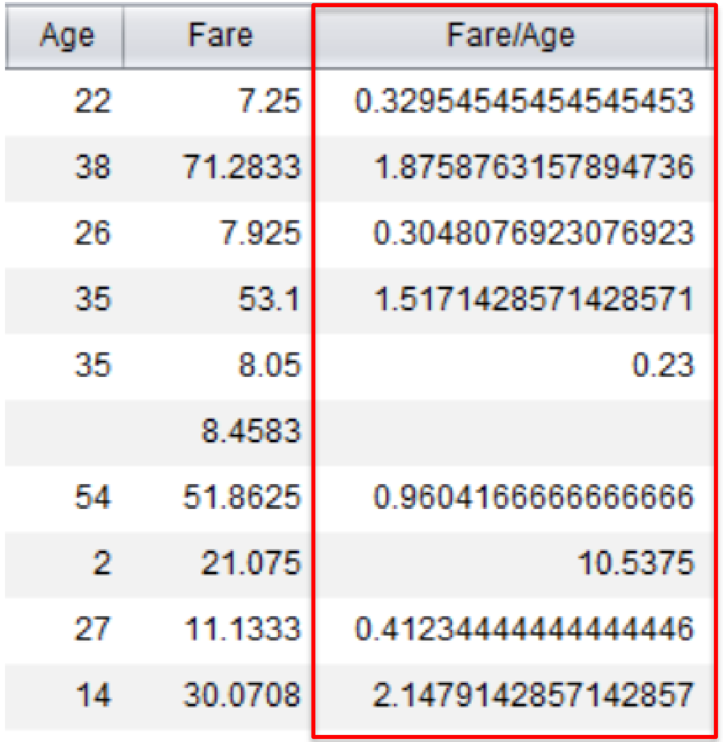
Exercise: derivative the ratio combination feature of two continuous variables in Titanic data, run and view the result.

Class exercise - ratio

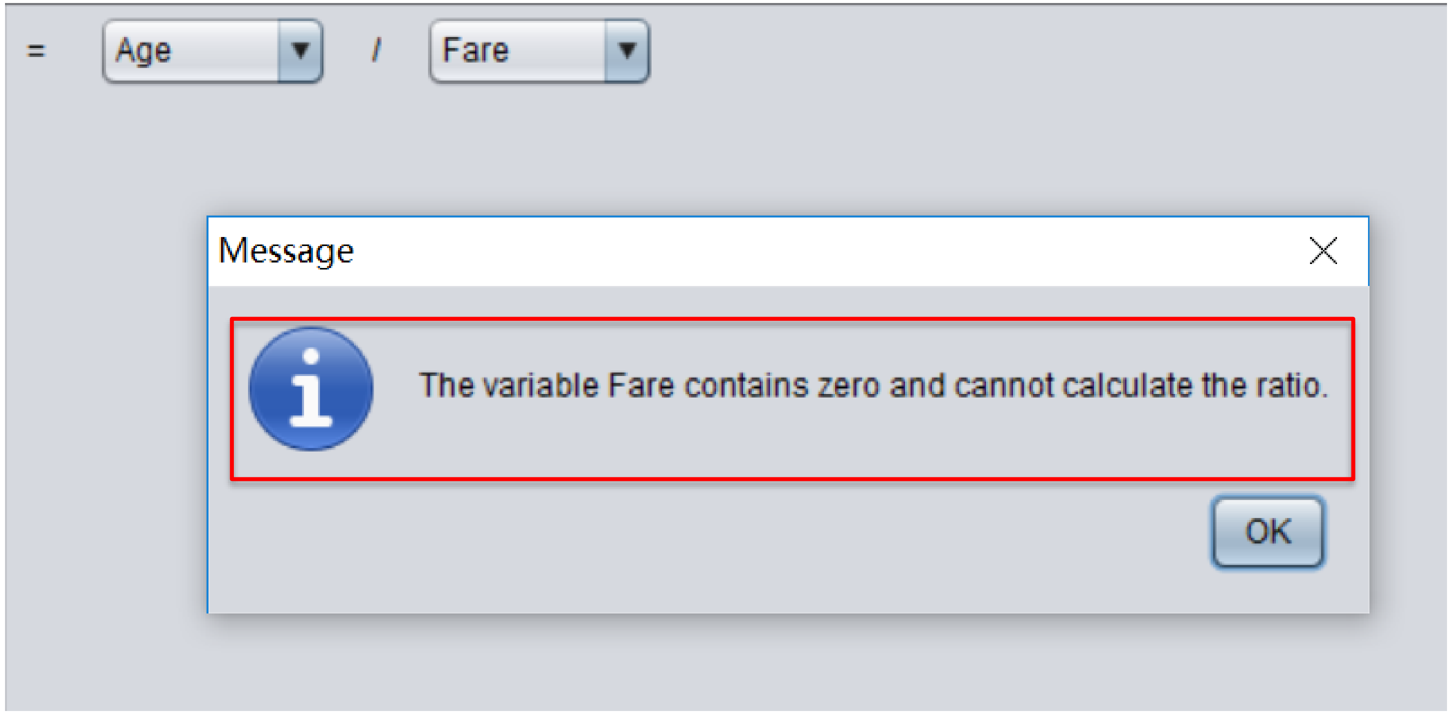
(1) Import data into YModel tool and add derivative variables.

(2) Divide "age" by "fare". The ratio cannot be calculated because "fare" contains 0.

(3) Divide "fare" by "age" to model and calculate.

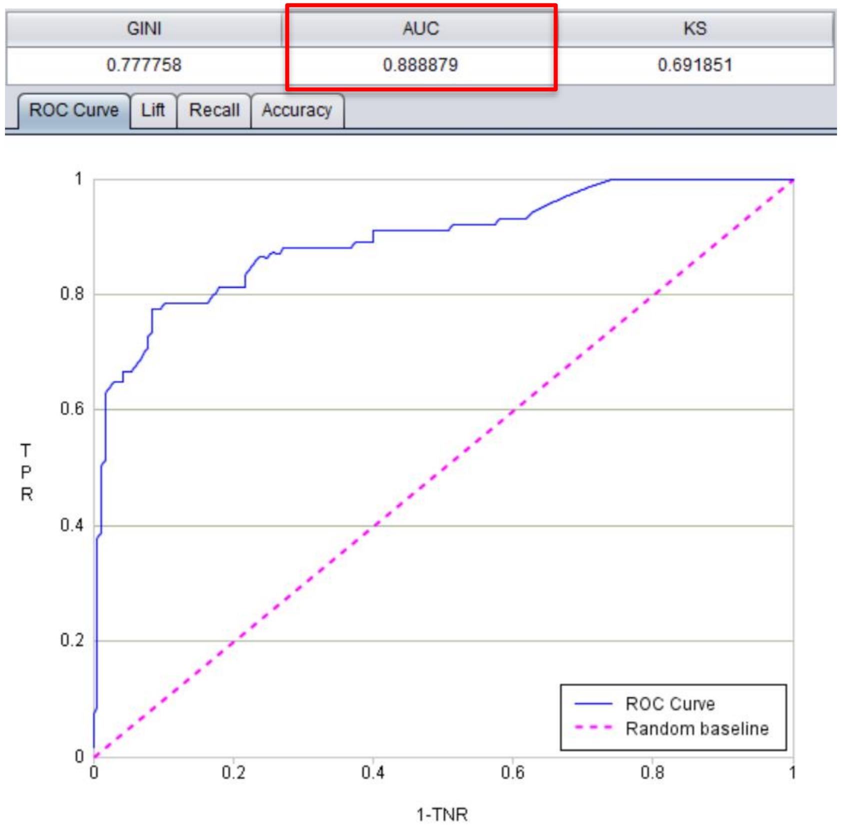


Age	Fare	Fare/Age
22	7.25	0.32954545454545453
38	71.2833	1.8758763157894736
26	7.925	0.3048076923076923
35	53.1	1.5171428571428571
35	8.05	0.23
	8.4583	
54	51.8625	0.9604166666666666
2	21.075	10.5375
27	11.1333	0.41234444444444446
14	30.0708	2.1479142857142857



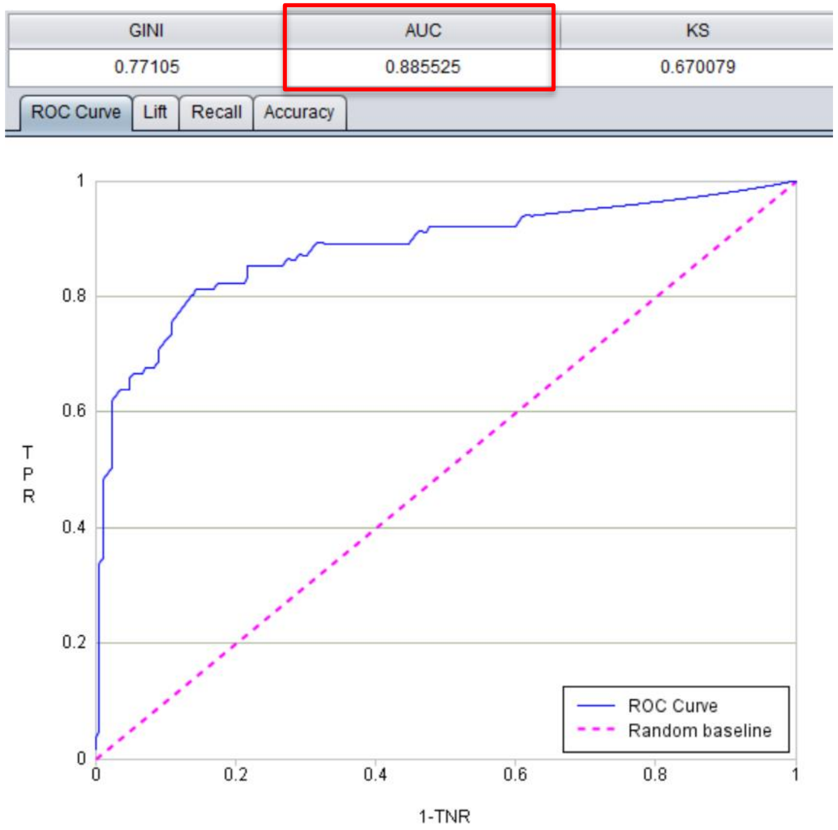
The variable Fare contains zero and cannot calculate the ratio.

Class exercise - ratio



GINI	AUC	KS
0.777758	0.888879	0.691851

Before the ratio transformation



GINI	AUC	KS
0.77105	0.885525	0.670079

After the ratio transformation

After dividing a numerical variable by another, AUC is slightly lower than before.

Date time variable

Date time variable

Date time variable is very important in data mining, which can extract a lot of useful information.

Common derivatives

For example, for the date variable, you can extract the month, season, week, holiday or not, and the days from now;
For the time variable, the morning, afternoon and night can be derived; for any two date variables, the interval days can be calculated, etc.

Time related feature extraction

Users can also extract a lot of feature variables based on date and time, such as the login information of APP users in the past 1 month and 3 months, and the operation situation in different time periods. Of course, how to derive should be understood according to the business significance, there is no unified approach.

Business scenario derivation

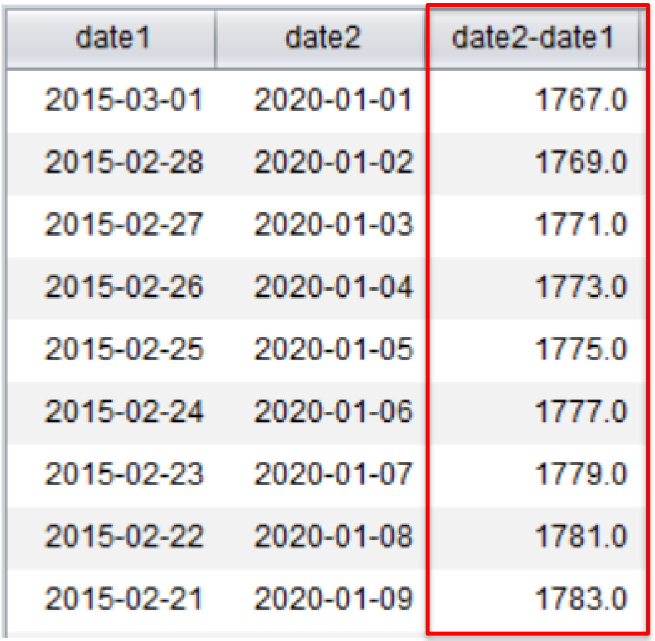
There are also some derived variables related to business scenarios, which need to be added by users based on business understanding. For example, the sales data of e-commerce “Black Friday" will be a special date, while the customer flow of cinemas may have a strong relationship with national public holidays and winter and summer holidays.

Note: Raqsoft YModel products will automatically derive common date time derived variables.

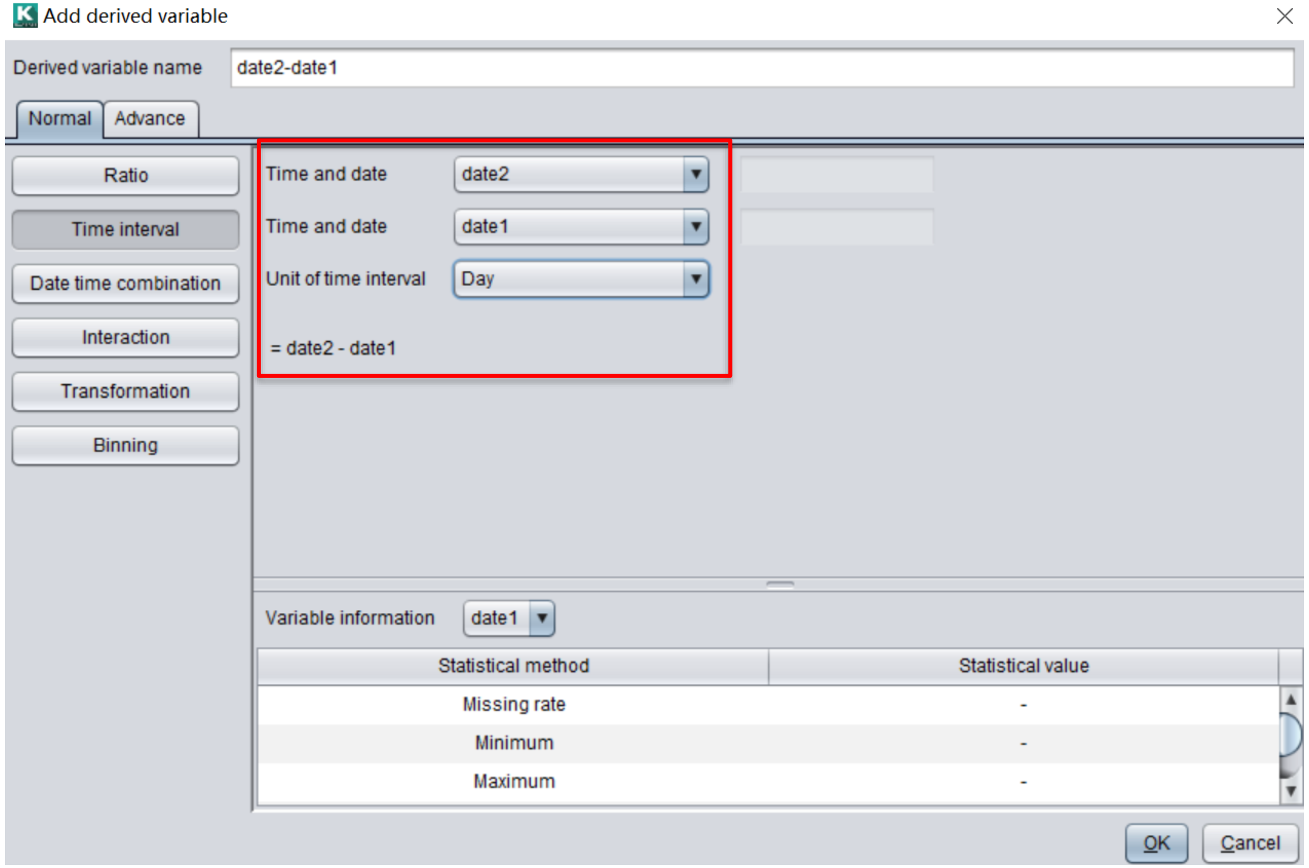
Class exercise - Time interval

Time interval refers to the subtraction of two time variables. For example, in the credit business, the user's repayment due date is subtracted from the actual repayment date to get the user's repayment habits. Some users like to prepay, some like to repay on the same day, and some are used to overdue.

Exercise: Date2 and date1 in catering_ sale.csv are subtracted and the interval is in days.



date1	date2	date2-date1
2015-03-01	2020-01-01	1767.0
2015-02-28	2020-01-02	1769.0
2015-02-27	2020-01-03	1771.0
2015-02-26	2020-01-04	1773.0
2015-02-25	2020-01-05	1775.0
2015-02-24	2020-01-06	1777.0
2015-02-23	2020-01-07	1779.0
2015-02-22	2020-01-08	1781.0
2015-02-21	2020-01-09	1783.0



Add derived variable

Derived variable name: date2-date1

Time and date: date2

Time and date: date1

Unit of time interval: Day

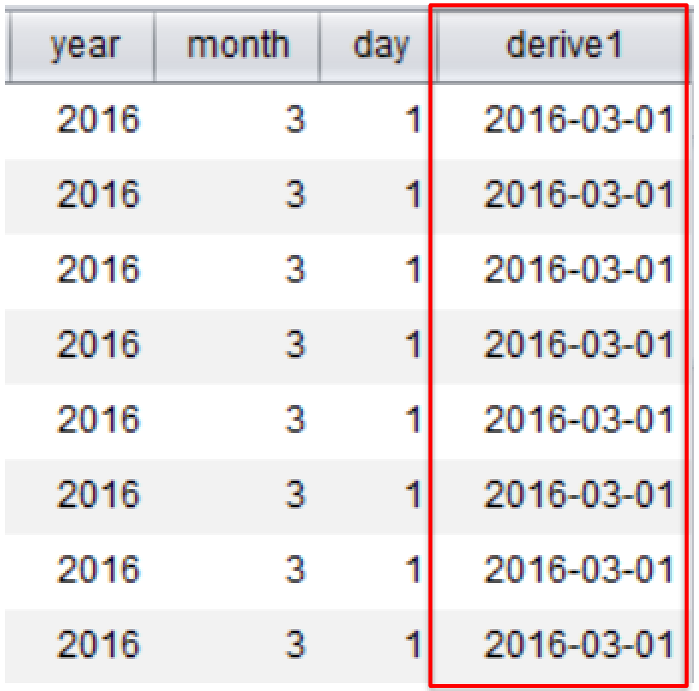
= date2 - date1

Variable information: date1

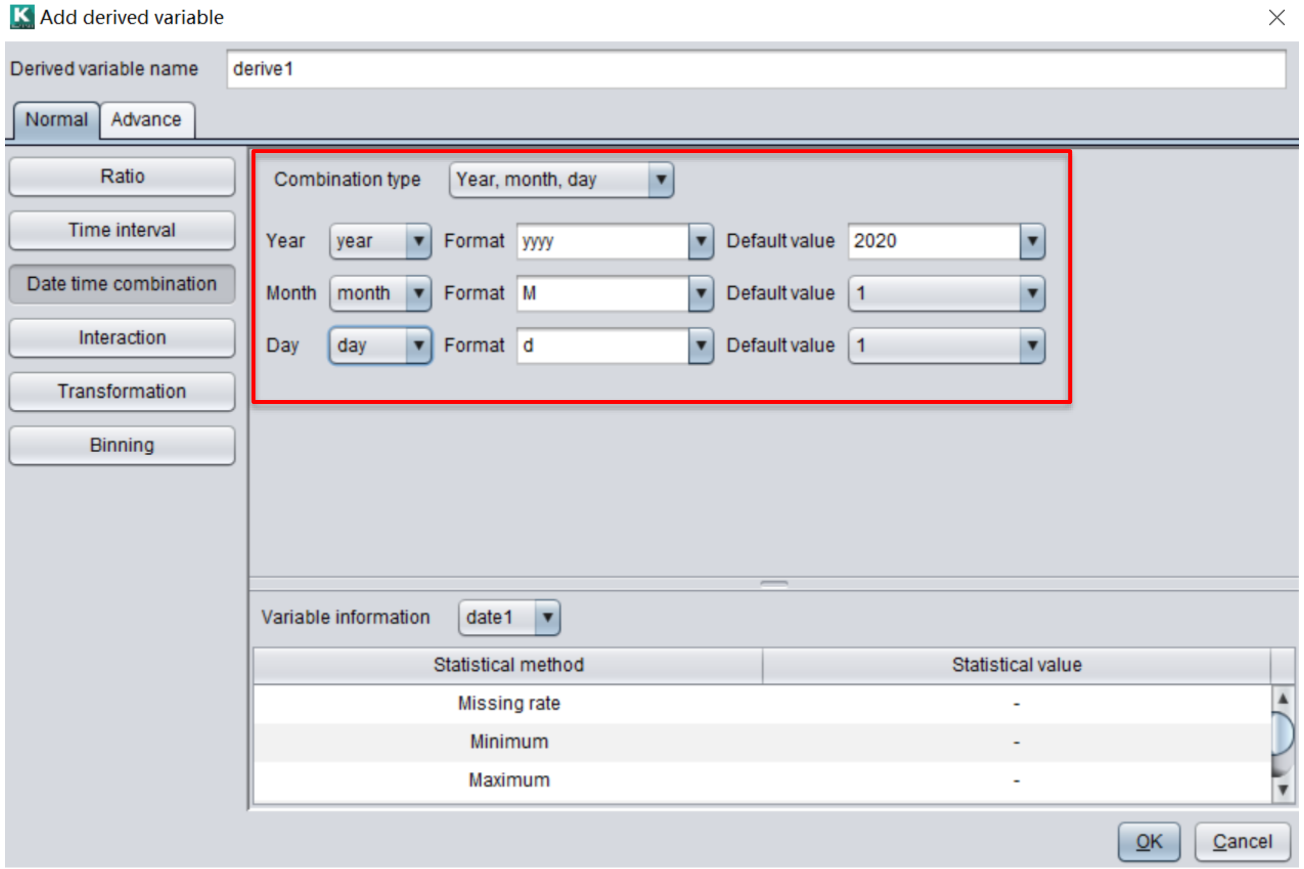
Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-

Class exercise - Time date combination

Time date combination can combine multiple time information variables into time variable,
For example, there are three variables "year", "month" and "day" in catering_ sale.csv. We can combine them into a time date variable.



year	month	day	derive1
2016	3	1	2016-03-01
2016	3	1	2016-03-01
2016	3	1	2016-03-01
2016	3	1	2016-03-01
2016	3	1	2016-03-01
2016	3	1	2016-03-01
2016	3	1	2016-03-01
2016	3	1	2016-03-01



Derived variable name: derive1

Combination type: Year, month, day

Year: year | Format: yyyy | Default value: 2020

Month: month | Format: M | Default value: 1

Day: day | Format: d | Default value: 1

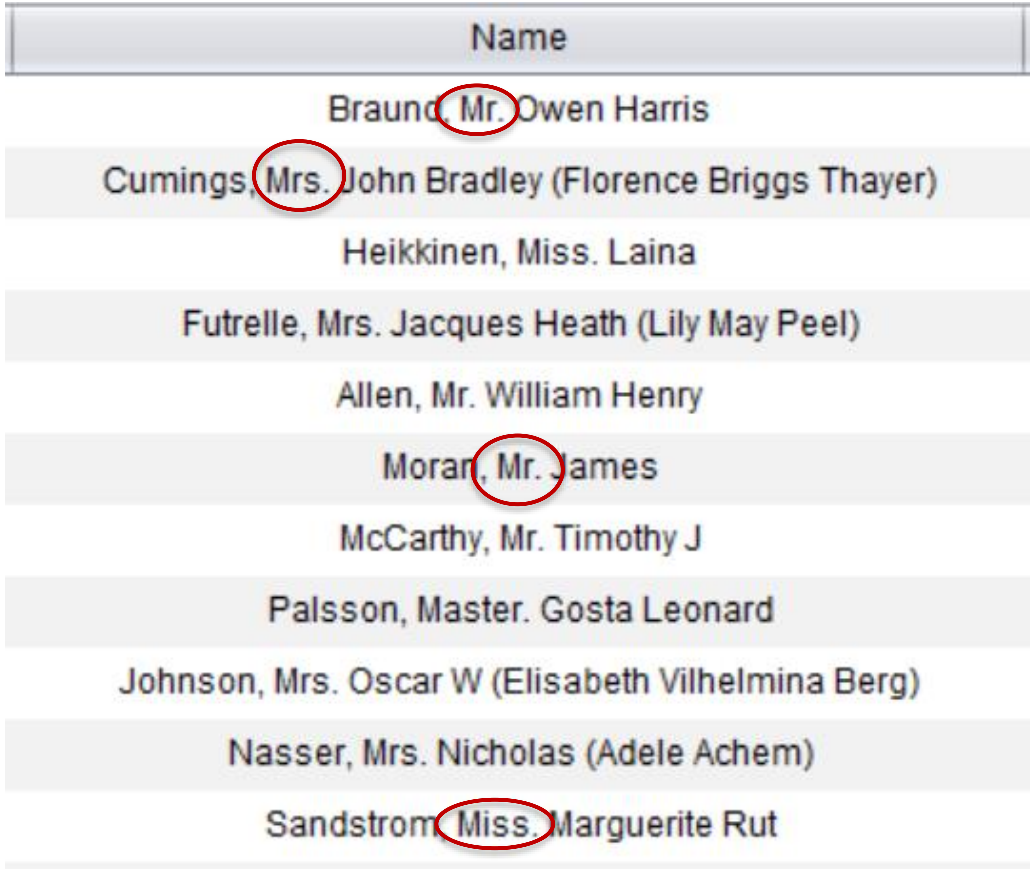
Variable information: date1

Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-

Other derivatives

The above only introduces the common methods of derived variable, and the adding methods of derived variable are far more than these, but they are usually derived by combining business understanding and data characteristics.

For example, in the Titanic data, there is a variable "name". Each passenger's name is different, so it's meaningless to model directly. But after careful observation, it is found that almost all passenger's name registrations contain Mr, Mrs ... We can extract these common appellations and generate new variables to participate in the modeling.

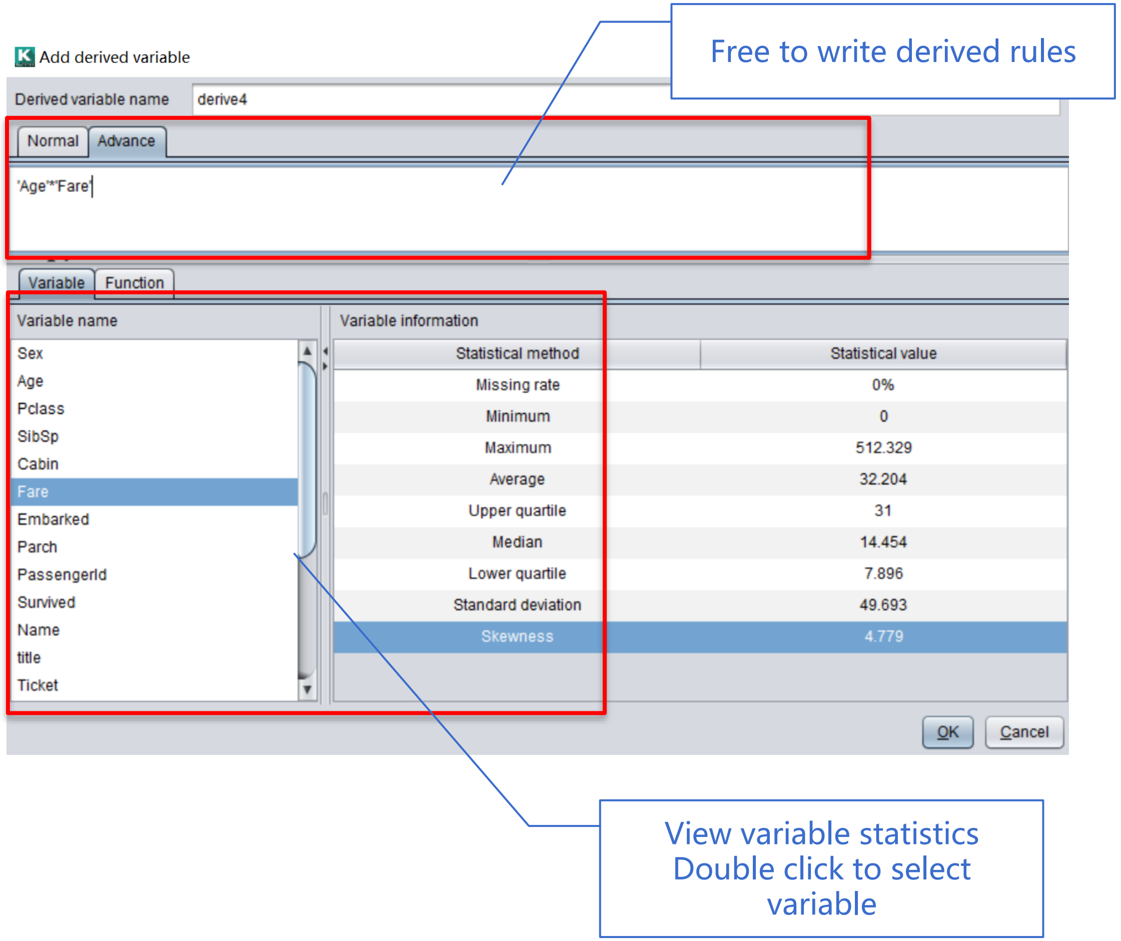


Name
Braund, Mr. Owen Harris
Cumings, Mrs. John Bradley (Florence Briggs Thayer)
Heikkinen, Miss. Laina
Futrelle, Mrs. Jacques Heath (Lily May Peel)
Allen, Mr. William Henry
Moran, Mr. James
McCarthy, Mr. Timothy J
Palsson, Master. Gosta Leonard
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
Nasser, Mrs. Nicholas (Adele Achem)
Sandstrom, Miss. Marguerite Rut

Class exercise - Other derivatives

Exercise: derive the "name" of the Titanic data, named "title", and see the derived effect.

You can use the advanced derived variable function of YModel tool to freely write derivative functions. The operation method is as follows:



Add derived variable

Derived variable name: derive4

'Age'*'Fare'

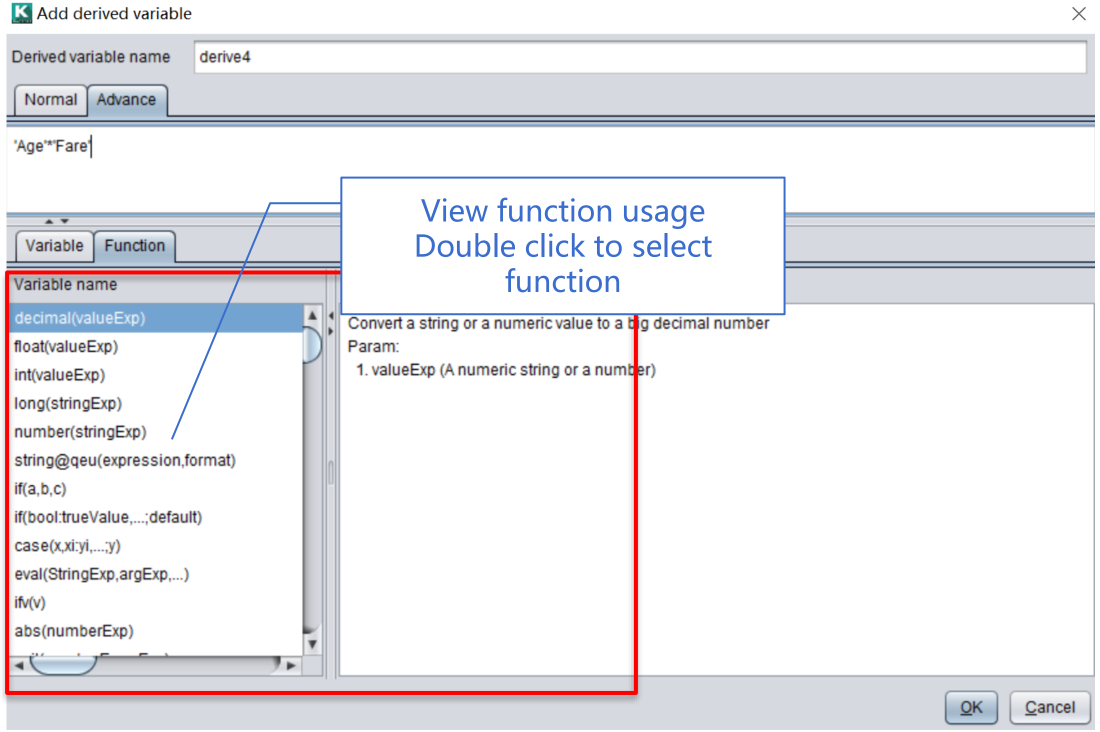
Variable name:
- Sex
- Age
- Pclass
- SibSp
- Cabin
- Fare
- Embarked
- Parch
- PassengerId
- Survived
- Name
- title
- Ticket

Variable information

Statistical method	Statistical value
Missing rate	0%
Minimum	0
Maximum	512.329
Average	32.204
Upper quartile	31
Median	14.454
Lower quartile	7.896
Standard deviation	49.693
Skewness	4.779

Free to write derived rules

View variable statistics
Double click to select variable



Add derived variable

Derived variable name: derive4

'Age'*'Fare'

Variable name:
- decimal(valueExp)
- float(valueExp)
- int(valueExp)
- long(stringExp)
- number(stringExp)
- string@qeu(expression,format)
- if(a,b,c)
- if(bool:trueValue,...;default)
- case(x,xi:yi,...;y)
- eval(StringExp,argExp,...)
- ifv(v)
- abs(numberExp)

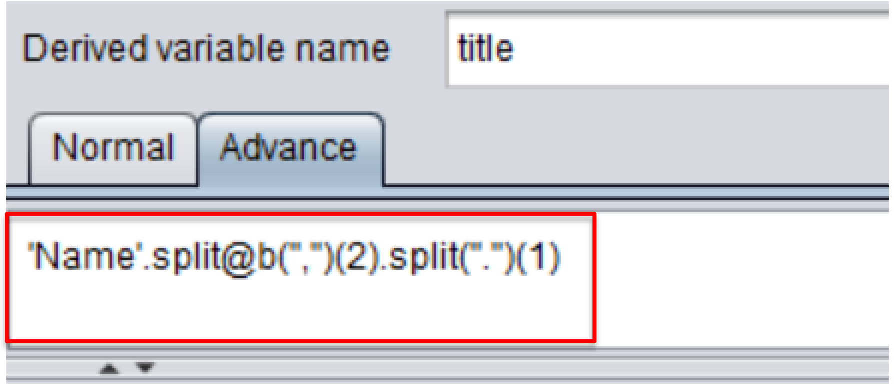
Convert a string or a numeric value to a big decimal number
Param:
1. valueExp (A numeric string or a number)

View function usage
Double click to select function

Class exercise - Other derivatives

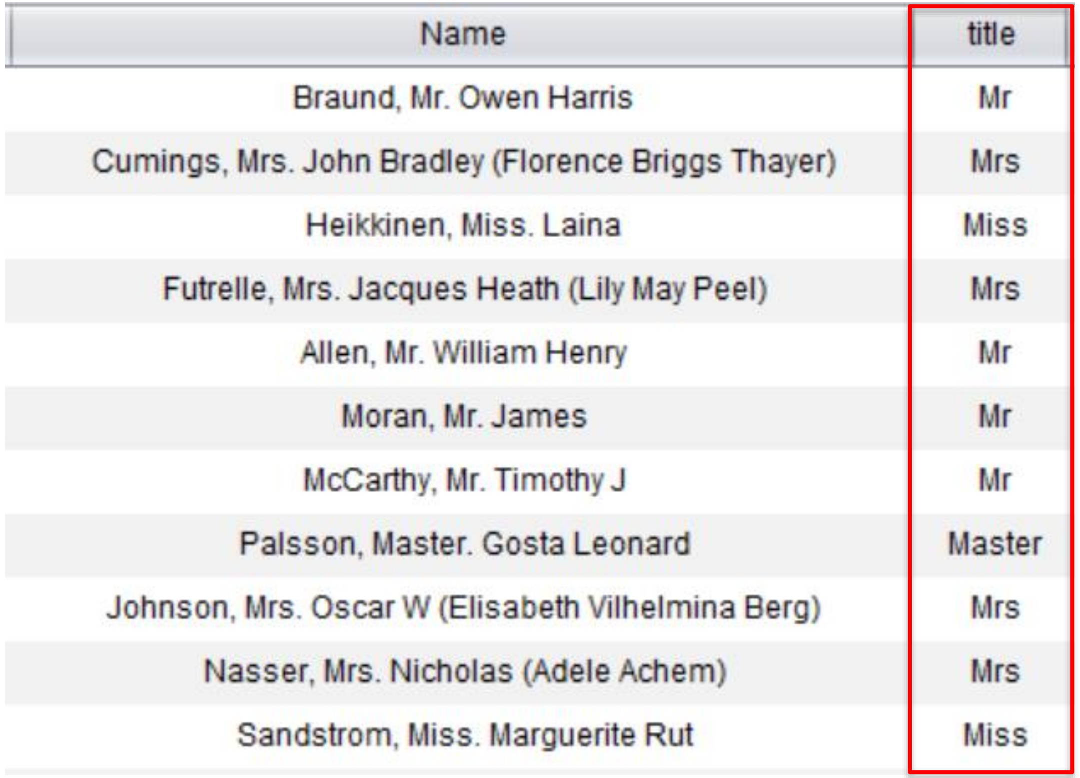
Exercise: derive the "name" of the Titanic data, named "title", and see the derived effect.

The derivative function is shown in the left figure:



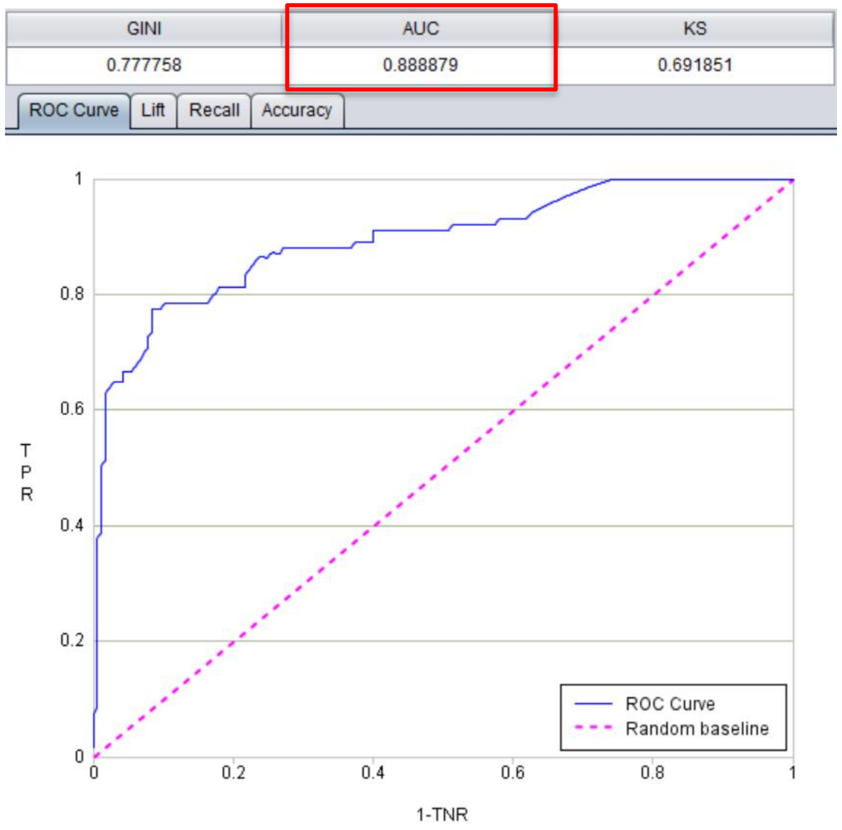
Derived variable name: title

```
'Name'.split@b(",")(2).split(".")(1)
```



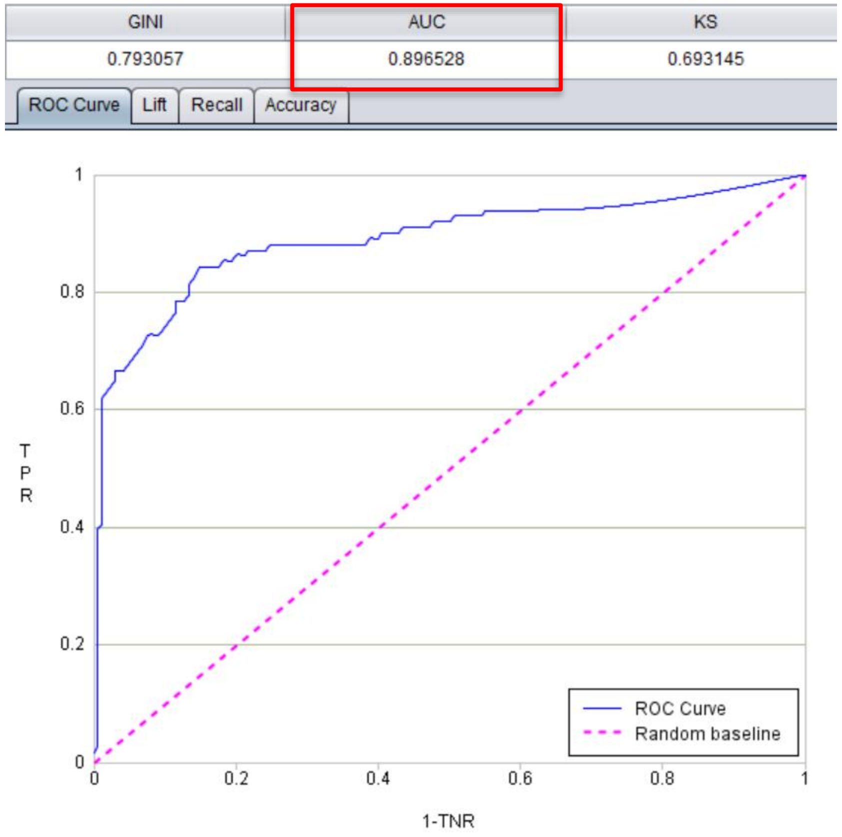
Name	title
Braund, Mr. Owen Harris	Mr
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	Mrs
Heikkinen, Miss. Laina	Miss
Futrelle, Mrs. Jacques Heath (Lily May Peel)	Mrs
Allen, Mr. William Henry	Mr
Moran, Mr. James	Mr
McCarthy, Mr. Timothy J	Mr
Palsson, Master. Gosta Leonard	Master
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	Mrs
Nasser, Mrs. Nicholas (Adele Achem)	Mrs
Sandstrom, Miss. Marguerite Rut	Miss

Class exercise - Other derivatives



GINI	AUC	KS
0.777758	0.888879	0.691851

Before adding



GINI	AUC	KS
0.793057	0.896528	0.693145

After adding

AUC slightly increased after adding "title" variable.

Conclusion

From the above several examples, we can see that the addition of derived variables can improve the effect of the model and also can reduce the effect of the model. We need to keep trying.

When adding derived variables, we must fully understand the business characteristics and consider the factors that affect the business objectives. For example, in the credit business, the user's default risk is often related to the repayment ability and willingness to repay. We can add derived variables from these two aspects. In the product purchase prediction, whether the user purchases is related to the user's purchasing ability, purchase demand, product promotion and other factors. In the health insurance scenario, the claim risk is related to the user's gender, age, physical condition, living habits and other factors.

Therefore, when adding the derived variables, it is necessary to consider what the business objectives are and what kind of variables may be derived from them, which is of great relevance to the objectives. The business significance of the fields is very important.

Algorithm selection and parameter tuning

In the chapter of modeling, we introduced some common algorithms. Different algorithms will have different performance in different data. Therefore, we can try different algorithms to optimize model performance. For the optimization of the same algorithm, we can adjust parameters to optimize model effect.

Exercise:

(1) model the Titanic data, and use the LogicClassification algorithm to observe the performance of the model.

(2) Try other algorithms and observe their performance.

(3) Change the parameters of an algorithm and observe the performance of the model.

In order to facilitate the readers to understand the meaning of parameters, the parameter introduction of common algorithms is arranged in the appendix for reference.

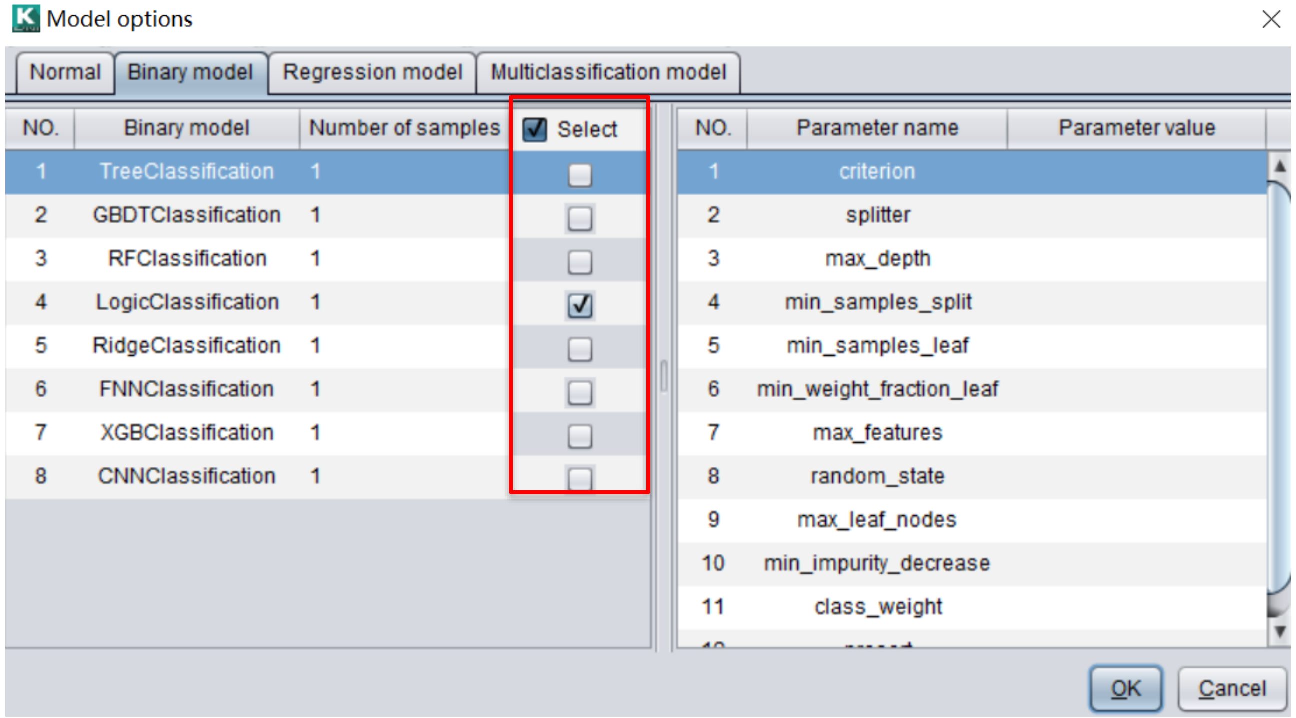
Class exercise - algorithm selection

We also use Raqsoft YModel tool to complete the exercise.

By default, YModel will calculate all algorithms in the algorithm library and automatically select the optimal algorithm or algorithm group . At the same time, it also supports the user to specify one or several algorithms for calculation (the algorithm specified by the user must be included in the YModel algorithm library).

Algorithm can be freely selected in model option menu.

Note: if you are not familiar with the algorithm or do not have specified requirements for business scenarios, it is recommended to use automatic calculation results.

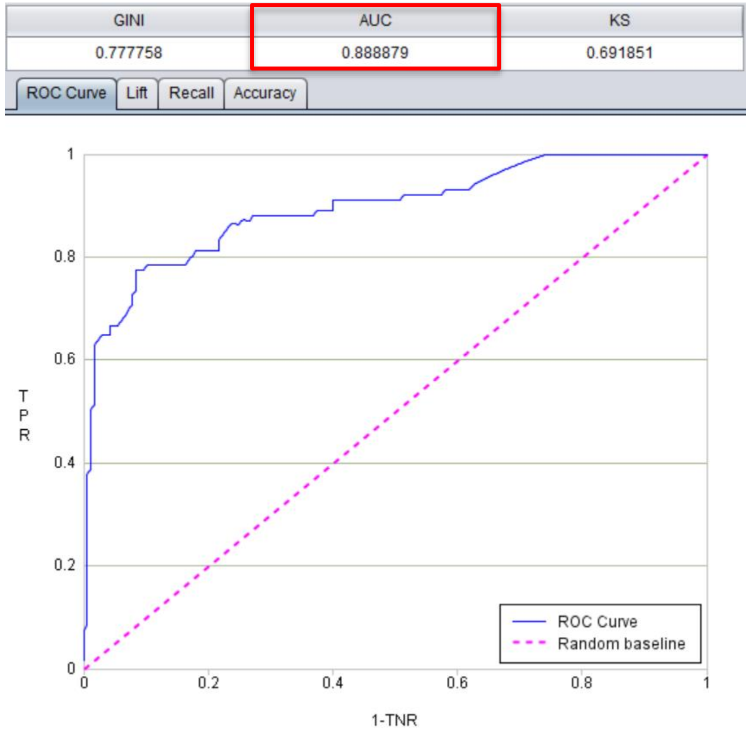


NO.	Binary model	Number of samples	☑ Select
1	TreeClassification	1	☐
2	GBDTClassification	1	☐
3	RFClassification	1	☐
4	LogicClassification	1	☑
5	RidgeClassification	1	☐
6	FNNClassification	1	☐
7	XGBClassification	1	☐
8	CNNClassification	1	☐

NO.	Parameter name	Parameter value
1	criterion	
2	splitter	
3	max_depth	
4	min_samples_split	
5	min_samples_leaf	
6	min_weight_fraction_leaf	
7	max_features	
8	random_state	
9	max_leaf_nodes	
10	min_impurity_decrease	
11	class_weight	

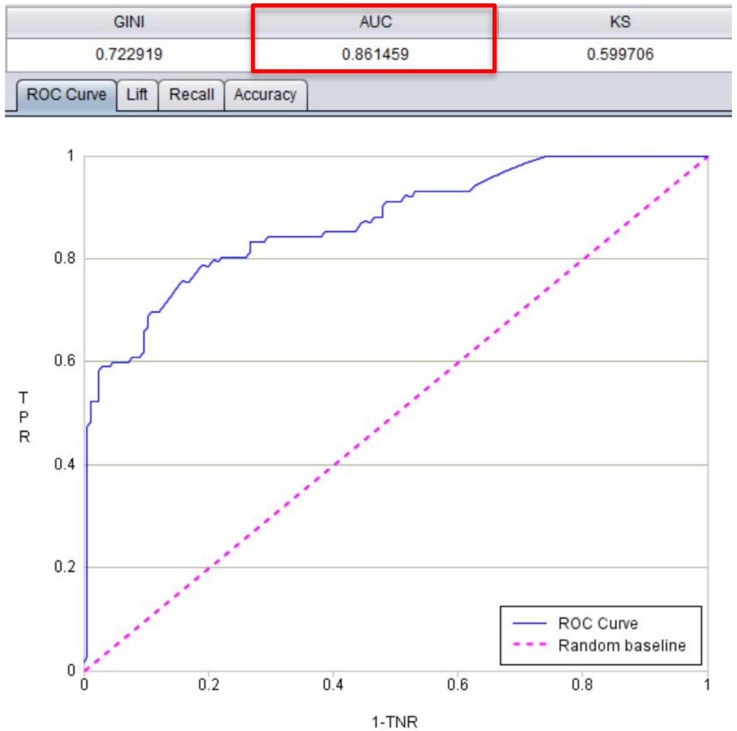
Class exercise - algorithm selection

For Titanic data, the comparison results between "logicClassification" and automatic modeling are as follows. Please practice other algorithms by yourself.



GINI	AUC	KS
0.777758	0.888879	0.691851

Automatic modeling



GINI	AUC	KS
0.722919	0.861459	0.599706

Manually specify algorithm modeling

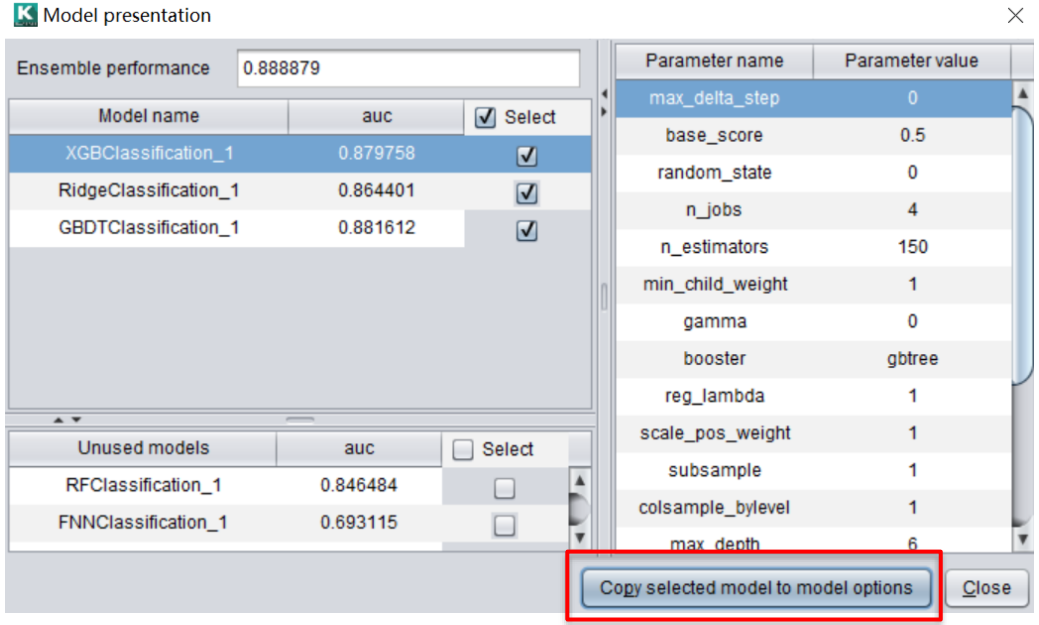
The effect of manually specifying algorithm model is not as good as that of automatically selecting algorithm.

Class exercise - parameter tuning

Using YModel products, users can further adjust parameters on the basis of automatic modeling to optimize the model and speed up the efficiency of parameter adjustment.

First, copy the algorithm and parameters automatically generated into the model options in the model presentation.

In the model option, parameters can be optimized on the basis of automatic modeling.

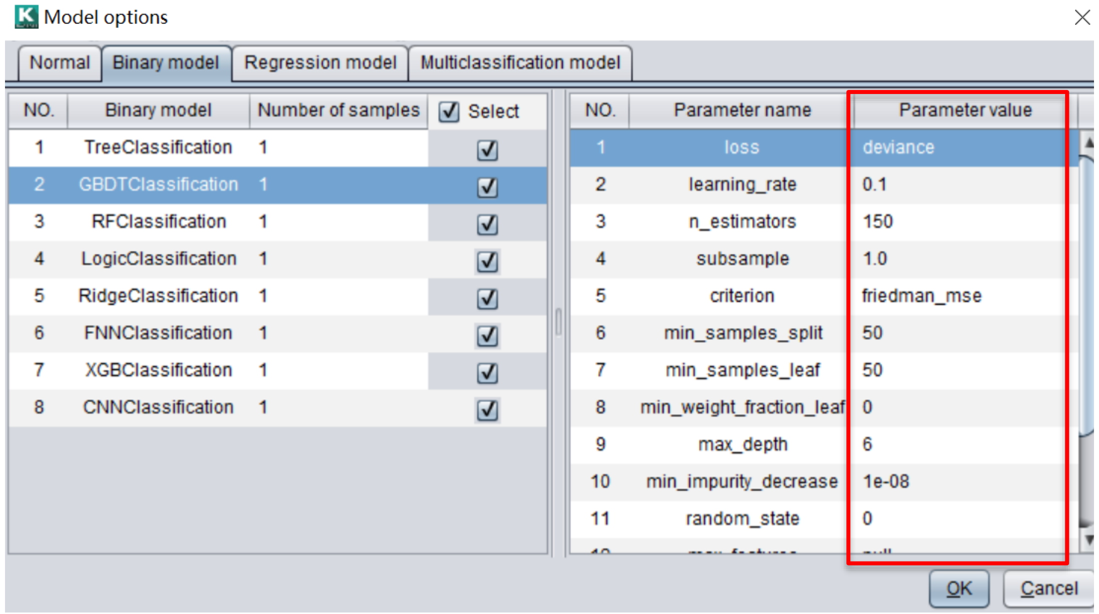


Ensemble performance: 0.888879

Model name	auc	☑ Select
XGBClassification_1	0.879758	☑
RidgeClassification_1	0.864401	☑
GBDTClassification_1	0.881612	☑

Unused models	auc	☐ Select
RFClassification_1	0.846484	☐
FNNClassification_1	0.693115	☐

Parameter name	Parameter value
max_delta_step	0
base_score	0.5
random_state	0
n_jobs	4
n_estimators	150
min_child_weight	1
gamma	0
booster	gbtree
reg_lambda	1
scale_pos_weight	1
subsample	1
colsample_bylevel	1
max_depth	6



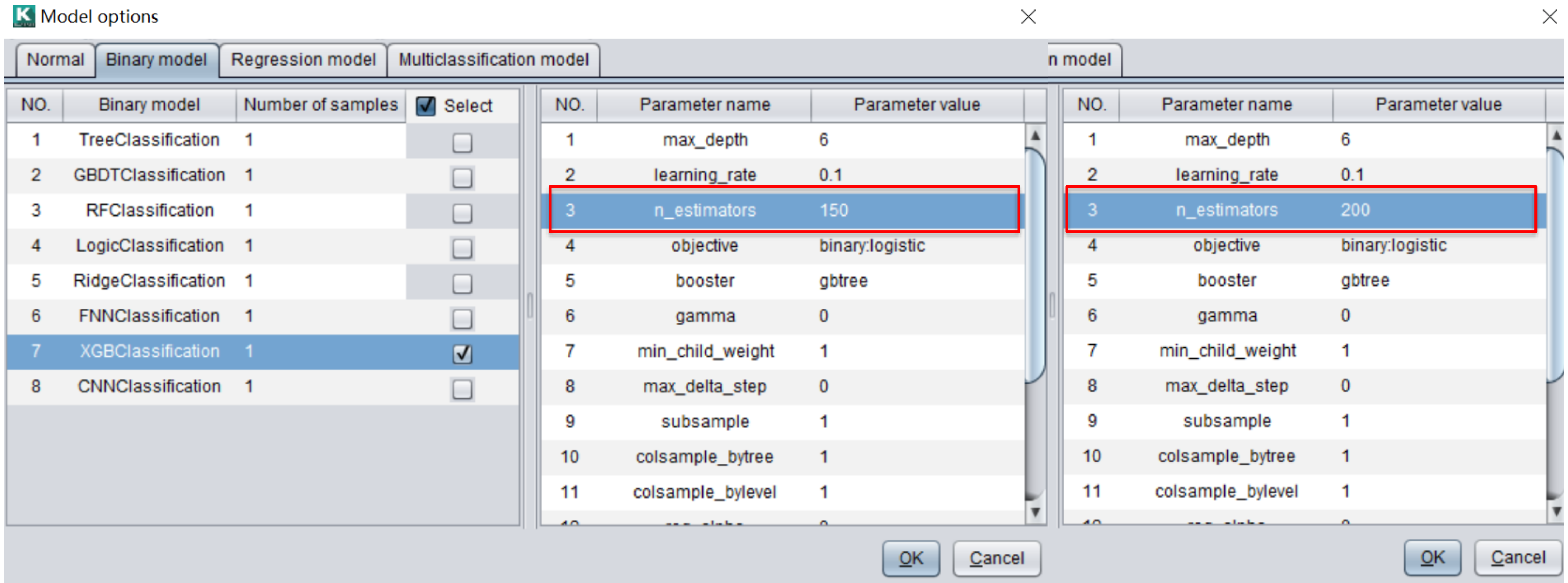
NO.	Binary model	Number of samples	☑ Select
1	TreeClassification	1	☑
2	GBDTClassification	1	☑
3	RFClassification	1	☑
4	LogicClassification	1	☑
5	RidgeClassification	1	☑
6	FNNClassification	1	☑
7	XGBClassification	1	☑
8	CNNClassification	1	☑

NO.	Parameter name	Parameter value
1	loss	deviance
2	learning_rate	0.1
3	n_estimators	150
4	subsample	1.0
5	criterion	friedman_mse
6	min_samples_split	50
7	min_samples_leaf	50
8	min_weight_fraction_leaf	0
9	max_depth	6
10	min_impurity_decrease	1e-08
11	random_state	0

Parameter value can be freely modified

Class exercise - parameter tuning

For example, based on the automatic modeling of Titanic data, the “n_estimators” of XGBC is changed from 150 to 200.

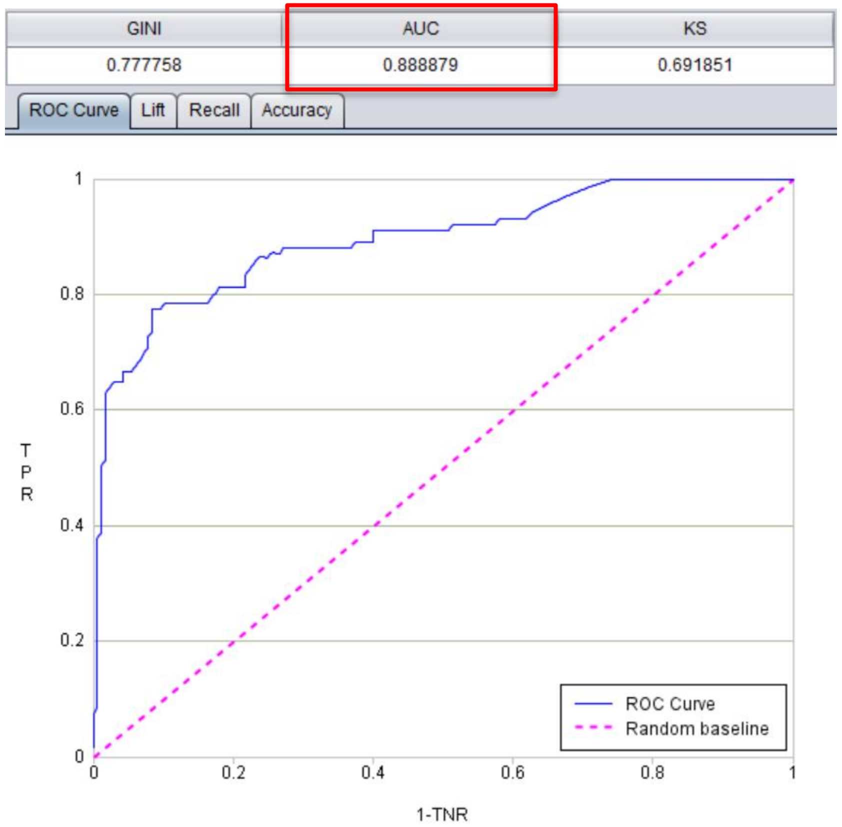


NO.	Binary model	Number of samples	☑ Select
1	TreeClassification	1	☐
2	GBDTClassification	1	☐
3	RFClassification	1	☐
4	LogicClassification	1	☐
5	RidgeClassification	1	☐
6	FNNClassification	1	☐
7	XGBClassification	1	☑
8	CNNClassification	1	☐

NO.	Parameter name	Parameter value
1	max_depth	6
2	learning_rate	0.1
3	n_estimators	150
4	objective	binary:logistic
5	booster	gbtree
6	gamma	0
7	min_child_weight	1
8	max_delta_step	0
9	subsample	1
10	colsample_bytree	1
11	colsample_bylevel	1

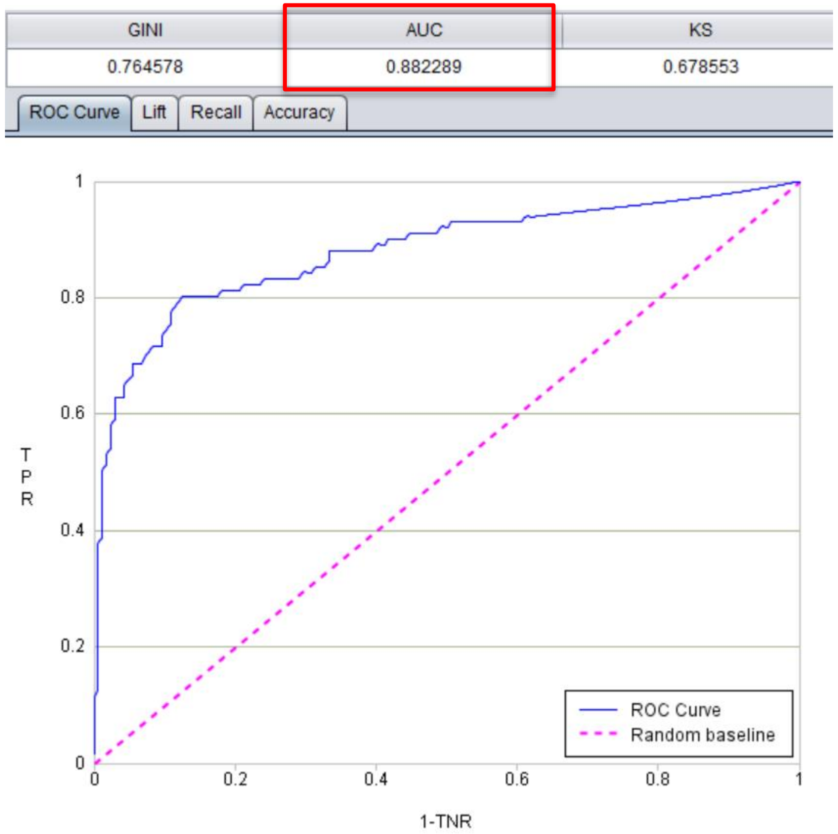
NO.	Parameter name	Parameter value
1	max_depth	6
2	learning_rate	0.1
3	n_estimators	200
4	objective	binary:logistic
5	booster	gbtree
6	gamma	0
7	min_child_weight	1
8	max_delta_step	0
9	subsample	1
10	colsample_bytree	1
11	colsample_bylevel	1

Class exercise - parameter tuning



GINI	AUC	KS
0.777758	0.888879	0.691851

Automatic parameter adjustment



GINI	AUC	KS
0.764578	0.882289	0.678553

Manual parameter adjustment

After manually modifying the parameters of the algorithm, the effect of the model decreases slightly.
Therefore, manual parameter adjustment needs rich experience and many attempts, and the efficiency of parameter adjustment is relatively low.

Appendix (common algorithm parameters) Linear regression

Parameter	Meaning
fit_intercept	If there is any truncation, if not, the straight line will cross the origin
normalize	Whether to normalize the data
copy_X	The default is true. When it is true, x will be copied, otherwise x will be overwritten
n_jobs	The number of cores used in the calculation. The default value is 1

Lasso parameter

Parameter	Meaning
alpha	Coefficient of regularization item, float, optional, default 1.0. When alpha is 0, the algorithm is equivalent to the ordinary least square method, which can be realized through linear regression, so it is not recommended to set alpha to 0
fit_intercept	Include intercept item or not, Boolean, optional, default to true
normalize	Standardize data or not, Boolean, optional, default false
copy_X	Boolean, optional, default true if true, x will be copied; otherwise, it may be overwritten
precompute	Whether to use the pre calculated Gram matrix to speed up the calculation, if set to 'auto', the machine will decide
max_iter	Int, optional, Max cycles
tol	Standard for stopping calculation, float type, default is 1e-4. Indicates that the calculation stops when the accuracy is met

Lasso parameter

Parameter	Meaning
warm_start	Warm start parameter, bool type. The default is false. If true, the last training result is used as initialization parameter, otherwise, the last training result is erased
positive	Bool, optional When set to true, force the coefficient to be positive
selection	str, default 'cyclic' if it is set to 'random', parameters will be updated randomly for each cycle, and will be updated successively according to the default setting.
random_state	Int, random seed, available only when selection is random

Ridge parameter

Parameter	Meaning
alpha	Coefficient of regularization item, float, optional, default 1.0. When alpha is 0, the algorithm is equivalent to the ordinary least square method, which can be realized through linear regression, so it is not recommended to set alpha to 0
fit_intercept	Include intercept item or not, Boolean, optional, default to true
normalize	Standardize data or not, Boolean, optional, default false
copy_X	Boolean, optional, default true if true, x will be copied; otherwise, it may be overwritten
solver	Optimization algorithm, auto：auto select svd：singular value decomposition method, which is more suitable for singular matrix calculation than Cholesky cholesky：Cholesky method sparse_cg：conjugate gradient method, suitable for big data calculation lsqr：least square method, sag：random average gradient descent method, good performance in big data

Ridge parameter

Parameter	**Meaning**
max_iter	Int, optional, Max cycles
tol	Standard for stopping calculation, float type, default is 1e-4. Indicates that the calculation stops when the accuracy is met
random_state	Int, random seed

Elastic Net parameter

Parameter	Meaning
alpha	Coefficient of regularization item, float, optional, default 1.0. When alpha is 0, the algorithm is equivalent to the ordinary least square method, which can be realized through linear regression, so it is not recommended to set alpha to 0
l1_ratio	Between 0 and 1, between L1 penalty and L2 penalty, if l1_ratio=1， then L1, if L1_ ratio = 1, then L2
fit_intercept	Include intercept item or not, Boolean, optional, default to true
normalize	Standardize data or not, Boolean, optional, default false
copy_X	Boolean, optional, default true if true, x will be copied; otherwise, it may be overwritten
precompute	Whether to use the pre calculated Gram matrix to speed up the calculation, if set to 'auto', the machine will decide
max_iter	Int, optional, Max cycles
tol	Standard for stopping calculation, float type, default is 1e-4. Indicates that the calculation stops when the accuracy is met

Elastic Net parameter

Parameter	Meaning
warm_start	Warm start parameter, bool type. The default is false. If true, the last training result is used as initialization parameter, otherwise, the last training result is erased
positive	Bool, optional When set to true, force the coefficient to be positive
selection	str, default 'cyclic' if it is set to 'random', parameters will be updated randomly for each cycle, and will be updated successively according to the default setting.
random_state	Int, random seed, available only when selection is random

logistic

Parameter	Meaning
penalty	Penalty item, STR type, optional parameters are L1 and L2, default is L2. Newton CG, sag and lbfgs only support L2 specification. L1 assumes that the parameters of the model satisfy the Laplace distribution, L2 assumes that the parameters of the model satisfy the Gaussian distribution.
dual	Dual or primitive method, bool type, default is false. Dual method is only used to solve L2 penalty term of iblinear. When the number of samples > sample characteristics, dual is usually set to false
tol	Standard for stopping calculation, float type, default is 1e-4. Indicates that the calculation stops when the accuracy is met
c	The reciprocal of the regularization coefficient λ, float type, default is 1.0. Must be a positive floating point number. The smaller the number, the stronger the regularization
fit_intercept	Include intercept item or not, Boolean, optional, default to true
intercept_scaling	Only useful when the regularization item is "liblinear" and fit_intercept is set to true. Float type, default is 1

logistic

Parameter	Meaning
class_weight	It represents the weight of various types in the classification model. It can be a dictionary or 'balanced' string. It is not entered by default, that is, it does not consider the weight, that is, it is none. If you choose to input, you can select balanced to let the class library calculate the type weight by itself, or you can input the weight of each type by yourself.
random_state	Random number seed, int type, optional parameter, default to none, only useful when the optimization algorithm is sag, liblinear
solver	There are five optional parameters, namely newton-cg,lbfgs,liblinear,sag,saga. The default is liblinear. The solver parameter determines our optimization method for the loss function of logistic regression. liblinear：using the open-source liblinear library to implement, using of the axis descent method to iteratively optimize the loss function. lbfgs：a kind of quasi Newton method, which uses the second derivative matrix of loss function, i.e. Hessian matrix, to iteratively optimize the loss function. newton-cg：it is also a kind of Newton method. It uses the second derivative matrix of loss function, namely Hessian matrix, to iteratively optimize the loss function. sag： random average gradient descent, which is a variation of gradient descent method. The difference between the gradient descent method and the general gradient descent method is that each iteration only uses a part of the sample to calculate the gradient, which is suitable for the case when there are many sample data. saga：a variation of linear convergence stochastic optimization algorithm.

logistic

Parameter	Meaning
max_iter	Algorithm convergence maximum number of iterations, int type, default is 100. Only useful when the optimization algorithm is Newton-CG, sag and lbfgs. The maximum number of iterations of the algorithm convergence
multi_class	Multi classification mode selection parameter, str type, optional parameters are ovr and multinomial, default is ovr
verbose	Log redundancy length, int type. The default is 0, the training process is not output. The result will be output occasionally when it is 1; when it is greater than 1, the result will be output for each sub model.
warm_start	Warm start parameter, bool type. The default is false. If true, the last training result is used as initialization parameter, otherwise, the last training result is erased
n_jobs	Parallel number. Int type, default is 1. When 1, use one core of CPU to run programs, and when 2, use two cores of CPU to run programs. When - 1, run the program with all CPU cores.

Decision tree parameters

Prameter	DecisionTreeClassifier	DecisionTreeRegressor
criterion	For feature selection criteria, "Gini" or "entropy" can be used. The former represents Gini coefficient and the latter represents information gain. Generally speaking, the default Gini coefficient can be used, that is, CART algorithm.	You can use "MSE" or "MAE". The former is the mean squared error, that is, the average of the sum of the squares of the errors. The latter is the average of the absolute error.
splitter	The selection criteria of feature points can be "best" or "random". The former finds the best partition point among all the partition points of the feature. The latter is to find the local optimal partition point randomly. The default "best" is suitable when the sample size is not large, and if the sample data size is very large, the decision tree construction recommends "random “.	
max_features	The maximum number of features considered in the partition. It is "None" by default, which means that all features are considered in the partition; if "log2" means that at most logN based on 2 features are considered in the partition; if "sqrt" or "auto" means that at most the squared root of N features are considered in the partition. If it is an integer, it represents the absolute number of features considered. If it is a floating-point number, it represents the feature percentage under consideration, that is, the number of features after the percentage rounding is considered. Where N is the total feature number of samples.	
max_depth	The maximum depth of the decision tree. It is not entered by default. If it is not entered, the depth of the subtree will not be limited. Generally speaking, this value can be ignored when there is little data or features. If the model has many samples and features, it is recommended to limit the maximum depth. The specific value depends on the distribution of data. Common values can be between 10-100.	

Decision tree parameters

Prameter	DecisionTreeClassifier	DecisionTreeRegressor
min_samples_split	The minimum number of samples required for node splitting, if the number of samples of a node is less than min_ samples_ split, it will not continue to try to select the best feature to divide. The default is 2.	
min_samples_leaf	The minimum sample number of leaf nodes. If the number of leaf nodes is less than the sample number, they will be pruned together with their sibling nodes. The default value is 1. You can enter the integer of the minimum number of samples or the percentage of the minimum number of samples in the total number of samples.	
min_weight_fraction_leaf	The minimum sample weight sum of leaf nodes, if less than this value, will be pruned together with brother nodes. The default value is 0, that is, the weight is not considered. Generally speaking, if there are many samples with missing values, or the distribution category deviation of the classification tree samples is large, the sample weight will be introduced, and this value should be paid attention to.	
max_leaf_nodes	Maximum leaf nodes. By limiting the maximum number of leaf nodes, over fitting can be prevented. The default value is "None", that is, the maximum number of leaf nodes is not limited. If the limit is added, the algorithm will establish the optimal decision tree within the maximum number of leaf nodes.	

Decision tree parameters

Prameter	DecisionTreeClassifier	DecisionTreeRegressor
class_weight	The weight of categories. It is mainly to prevent too many samples of some categories in the training set, which results in the decision tree leaning too much to these categories. By default, "None", you can specify the weight of each sample by yourself, or use "balanced". If you use "balanced", the algorithm will calculate the weight by itself, and the sample weight corresponding to the category with small sample size will be higher.	Not applicable to regression trees
min_impurity_split	This value limits the growth of decision tree. If the impurity (Gini coefficient, information gain, mean square deviation, absolute difference) of a node is less than this threshold, the node will not generate a child node.	
presort	Whether to pre sort the data or not. Boolean value. Default false: do not sort.	

Random forest parameters

Parameter	Meaning
n_estimators	The number of decision trees. If n_estimators is too small, it is easy to under fit; If it is too large, it can not significantly enhance the model. So choose the right number of n_ estimators.
bootstrp	Whether to use put back sampling for sample set to build tree, true means yes, and the default value is true.
oob_score	Whether to evaluate the quality of the model with out of bag samples. True means yes, the default value is false.

Random forest parameters

Parameters	Meaning
criterion	For feature selection criteria, "Gini" or "entropy" can be used. The former represents Gini coefficient and the latter represents information gain. Generally speaking, the default Gini coefficient can be used, that is, CART algorithm. You can use "MSE" or "MAE “ for regression model. The former is the mean squared error, that is, the average of the sum of the squares of the errors. The latter is the average of the absolute error.
max_features	The maximum number of features considered in the partition. It is "None" by default, which means that all features are considered in the partition; if "log2" means that at most logN based on 2 features are considered in the partition; if "sqrt" or "auto" means that at most the squared root of N features are considered in the partition. If it is an integer, it represents the absolute number of features considered. If it is a floating-point number, it represents the feature percentage under consideration, that is, the number of features after the percentage rounding is considered. Where N is the total feature number of samples.
max_depth	The maximum depth of the decision tree. It is not entered by default. If it is not entered, the depth of the subtree will not be limited. Generally speaking, this value can be ignored when there is little data or features. If the model has many samples and features, it is recommended to limit the maximum depth. The specific value depends on the distribution of data. Common values can be between 10-100.
min_samples_leaf	The minimum sample number of leaf nodes. If the number of leaf nodes is less than the sample number, they will be pruned together with their sibling nodes. The default value is 1. You can enter the integer of the minimum number of samples or the percentage of the minimum number of samples in the total number of samples.

Random forest parameters

Parameters	Meaning
min_samples_split	The minimum number of samples required for node splitting, if the number of samples of a node is less than min_samples_ split, it will not continue to try to select the best feature to divide. The default is 2.
max_leaf_nodes	Maximum leaf nodes. By limiting the maximum number of leaf nodes, over fitting can be prevented. The default value is "None", that is, the maximum number of leaf nodes is not limited. If the limit is added, the algorithm will establish the optimal decision tree within the maximum number of leaf nodes.
min_impurity_decrease	This value limits the growth of decision tree. If the impurity (Gini coefficient, information gain, mean square deviation, absolute difference) of a node is less than this threshold, the node will not generate a child node.
min_samples_leaf	The minimum sample number of leaf nodes. If the number of leaf nodes is less than the sample number, they will be pruned together with their sibling nodes. The default value is 1. You can enter the integer of the minimum number of samples or the percentage of the minimum number of samples in the total number of samples.
min_weight_fraction_leaf	The minimum sample weight sum of leaf nodes, if less than this value, will be pruned together with brother nodes. The default value is 0, that is, the weight is not considered. Generally speaking, if there are many samples with missing values, or the distribution category deviation of the classification tree samples is large, the sample weight will be introduced, and this value should be paid attention to.

GBDT parameters

Parameters	Meaning
n_estimators	The maximum number of weak learners.
learning_rate	Step size, i.e. the weight reduction coefficient a of each learner, is one of the gbdt regularization methods
subsample	Subsampling, value (0,1]. It is also one of gbdt regularization methods to decide whether to sample the original dataset and the proportion of sampling.
init	Weak learner at initialization time. If not set, the default is used.
loss	Loss function, optional {'ls'-square loss function，'lad’ - absolute loss function, 'huber'-huber loss function,'quantile'-quantile loss function}，default 'ls'
alpha	When using "Huber" and "quantile", you need to specify the corresponding value. The default value is 0.9. If there are many noise points, the quantile value can be reduced appropriately

GBDT parameters

Parameters	Meaning
criterion	The decision tree node search criterion for the optimal segmentation point. Default is "friedman_mse"， "mse “ and 'mae “ are optional.
max_features	The maximum number of features considered in the partition. It is "None" by default, which means that all features are considered in the partition; if "log2" means that at most logN based on 2 features are considered in the partition; if "sqrt" or "auto" means that at most the squared root of N features are considered in the partition. If it is an integer, it represents the absolute number of features considered. If it is a floating-point number, it represents the feature percentage under consideration, that is, the number of features after the percentage rounding is considered. Where N is the total feature number of samples.
max_depth	The maximum depth of the decision tree. It is not entered by default. If it is not entered, the depth of the subtree will not be limited. Generally speaking, this value can be ignored when there is little data or features. If the model has many samples and features, it is recommended to limit the maximum depth. The specific value depends on the distribution of data. Common values can be between 10-100.
min_samples_leaf	The minimum sample number of leaf nodes. If the number of leaf nodes is less than the sample number, they will be pruned together with their sibling nodes. The default value is 1. You can enter the integer of the minimum number of samples or the percentage of the minimum number of samples in the total number of samples.

GBDT parameters

Parameter	Meaning
min_samples_split	The minimum number of samples required for node splitting, if the number of samples of a node is less than min_ samples_ split, it will not continue to try to select the best feature to divide. The default is 2.
min_samples_leaf	The minimum sample number of leaf nodes. If the number of leaf nodes is less than the sample number, they will be pruned together with their sibling nodes. The default value is 1. You can enter the integer of the minimum number of samples or the percentage of the minimum number of samples in the total number of samples.
min_impurity_split	Minimum impurity of node division
max_leaf_nodes	Maximum leaf nodes. By limiting the maximum number of leaf nodes, over fitting can be prevented. The default value is "None", that is, the maximum number of leaf nodes is not limited. If the limit is added, the algorithm will establish the optimal decision tree within the maximum number of leaf nodes.
presort	Whether to sort the data in advance to speed up the search of the optimal segmentation point. The default is pre sorting. If the data is sparse, it will not be pre sorted. In addition, the sparse data cannot be set to true.
validationfraction	The proportion of validation data reserved for early stop. Can only be used when n_iter_no_change is set.
n_iter_no_change	Used to decide whether to use early stop to terminate training when the verification score does not improve.

XGB parameters

Parameter	Classification	Regression
objective	Learning objective Binary： binary:logistic Return prediction probability; binary:logitraw Return the score before logistic transformation; binary:hinge Return 0 or 1 classification instead of probability Multi： multi：softmax num_class=n return class multi：softprob num_class=n return probability	Learning objective reg:linear (default) reg:logistic
eval_metric	(default error)： auc--Area under ROC curve error--error rate（Binary） merror--error rate（Multi） logloss--negative log likelihood function（Binary） mlogloss--negative log likelihood function（Multi）	(default rmse)： rmse--root mean square error mae--mean absolute error
seed	[default 0] Random seed, which can be set to reproduce the results of random data, can also be used to adjust parameters	

XGB parameters

Parameter	Meaning
eta	[default 0.3] Similar to the learning rate of GBM, the stability of the model can be improved by reducing the weight of each step.
min_child_weight	Minimum leaf node sample weight sum.
max_depth	[default 6] Same as the parameters in GBM, it is the maximum depth of the tree, which is used to avoid over fitting.
max_leaf_nodes	Maximum number of leaf nodes
gamma	[default 0] The minimum loss function reduction required for node splitting. The larger the value of this parameter, the more conservative the algorithm is.
max_delta_step	[default 0] This parameter limits the maximum step size for each tree weight change. If the value of this parameter is 0, it means there is no constraint. If it is given a positive value, it makes the algorithm more conservative. Usually, this parameter does not need to be set. But when the samples of each category are very unbalanced, it is very helpful for logical regression.

XGB parameters

Parameter	Meaning
subsample	[default 1] Same as subsample parameter in GBM. This parameter controls the proportion of random samples per tree. Reducing the value of this parameter will make the algorithm more conservative and avoid over fitting. However, if this value is set too small, it can cause under fitting.
colsample_bytree	[default 1] Similar to max_features parameter in GBM. Used to control the proportion of random sampling columns per tree(each column is a feature).
colsample_bylevel	[default 1] Sample scale of columns at each horizontal level when splitting
lambda	[default 1] L2 regular term, similar to ridge regression
alpha	[default 1] L1 regular term, similar to lasso regression, can be applied in the case of high dimensions, making the algorithm faster.
scale_pos_weight	[default 1] Adjust positive and negative sample balance, negative sample / positive sample

Glossary – for reference

derived variable
binning
Equi-width binning
Equi-frequency binning
transformation
interaction
ratio
mean encoding
odds encoding
log-odds encoding
mean encoding
tangent
arc tangent
hyperbolic tangent
Box-Cox

Chapter 7 Comprehensive cases

7.1 Classification model case

7.2 Regression model case

Data presentation

Titanic survivor prediction is a classic game on kaggle. This section introduces the data mining process based on this data. (It’ s oriented for beginners, masters please bypass)

“titanic_train.csv” :
There are 891 samples and 12 variables in the training set (with target variable).

“titanic_test.csv” :
There are 418 samples and 11 variables in the set to be tested (without target variable).

Analysis objectives:

1. Find out the factors that affect the survival of passengers.
2. Build model according to the training set and predict with the test set.

No.	Variable	Description
1	PassengerId	Passenger ID
2	Survived	Survived or not
3	Pclass	Class of ticket
4	Name	Passenger name
5	Sex	Passenger gender
6	Age	Passenger age
7	SibSp	Number of siblings and spouse of passenger
8	Parch	Number of parents and children of passenger
9	Ticket	Ticket number
10	Fare	Fare price
11	Cabin	Cabin
12	Embarked	Embarked harbor

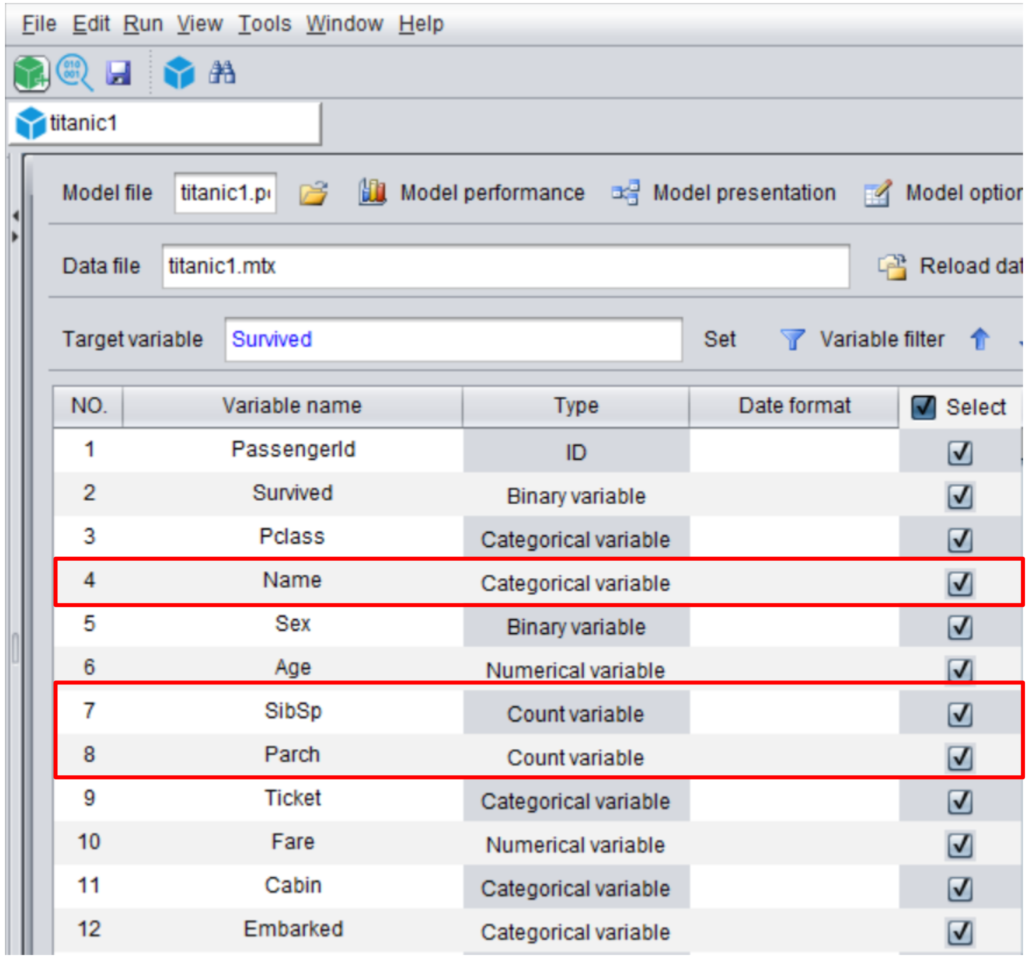
Data dictionary

Now that we have a general understanding of the data, let's explore and preprocess the dataset to try to find out the factors that affect the survival of passengers.

Data exploration

1. Check variable type according to data dictionary

The figure on the right is the variable type automatically recognized by Raqsoft YModel tool. “Name” is recognized as ID because it has no duplicate value. Like “Passengerid”, it is considered as the unique identification of each record. However, we need to extract some information from name, so we change it to categorical variable. “Sibsp” and “parch” represent the number of family members, but are recognized as categorical variables, so they are changed to count variables.



Model file titanic1.p

Data file titanic1.mtx

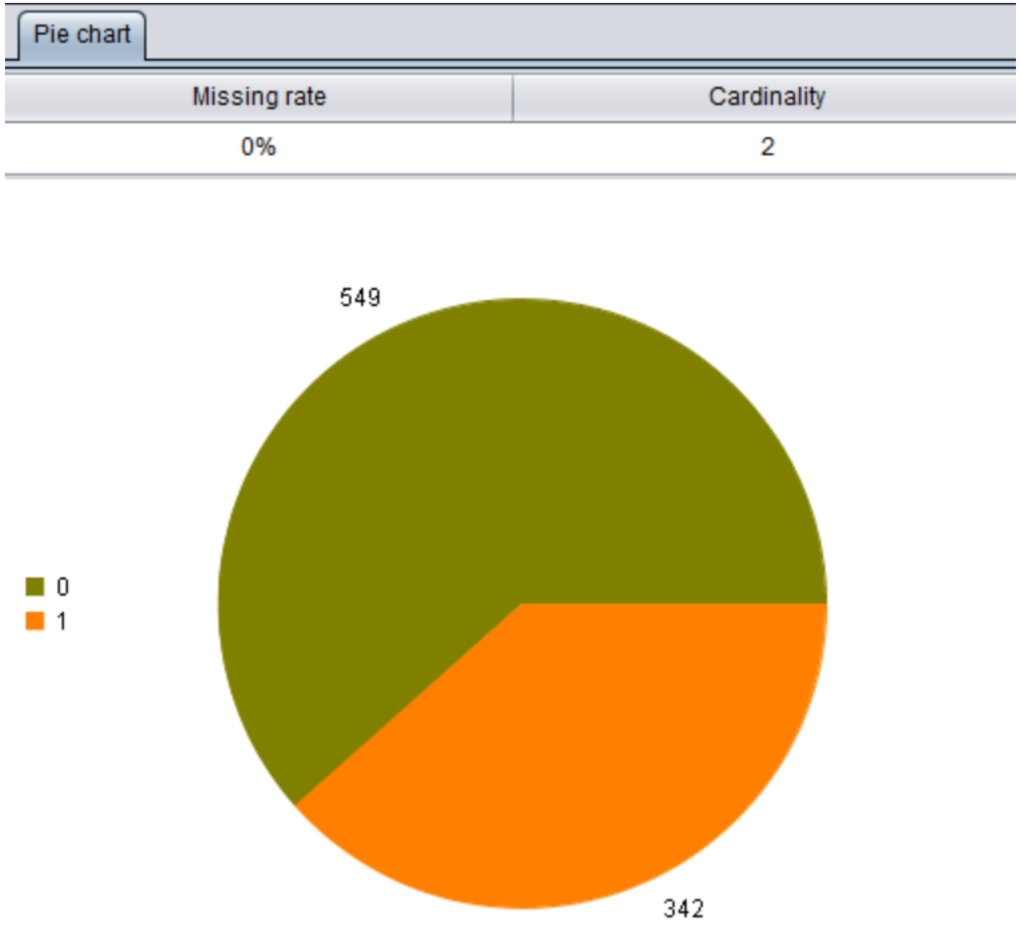
Target variable Survived

NO.	Variable name	Type	Date format	☑ Select
1	PassengerId	ID		☑
2	Survived	Binary variable		☑
3	Pclass	Categorical variable		☑
4	Name	Categorical variable		☑
5	Sex	Binary variable		☑
6	Age	Numerical variable		☑
7	SibSp	Count variable		☑
8	Parch	Count variable		☑
9	Ticket	Categorical variable		☑
10	Fare	Numerical variable		☑
11	Cabin	Categorical variable		☑
12	Embarked	Categorical variable		☑

Data exploration - distribution analysis

2. Target variable “Survived”

There are two categories 1 and 0 for survival, where 1 represents survival and 0 represents death. There is no missing value and it is the target variable of this modeling.



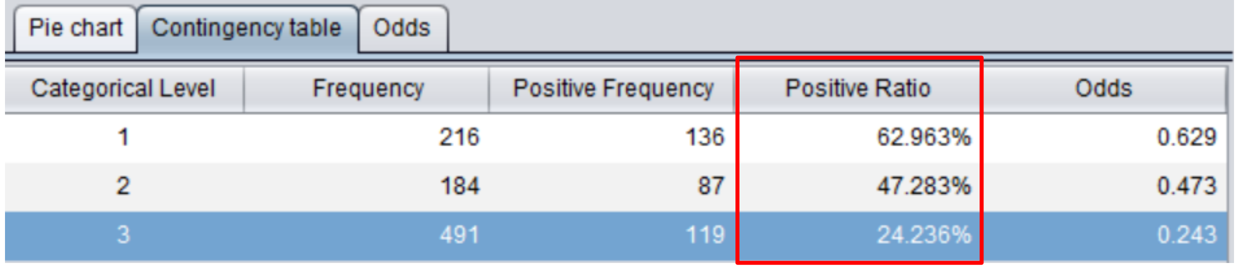
Missing rate	Cardinality
0%	2

Data exploration - distribution analysis

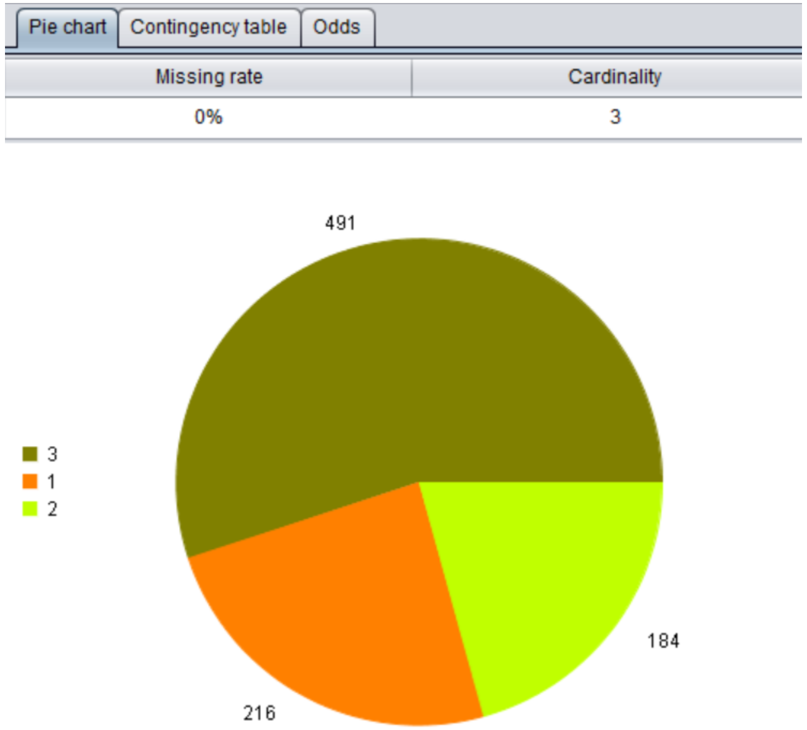
3. Pclass

There are three classes of tickets, namely 1, 2 and 3, among which the number of people in class 1 and 2 is small, the number of people in class 3 is large and there is no missing value.

After grouping by target variable and looking at the survival rate, it can be found that the higher the ticket grade is, the higher the survival rate is. (It’ s very important to work hard to make money!)



Categorical Level	Frequency	Positive Frequency	Positive Ratio	Odds
1	216	136	62.963%	0.629
2	184	87	47.283%	0.473
3	491	119	24.236%	0.243

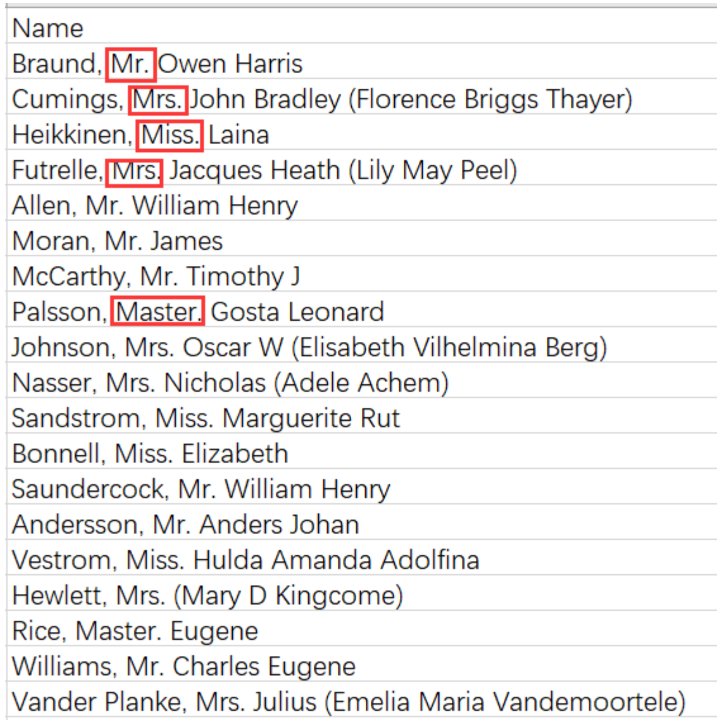


Missing rate	Cardinality
0%	3

Data exploration

3. Name

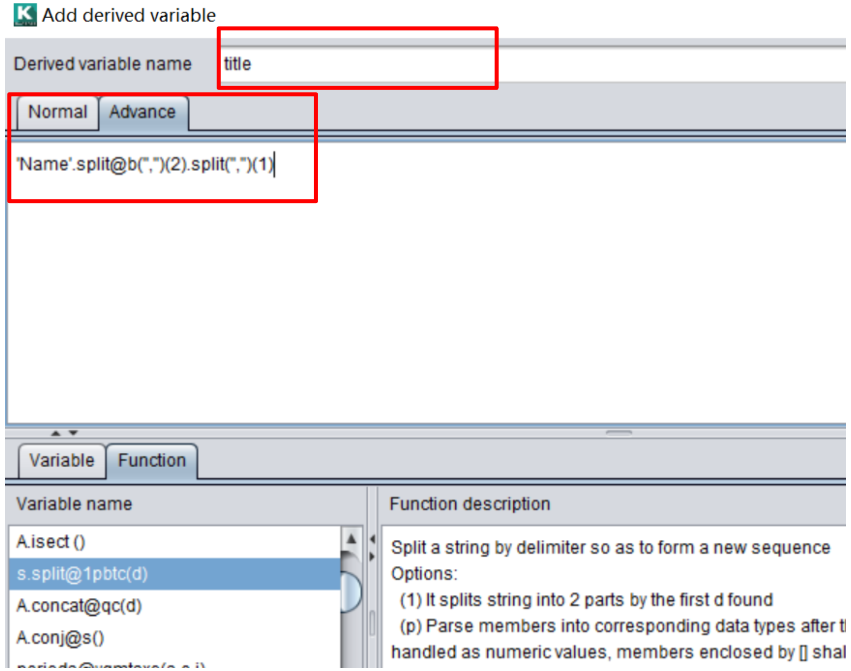
Name, it is found that it contains some information, such as Mr. , Miss, etc. , which is extracted to see if it is useful.



Name
Braund, Mr. Owen Harris
Cumings, Mrs. John Bradley (Florence Briggs Thayer)
Heikkinen, Miss. Laina
Futrelle, Mrs. Jacques Heath (Lily May Peel)
Allen, Mr. William Henry
Moran, Mr. James
McCarthy, Mr. Timothy J
Palsson, Master. Gosta Leonard
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
Nasser, Mrs. Nicholas (Adele Achem)
Sandstrom, Miss. Marguerite Rut
Bonnell, Miss. Elizabeth
Saundercock, Mr. William Henry
Andersson, Mr. Anders Johan
Vestrom, Miss. Hulda Amanda Adolfina
Hewlett, Mrs. (Mary D Kingcome)
Rice, Master. Eugene
Williams, Mr. Charles Eugene
Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)

Data preprocessing - generate derived variables

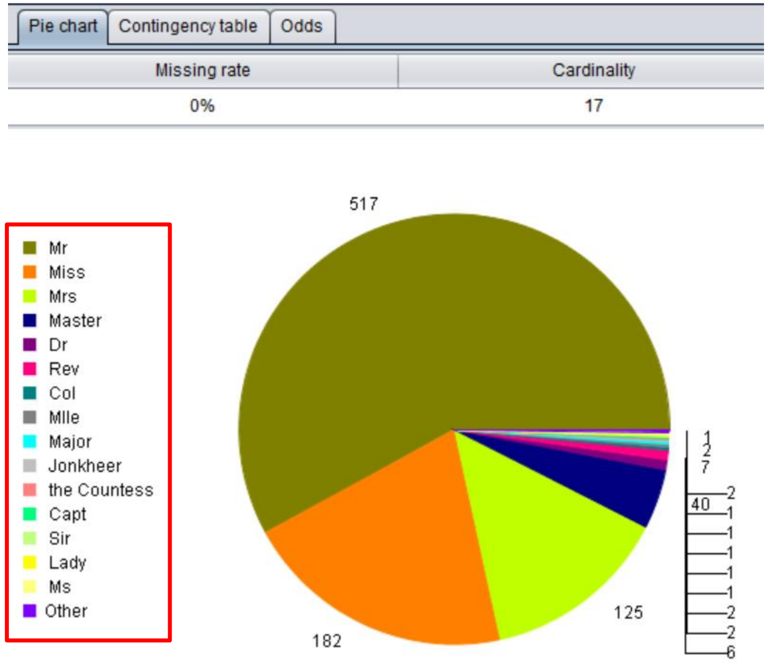
Use variable “Name” to generate variable “title”



Derived variable name: title

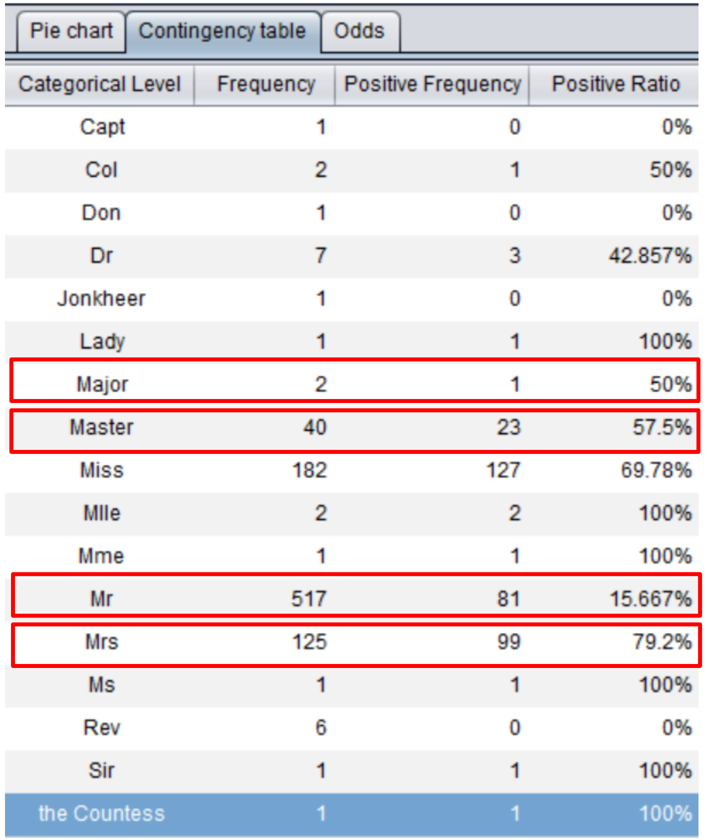
```
'Name'.split@b(",")(2).split(",")(1)
```

Variable name	Function description
A.isect ()	Split a string by delimiter so as to form a new sequence Options: (1) It splits string into 2 parts by the first d found (p) Parse members into corresponding data types after t handled as numeric values, members enclosed by [] shal
s.split@1pbtc(d)	
A.concat@qc(d)	
A.conj@s()	



Missing rate	Cardinality
0%	17

- 🟫 Mr
- 🟧 Miss
- 🟩 Mrs
- 🟦 Master
- 🟪 Dr
- 🟥 Rev
- 🟦 Col
- ⬜ Mlle
- 🟦 Major
- ⬜ Jonkheer
- 🟥 the Countess
- 🟩 Capt
- 🟩 Sir
- 🟨 Lady
- 🟨 Ms
- 🟪 Other


Pie chart
Contingency table
Odds

Categorical Level	Frequency	Positive Frequency	Positive Ratio
Capt	1	0	0%
Col	2	1	50%
Don	1	0	0%
Dr	7	3	42.857%
Jonkheer	1	0	0%
Lady	1	1	100%
Major	2	1	50%
Master	40	23	57.5%
Miss	182	127	69.78%
Mlle	2	2	100%
Mme	1	1	100%
Mr	517	81	15.667%
Mrs	125	99	79.2%
Ms	1	1	100%
Rev	6	0	0%
Sir	1	1	100%
the Countess	1	1	100%

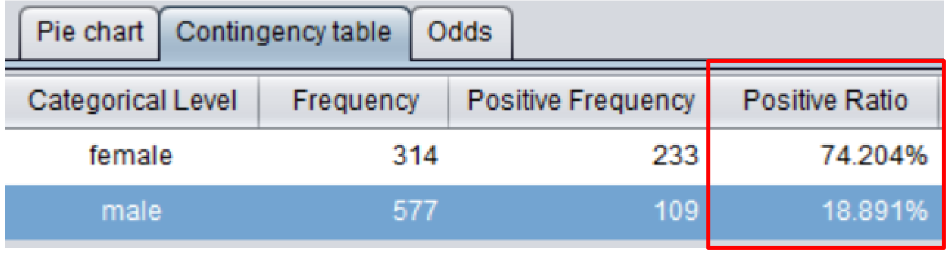
Extract the information of the names and check the statistics after grouping, it is found that the survival rate of Miss, Mrs. and master is very high, but the survival rate of Mr. is very low, indicating that this variable is useful.

Data exploration - distribution analysis

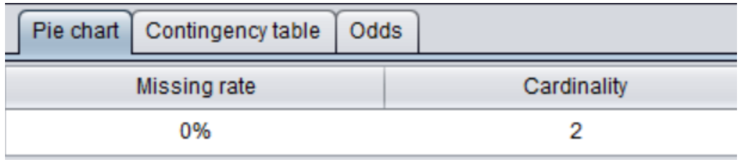
4. Sex

Gender, there are two categories of men and women, most of which are men. No missing values.

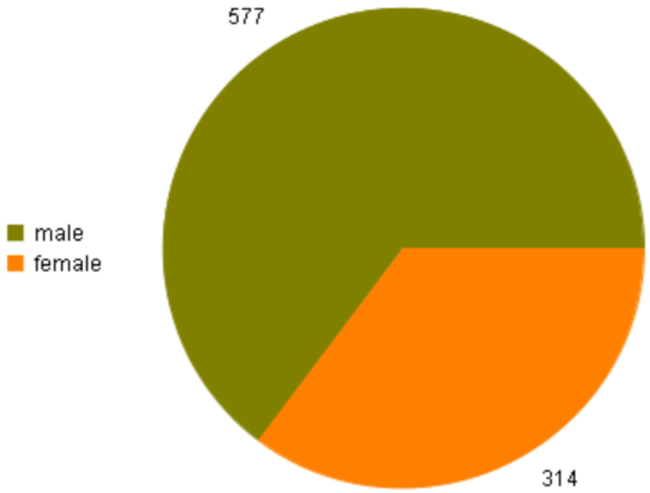
After grouping, the survival rate of women was much higher than that of men. It shows that this variable is very important.



Categorical Level	Frequency	Positive Frequency	Positive Ratio
female	314	233	74.204%
male	577	109	18.891%

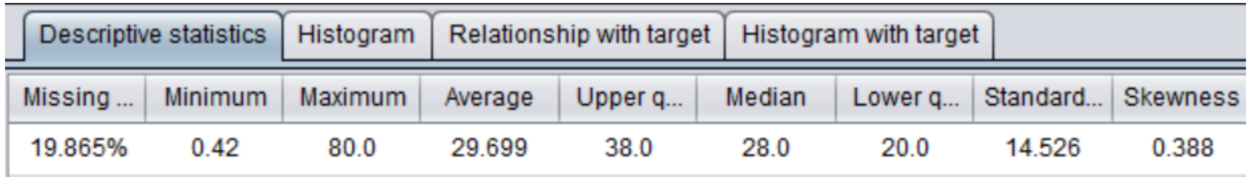


Missing rate	Cardinality
0%	2



Data exploration: distribution analysis, contrastive analysis and statistical analysis

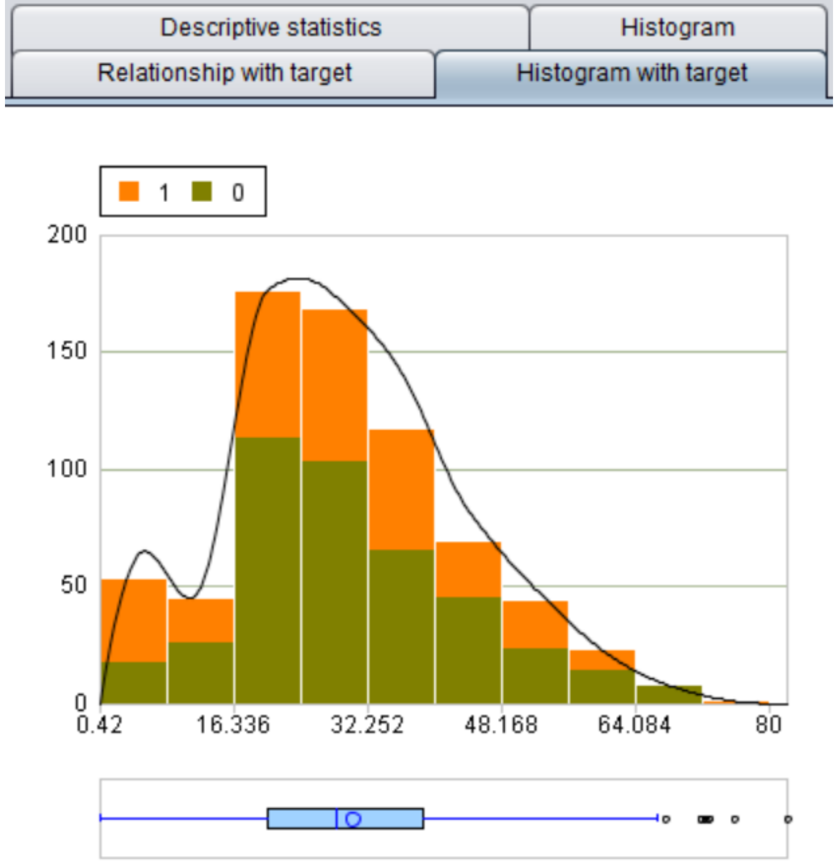
5. Age



Missing ...	Minimum	Maximum	Average	Upper q...	Median	Lower q...	Standard...	Skewness
19.865%	0.42	80.0	29.699	38.0	28.0	20.0	14.526	0.388

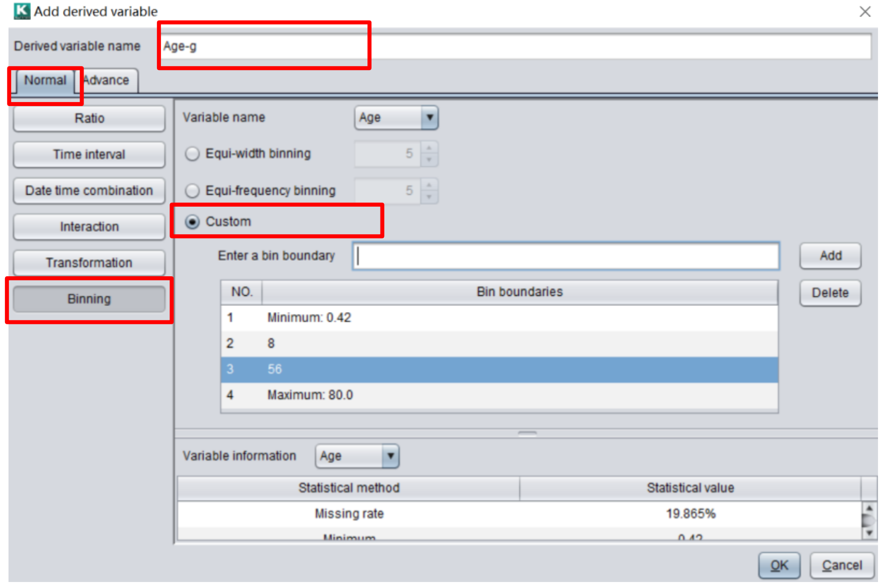
Age, the youngest is only 0.42, the oldest is 80. The missing rate is 19.865%. The intelligent modeling tool will automatically fill in the missing value, which does not need to be processed.

From the frequency distribution chart of grouped target, we can see that the survival probability of children under 8 years old is very large, that of middle-aged and old people over 56 years old is very small, and that of young people has little change. According to this, it can be divided into three groups: 0-8 years old, 9-56 years old, and over 57 years old, generating the derived variable age_g.



Data preprocessing - generate derived variable, continuous variable discretization

Use the variable "Age" to generate the derived variable Age_g.



Derived variable name: Age-g

Variable name: Age

○ Equi-width binning 5

○ Equi-frequency binning 5

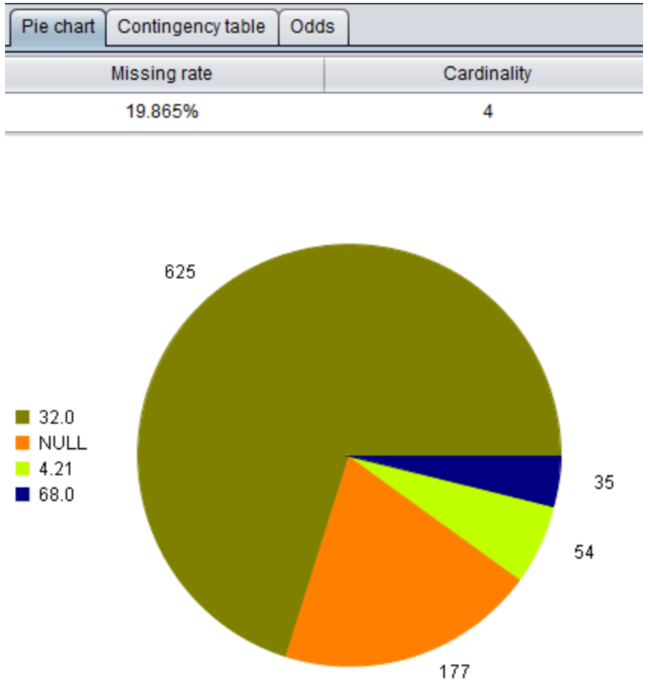
● Custom

Enter a bin boundary:

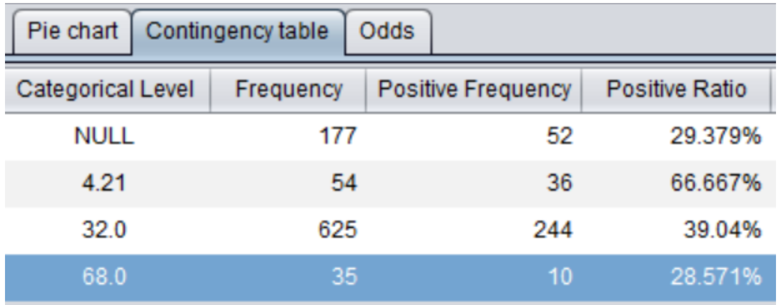
NO.	Bin boundaries
1	Minimum: 0.42
2	8
3	56
4	Maximum: 80.0

Variable information: Age

Statistical method	Statistical value
Missing rate	19.865%
Minimum	0.42



Missing rate	Cardinality
19.865%	4


Pie chart
Contingency table
Odds

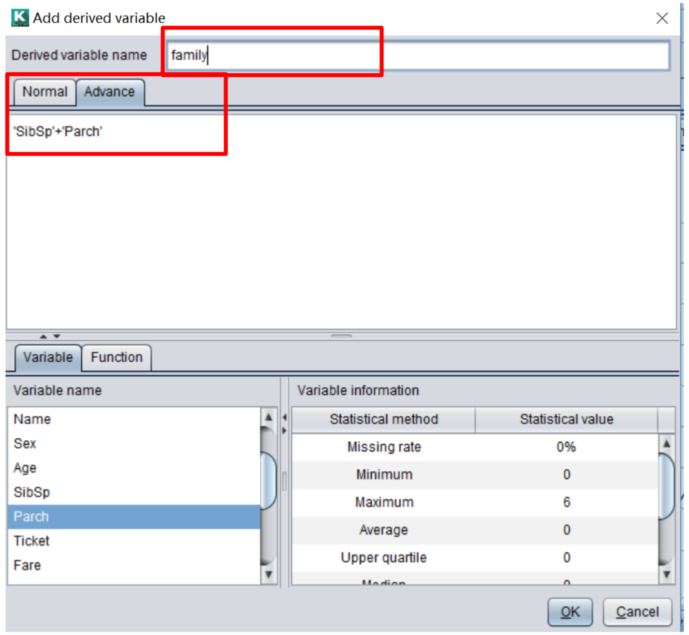
Categorical Level	Frequency	Positive Frequency	Positive Ratio
NULL	177	52	29.379%
4.21	54	36	66.667%
32.0	625	244	39.04%
68.0	35	10	28.571%

The missing rate of the derived variable Age_g inherits the missing rate of "Age", and YModel will fill the missing value intelligently, which does not need to be dealt with separately. Looking at the results of grouping statistics, the survival rate of children is high, the survival rate of middle-aged and old people is low, and the survival rate of young and middle-aged people is between the two, which is an important variable.

Data preprocessing - generate derived variable, variable interaction

6. Using “Sibsp”, “Parch” to generate derived variable “family”.

The number of siblings and spouses, the number of parents and children(no missing value) are all number of family members. Add these two variables to form the “family” variable.



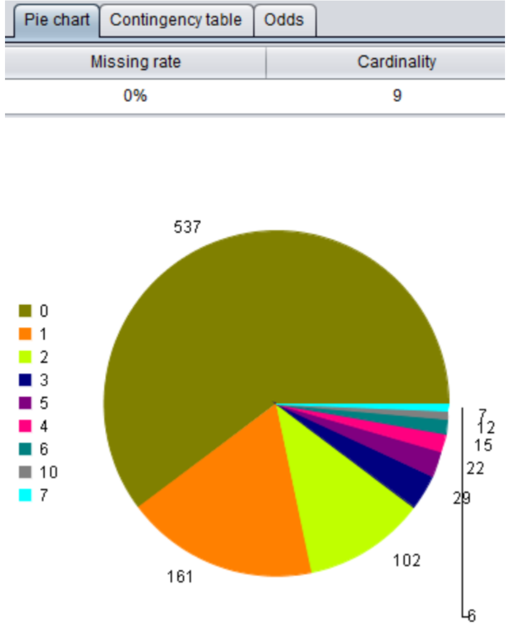
Derived variable name: family

'SibSp'+'Parch'

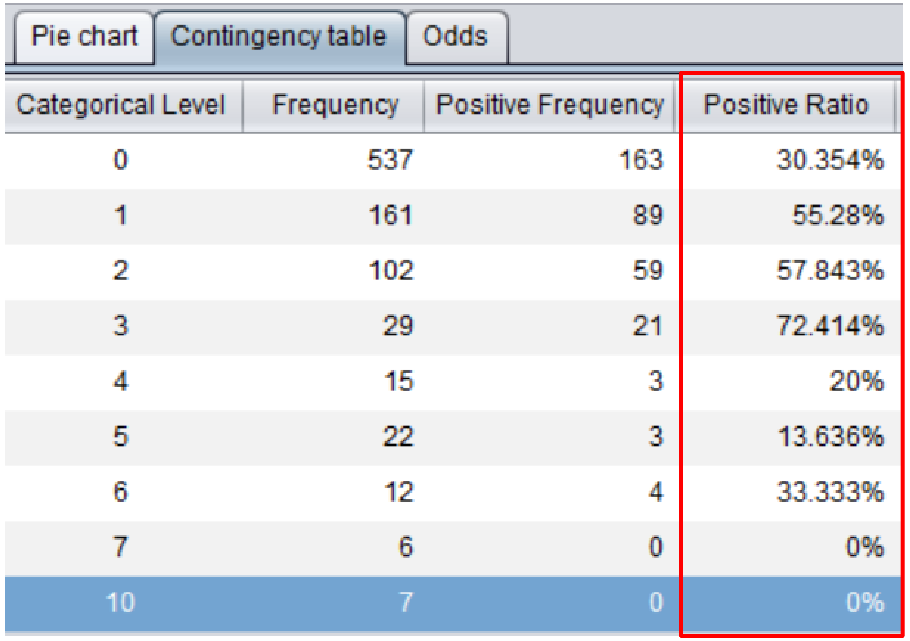
Variable name:
- Name
- Sex
- Age
- SibSp
- Parch
- Ticket
- Fare

Variable information:

Statistical method	Statistical value
Missing rate	0%
Minimum	0
Maximum	6
Average	0
Upper quartile	0
Median	0



Missing rate	Cardinality
0%	9


Pie chart
Contingency table
Odds

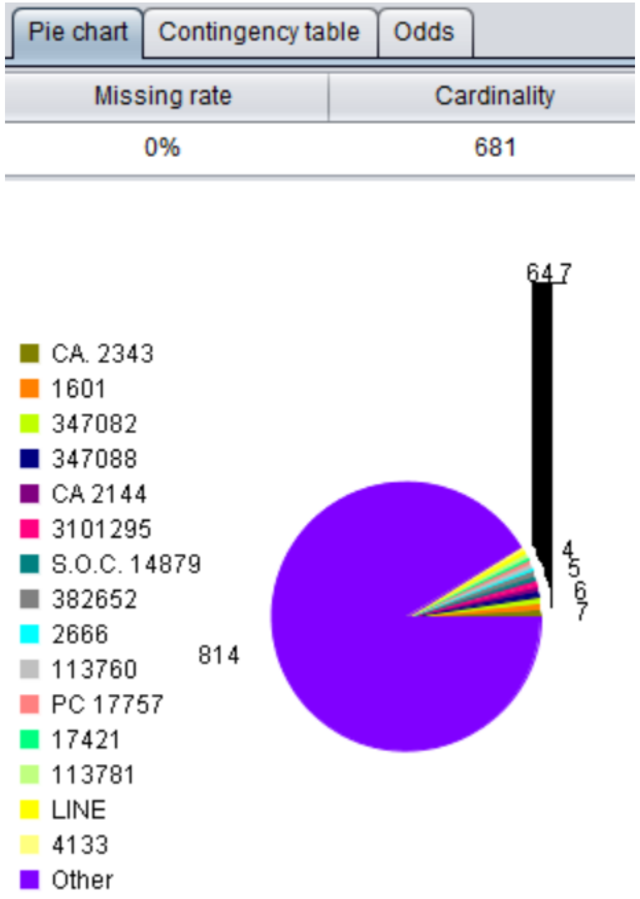
Categorical Level	Frequency	Positive Frequency	Positive Ratio
0	537	163	30.354%
1	161	89	55.28%
2	102	59	57.843%
3	29	21	72.414%
4	15	3	20%
5	22	3	13.636%
6	12	4	33.333%
7	6	0	0%
10	7	0	0%

When observing the generated derivative variable family, single people accounted for the majority, but the survival rate was only 30.354%. When the number of family members was 1-3, family relationship would help them to be saved, but when the number was more than 3, family members would be concerned about each other, resulting in the decline of survival rate. This variable is also an important variable.

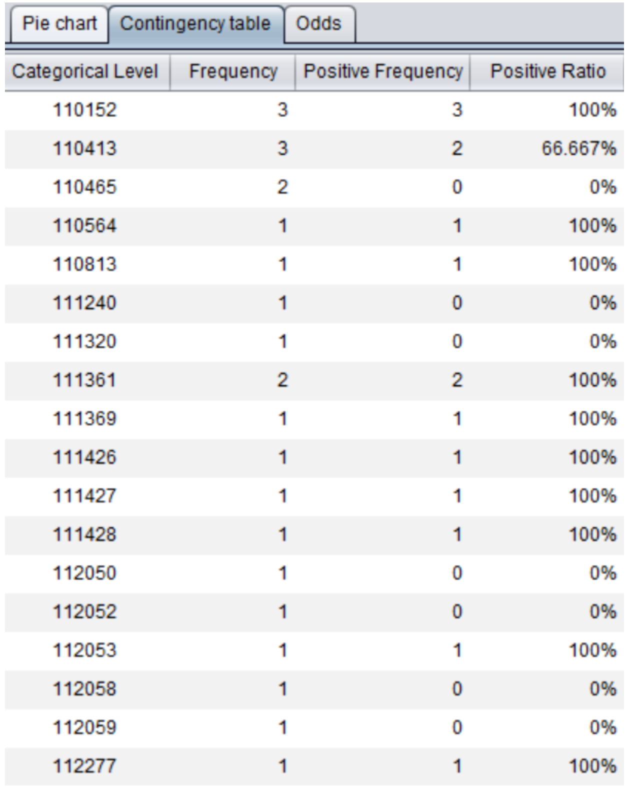
Data exploration - distribution analysis

7. Ticket

There are too many categories for ticket numbers. To view pie chart and grouping statistics, there is not too much information provided. So this variable can be discarded.



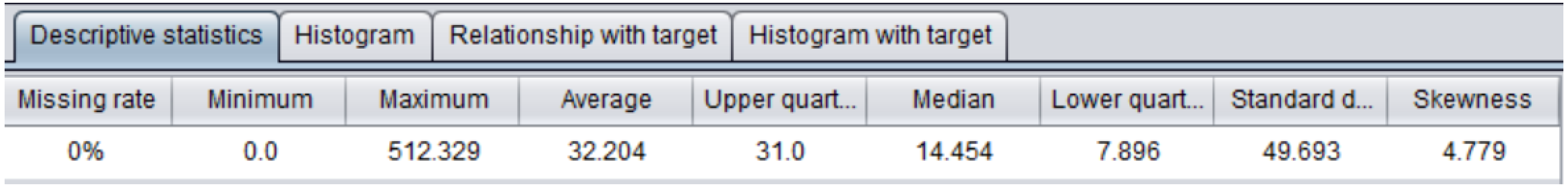
Missing rate	Cardinality
0%	681



Categorical Level	Frequency	Positive Frequency	Positive Ratio
110152	3	3	100%
110413	3	2	66.667%
110465	2	0	0%
110564	1	1	100%
110813	1	1	100%
111240	1	0	0%
111320	1	0	0%
111361	2	2	100%
111369	1	1	100%
111426	1	1	100%
111427	1	1	100%
111428	1	1	100%
112050	1	0	0%
112052	1	0	0%
112053	1	1	100%
112058	1	0	0%
112059	1	0	0%
112277	1	1	100%

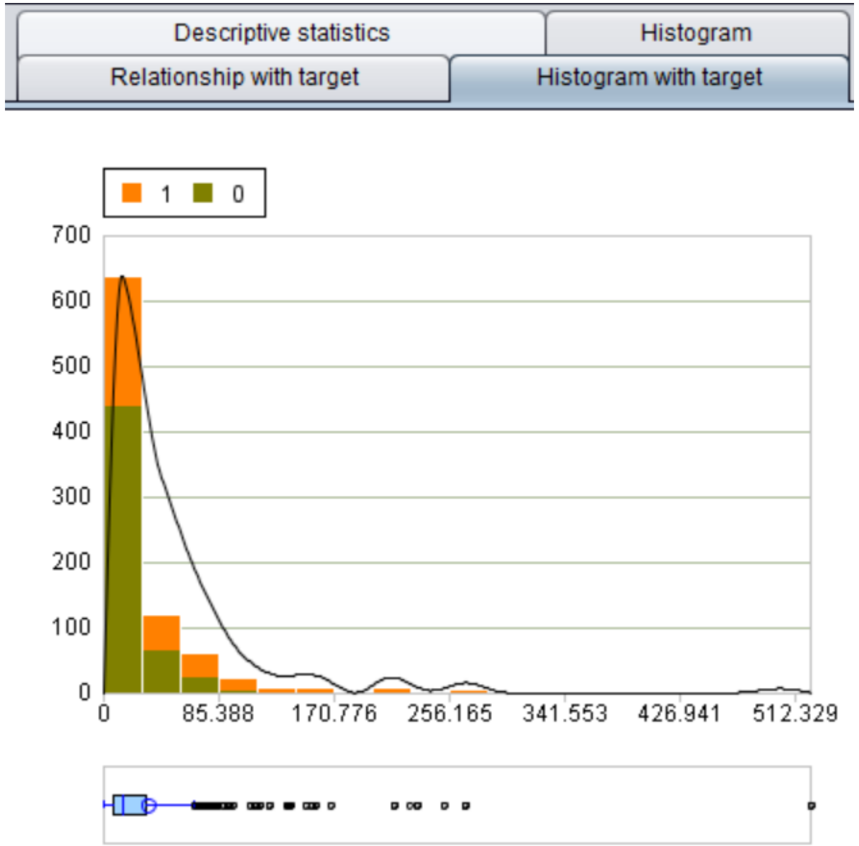
Data exploration: distribution analysis, contrastive analysis and statistical analysis

8. Fare



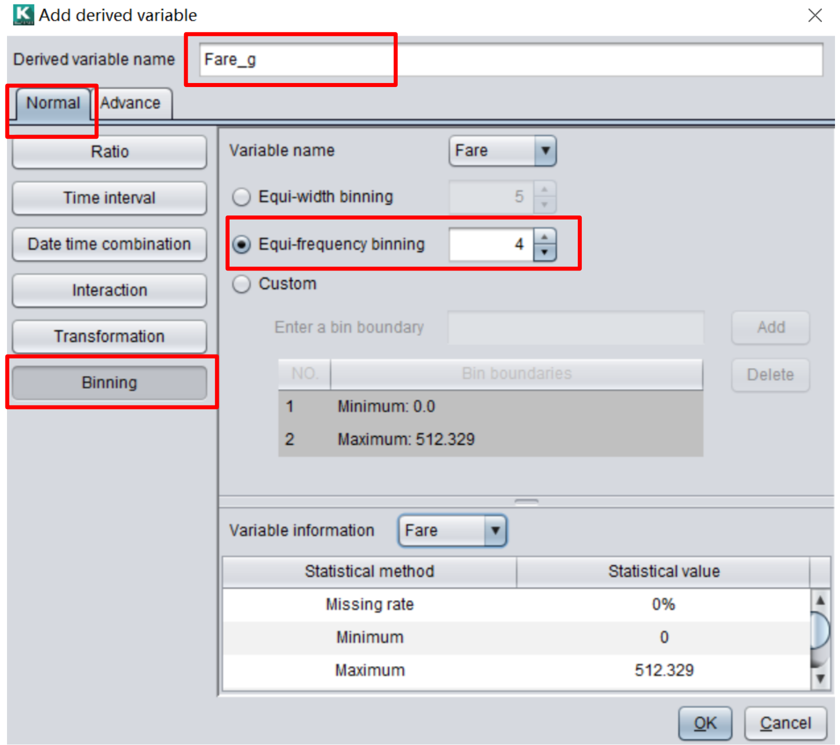
Missing rate	Minimum	Maximum	Average	Upper quart...	Median	Lower quart...	Standard d...	Skewness
0%	0.0	512.329	32.204	31.0	14.454	7.896	49.693	4.779

Fare, minimum 0, maximum 512.329. The skewness is 4.779, serious right deviation, no missing value. From the distribution diagram, we can see that the higher the ticket price is, the larger the proportion of survival is. Therefore, we can use the method of equal frequency grouping to discretize it into four groups.



Data preprocessing - generate derived variable, continuous variable discretization

The variable “Fare” is equi-frequency discretized into four groups, and generate derivative variable “Fare_g” .



Derived variable name: Fare_g

Variable name: Fare

○ Equi-width binning 5

● Equi-frequency binning 4

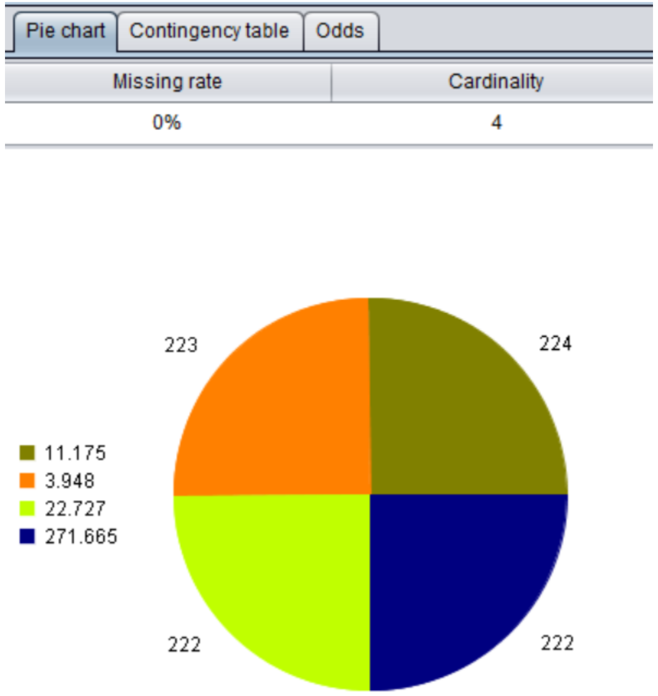
○ Custom

Enter a bin boundary

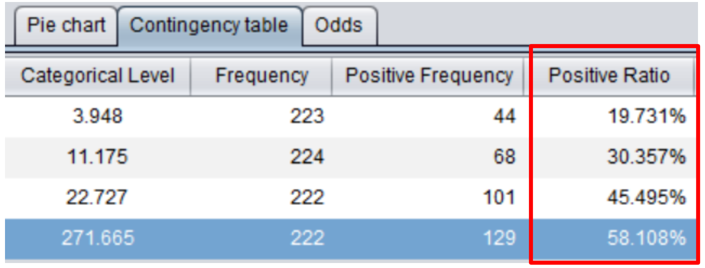
NO.	Bin boundaries
1	Minimum: 0.0
2	Maximum: 512.329

Variable information: Fare

Statistical method	Statistical value
Missing rate	0%
Minimum	0
Maximum	512.329



Missing rate	Cardinality
0%	4


Pie chart
Contingency table
Odds

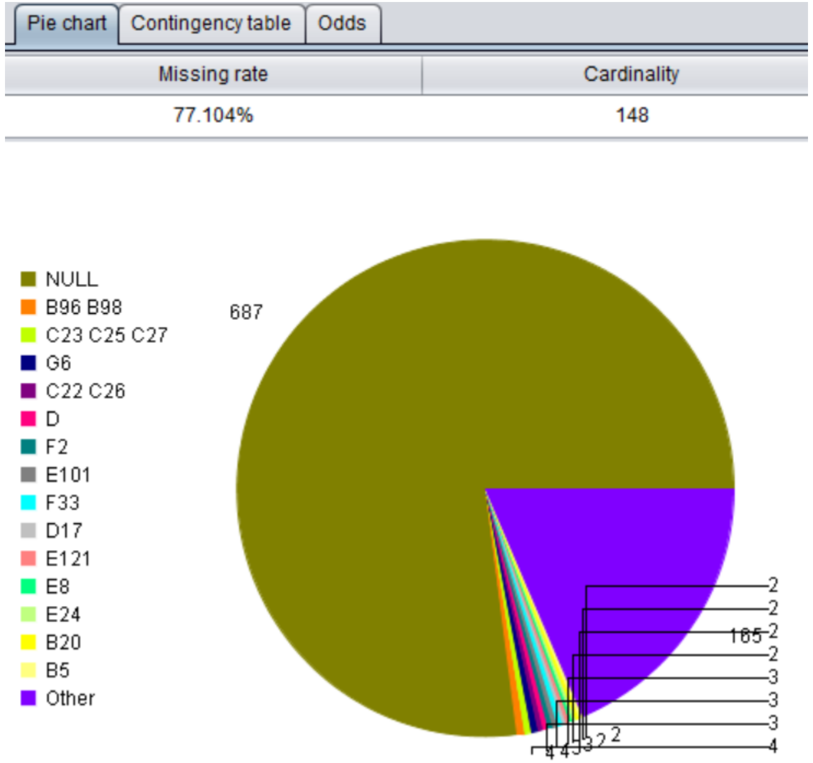
Categorical Level	Frequency	Positive Frequency	Positive Ratio
3.948	223	44	19.731%
11.175	224	68	30.357%
22.727	222	101	45.495%
271.665	222	129	58.108%

By observing the generated derivative variable “Fare_g” , the group with low fare price is only less than 20% the group with high fare price reaches 58% . It shows that this variable can distinguish the target variable well and is an important variable.

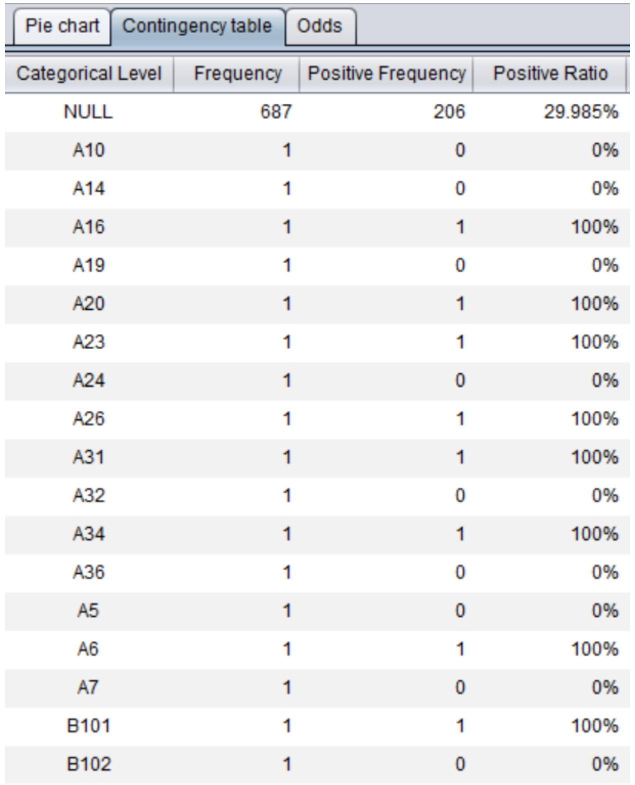
Data exploration——missing value analysis and distribution analysis

9.Cabin

There are many classifications for cabin number, and the missing rate is higher than 77%. It seems that this variable is useless, but we can extract whether this variable is missing as a message, that is, missing is 1, not missing is 0. This is also a way to extract missing value information.



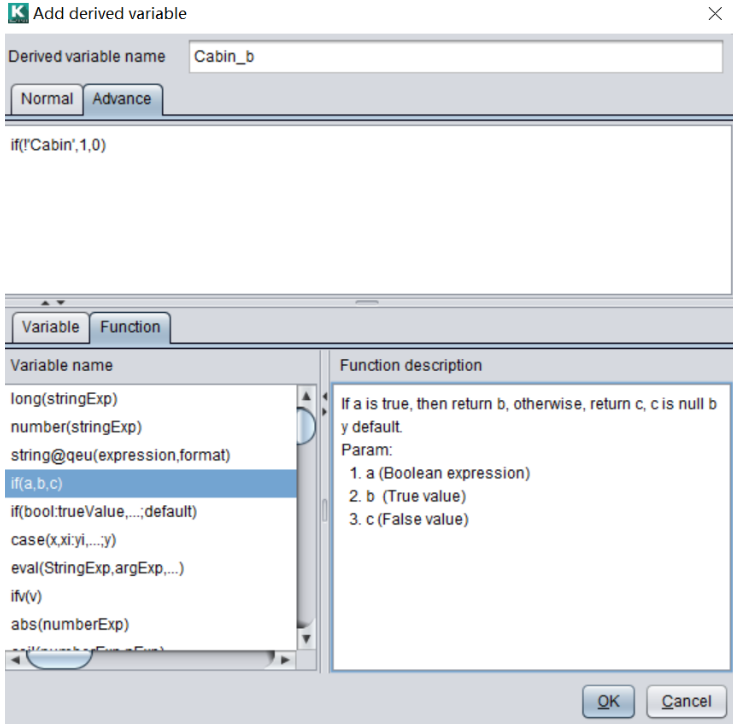
Missing rate	Cardinality
77.104%	148



Categorical Level	Frequency	Positive Frequency	Positive Ratio
NULL	687	206	29.985%
A10	1	0	0%
A14	1	0	0%
A16	1	1	100%
A19	1	0	0%
A20	1	1	100%
A23	1	1	100%
A24	1	0	0%
A26	1	1	100%
A31	1	1	100%
A32	1	0	0%
A34	1	1	100%
A36	1	0	0%
A5	1	0	0%
A6	1	1	100%
A7	1	0	0%
B101	1	1	100%
B102	1	0	0%

Data preprocessing——generate derived variable and extract missing value information

Extract the missing value information of cabin. The value is 1 if cabin info is missing, and the value is 0 if not missing. Generate the derived variable “Cabin_b” .



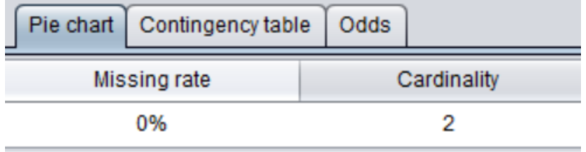
Derived variable name: Cabin_b

```
if(!'Cabin',1,0)
```

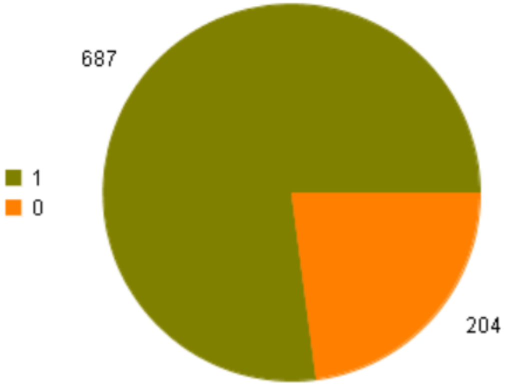
Variable name
- long(stringExp)
- number(stringExp)
- string@qeu(expression,format)
- if(a,b,c)
- if(bool:trueValue,...;default)
- case(x,xi:yi,...;y)
- eval(StringExp,argExp,...)
- ifv(v)
- abs(numberExp)

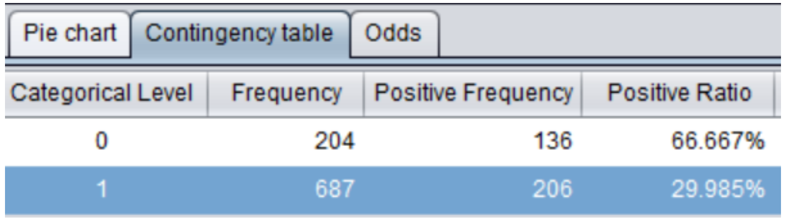
Function description

If a is true, then return b, otherwise, return c, c is null by default.
Param:
1. a (Boolean expression)
2. b (True value)
3. c (False value)


Pie chart
Contingency table
Odds

Missing rate	Cardinality
0%	2




Pie chart
Contingency table
Odds

Categorical Level	Frequency	Positive Frequency	Positive Ratio
0	204	136	66.667%
1	687	206	29.985%

Note: for the processing of missing value, Yiming intelligent modeling product will be automatically completed, and the user does not need to operate. Here, it is only used for teaching demonstration.

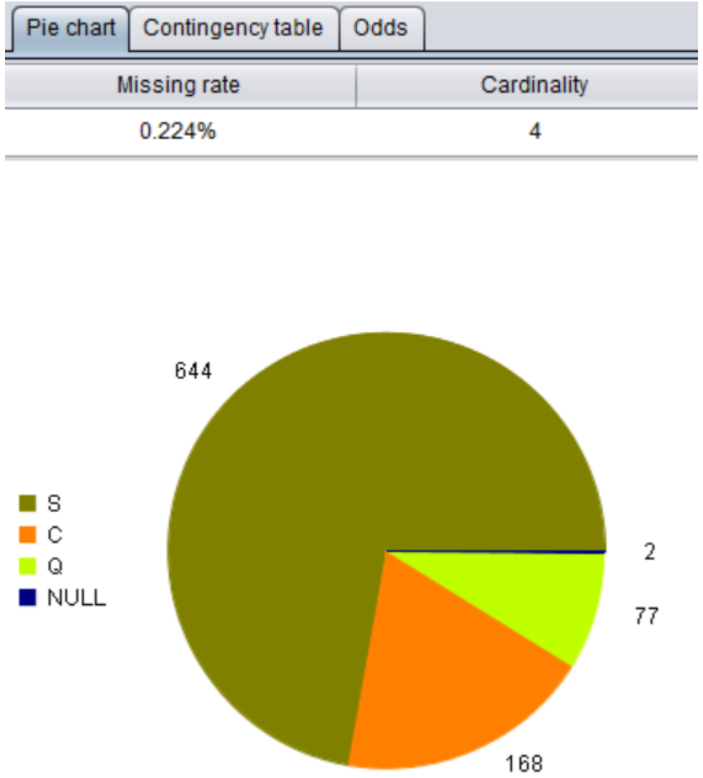
Observing the derived variable “Cabin_b” , the statistical information of grouping shows that the survival rate of cabin not missing, that is, cabin = 0, is very high, reaching 2 / 3, while the survival rate of missing is less than 30%. We can speculate boldly that only a good cabin has a number, which is equivalent to VIP. Other cabins do not have a number (still need to earn more money).

Data exploration: distribution analysis, contrastive analysis and missing value analysis

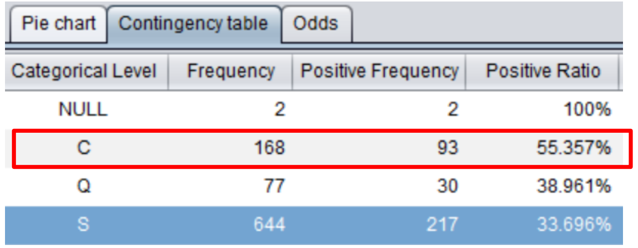
10.Embarked

There are three types of embarked ports, among them S is the majority, and C and Q is less, and two missing values. YModel will automatically process them without further handling.

Intuitively speaking, there should be no relationship between embarked port and survival, but the data tells us that the survival rate of passengers embarked at port C is significantly higher than that at other ports, so sometimes intuition is not reliable.



Missing rate	Cardinality
0.224%	4

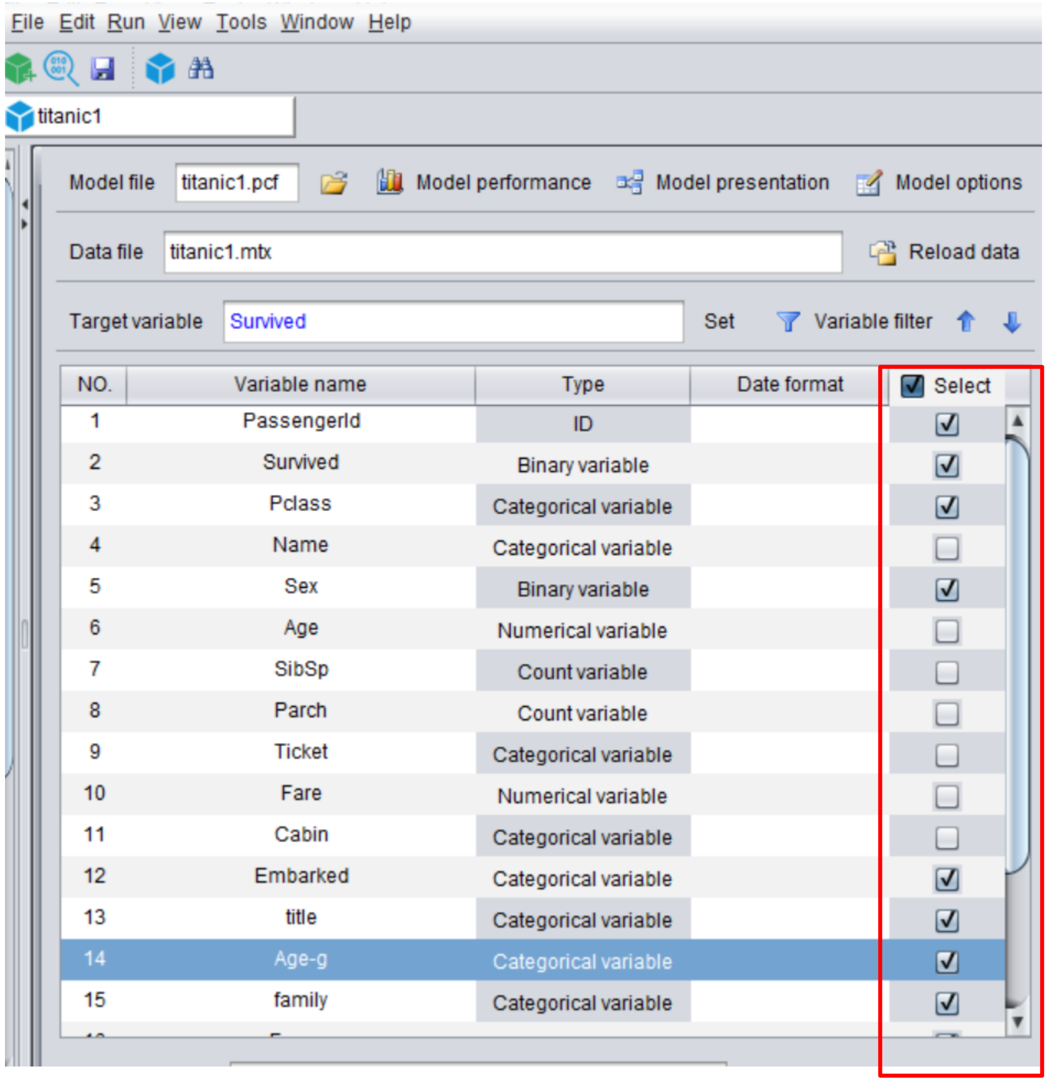


Categorical Level	Frequency	Positive Frequency	Positive Ratio
NULL	2	2	100%
C	168	93	55.357%
Q	77	30	38.961%
S	644	217	33.696%

Data preprocessing - variable selection

11. Remove irrelevant variables and keep useful variables

1. “Passengerid” , the unique identification of passengers, useless, removed;
2. “Name” , the title variable is extracted, useless, removed;
3. “Age” , the Age_g variable is generated, no need, removed;
4. “SibSp” and “Parch” , generate family variable, , no need, removed;
5. “Ticket” , too many categories, useless, removed;
6. “Fare” , the fare_g variable is generated, no need, and removed;
7. “Cabin” , the cabin_g variable is generated , no need, and removed



Model file titanic1.pcf

Data file titanic1.mtx

Target variable Survived

NO.	Variable name	Type	Date format	☑ Select
1	PassengerId	ID		☑
2	Survived	Binary variable		☑
3	Pclass	Categorical variable		☑
4	Name	Categorical variable		☐
5	Sex	Binary variable		☑
6	Age	Numerical variable		☐
7	SibSp	Count variable		☐
8	Parch	Count variable		☐
9	Ticket	Categorical variable		☐
10	Fare	Numerical variable		☐
11	Cabin	Categorical variable		☐
12	Embarked	Categorical variable		☑
13	title	Categorical variable		☑
14	Age-g	Categorical variable		☑
15	family	Categorical variable		☑

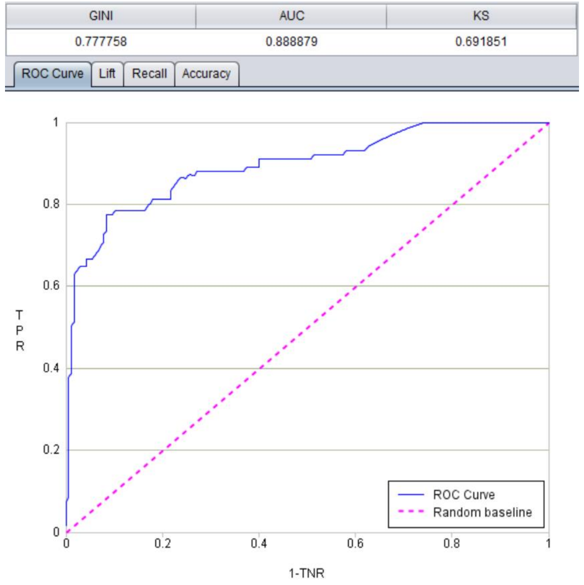
Note: Raqsoft YModel product will automatically eliminate the variables that are useless for modeling, and the user does not need to operate. Here, it is only used for teaching demonstration.

Summary of exploration and preprocessing methods

No	Variable name	Exploration content	Exploration result	Preprocessing content	Preprocessing result
1	PassengerId	ID variable，useless information	Useless variable，removed		
2	Survived	Distribution analysis	The proportion of positive and negative samples is close to 3:5		
3	Pclass	Distribution analysis，Grouping statistics	The higher the level, the lower the survival rate		
4	Name	Content analysis	Title information can be extracted	Extract valuable information	Generate derived variable title
5	Sex	Distribution analysis，Grouping statistics	Women's survival rate is much higher than men's		
6	Age	Missing value analysis, distribution analysis	High survival rate of children and low survival rate of the elderly	Discretization of continuous variable	Generate derived variable Age_g
7	SibSp	Meaning analysis	Siblings，spouse，children，parents are all family	Variable interaction	Generate derived variable family
8	Parch				
9	Ticket	Distribution analysis	Useless information，removed		
10	Fare	Distribution analysis，Grouping statistics	Serious skew, the higher the fare, the higher the survival rate	Discretization of continuous variable	Generate derived variable Fare_g
11	Cabin	Missing value analysis, distribution analysis	The missing rate is very high, and useful information may exist	Extract missing value information	Generate derived variable Cabin_b
12	Embarked	Missing value analysis, distribution analysis	High survival rate of passengers embarked in port c		

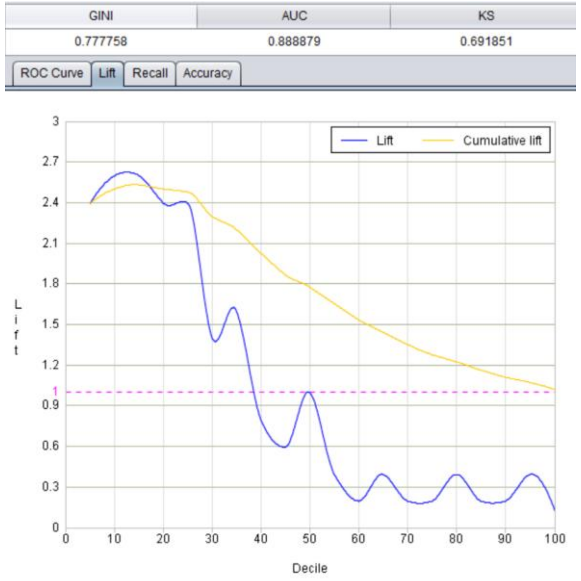
Note: for the above variables, only the derived variables in the yellow part need to be added manually, and the rest of the analysis will be processed automatically by Raqsoft YModel software .

Since YModel tool includes data preprocessing and modeling process, we can also directly model with the original data without the above data exploration and preprocessing operations. The results of model 1 are as follows:



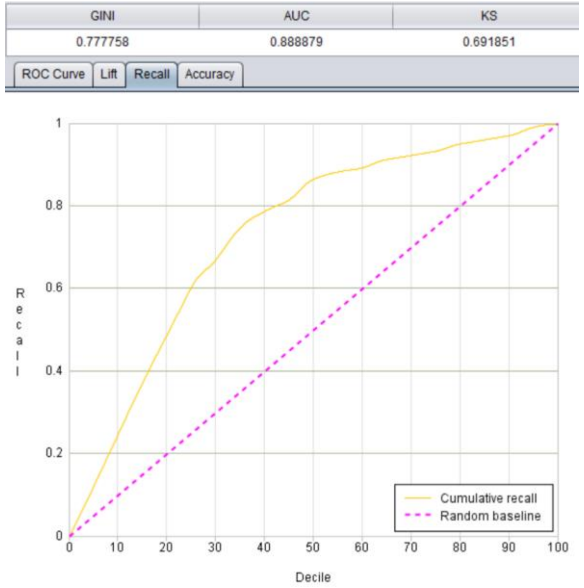
GINI	AUC	KS
0.777758	0.888879	0.691851

ROC curve



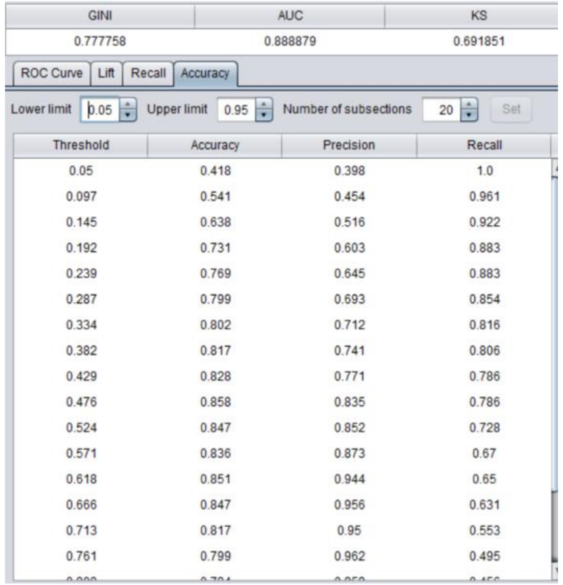
GINI	AUC	KS
0.777758	0.888879	0.691851

Lift chart



GINI	AUC	KS
0.777758	0.888879	0.691851

Recall chart



GINI	AUC	KS
0.777758	0.888879	0.691851

Lower limit 0.05 Upper limit 0.95 Number of subsections 20

Threshold	Accuracy	Precision	Recall
0.05	0.418	0.398	1.0
0.097	0.541	0.454	0.961
0.145	0.638	0.516	0.922
0.192	0.731	0.603	0.883
0.239	0.769	0.645	0.883
0.287	0.799	0.693	0.854
0.334	0.802	0.712	0.816
0.382	0.817	0.741	0.806
0.429	0.828	0.771	0.786
0.476	0.858	0.835	0.786
0.524	0.847	0.852	0.728
0.571	0.836	0.873	0.67
0.618	0.851	0.944	0.65
0.666	0.847	0.956	0.631
0.713	0.817	0.95	0.553
0.761	0.799	0.962	0.495

Accuracy table

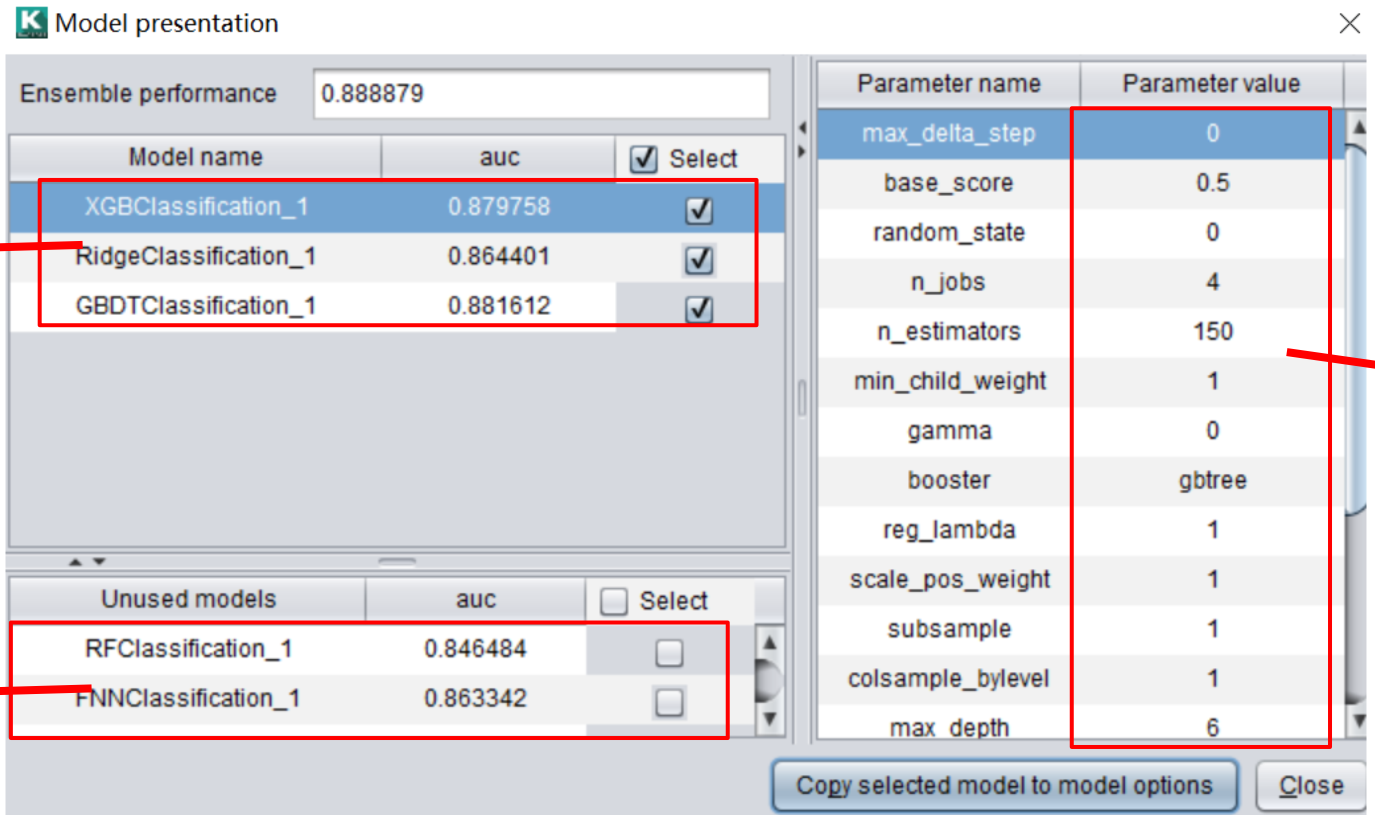
From the evaluation index, we can see that the performance of the model is very good, GINI=0.7777，AUC=0.8889, the model is acceptable.
Observing the accuracy table, we can see that when the threshold value is 0.476, the accuracy reaches its highest 0.858. It indicates that the prediction accuracy is the highest when passengers with a prediction probability greater than 0.476 are regarded as survivors and passengers with a prediction probability less than 0.476 as victims. The precision is 0.835, which means 83.5% of the passengers predicted to be survivors are indeed survivors. The recall was 0.786, indicating that the passengers predicted to be survivors are 78.6% of all survivors.

Which models are used for modeling:

Selected models

Unused models

Model parameters



Ensemble performance: 0.888879

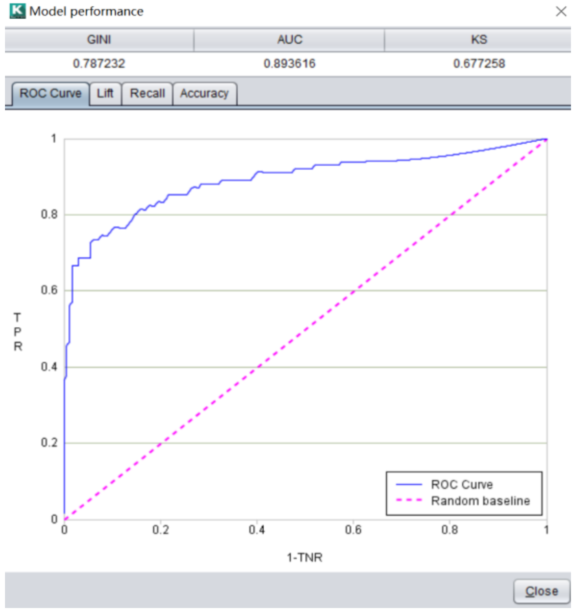
Model name	auc	☑ Select
XGBClassification_1	0.879758	☑
RidgeClassification_1	0.864401	☑
GBDTClassification_1	0.881612	☑

Unused models	auc	☐ Select
RFClassification_1	0.846484	☐
FNNClassification_1	0.863342	☐

Parameter name	Parameter value
max_delta_step	0
base_score	0.5
random_state	0
n_jobs	4
n_estimators	150
min_child_weight	1
gamma	0
booster	gbtree
reg_lambda	1
scale_pos_weight	1
subsample	1
colsample_bylevel	1
max_depth	6

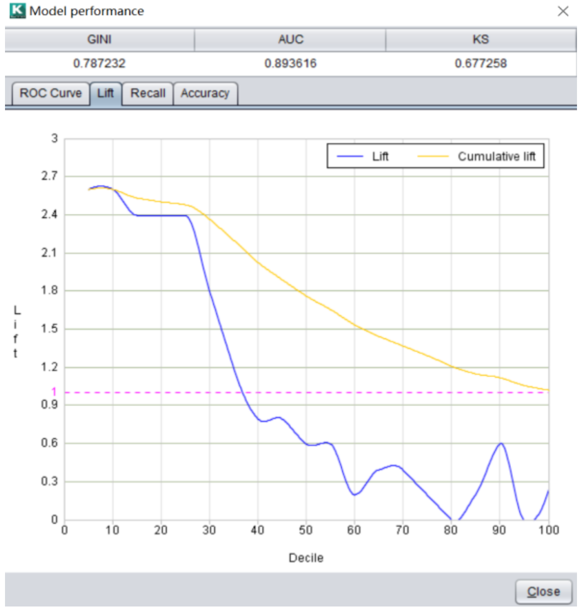
In this modeling, XGB, Ridge and GBDT are selected as three classification models, and the optimal combination model is obtained by combining these three models. Model parameters are automatically selected by YModel. Data mining experts can select "copy selected model to model option" to modify parameters and re model.

According to the results of data exploration in the previous section, the yellow part 4, 7, 8 and 10 items in page p339 were derived manually and the model 2 was established. The results are as follows:



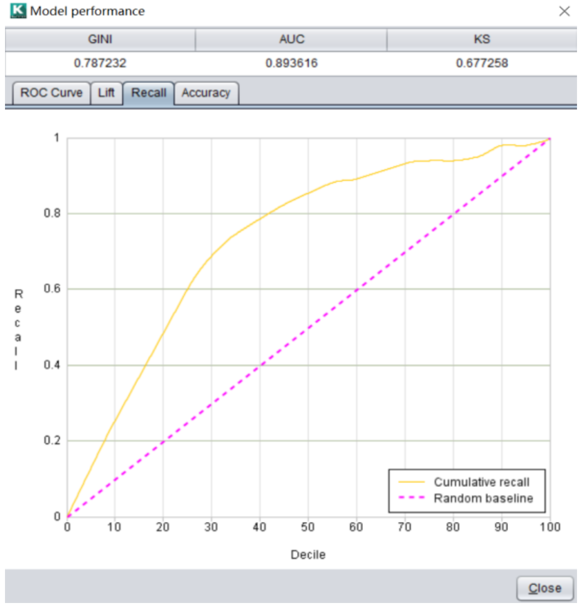
GINI	AUC	KS
0.787232	0.893616	0.677258

ROC curve



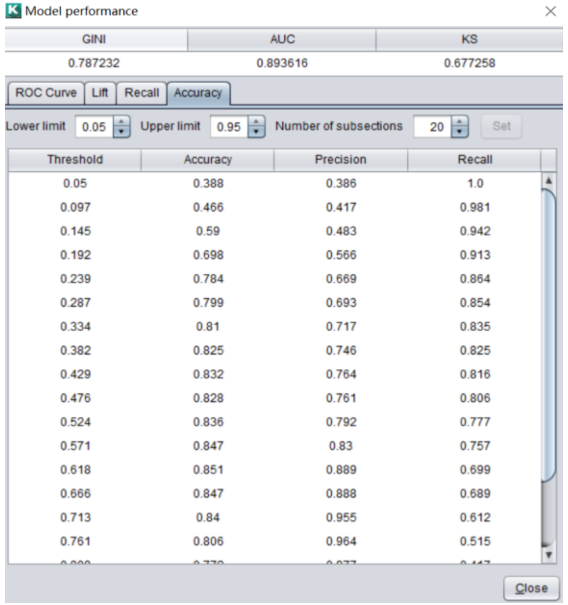
GINI	AUC	KS
0.787232	0.893616	0.677258

Lift chart



GINI	AUC	KS
0.787232	0.893616	0.677258

Recall chart



GINI	AUC	KS
0.787232	0.893616	0.677258

Lower limit 0.05 Upper limit 0.95 Number of subsections 20

Threshold	Accuracy	Precision	Recall
0.05	0.388	0.386	1.0
0.097	0.466	0.417	0.981
0.145	0.59	0.483	0.942
0.192	0.698	0.566	0.913
0.239	0.784	0.669	0.864
0.287	0.799	0.693	0.854
0.334	0.81	0.717	0.835
0.382	0.825	0.746	0.825
0.429	0.832	0.764	0.816
0.476	0.828	0.761	0.806
0.524	0.836	0.792	0.777
0.571	0.847	0.83	0.757
0.618	0.851	0.889	0.699
0.666	0.847	0.888	0.689
0.713	0.84	0.955	0.612
0.761	0.806	0.964	0.515

Accuracy table

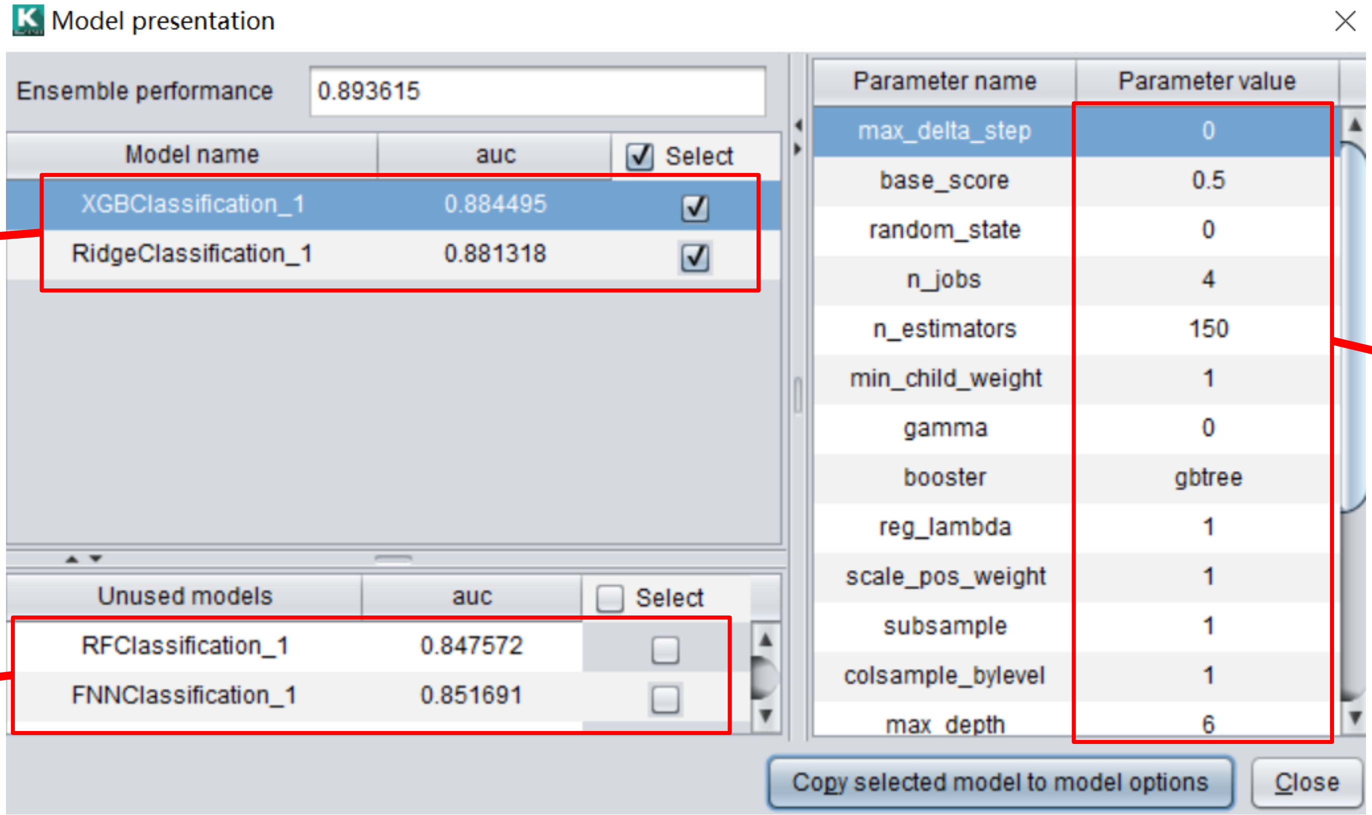
From the evaluation index, Gini = 0.7872, AUC = 0.8936, the overall performance of the model slightly improved.

Observing the accuracy table, we can see that when the threshold value is 0.618, the accuracy reaches its highest 0.851. It indicates that the prediction accuracy is the highest when passengers with a prediction probability greater than 0.618 are regarded as survivors and passengers with a prediction probability less than 0.618 as victims. The precision is 0.889, which means 88.9% of the passengers predicted to be survivors are indeed survivors. The recall was 0.699, indicating that the passengers predicted to be survivors are 69.9% of all survivors.

Which models are used for modeling:

Selected models

Unused models



Model presentation

Ensemble performance 0.893615

Model name	auc	☑ Select
XGBClassification_1	0.884495	☑
RidgeClassification_1	0.881318	☑

Unused models	auc	☐ Select
RFClassification_1	0.847572	☐
FNNClassification_1	0.851691	☐

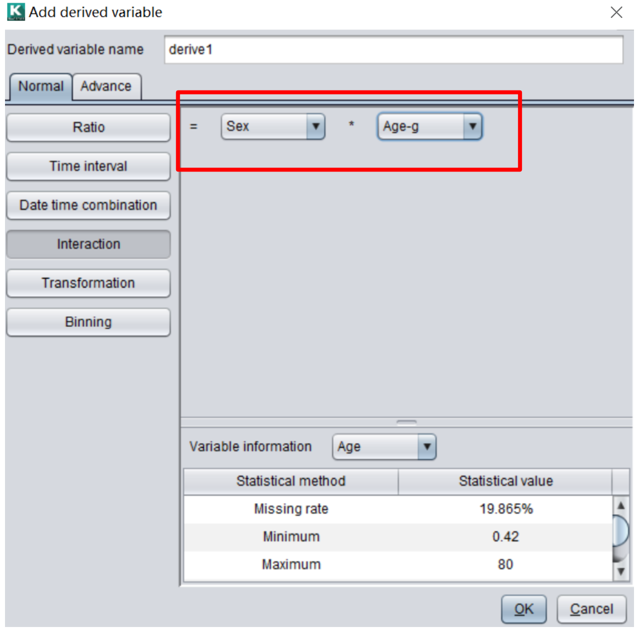
Parameter name	Parameter value
max_delta_step	0
base_score	0.5
random_state	0
n_jobs	4
n_estimators	150
min_child_weight	1
gamma	0
booster	gbtree
reg_lambda	1
scale_pos_weight	1
subsample	1
colsample_bylevel	1
max_depth	6

Model parameters

In this modeling, XGB and Ridge are selected as two classification models, and the optimal combination model is obtained by combining these two models. Model parameters are automatically selected by intelligent modeling tool. Data mining experts can select "copy selected model to model option" to modify parameters and re model.

After the model is built, the importance of the variables will be returned, and the derived variables can be generated by interacting the high importance variables to continue to optimize the model,

For example,derive1=Sex*Age_g

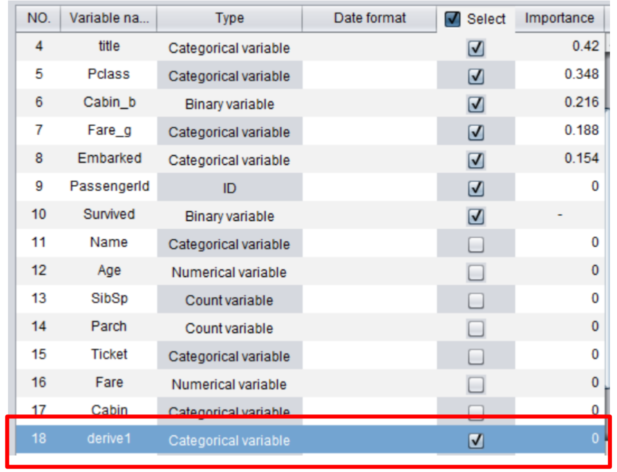


Derived variable name: derive1

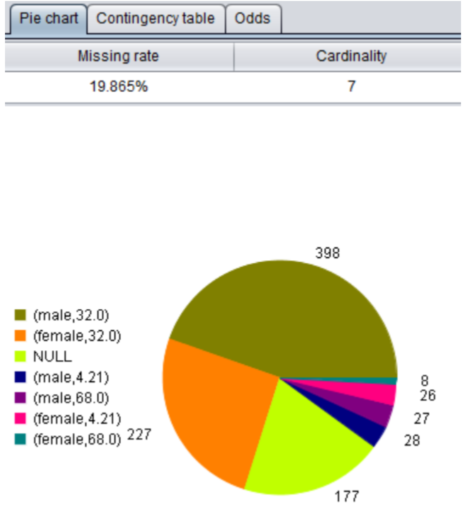
= Sex * Age-g

Variable information: Age

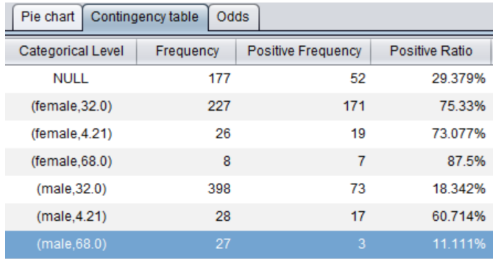
Statistical method	Statistical value
Missing rate	19.865%
Minimum	0.42
Maximum	80



NO.	Variable na...	Type	Date format	☑ Select	Importance
4	title	Categorical variable		☑	0.42
5	Pclass	Categorical variable		☑	0.348
6	Cabin_b	Binary variable		☑	0.216
7	Fare_g	Categorical variable		☑	0.188
8	Embarked	Categorical variable		☑	0.154
9	PassengerId	ID		☑	0
10	Survived	Binary variable		☑	-
11	Name	Categorical variable		☐	0
12	Age	Numerical variable		☐	0
13	SibSp	Count variable		☐	0
14	Parch	Count variable		☐	0
15	Ticket	Categorical variable		☐	0
16	Fare	Numerical variable		☐	0
17	Cabin	Categorical variable		☐	0
18	derive1	Categorical variable		☑	0



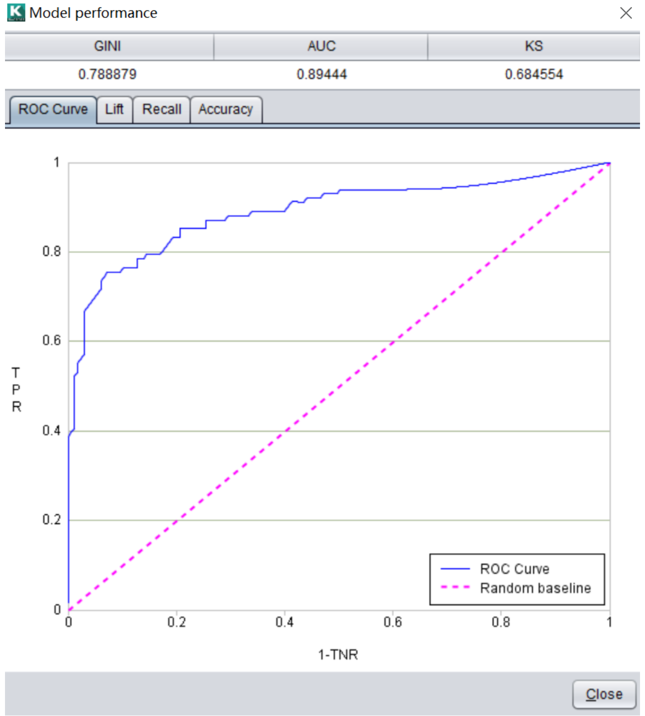
Missing rate	Cardinality
19.865%	7



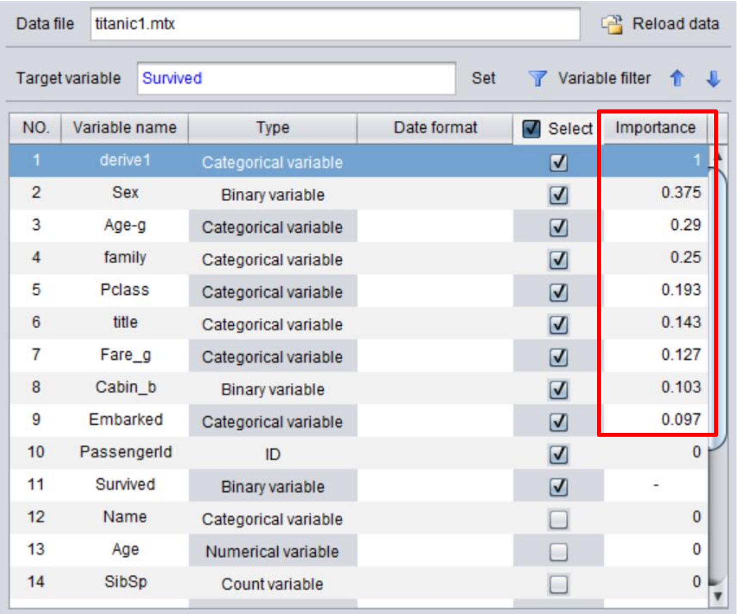
Categorical Level	Frequency	Positive Frequency	Positive Ratio
NULL	177	52	29.379%
(female,32.0)	227	171	75.33%
(female,4.21)	26	19	73.077%
(female,68.0)	8	7	87.5%
(male,32.0)	398	73	18.342%
(male,4.21)	28	17	60.714%
(male,68.0)	27	3	11.111%

Other interaction variables can also be added, but it should be noted that when there are many categories of a variable (for example, there are 9 categories in family), it is not suitable to continue interaction. Because the number of categories after interaction is the product of the number of categories of two variables (for example, the number of categories in sex * family is 18), too many categories will affect the model effect. It is recommended that the family be further divided into three categories (i.e. 0,1-3, 3 or more) according to the number of members, and then interact.

For model 2, the model performance after adding the interactive derived variable, derive1:


Model performance

GINI	AUC	KS
0.788879	0.89444	0.684554


Data file: titanic1.mtx

Target variable: Survived

NO.	Variable name	Type	Date format	☑ Select	Importance
1	derive1	Categorical variable		☑	1
2	Sex	Binary variable		☑	0.375
3	Age-g	Categorical variable		☑	0.29
4	family	Categorical variable		☑	0.25
5	Pclass	Categorical variable		☑	0.193
6	title	Categorical variable		☑	0.143
7	Fare_g	Categorical variable		☑	0.127
8	Cabin_b	Binary variable		☑	0.103
9	Embarked	Categorical variable		☑	0.097
10	PassengerId	ID		☑	0
11	Survived	Binary variable		☑	-
12	Name	Categorical variable		☐	0
13	Age	Numerical variable		☐	0
14	SibSp	Count variable		☐	0

	AUC	GINI
Before adding a derived variable	0.893616	0.787232
After adding a derived variable	0.89444	0.788879

After adding the derived variable, the performance of the model is better than that of the original model. Moreover, according to the importance of variables, derived1 has become the most important variable, indicating that the newly added derived variable is useful.

Model evaluation and conclusions

From the analysis of the results of model 1 and model 2, it can be seen that the performance of the two models is almost the same. The precision of the model built directly from the original data is slightly poor but the recall rate is high; For modeling after data exploration and feature extraction, the precision of the model is improved, but the recall rate is decreased. How to choose depends on business requirements.

Conclusions:

Data exploration and analysis should be combined with the objectives and data characteristics.

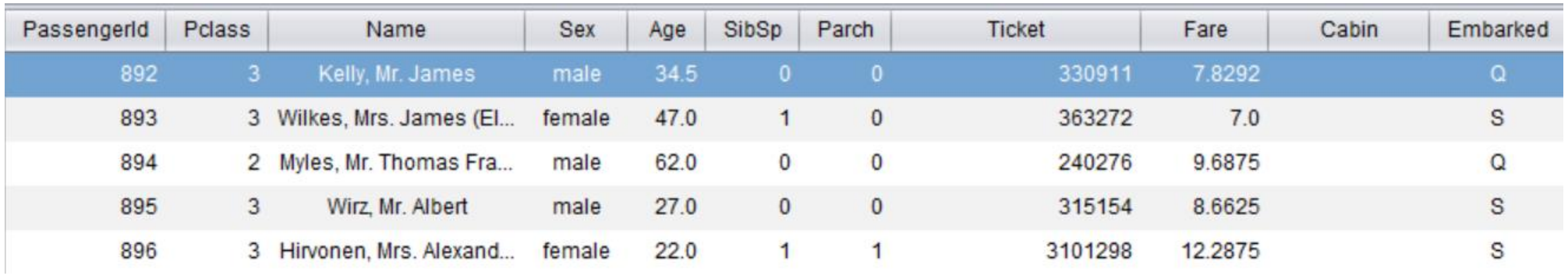
How to evaluate and select models needs to integrate business objectives.

Interactive derivation of important variables is a common method of model optimization.

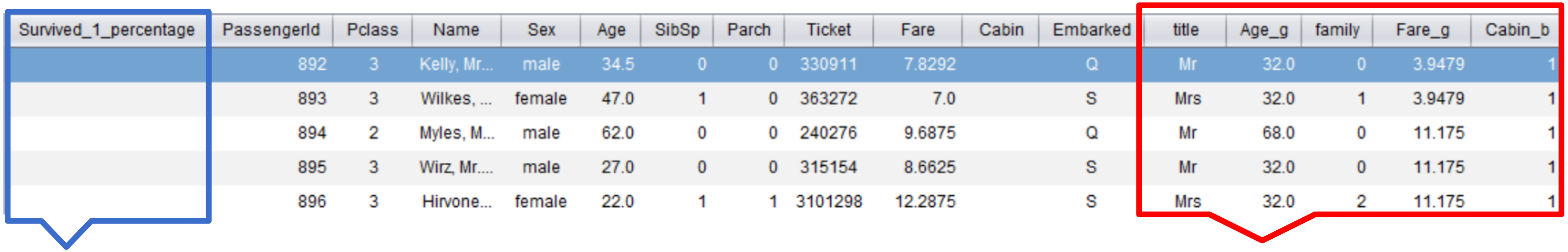
Model application

After the model is built, it is used to predict on the test set.

YModel will automatically process the test set and generate derived variables, as shown in the following figure:



PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
893	3	Wilkes, Mrs. James (El...	female	47.0	1	0	363272	7.0		S
894	2	Myles, Mr. Thomas Fra...	male	62.0	0	0	240276	9.6875		Q
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625		S
896	3	Hirvonen, Mrs. Alexand...	female	22.0	1	1	3101298	12.2875		S



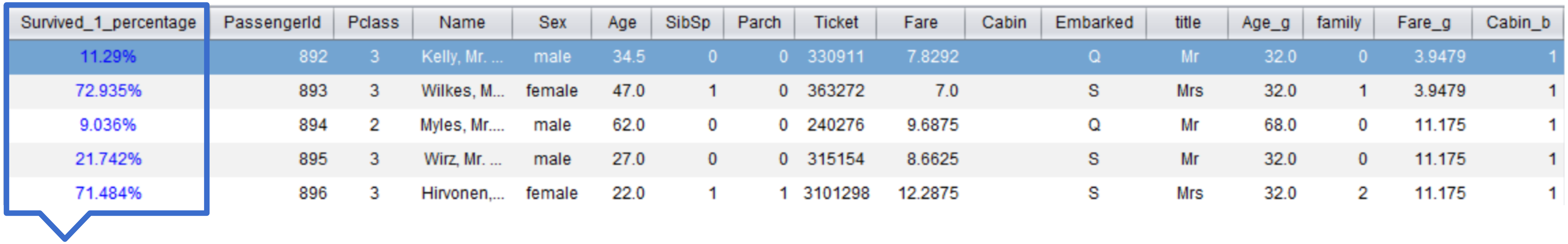
Survived_1_percentage	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title	Age_g	family	Fare_g	Cabin_b
	892	3	Kelly, Mr...	male	34.5	0	0	330911	7.8292		Q	Mr	32.0	0	3.9479	1
	893	3	Wilkes, ...	female	47.0	1	0	363272	7.0		S	Mrs	32.0	1	3.9479	1
	894	2	Myles, M...	male	62.0	0	0	240276	9.6875		Q	Mr	68.0	0	11.175	1
	895	3	Wirz, Mr....	male	27.0	0	0	315154	8.6625		S	Mr	32.0	0	11.175	1
	896	3	Hirvone...	female	22.0	1	1	3101298	12.2875		S	Mrs	32.0	2	11.175	1

Prediction result column

Added derived variables

Model application

The prediction results of the model are as follows:



Survived_1_percentage	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	title	Age_g	family	Fare_g	Cabin_b
11.29%	892	3	Kelly, Mr. ...	male	34.5	0	0	330911	7.8292		Q	Mr	32.0	0	3.9479	1
72.935%	893	3	Wilkes, M...	female	47.0	1	0	363272	7.0		S	Mrs	32.0	1	3.9479	1
9.036%	894	2	Myles, Mr....	male	62.0	0	0	240276	9.6875		Q	Mr	68.0	0	11.175	1
21.742%	895	3	Wirz, Mr. ...	male	27.0	0	0	315154	8.6625		S	Mr	32.0	0	11.175	1
71.484%	896	3	Hirvonen,...	female	22.0	1	1	3101298	12.2875		S	Mrs	32.0	2	11.175	1

The predicted result is the probability of survival (survived = 1). For example, the first predicted result is 11.29%, which means that the passenger only has a chance of 11.29% to survive.

Chapter 7 Comprehensive cases

Data presentation

House_ prices prediction dataset comes from kaggle, which is a dataset for regression prediction. This section uses this data to introduce the data mining process of regression prediction. （It’ s oriented for beginners, masters please bypass)

house_ prices_ train.csv : there are 1460 records and 81 variables in the training set (with target variable).
house_ prices_ test.csv : there are 1459 records and 80 variables in the set to be tested (no target variable).

Analysis objectives

1. Find out the factors affecting the house price.
2. Build model according to the training set, predict the data of the test set.

Data presentation

There are many variables in the data. The dataset dictionary is organized into CSV file for storage, as shown in the following figure:

There are 81 variables in total. Statistics by category are as follows:

ID：1

Binary variable: 4

Categorical variable: 42

Count variable: 9

Numeric variable: 20 (including target variable)

Time date: 5

Variable_Name	Variable_Type	Variable_Description
Id	ID	
Alley	Binary	Type of alley
CentralAir	Binary	Central air conditioning or not
Street	Binary	Street type
Utilities	Binary	Utilities type
BldgType	Categorical	Building type
BsmtCond	Categorical	Basement condition
BsmtExposure	Categorical	Basement exposure
BsmtFinType1	Categorical	Basement finish type
BsmtFinType2	Categorical	Basement finish type2
BsmtQual	Categorical	Basement quality
Condition1	Categorical	Condition1
Condition2	Categorical	Condition2
……	……	……

Data presentation

4 binary variables

Alley	Binary	Alley type
CentralAir	Binary	Central air or not
Street	Binary	Street type
Utilities	Binary	Utilities type

5 time date variables

GarageYrBlt	Time date	Garage year built
MoSold	Time date	Month sold
YearBuilt	Time date	Year built
YearRemodAdd	Time date	Year of remodeling
YrSold	Time date	Year sold

9 count variables

BsmtFullBath	Count	Basement full bath
BsmtHalfBath	Count	Basement half bath
FullBath	Count	Full bath
HalfBath	Count	Half bath
BedroomAbvGr	Count	Bedrooms
Fireplaces	Count	Fireplaces
GarageCars	Count	Garage cars
KitchenAbvGr	Count	Kitchens
TotRmsAbvGrd	Count	Total rooms

20 numeric variables

1stFlrSF	Numeric	1st floor area
2ndFlrSF	Numeric	2nd floor area
3SsnPorch	Numeric	3 season porch area
BsmtFinSF1	Numeric	Basement finish area1
BsmtFinSF2	Numeric	Basement finish area2
BsmtUnfSF	Numeric	Unfinished basement area
EnclosedPorch	Numeric	Enclosed porch area
GarageArea	Numeric	Garage area
GrLivArea	Numeric	Living area
LotArea	Numeric	Lot area
LotFrontage	Numeric	Lot Frontage
LowQualFinSF	Numeric	Low quality finish area
MasVnrArea	Numeric	Mas Vnr area
MiscVal	Numeric	Misc value
OpenPorchSF	Numeric	Open porch area
PoolArea	Numeric	Pool area
SalePrice	Numeric	Sale price，target variable
ScreenPorch	Numeric	Screen porch area
TotalBsmtSF	Numeric	Total basement area
WoodDeckSF	Numeric	Wood deck area

The remaining are 42 categorical variables

These variables may affect the house price. According to the significance of the variables, they can be divided into the following categories: the location of the house, the style of the house, the decoration of the house, the basement, the living area, the garage, the construction time, etc. then we can determine the importance of these variables according to the variables we consider when buying the house, or we can take these as the basis for data exploration and preprocessing.

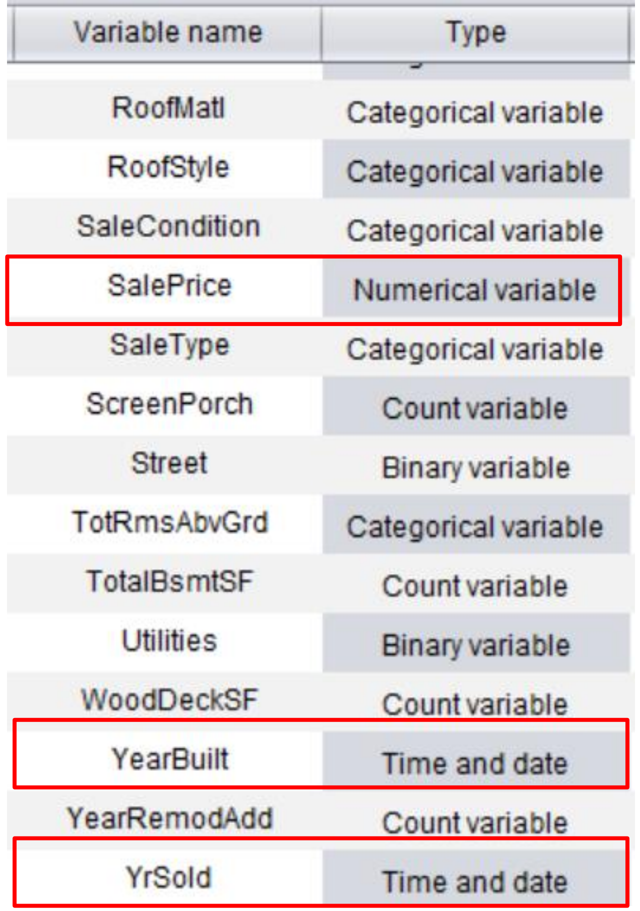
There are many variables, we can not analyze every variable here, so we have to choose some to analyze, mainly introducing the method of data mining, interested students can deeply analyze.

Data exploration

1. Check variable type according to data dictionary

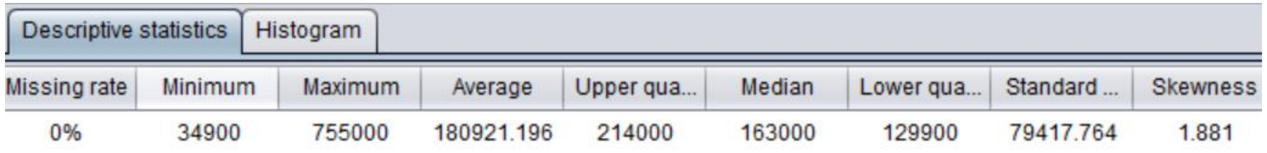
Modify the automatically recognized variable types according to the sorted data dictionary, such as
Sale price: count variable - numerical variable
Year built: count variable - time date
Fireplaces: categorical variable - count variable
……

Modeling tools will mistakenly recognize some numerical count variables as categorical variables, some numerical variables as count variables, and years as count variables, which need to be changed.



Variable name	Type
RoofMatl	Categorical variable
RoofStyle	Categorical variable
SaleCondition	Categorical variable
SalePrice	Numerical variable
SaleType	Categorical variable
ScreenPorch	Count variable
Street	Binary variable
TotRmsAbvGrd	Categorical variable
TotalBsmtSF	Count variable
Utilities	Binary variable
WoodDeckSF	Count variable
YearBuilt	Time and date
YearRemodAdd	Count variable
YrSold	Time and date

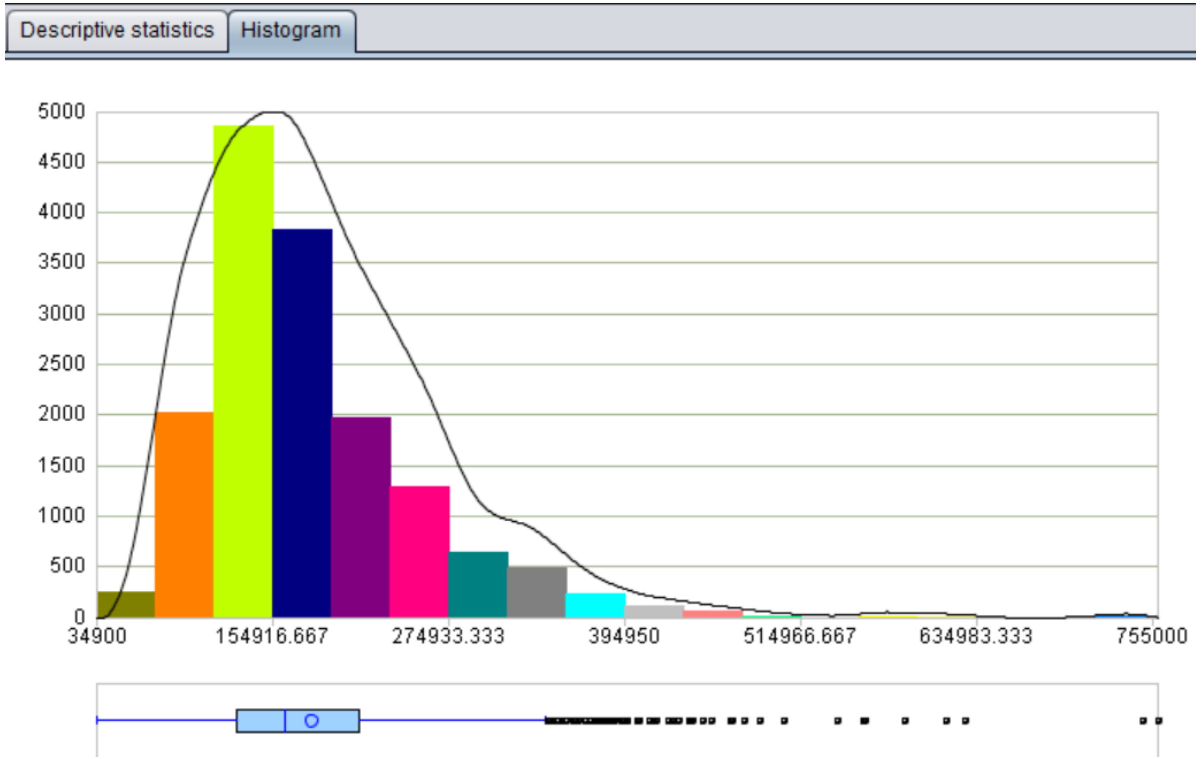
Data exploration - distribution, statistics analysis



Missing rate	Minimum	Maximum	Average	Upper qua...	Median	Lower qua...	Standard ...	Skewness
0%	34900	755000	180921.196	214000	163000	129900	79417.764	1.881

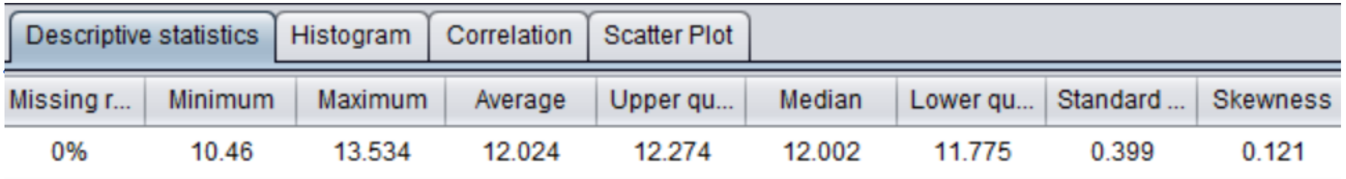
2. Target variable “SalePrice”

House price, the minimum value is 34900, the maximum value is 755000, and the skewness is 1.881; observing the distribution chart, it is found that the variable is on the right-skewed, and can be logarithmically transformed to make it normal distribution.

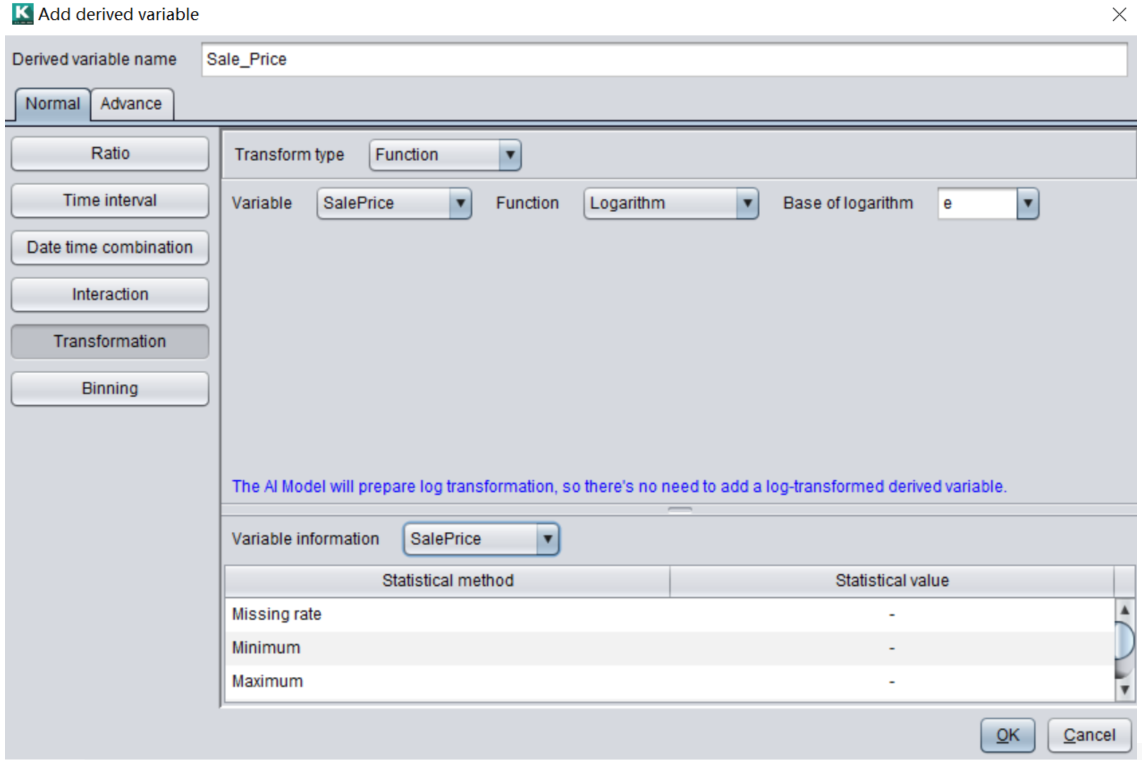


Data preprocessing - Data Rectification

Using logarithm transformation to rectify the deviation of saleprice


Descriptive statistics
Histogram
Correlation
Scatter Plot

Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	10.46	13.534	12.024	12.274	12.002	11.775	0.399	0.121



Derived variable name: Sale_Price

Transform type: Function

Variable: SalePrice

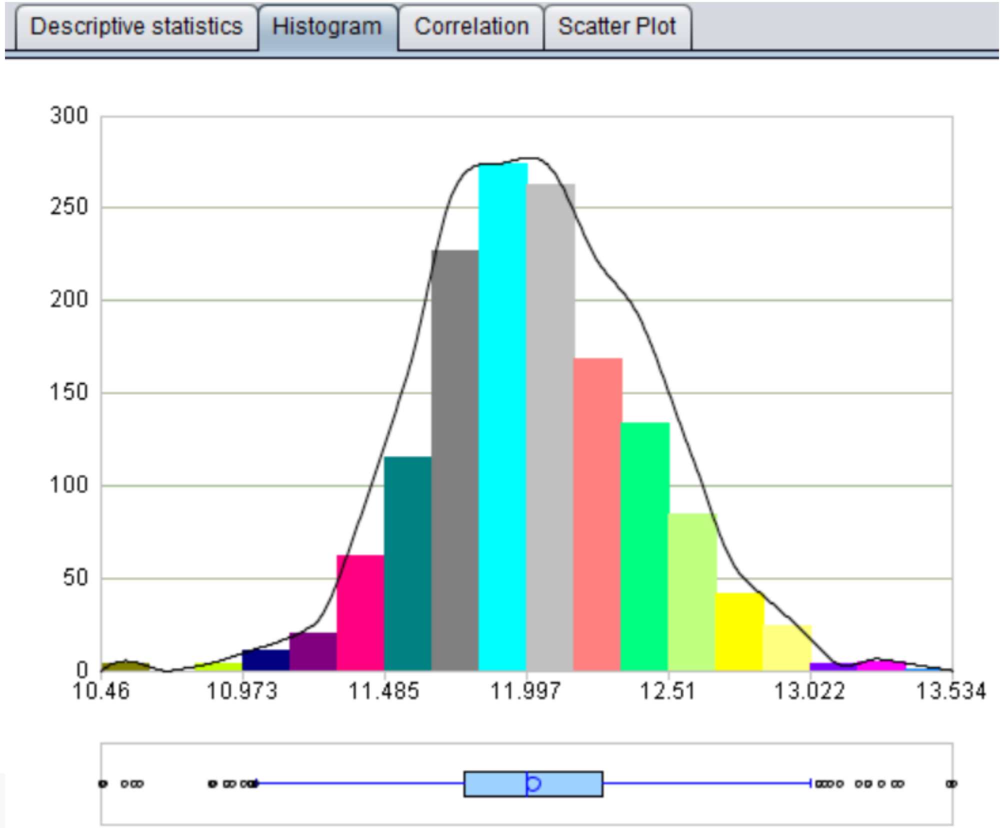
Function: Logarithm

Base of logarithm: e

The AI Model will prepare log transformation, so there's no need to add a log-transformed derived variable.

Variable information: SalePrice

Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-

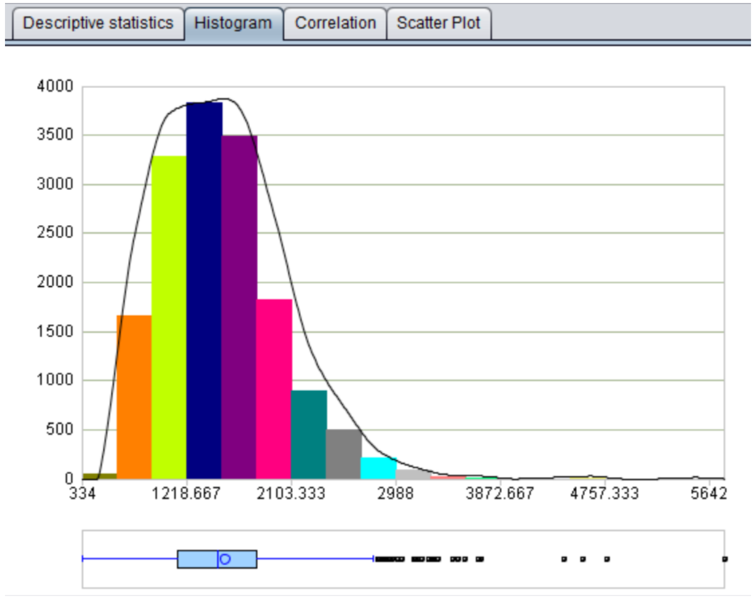


After the transformation, the variable is basically normal distribution, which is conducive to model building.

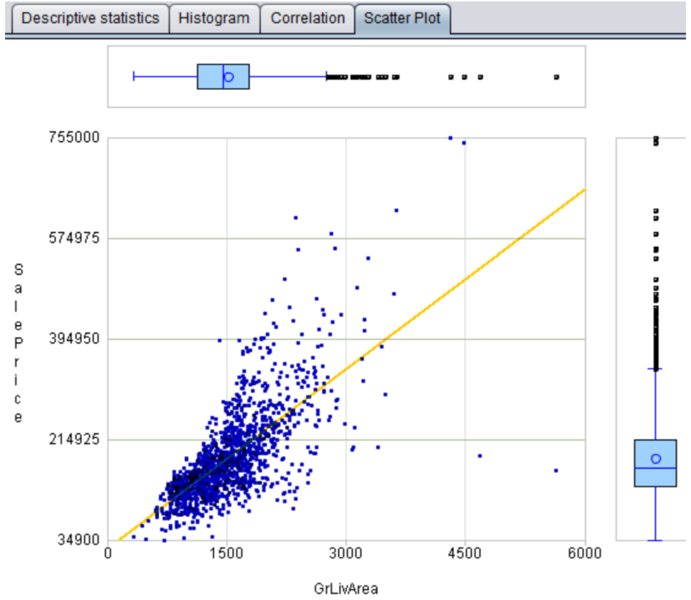
Note: YModel will automatically rectify and normalize the numerical variables. It does not need to be completed manually, as long as the variable type is set correctly. This is just to remind the reader that the data needs to be rectified by logarithmic transformation.

Data exploration - distribution analysis, statistics analysis, correlation analysis

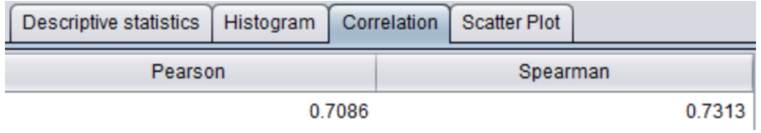
The same method can be used for characteristic variables similar to target variable distribution, such as living area GrLivArea.



Frequency distribution chart



Single factor scatter plot



Pearson	Spearman
0.7086	0.7313

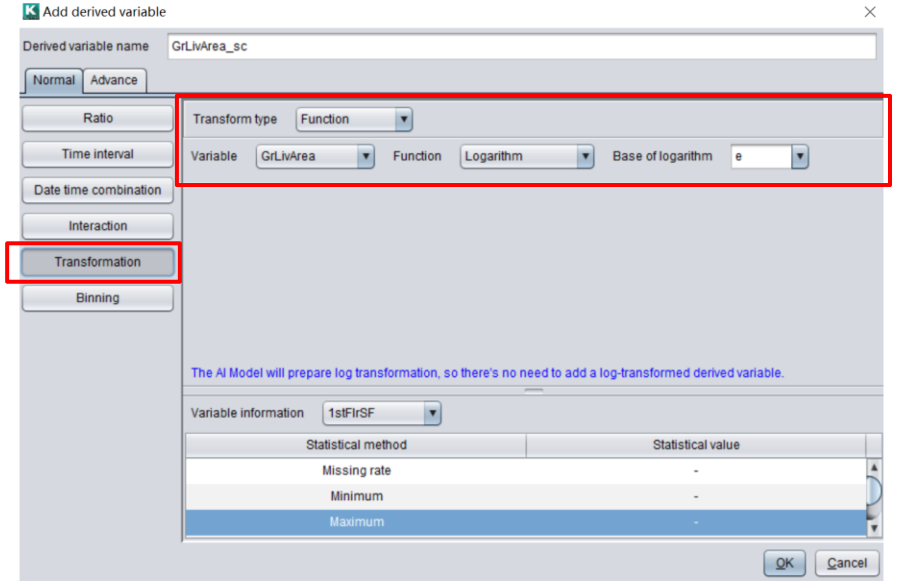
Correlation coefficient of target variable

Observation shows that we can not only see the distribution of the data, but also see the correlation coefficient with the target variable and single factor scatter plot. Both correlation coefficients are greater than 0.7, indicating that the variable is highly correlated with the house price. On the scatter diagram, we can also see that the larger the living area is, the higher the house price is, so the importance of the variable can be seen.

Data preprocessing - Data Rectification

In this data, there are variables similar to the GrLivArea distribution, including the basement area TotalBsmtSF , the first floor area 1stFlrSF, LotFrontage, etc., which can be processed in the same way.

Rectification of living area GrLivArea by logarithmic transformation.



Add derived variable

Derived variable name: GrLivArea_sc

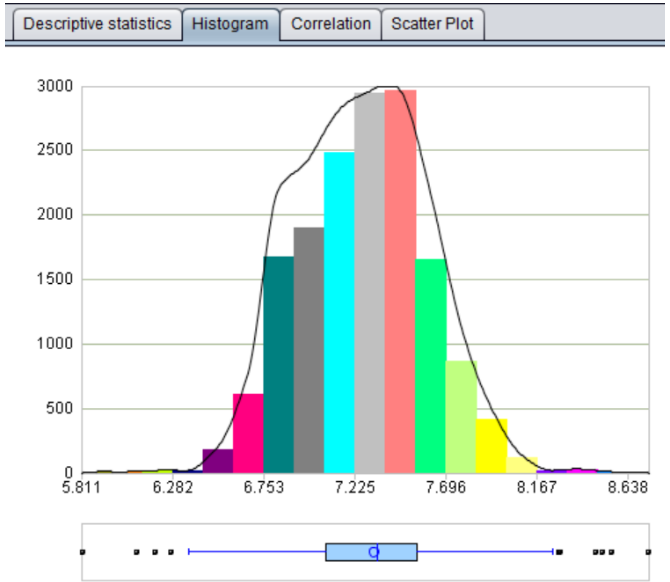
Transform type: Function

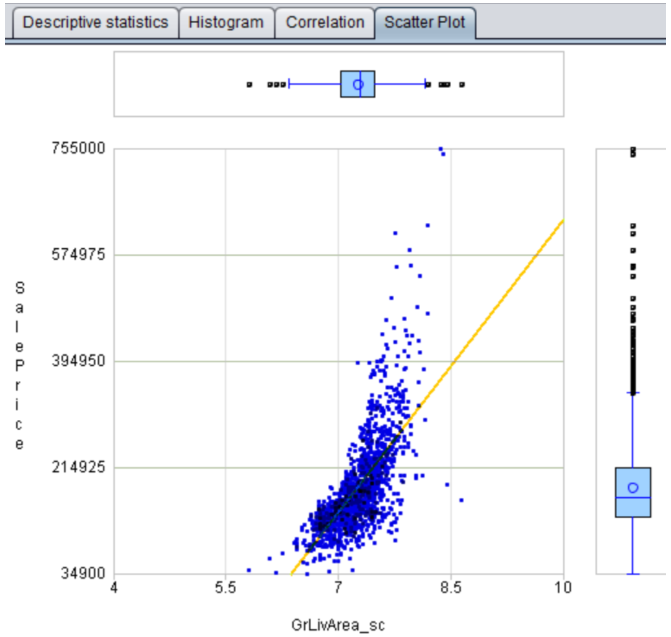
Variable: GrLivArea Function: Logarithm Base of logarithm: e

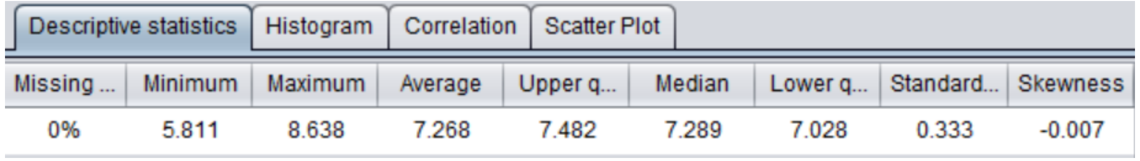
The AI Model will prepare log transformation, so there's no need to add a log-transformed derived variable.

Variable information: 1stFlrSF

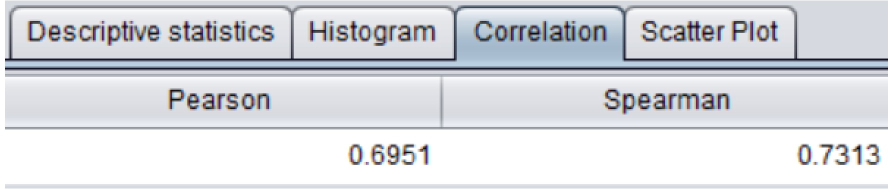
Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-






Descriptive statistics
Histogram
Correlation
Scatter Plot

Missing ...	Minimum	Maximum	Average	Upper q...	Median	Lower q...	Standard...	Skewness
0%	5.811	8.638	7.268	7.482	7.289	7.028	0.333	-0.007

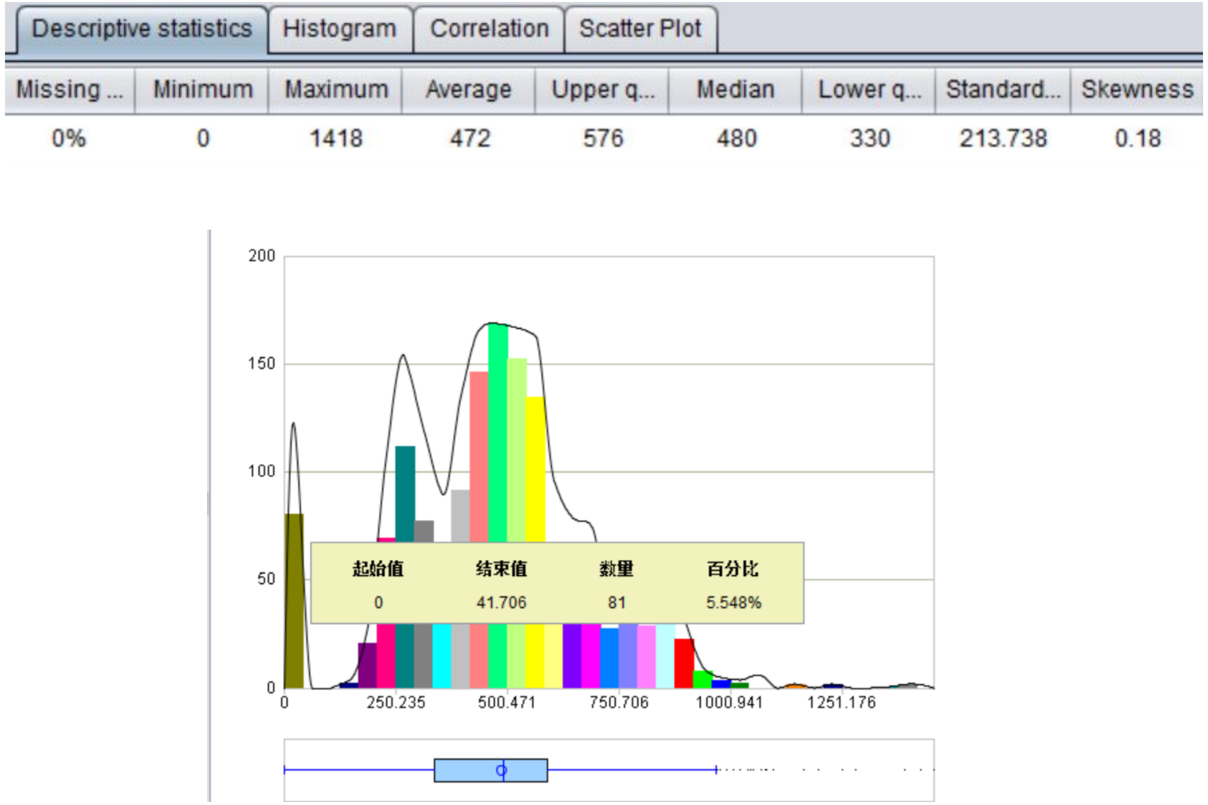

Descriptive statistics
Histogram
Correlation
Scatter Plot

Pearson	Spearman
0.6951	0.7313

After data rectification, the skewness becomes -0.007, almost no skew, and the correlation coefficient and single factor scatter diagram are also very high, which shows that this variable is more favorable for modeling than the original variable.

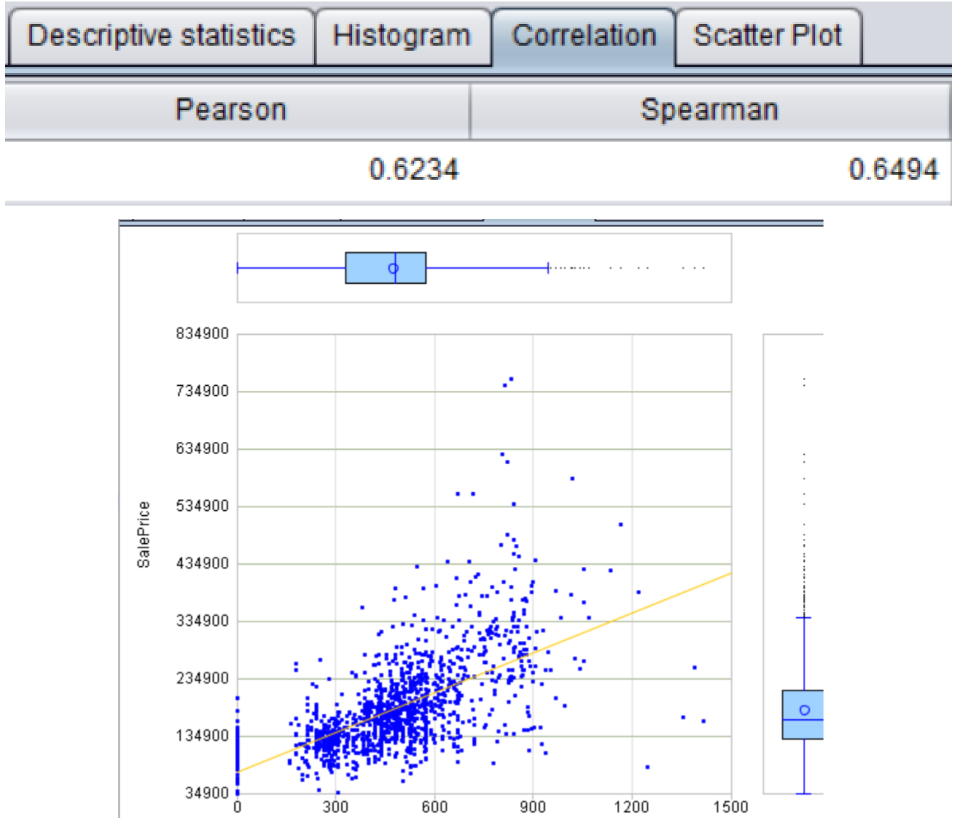
Data exploration - distribution analysis, statistics analysis, correlation analysis

GarageArea



Missing ...	Minimum	Maximum	Average	Upper q...	Median	Lower q...	Standard...	Skewness
0%	0	1418	472	576	480	330	213.738	0.18

起始值	结束值	数量	百分比
0	41.706	81	5.548%

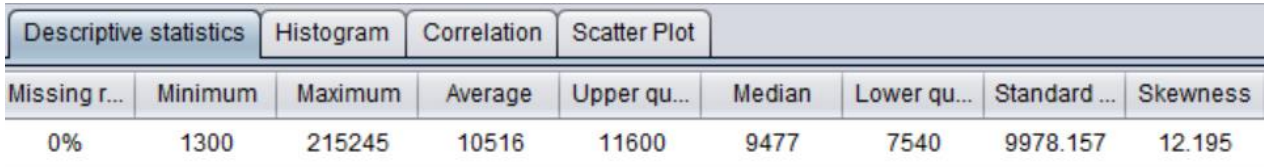


Pearson	Spearman
0.6234	0.6494

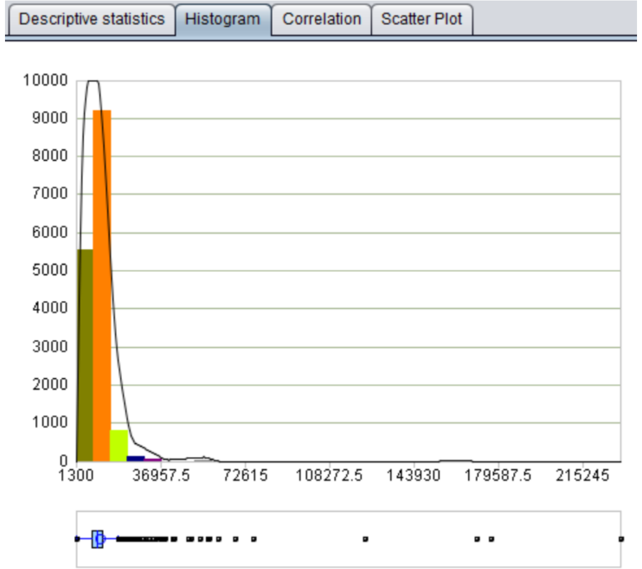
Observing the garage area, we found that the garage area of 81 houses is 0, that is to say, these houses have no garage, the overall distribution is similar to the normal distribution, and the correlation coefficient and the single factor scatter diagram shows that the variable is positively correlated with the house price, which does not need to be dealt with.

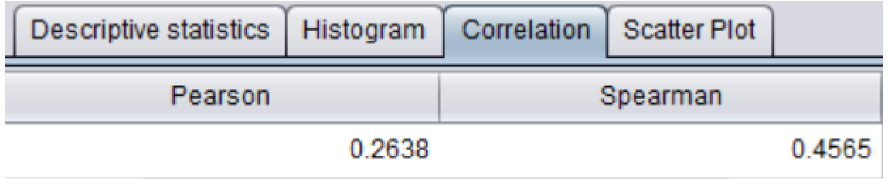
Data exploration - distribution analysis, statistics analysis, correlation analysis

LotArea

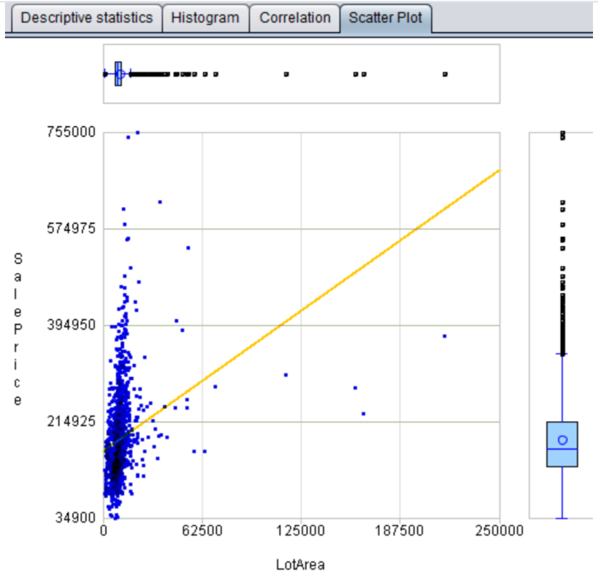


Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	1300	215245	10516	11600	9477	7540	9978.157	12.195





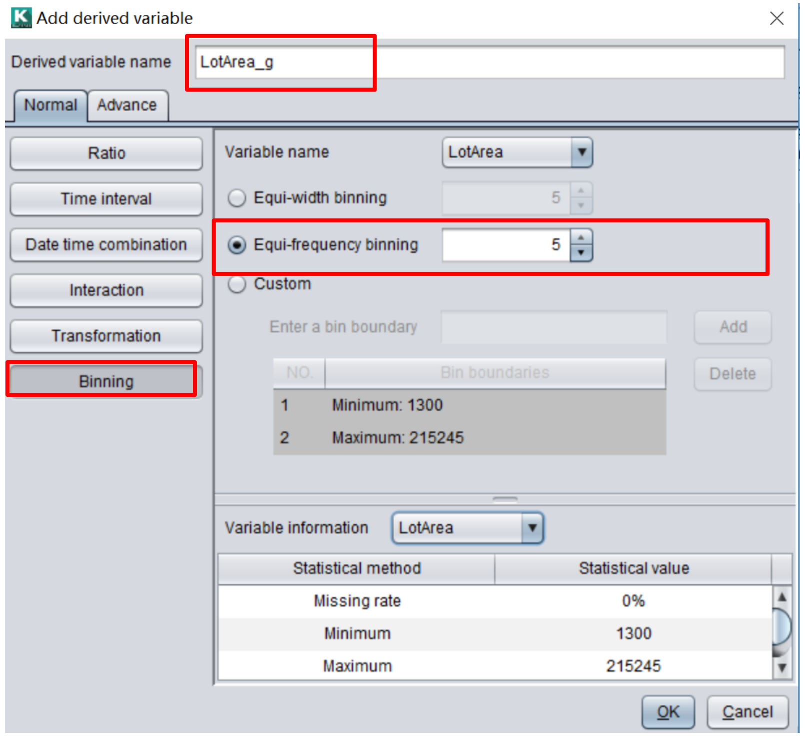
Pearson	Spearman
0.2638	0.4565



It is found that the skewness of LotArea is 12.195, which is too large to be corrected. Therefore, we can consider using the method of equal frequency binning to discretize it into categorical variable, and then observe two correlation coefficients, Pearson coefficient is only 0.2638, Spearman coefficient is 0.4565, which indicates that the variable is not linearly related to house price, but still presents a monotonous trend, that is, the larger the lot area, the higher the house price, but it is not the linear relationship.

Data preprocessing - continuous data discretization

Using the method of equal frequency binning to discretize LotArea.



Derived variable name: LotArea_g

Variable name: LotArea

○ Equi-width binning 5

● Equi-frequency binning 5

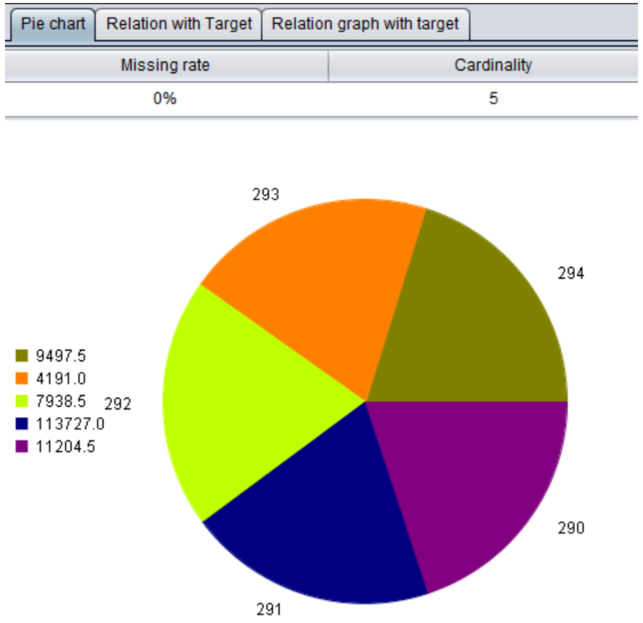
○ Custom

Enter a bin boundary

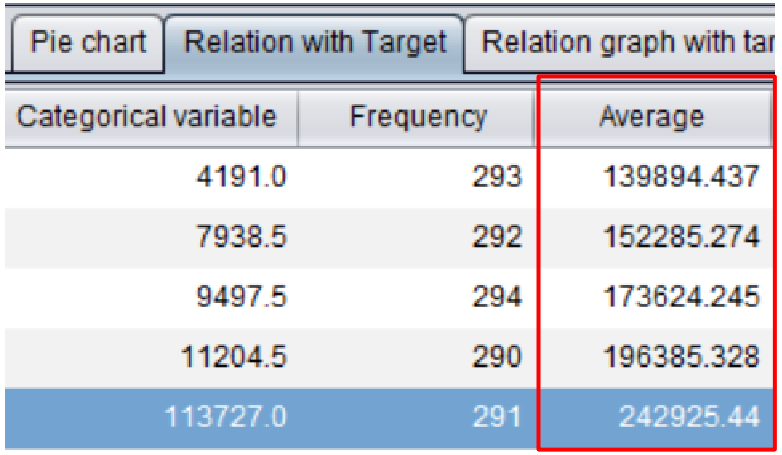
NO.	Bin boundaries
1	Minimum: 1300
2	Maximum: 215245

Variable information: LotArea

Statistical method	Statistical value
Missing rate	0%
Minimum	1300
Maximum	215245



Missing rate	Cardinality
0%	5

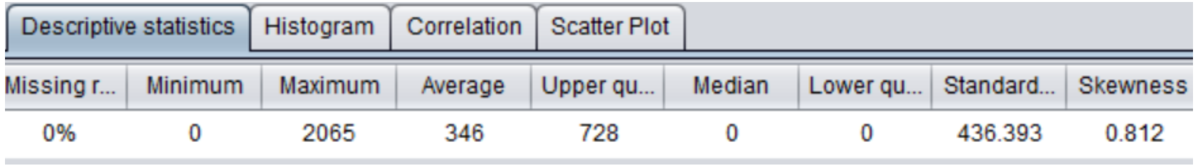


Categorical variable	Frequency	Average
4191.0	293	139894.437
7938.5	292	152285.274
9497.5	294	173624.245
11204.5	290	196385.328
113727.0	291	242925.44

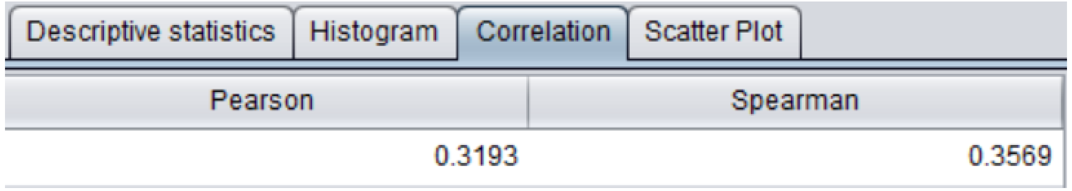
After equal frequency binning, observe the statistics after grouping. With the increase of lot area, the average value of house price rises obviously. This variable should be effective.

Data exploration - distribution analysis, statistics analysis, correlation analysis

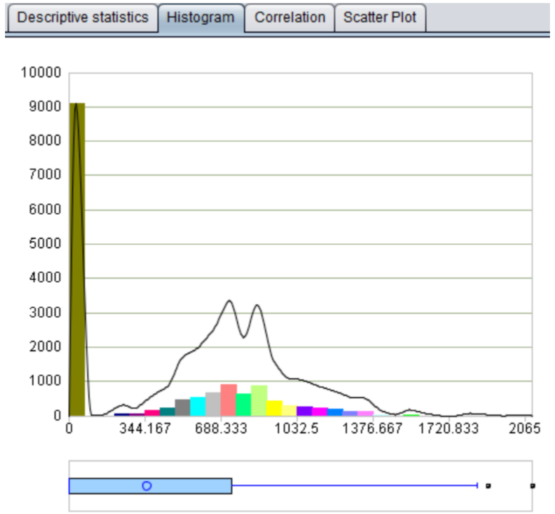
2ndFlrSF

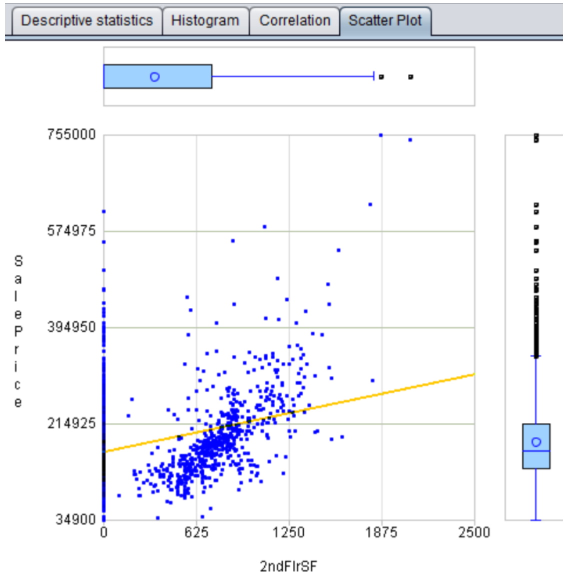


Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard...	Skewness
0%	0	2065	346	728	0	0	436.393	0.812



Pearson	Spearman
0.3193	0.3569



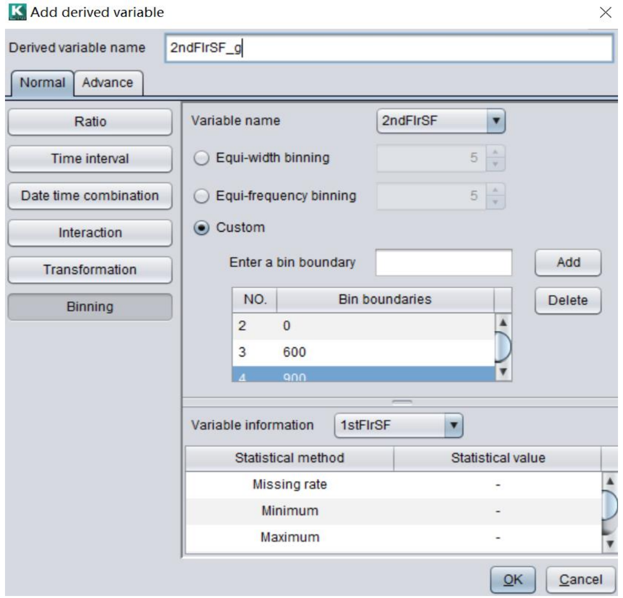


The skewness of 2ndFlrSF is 0.812, which is not large. However, after observing the distribution map, we can see that a considerable part of the second floor area is 0, that is, the bungalow. In this case, we need to consider binning, and distinguish the bungalow and the second floor. From the scatter diagram, we can see that the house price increases with the area of the second floor, so it is reasonable to divide it into four parts, That is, 0,0~600,601~900, more than 900.

Data preprocessing - continuous data discretization

The variables similar to the distribution of 2ndFlrSF include WoodDeckSF and so on. We can start binning according to different data distribution.

Discretization of 2ndFlrSF by using the method of custom binning.



Derived variable name: 2ndFlrSF_g

Variable name: 2ndFlrSF

○ Equi-width binning 5

○ Equi-frequency binning 5

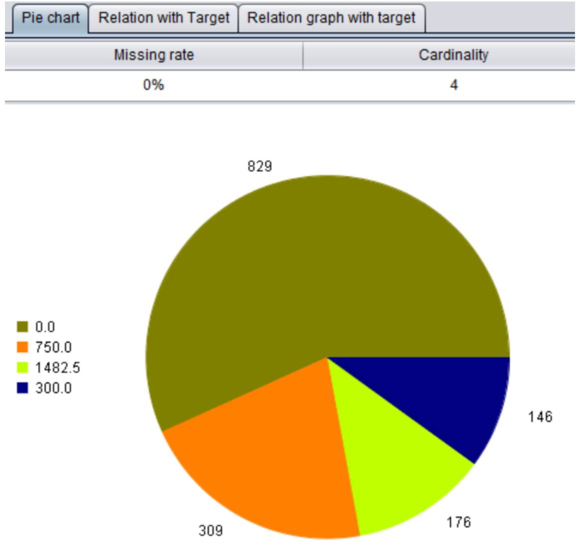
● Custom

Enter a bin boundary:

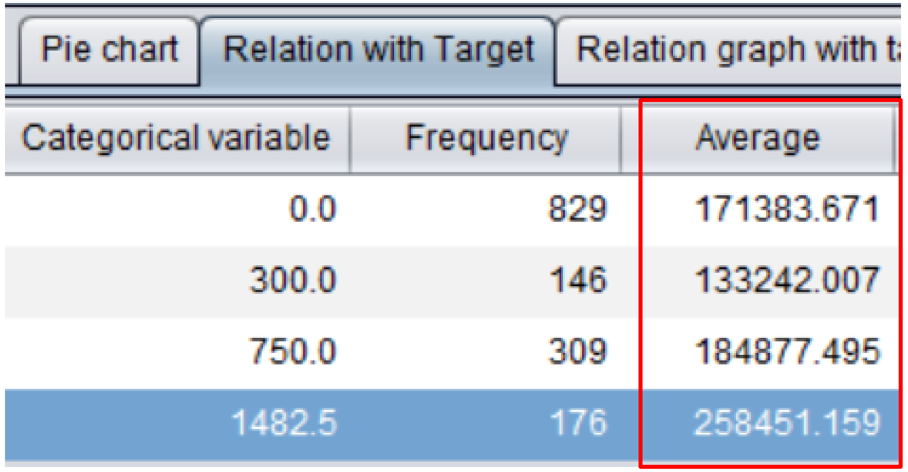
NO.	Bin boundaries
2	0
3	600
4	900

Variable information: 1stFlrSF

Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-



Missing rate	Cardinality
0%	4

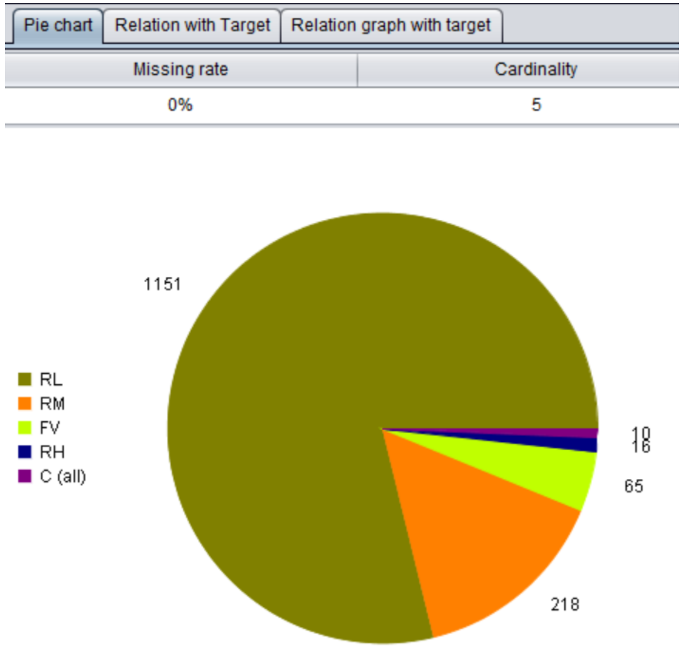

Pie chart
Relation with Target
Relation graph with t

Categorical variable	Frequency	Average
0.0	829	171383.671
300.0	146	133242.007
750.0	309	184877.495
1482.5	176	258451.159

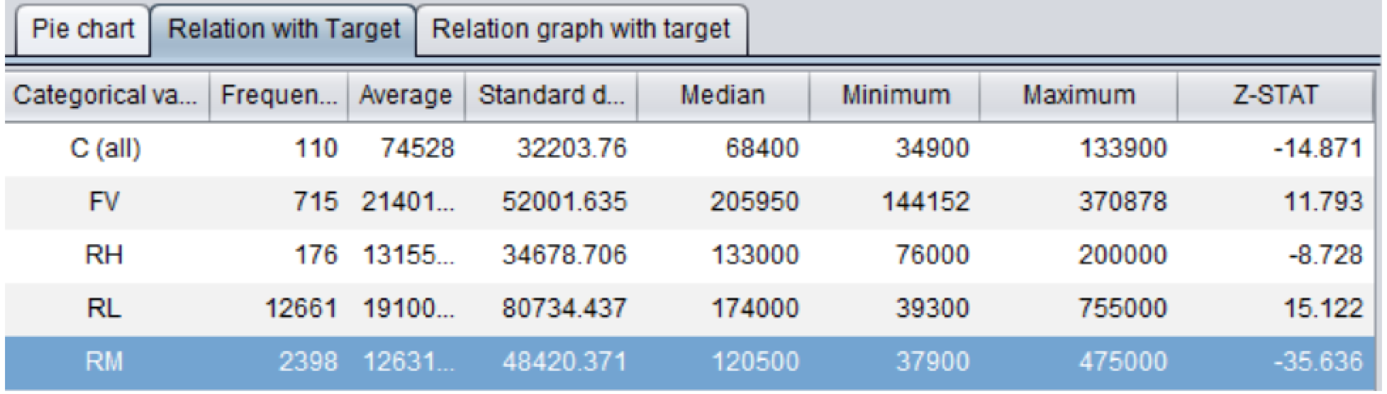
After custom binning, observe the statistics after grouping, the house price of bungalow should consider other aspects more, and the second floor well reflects that the larger the area, the higher the house price.

Data exploration - distribution analysis

MSZoning



Missing rate	Cardinality
0%	5

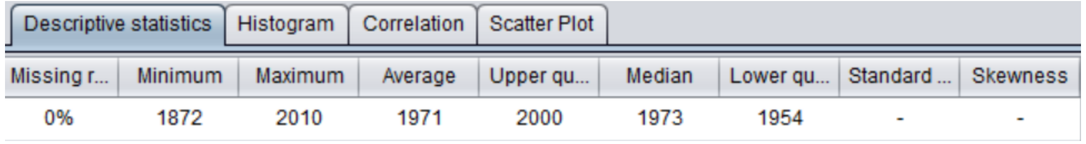

Pie chart
Relation with Target
Relation graph with target

Categorical va...	Frequen...	Average	Standard d...	Median	Minimum	Maximum	Z-STAT
C (all)	110	74528	32203.76	68400	34900	133900	-14.871
FV	715	21401...	52001.635	205950	144152	370878	11.793
RH	176	13155...	34678.706	133000	76000	200000	-8.728
RL	12661	19100...	80734.437	174000	39300	755000	15.122
RM	2398	12631...	48420.371	120500	37900	475000	-35.636

The first thing we think about when we buy a house is location. Observing the houses of MSZoning and RL is the most, followed by RM, RH, FV and C (all). Observing the average value of house prices after grouping, we can find that the price difference of each zone is very obvious, so like our intuition, the location is very important. For such categorical variable, one hot coding is usually used for preprocessing, which will be done by the YModel tool preprocesses, so we don't need to complete it manually.

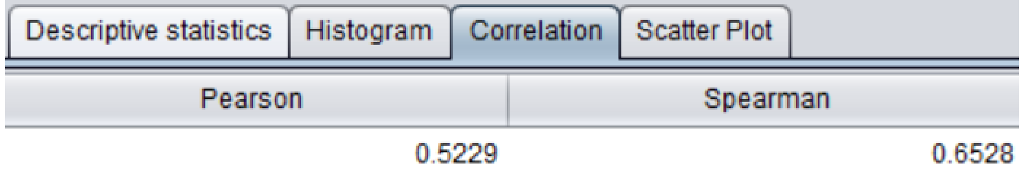
Data exploration - distribution analysis, statistics analysis, correlation analysis

YearBuilt



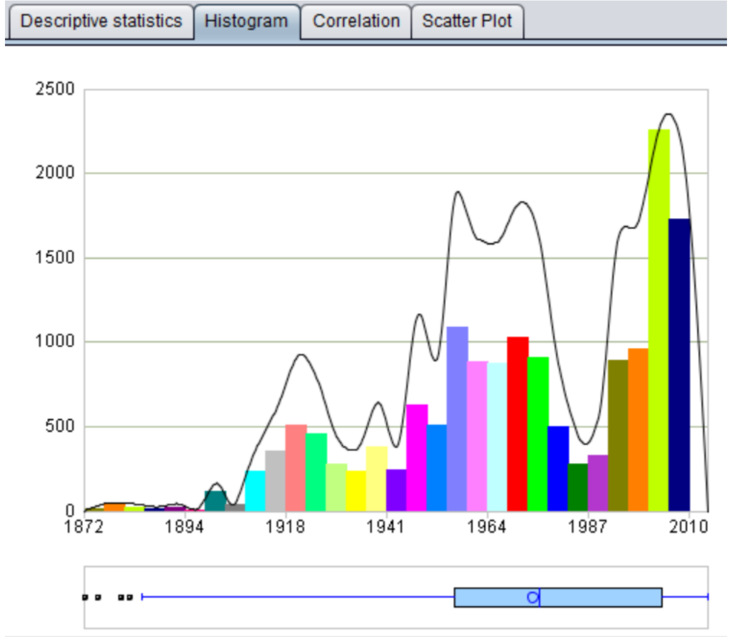
Descriptive statistics
Histogram
Correlation
Scatter Plot

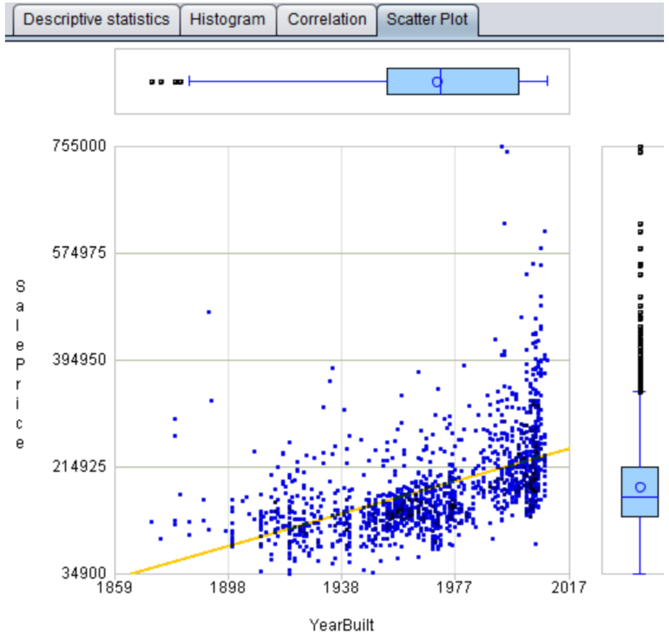
Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	1872	2010	1971	2000	1973	1954	-	-



Descriptive statistics
Histogram
Correlation
Scatter Plot

Pearson	Spearman
0.5229	0.6528

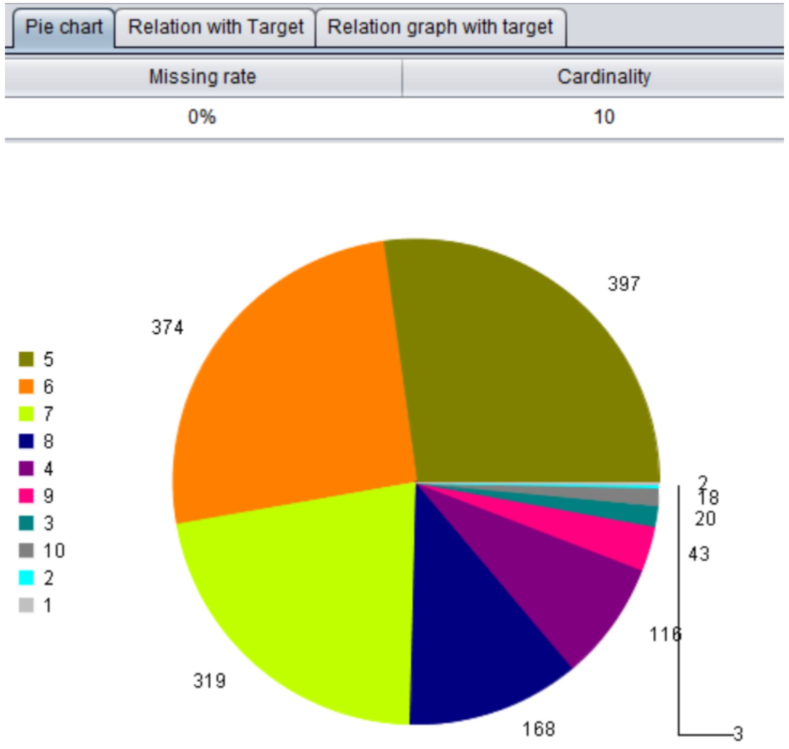




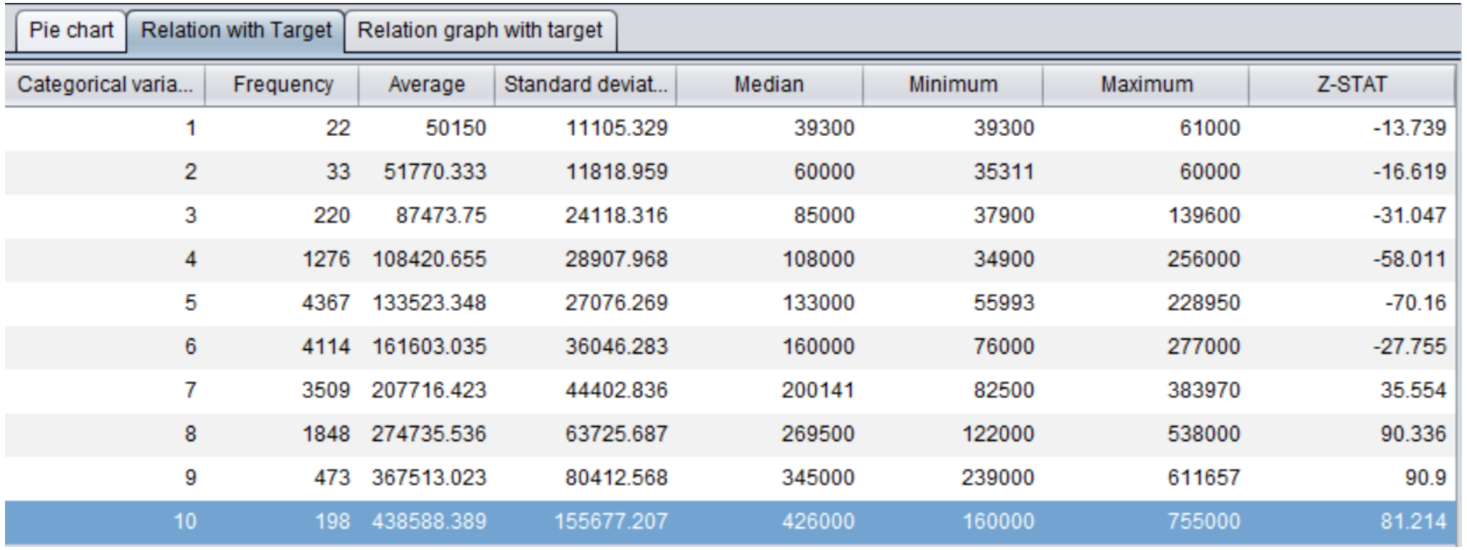
Of course, we need to consider the year built when we buy a house. The old house is usually not as expensive as the new one. We can see that the earliest house was built in 1872 (with a long history), but both the correlation coefficient and the scatter chart show that the age and the house price are basically positively related.

Data exploration - distribution analysis

OverallQual



Missing rate	Cardinality
0%	10

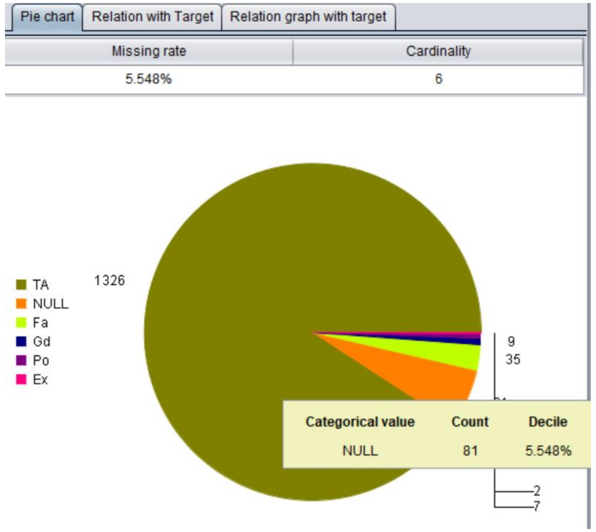

Pie chart
Relation with Target
Relation graph with target

Categorical varia...	Frequency	Average	Standard deviat...	Median	Minimum	Maximum	Z-STAT
1	22	50150	11105.329	39300	39300	61000	-13.739
2	33	51770.333	11818.959	60000	35311	60000	-16.619
3	220	87473.75	24118.316	85000	37900	139600	-31.047
4	1276	108420.655	28907.968	108000	34900	256000	-58.011
5	4367	133523.348	27076.269	133000	55993	228950	-70.16
6	4114	161603.035	36046.283	160000	76000	277000	-27.755
7	3509	207716.423	44402.836	200141	82500	383970	35.554
8	1848	274735.536	63725.687	269500	122000	538000	90.336
9	473	367513.023	80412.568	345000	239000	611657	90.9
10	198	438588.389	155677.207	426000	160000	755000	81.214

This variable is very important intuitively. It gives the house 10 grades. The higher the grade is, the better the quality is. Of course, the higher the house price is. It's easy to find this relationship by looking at the grouping target statistics.

Data exploration - distribution analysis, missing value analysis

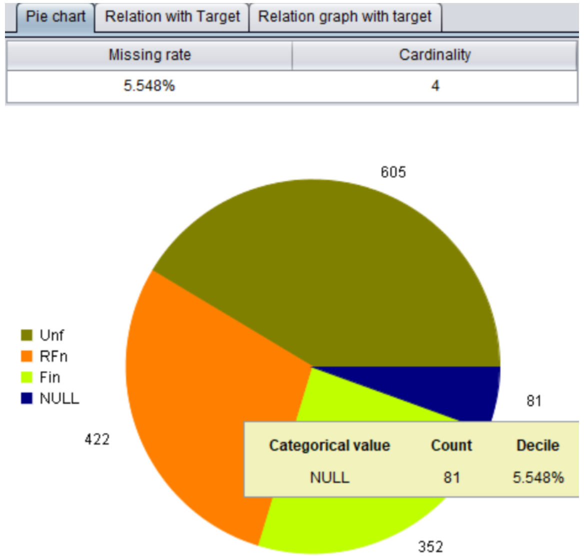
GarageX



Missing rate	Cardinality
5.548%	6

Categorical value	Count	Decile
NULL	81	5.548%

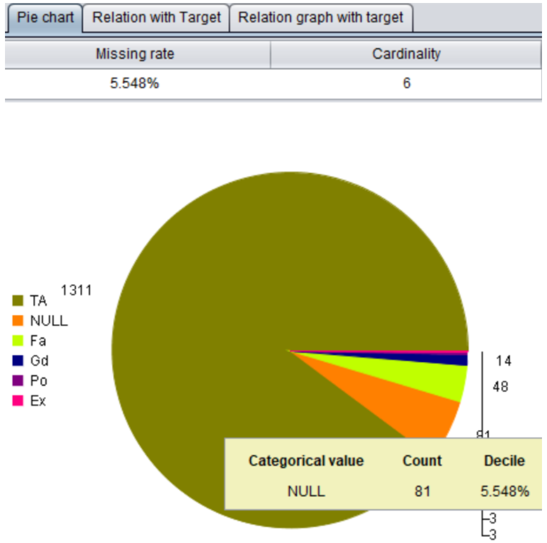
GarageCond



Missing rate	Cardinality
5.548%	4

Categorical value	Count	Decile
NULL	81	5.548%

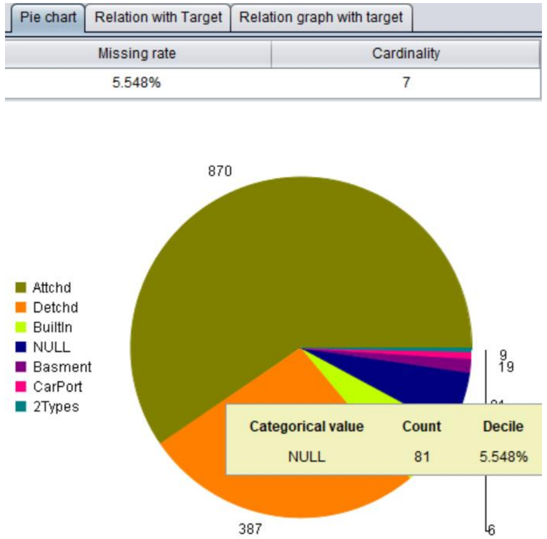
GarageFinish



Missing rate	Cardinality
5.548%	6

Categorical value	Count	Decile
NULL	81	5.548%

GarageQual

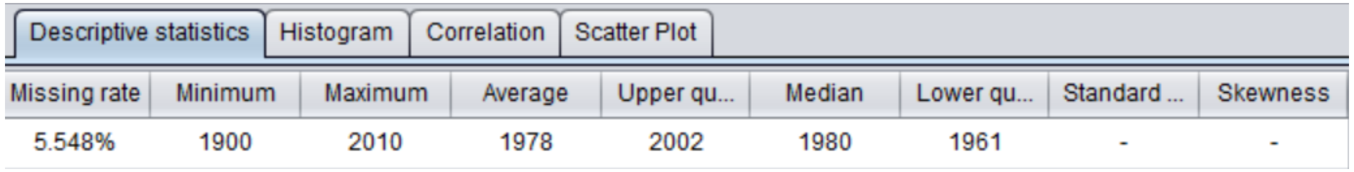


Missing rate	Cardinality
5.548%	7

Categorical value	Count	Decile
NULL	81	5.548%

GarageType

GarageYrBlt

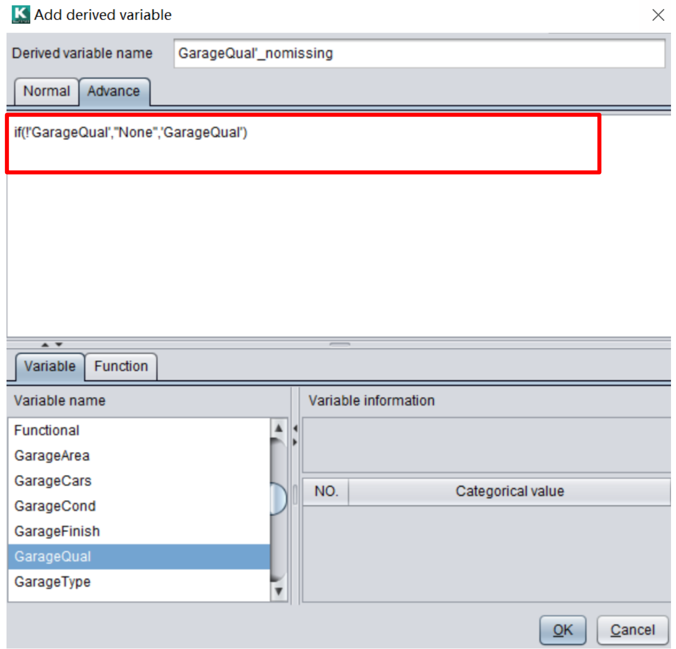


Missing rate	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
5.548%	1900	2010	1978	2002	1980	1961	-	-

When analyzing the garage area, we found that there are 81 houses without a garage, so some attributes of the garage are missing. Therefore, for these variables, the missing value of the categorical variable is set to none, which becomes a separate category. For the year, we set them to the earliest value of the garage, that is 1900.

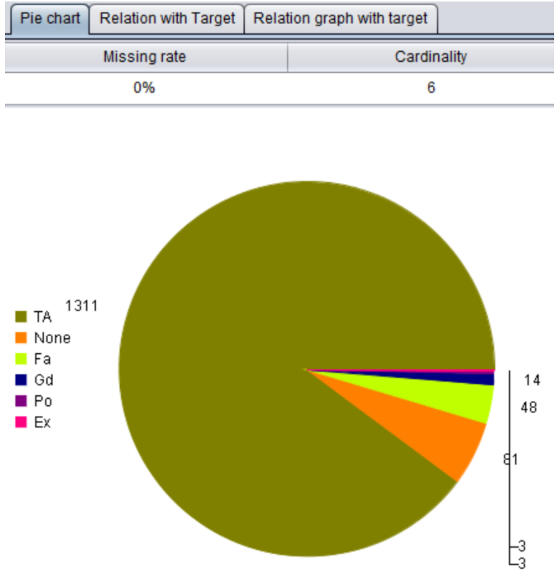
Data preprocessing - missing value processing

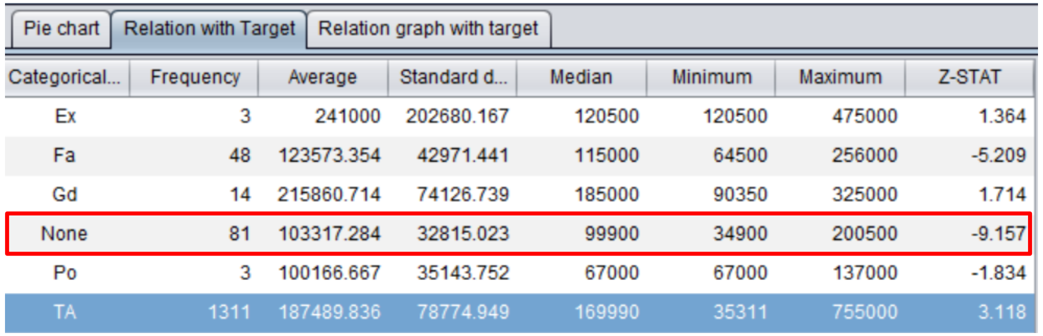
Take GarageQual as an example to add the derived variable GarageQual_nomissing, fill in missing value.


Add derived variable

Derived variable name: GarageQual'_nomissing

```
if(!'GarageQual',"None",'GarageQual')
```


Missing rate	Cardinality
0%	6



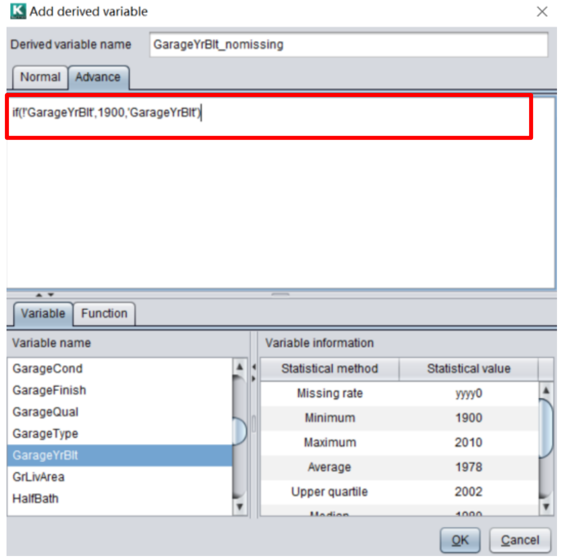
Categorical...	Frequency	Average	Standard d...	Median	Minimum	Maximum	Z-STAT
Ex	3	241000	202680.167	120500	120500	475000	1.364
Fa	48	123573.354	42971.441	115000	64500	256000	-5.209
Gd	14	215860.714	74126.739	185000	90350	325000	1.714
None	81	103317.284	32815.023	99900	34900	200500	-9.157
Po	3	100166.667	35143.752	67000	67000	137000	-1.834
TA	1311	187489.836	78774.949	169990	35311	755000	3.118

The missing rate of the garage series of this data is not very large, and the YModel will automatically process this data. Here, we just remind you to fill in the missing value according to the characteristics of the data.

Data preprocessing - missing value processing

Variables similar to garage series include PoolX series, FirePlaceX series, BsmtX series, which can be filled in the same way.

For GarageYrBlt, add derived variable GarageYrBlt _nomissing, use 1900 to fill in missing values.

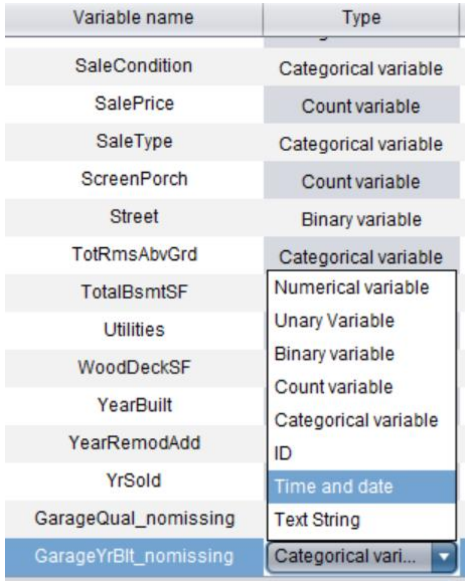


Add derived variable

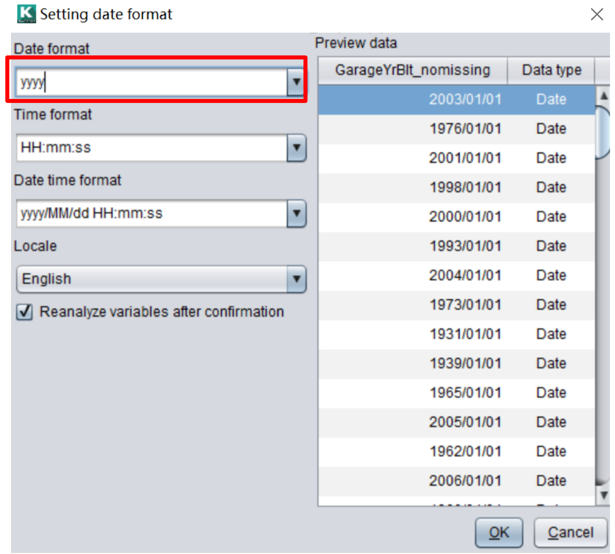
Derived variable name: GarageYrBlt_nomissing

```
if(!'GarageYrBlt',1900,'GarageYrBlt')
```

Statistical method	Statistical value
Missing rate	yyyy0
Minimum	1900
Maximum	2010
Average	1978
Upper quartile	2002



Variable name	Type
SaleCondition	Categorical variable
SalePrice	Count variable
SaleType	Categorical variable
ScreenPorch	Count variable
Street	Binary variable
TotRmsAbvGrd	Categorical variable
TotalBsmtSF	
Utilities	
WoodDeckSF	
YearBuilt	
YearRemodAdd	
YrSold	
GarageQual_nomissing	
GarageYrBlt_nomissing	Categorical vari...



Setting date format

Date format: yyyy

Time format: HH:mm:ss

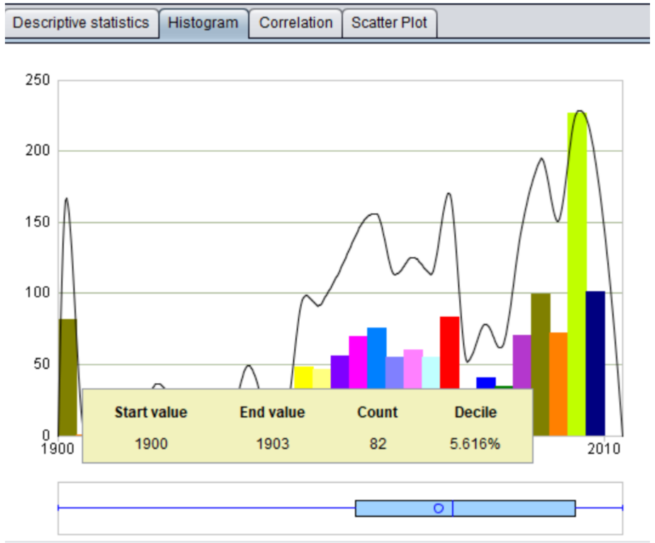
Date time format: yyyy/MM/dd HH:mm:ss

Locale: English

☑ Reanalyze variables after confirmation

Preview data

GarageYrBlt_nomissing	Data type
2003/01/01	Date
1976/01/01	Date
2001/01/01	Date
1998/01/01	Date
2000/01/01	Date
1993/01/01	Date
2004/01/01	Date
1973/01/01	Date
1931/01/01	Date
1939/01/01	Date
1965/01/01	Date
2005/01/01	Date
1962/01/01	Date
2006/01/01	Date

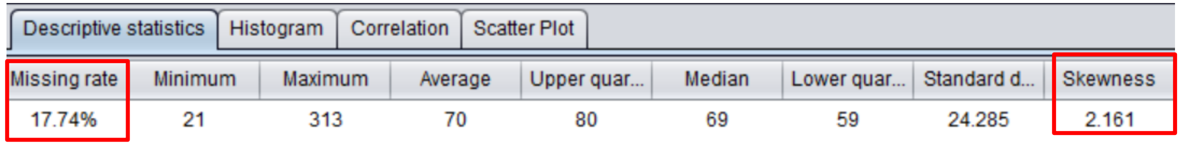


Start value	End value	Count	Decile
1900	1903	82	5.616%

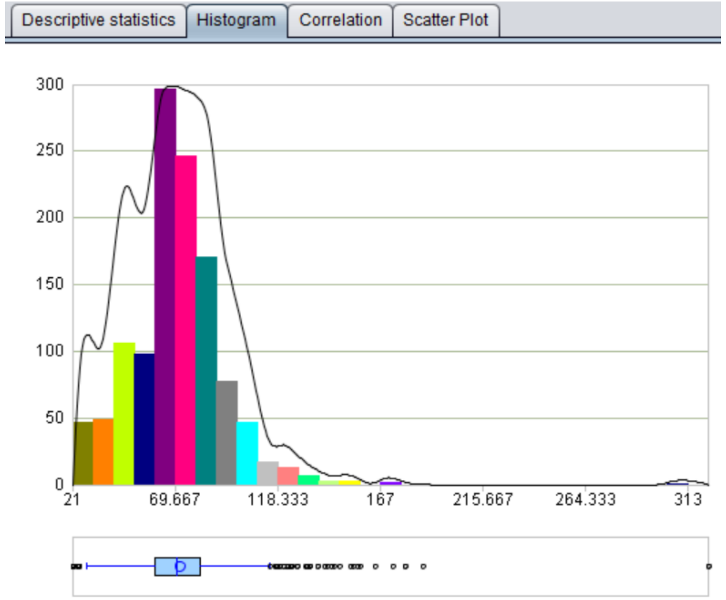
For the missing values of time and date variables, YModel has a special way to fill the missing values. We simply fill in 1900 here.
Note: after filling in, we notice that the new derived variable has become a categorical variable, and we need to change it to the time date type. After filling it up, we found that there were more garages in 1900. The missing value is gone.

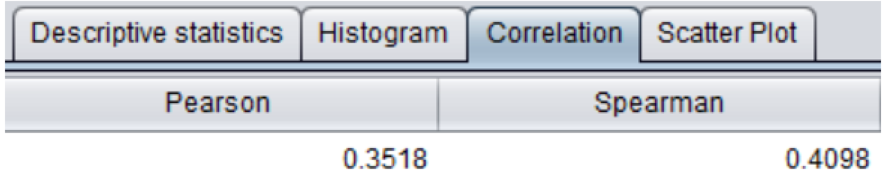
Data exploration - distribution analysis, missing value analysis

LotFrontage

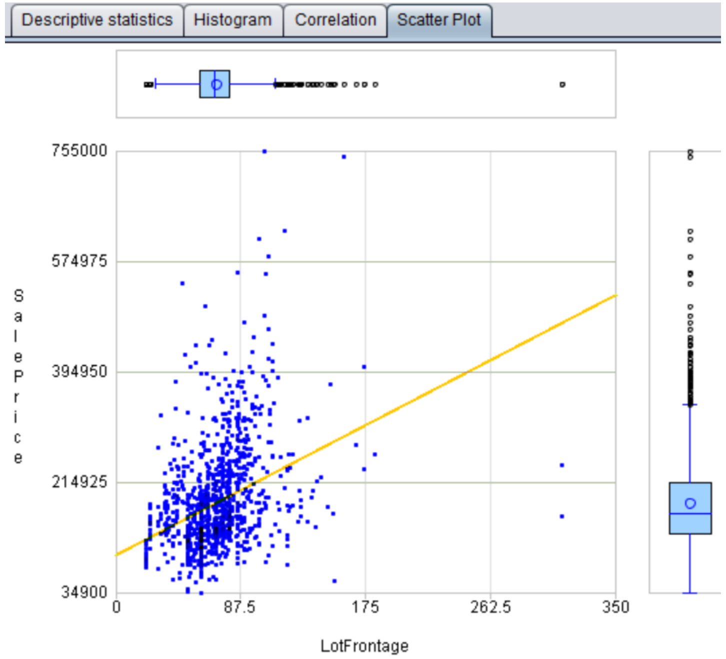


Missing rate	Minimum	Maximum	Average	Upper quar...	Median	Lower quar...	Standard d...	Skewness
17.74%	21	313	70	80	69	59	24.285	2.161





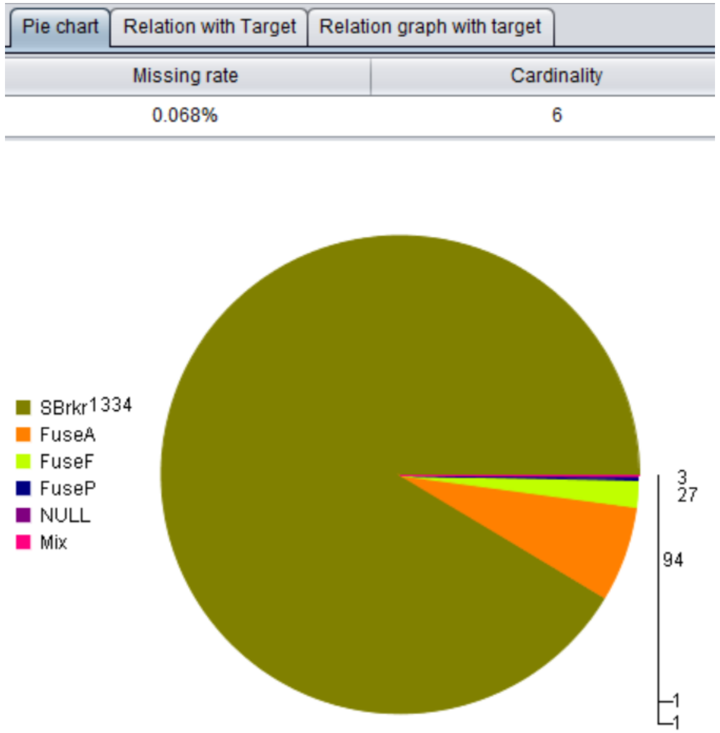
Pearson	Spearman
0.3518	0.4098



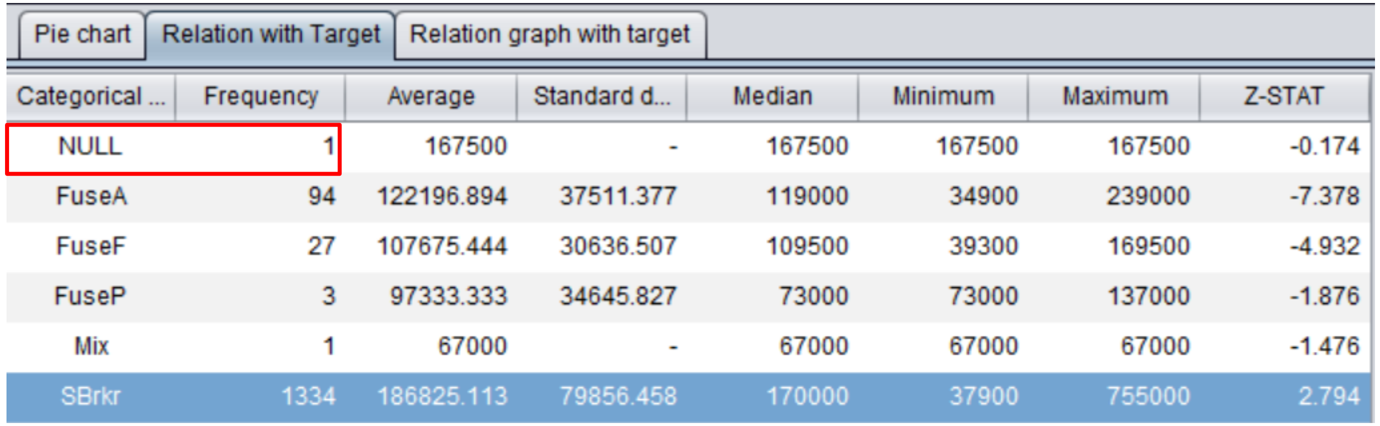
It is common to fill the missing value of this distribution. Because of its large skewness, it can be filled with median. The YModel tool will automatically handle this missing value, so we will not deal with it here.

Data exploration - distribution analysis, missing value analysis

Electrical



Missing rate	Cardinality
0.068%	6

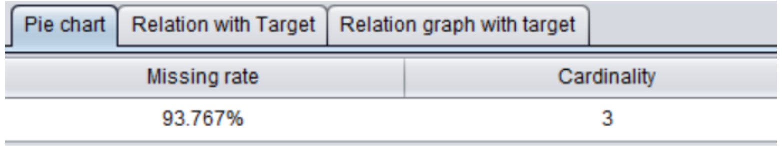


Categorical ...	Frequency	Average	Standard d...	Median	Minimum	Maximum	Z-STAT
NULL	1	167500	-	167500	167500	167500	-0.174
FuseA	94	122196.894	37511.377	119000	34900	239000	-7.378
FuseF	27	107675.444	30636.507	109500	39300	169500	-4.932
FuseP	3	97333.333	34645.827	73000	73000	137000	-1.876
Mix	1	67000	-	67000	67000	67000	-1.476
SBrkr	1334	186825.113	79856.458	170000	37900	755000	2.794

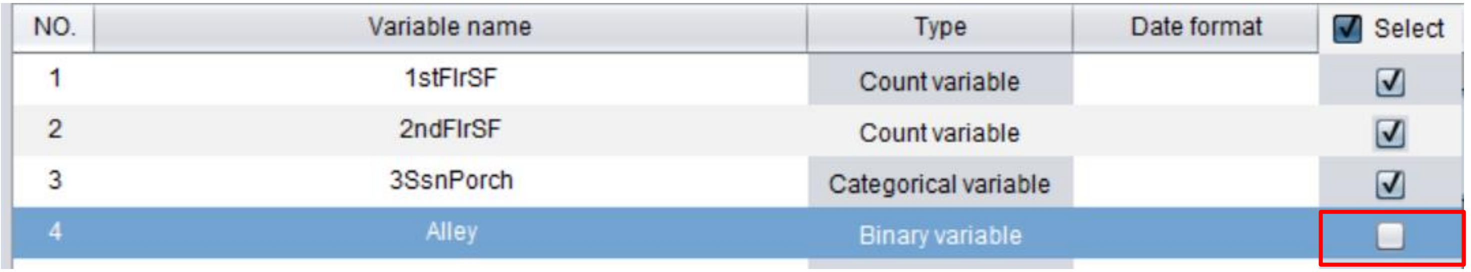
In this case, the missing values are very little. We can choose to delete this record or use mode to fill in the missing value. YModel usually use mode to fill in.

Data exploration - missing value analysis

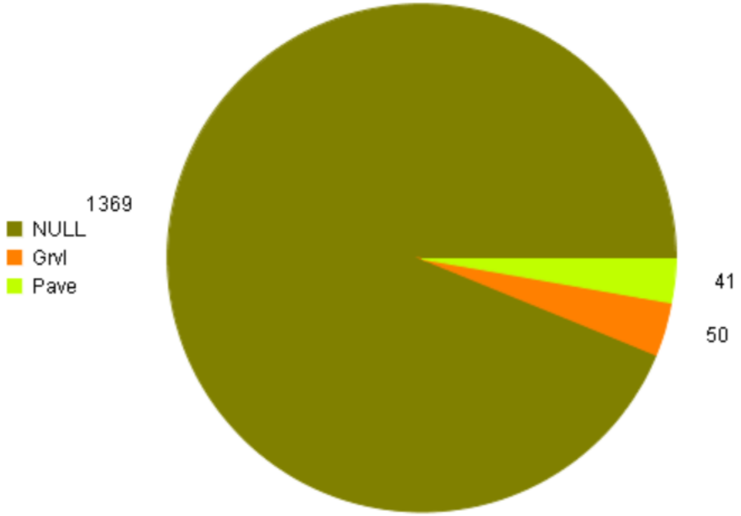
Alley



Missing rate	Cardinality
93.767%	3



NO.	Variable name	Type	Date format	☑ Select
1	1stFlrSF	Count variable		☑
2	2ndFlrSF	Count variable		☑
3	3SsnPorch	Categorical variable		☑
4	Alley	Binary variable		☐

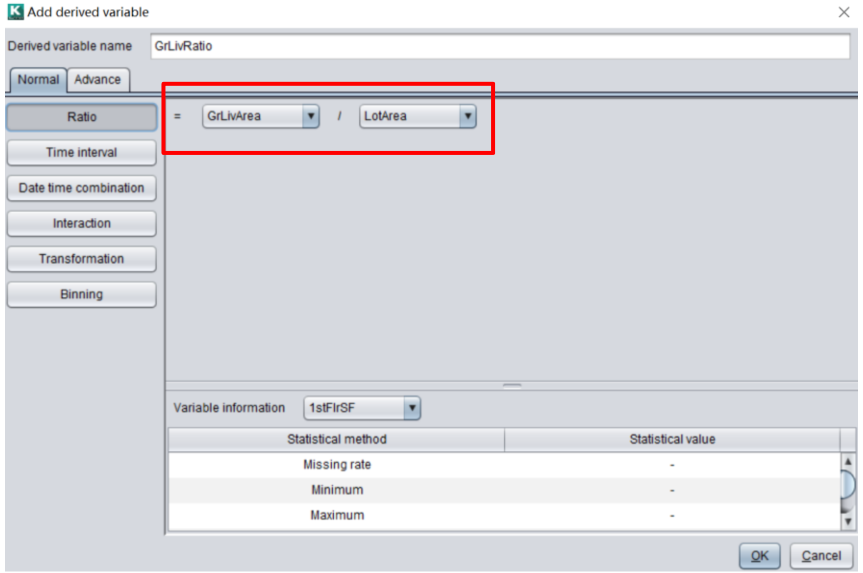


There are two ways to deal with the variable with a large number of missing, the first is to directly discard it; the second is to treat the missing value as another category and set it as none. Because we seldom consider that the type of alley is asphalt or other when we buy a house, so we can delete it.

The missing value filling introduced here is a simple filling method, which is also a relatively effective method. And the YModel tool has the way of intelligent filling. It is more effective and targeted to fill the missing value by building the model.

Data preprocessing - add derived variables, variable interaction

Use GrLivArea / LotArea to generate living area proportion.

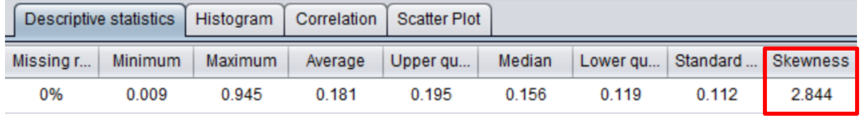


Derived variable name: GrLivRatio

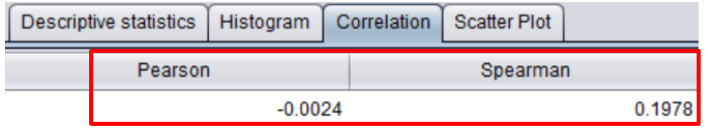
= GrLivArea / LotArea

Variable information: 1stFlrSF

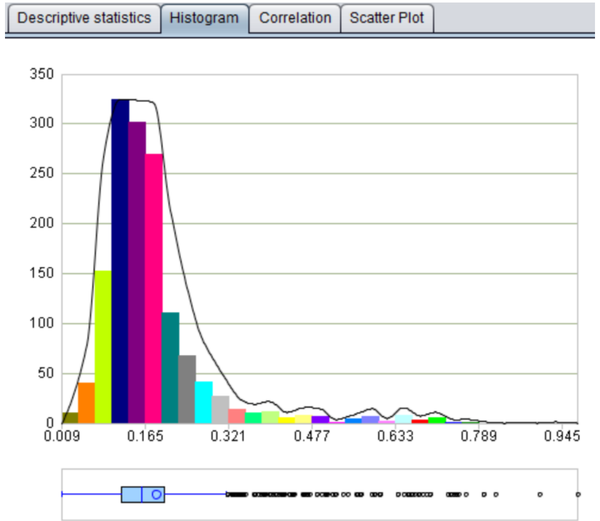
Statistical method	Statistical value
Missing rate	-
Minimum	-
Maximum	-

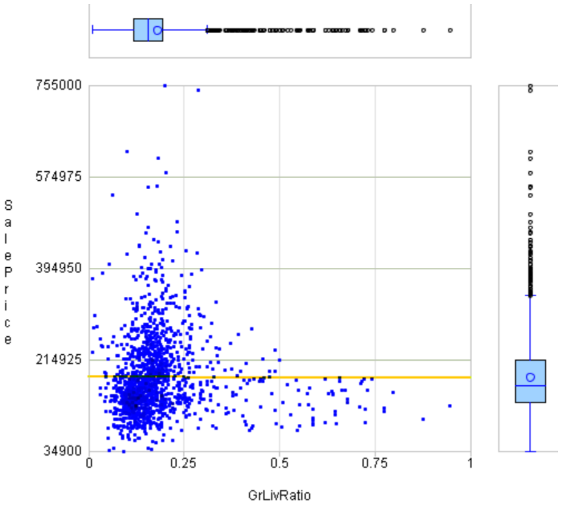

Descriptive statistics
Histogram
Correlation
Scatter Plot

Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	0.009	0.945	0.181	0.195	0.156	0.119	0.112	2.844


Descriptive statistics
Histogram
Correlation
Scatter Plot

Pearson	Spearman
-0.0024	0.1978

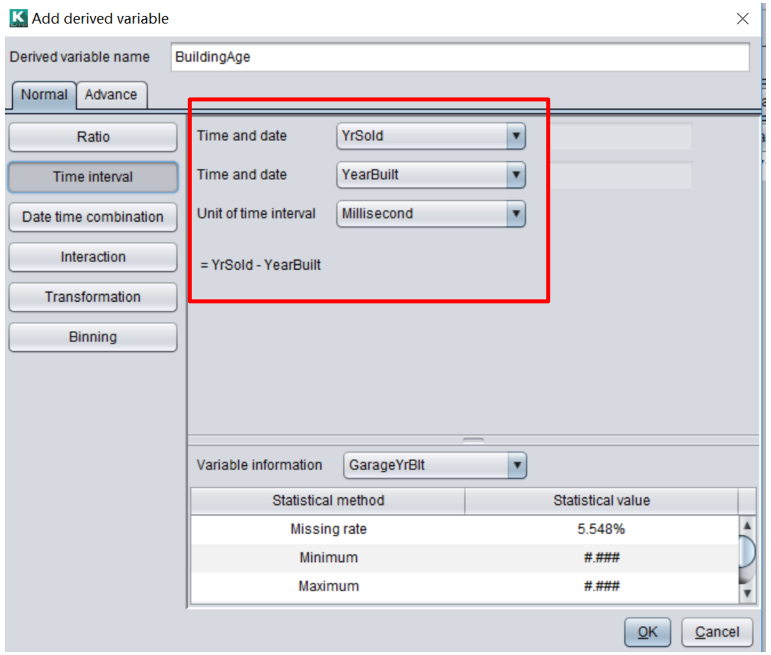




Observing this derived variable, we found that it is seriously biased and should be corrected, and the correlation coefficient is very small. The scatter diagram also shows that the correlation degree is very low. This variable is not very useful at preliminary inference. Here is mainly to introduce that the ratio method can be used between two variables to generate the derived variable, which is temporarily not available in the YModel tool.

Data preprocessing - add derived variables, calculate time interval

Calculate the time interval between YrSold and YearBuilt as "BuildingAge".



Add derived variable

Derived variable name: BuildingAge

Time and date: YrSold

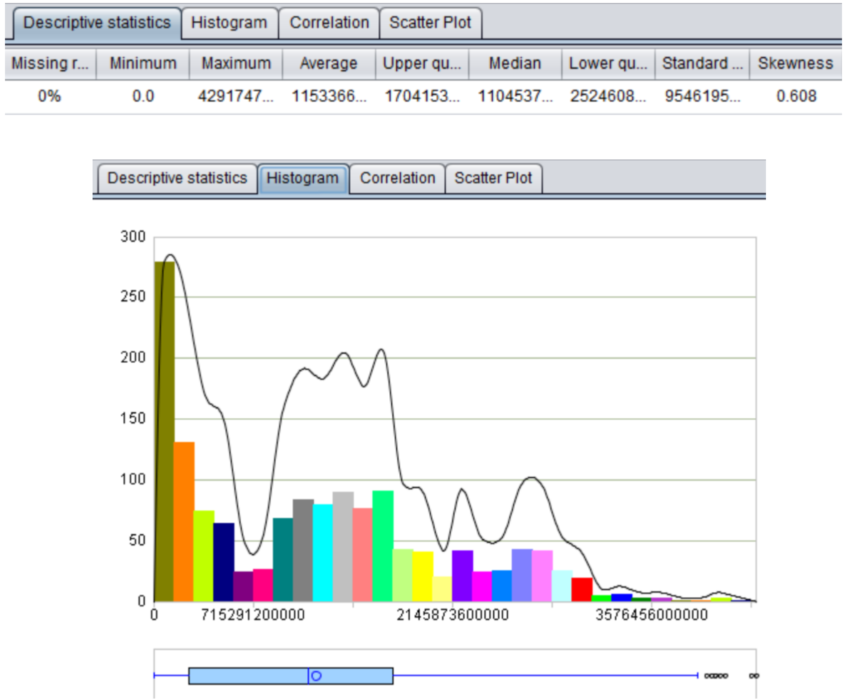
Time and date: YearBuilt

Unit of time interval: Millisecond

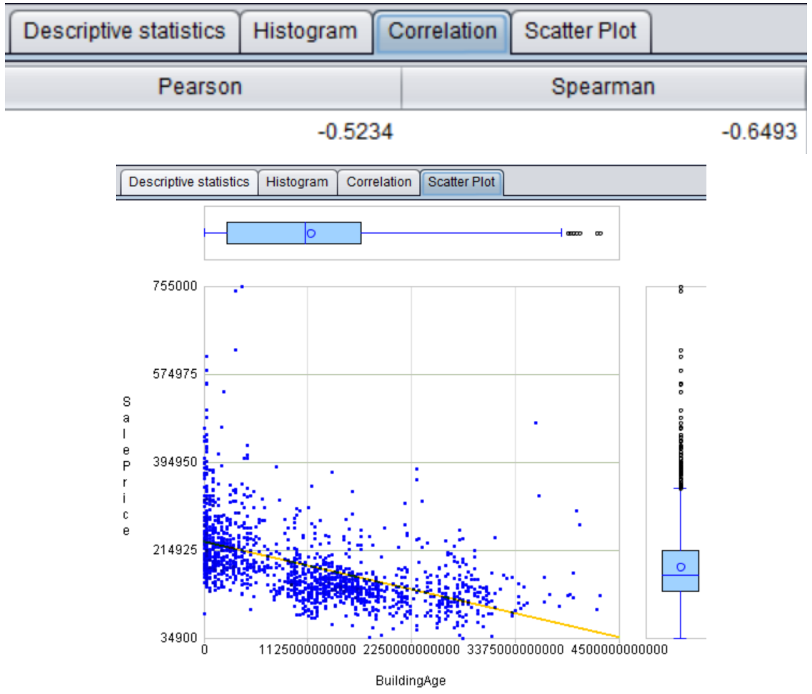
= YrSold - YearBuilt

Variable information: GarageYrBlt

Statistical method	Statistical value
Missing rate	5.548%
Minimum	#.###
Maximum	#.###



Missing r...	Minimum	Maximum	Average	Upper qu...	Median	Lower qu...	Standard ...	Skewness
0%	0.0	4291747...	1153366...	1704153...	1104537...	2524608...	9546195...	0.608



Pearson	Spearman
-0.5234	-0.6493

Observe this derived variable, the minimum value is 0, the maximum value is 136.09, the correlation coefficient is less than 0, but the absolute value is greater than 0.5, which indicates that it is negatively correlated with the house price. The scatter chart shows the same rule, so this variable should be relatively important, but it has the same role as Yearbuilt. Here it's just to introduce the method, and it is not necessary to do this step in the real preprocessing.

Data preprocessing - remove redundant variables

After preprocessing, we remove the following redundant variables:

1. LotArea：already generated derived variable LotArea_g
2. 2ndFlrSF： already generated derived variable 2ndFlrSF_g
3.GarageYrBlt： already generated derived variable GarageYrBlt_nomissing
4. Alley： the missing rate is too high and has little to do with the house price, so abandon it

If the derived variable is generated by the same method above, the original variable can be removed;
If we are not sure whether the generated derived variables are more conducive to modeling than the original variables and time allows, it is also possible to retain the original variables and hand them to YModel for processing and modeling.

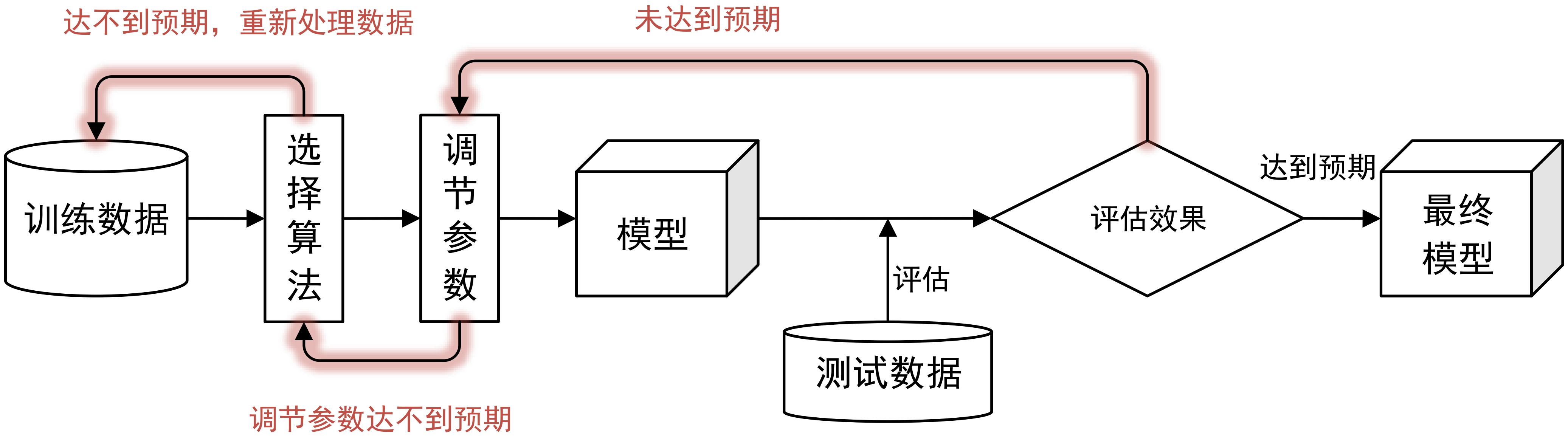
Summary of data exploration and preprocessing

No	Variable	Exploration content	Exploration result	Preprocessing content	Preprocessing result
1	SalePrice	Distribution analysis, statistics analysis	Skew, need to be corrected	Logarithmic correction	no skew
2	GrLivArea	Distribution analysis, statistics analysis, correlation analysis	Skew, positive correlation with SalePrice	Logarithmic correction	no skew
3	GarageArea	Distribution analysis, statistics analysis, correlation analysis	Some houses have no garage		
4	LotArea	Distribution analysis, statistics analysis, correlation analysis	Severe skewness	Equifrequency discretization	Generate component variable
5	2ndFlrSF	Distribution analysis, statistics analysis, correlation analysis	Some houses have no second floor	Custom discretization	Generate component variable
6	MSZoning	Distribution analysis, group statistics	Close relationship between zoning and house price		
7	YearBuilt	Distribution analysis, statistics analysis, correlation analysis	positive correlation with SalePrice		
8	OverallQual	Distribution analysis, group statistics	The overall quality of the house is closely related to the house price		
9	GarageX	Distribution analysis, missing value analysis	Missing because there is no garage	Missing value as new category	Derived no missing value variable
10	GarageYrBlt	Distribution analysis, missing value analysis	Missing because there is no garage	Fill with earliest year	Derived no missing value variable
11	LotFrontage	Distribution analysis, missing value analysis	Skew, need to fill in missing values		
12	Alley	Distribution analysis, missing value analysis	Too many missing values and unimportant , discard		
13	GrLivArea/LotArea			Variable interaction	Derived ratio variable
14	YrSold-YearBuilt			Variable interaction	Derived time interval

***Note**: only the yellow and blue parts of the above variable preprocessing need to be done manually. The variable interaction is mainly to introduce the data preprocessing method, and the generated two derived variables are not necessarily effective for house price prediction. If we write our own code to do these preprocessing it will be very troublesome, with modeling tool it is very simple*

Modeling and evaluation

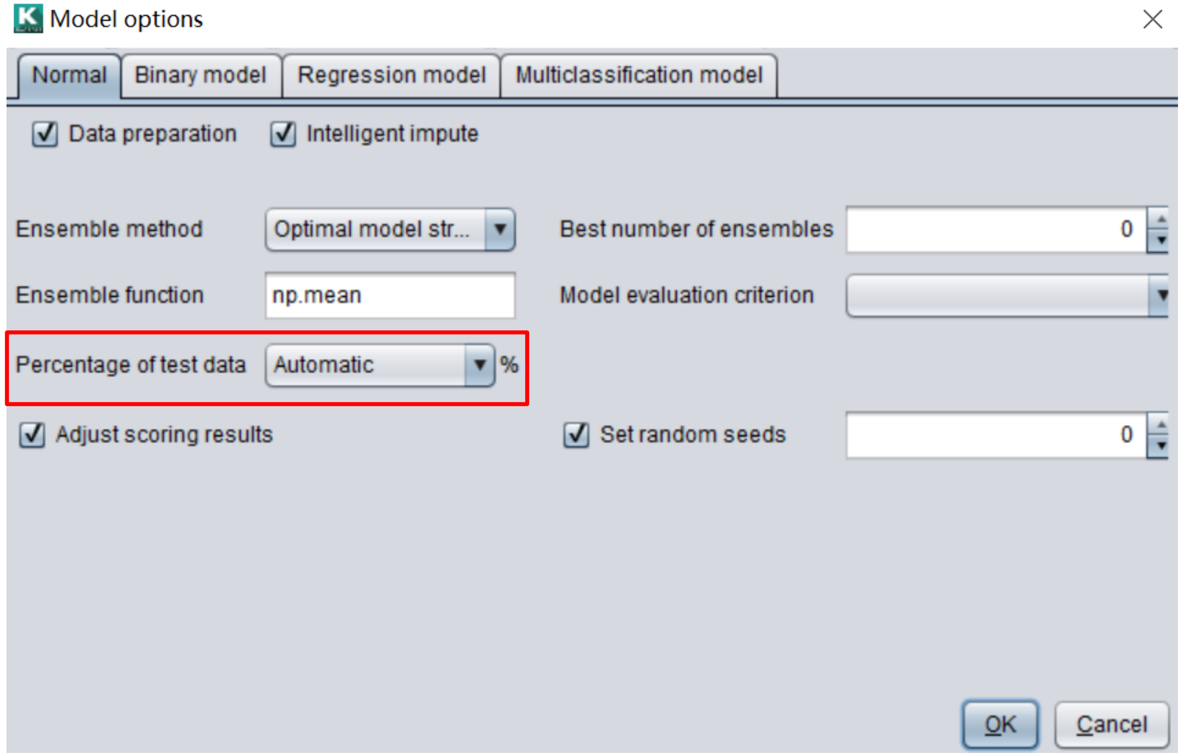
After data preprocessing, modeling can be performed. The modeling process is as follows:



The modeling process is very complex, we give it to YModel, which includes data exploration, data preprocessing, modeling, evaluation and other modules. When modeling, it will automatically cut the data into training set and test set, and model on the training set and evaluate on the test set. It also includes intelligent modeling methods such as parameter adjustment and algorithm fusion. Users can build an ideal model at a low cost.

Modeling and evaluation

Model selection and tuning parameters are too complex. We just need to use the default options, as in the following figure:



☑ Data preparation ☑ Intelligent impute

Ensemble method: Optimal model str...

Best number of ensembles: 0

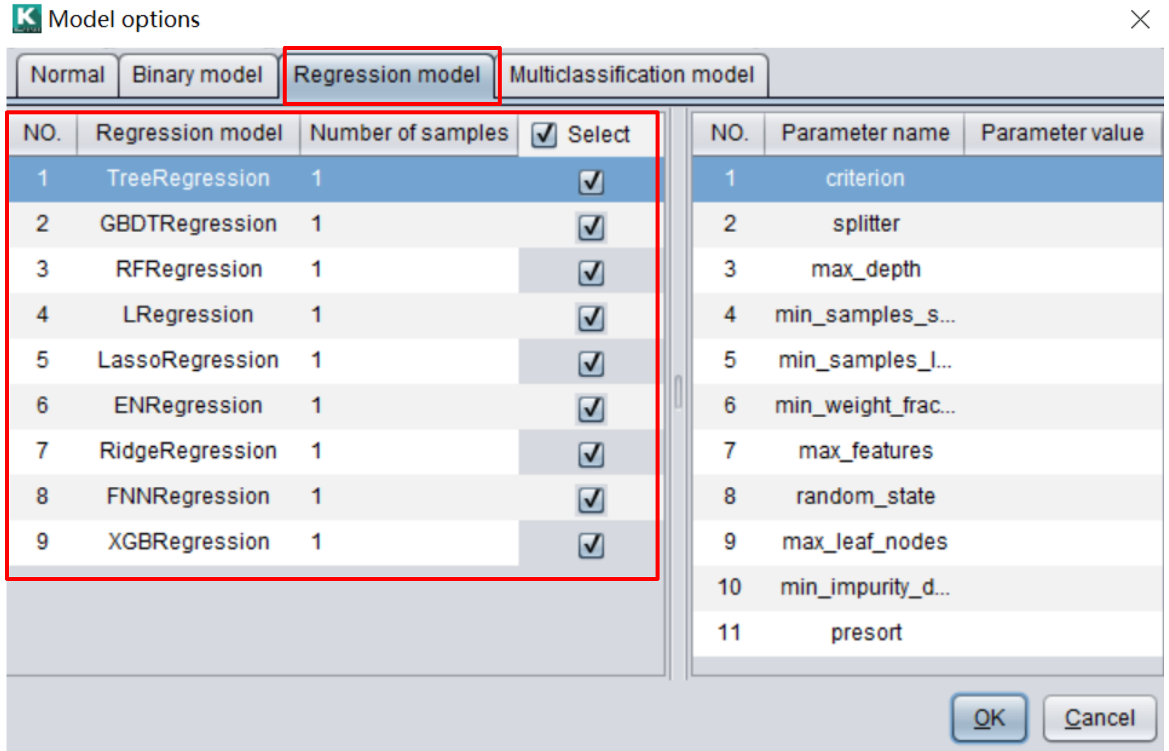
Ensemble function: np.mean

Model evaluation criterion:

Percentage of test data: Automatic %

☑ Adjust scoring results

☑ Set random seeds: 0

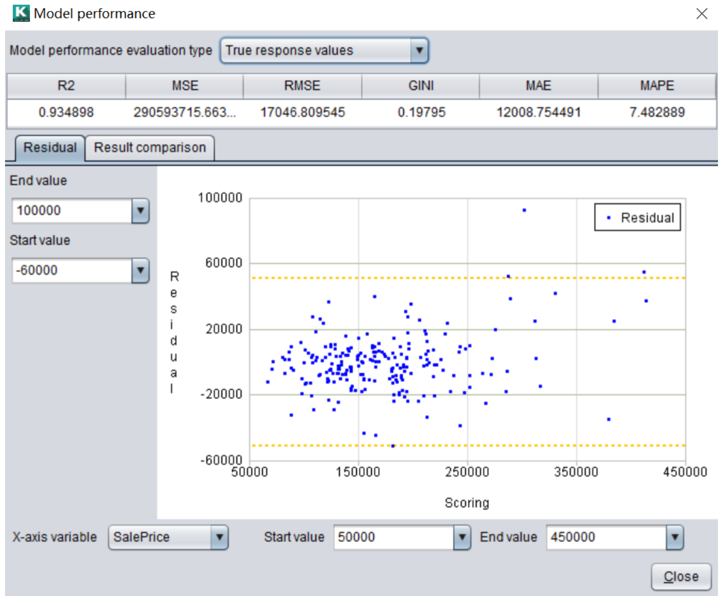


NO.	Regression model	Number of samples	☑ Select
1	TreeRegression	1	☑
2	GBDTRegression	1	☑
3	RFRegression	1	☑
4	LRegression	1	☑
5	LassoRegression	1	☑
6	ENRegression	1	☑
7	RidgeRegression	1	☑
8	FNNRegression	1	☑
9	XGBRegression	1	☑

NO.	Parameter name	Parameter value
1	criterion	
2	splitter	
3	max_depth	
4	min_samples_s...	
5	min_samples_l...	
6	min_weight_frac...	
7	max_features	
8	random_state	
9	max_leaf_nodes	
10	min_impurity_d...	
11	presort	

YModel will help us to build the model quickly, use the appropriate method to avoid over fitting, and calculate the evaluation index on the test set, which is convenient for us to evaluate.

Model performance



Model performance

Model performance evaluation type: True response values

R2	MSE	RMSE	GINI	MAE	MAPE
0.934898	290593715.663...	17046.809545	0.19795	12008.754491	7.482889

End value: 100000

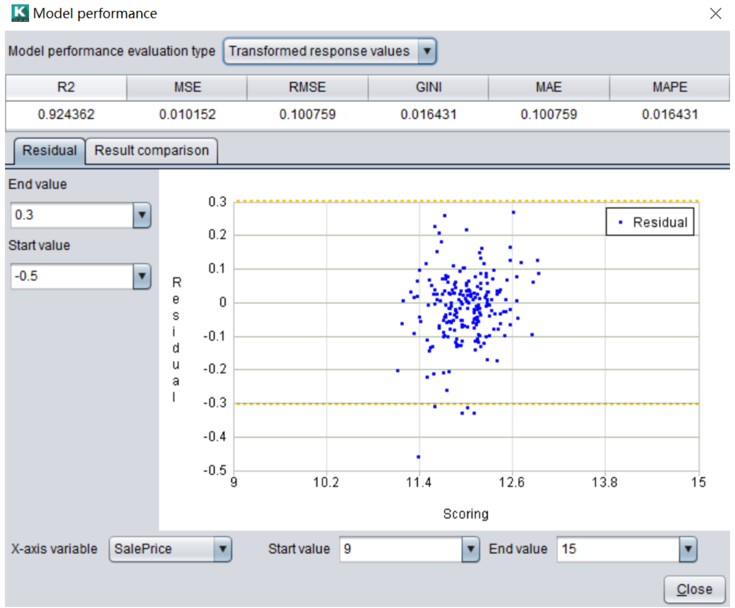
Start value: -60000

X-axis variable: SalePrice

Start value: 50000

End value: 450000

True value performance



Model performance

Model performance evaluation type: Transformed response values

R2	MSE	RMSE	GINI	MAE	MAPE
0.924362	0.010152	0.100759	0.016431	0.100759	0.016431

End value: 0.3

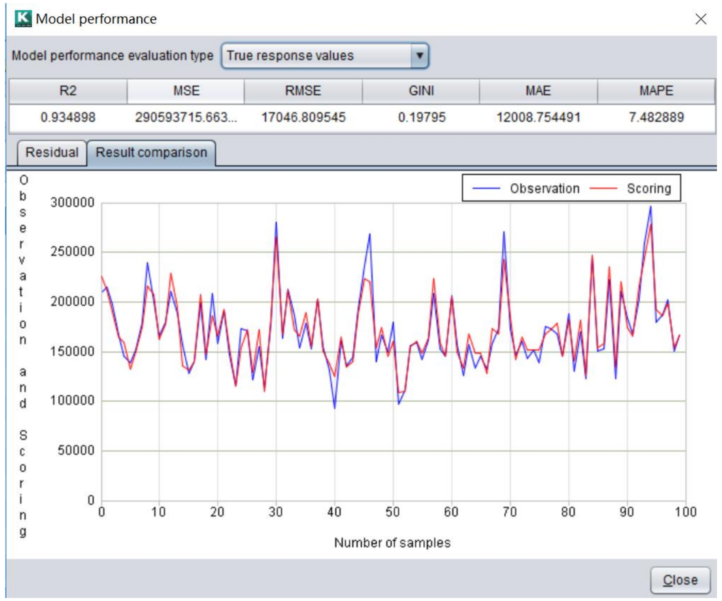
Start value: -0.5

X-axis variable: SalePrice

Start value: 9

End value: 15

Transformed value performance



Model performance

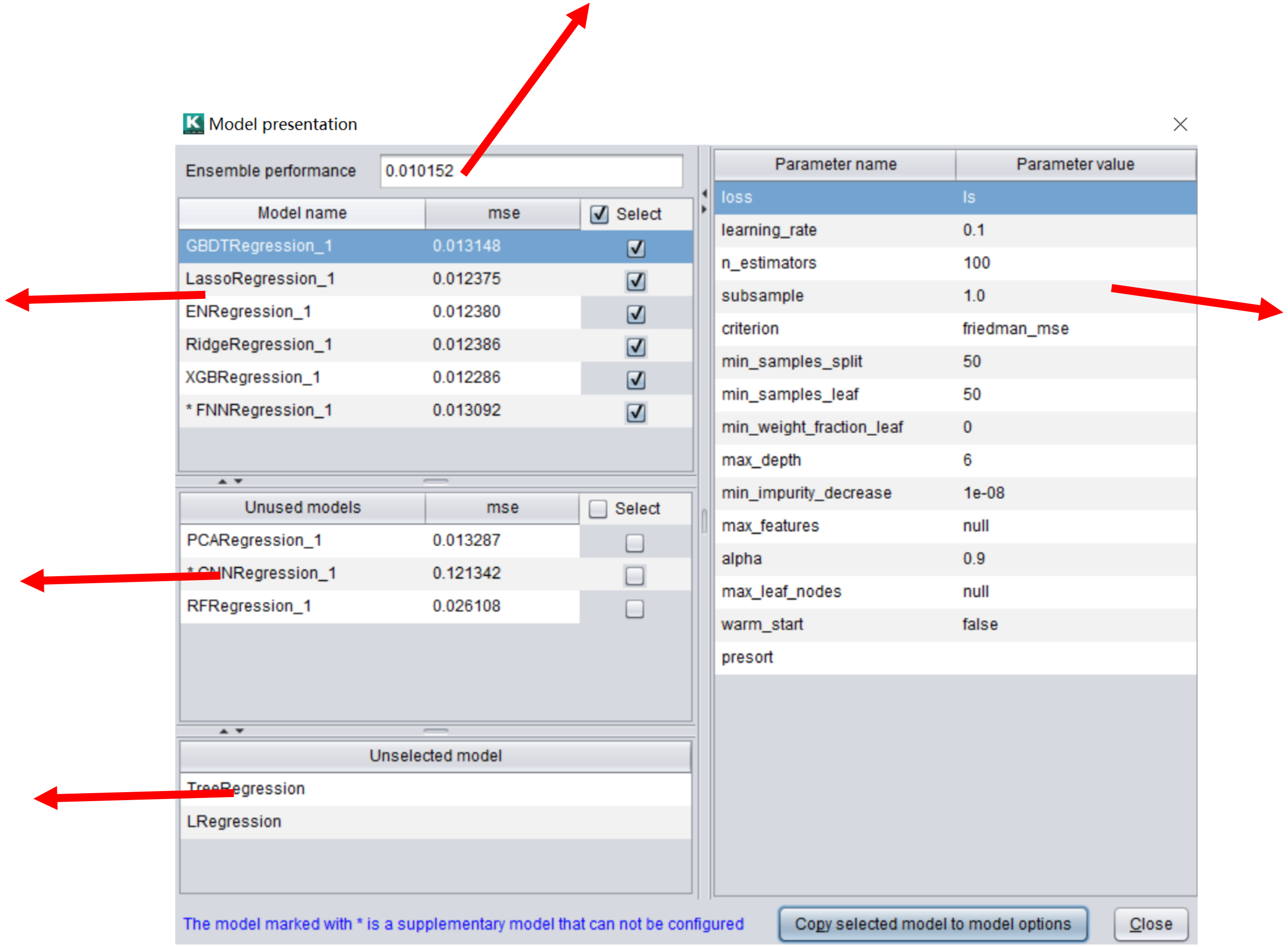
Model performance evaluation type: True response values

R2	MSE	RMSE	GINI	MAE	MAPE
0.934898	290593715.663...	17046.809545	0.19795	12008.754491	7.482889

Results comparison chart

From the perspective of evaluation indexes, the performance of the model is very good. The R^2 of the true value and transformed value is greater than 0.9, which means that the model explains more than 90% of the uncertainty. RMSE of the true value is 17046.81, which means that the average difference between the predicted house price and the real house price is 17046.81 , then look at the RMSE = 0.1, which means that the difference between the transformed house price and the predicted result of the model is 1%. Then look at the residual chart, the residual chart of the true value is basically randomly distributed around 0, especially if you are picky, there is a slight trend that the higher the house price, the greater the deviation, but the residual chart after the transform is basically randomly distributed around 0. By looking at the result comparison chart, we can also feel that the prediction results are relatively accurate and can be used.

Which models are used for modeling:



Transformed MSE

Used model

Unused model

Unselected model

Model parameter

Model presentation

Ensemble performance: 0.010152

Model name	mse	☑ Select
GBDTRegression_1	0.013148	☑
LassoRegression_1	0.012375	☑
ENRegression_1	0.012380	☑
RidgeRegression_1	0.012386	☑
XGBRegression_1	0.012286	☑
* FNNRegression_1	0.013092	☑

Unused models	mse	☐ Select
PCARegression_1	0.013287	☐
* CNNRegression_1	0.121342	☐
RFRegression_1	0.026108	☐

Unselected model
TreeRegression
LRegression

Parameter name	Parameter value
loss	ls
learning_rate	0.1
n_estimators	100
subsample	1.0
criterion	friedman_mse
min_samples_split	50
min_samples_leaf	50
min_weight_fraction_leaf	0
max_depth	6
min_impurity_decrease	1e-08
max_features	null
alpha	0.9
max_leaf_nodes	null
warm_start	false
presort	

The model marked with * is a supplementary model that can not be configured

In this modeling, six regression models are selected and used to get the optimal combination model. Model parameters are automatically selected by YModel. Data mining experts can select "copy selected model to model option" to modify parameters and re model.

Variable importance

The results of variable importance in descending order：

GirLivArea、 MSZoning 、OverallCond、YearBuilt……

Among them, the most important derived variable is

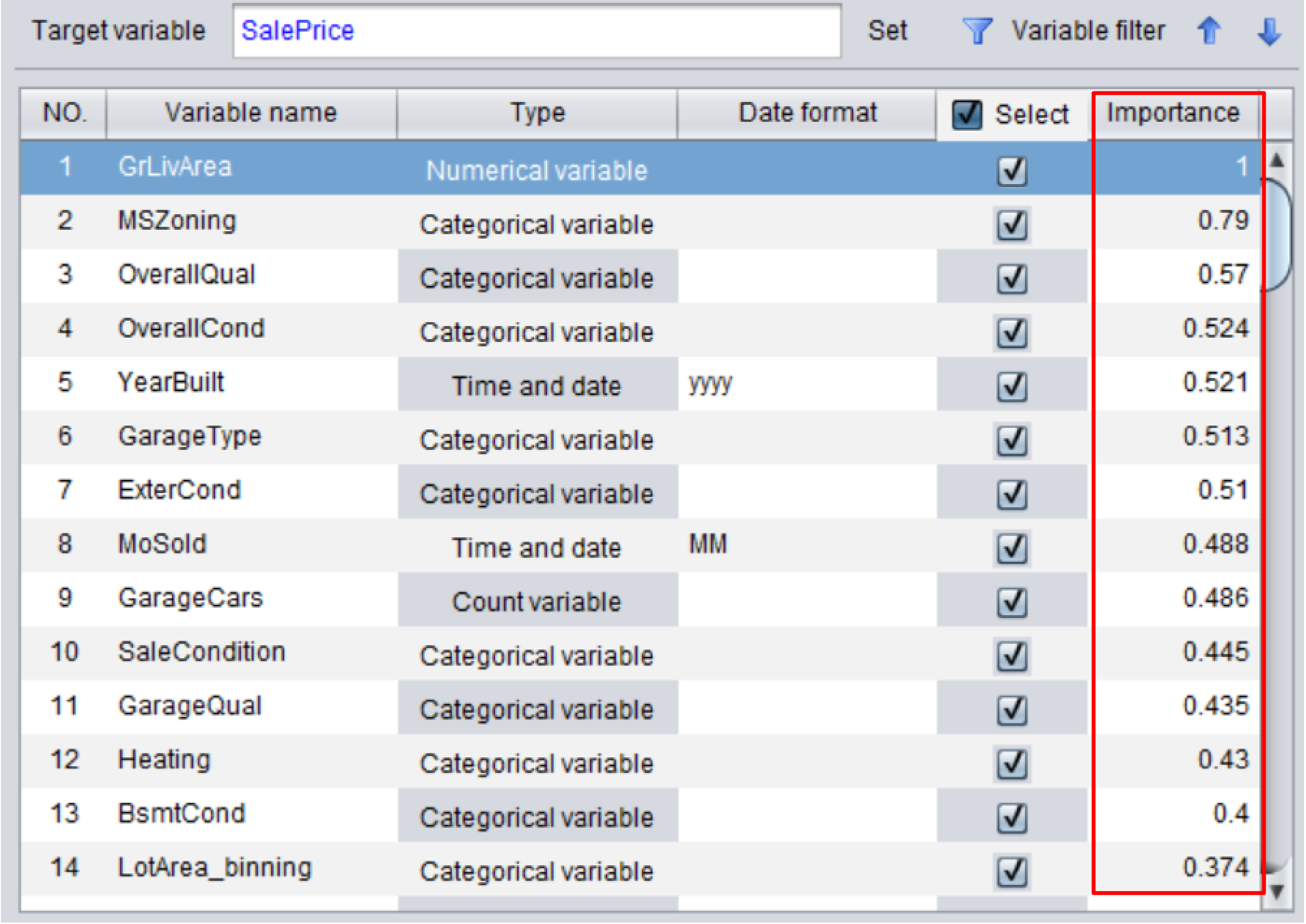
LotArea_binning, 14th，followed by GarageYrBlt_nomissing,

The importance of BuildingAge、2ndFlrSF_g、GrLivRatio are

all low. Here we mainly introduce the methods of

exploration and preprocessing. In this data, these derived

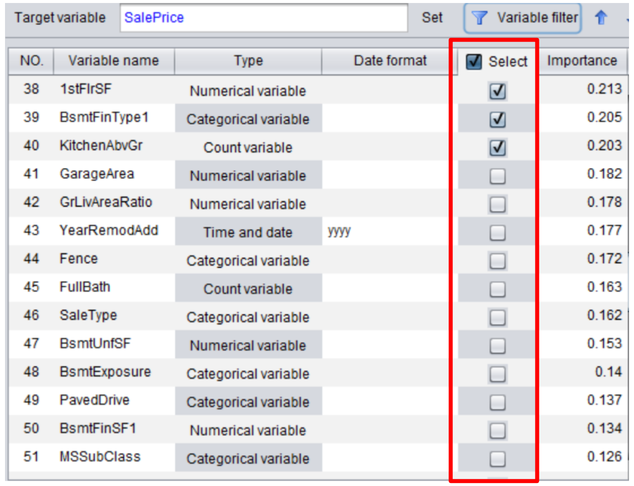
variables are unnecessary.



Target variable: SalePrice

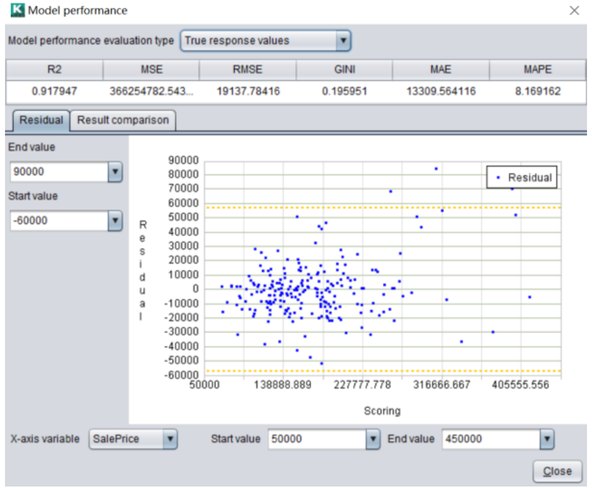
NO.	Variable name	Type	Date format	☑ Select	Importance
1	GrLivArea	Numerical variable		☑	1
2	MSZoning	Categorical variable		☑	0.79
3	OverallQual	Categorical variable		☑	0.57
4	OverallCond	Categorical variable		☑	0.524
5	YearBuilt	Time and date	yyyy	☑	0.521
6	GarageType	Categorical variable		☑	0.513
7	ExterCond	Categorical variable		☑	0.51
8	MoSold	Time and date	MM	☑	0.488
9	GarageCars	Count variable		☑	0.486
10	SaleCondition	Categorical variable		☑	0.445
11	GarageQual	Categorical variable		☑	0.435
12	Heating	Categorical variable		☑	0.43
13	BsmtCond	Categorical variable		☑	0.4
14	LotArea_binning	Categorical variable		☑	0.374

We can use the high importance derived variables interaction to add new derived variables to improve the performance of the model. Interested students can try it on their own. Here we remove the low importance variables and build a model again to see the effect.



Target variable: SalePrice

NO.	Variable name	Type	Date format	☑ Select	Importance
38	1stFlrSF	Numerical variable		☑	0.213
39	BsmtFinType1	Categorical variable		☑	0.205
40	KitchenAbvGr	Count variable		☑	0.203
41	GarageArea	Numerical variable		☐	0.182
42	GrLivAreaRatio	Numerical variable		☐	0.178
43	YearRemodAdd	Time and date	yyyy	☐	0.177
44	Fence	Categorical variable		☐	0.172
45	FullBath	Count variable		☐	0.163
46	SaleType	Categorical variable		☐	0.162
47	BsmtUnfSF	Numerical variable		☐	0.153
48	BsmtExposure	Categorical variable		☐	0.14
49	PavedDrive	Categorical variable		☐	0.137
50	BsmtFinSF1	Numerical variable		☐	0.134
51	MSSubClass	Categorical variable		☐	0.126



Model performance evaluation type: True response values

R2	MSE	RMSE	GINI	MAE	MAPE
0.917947	366254782.543...	19137.78416	0.195951	13309.564116	8.169162

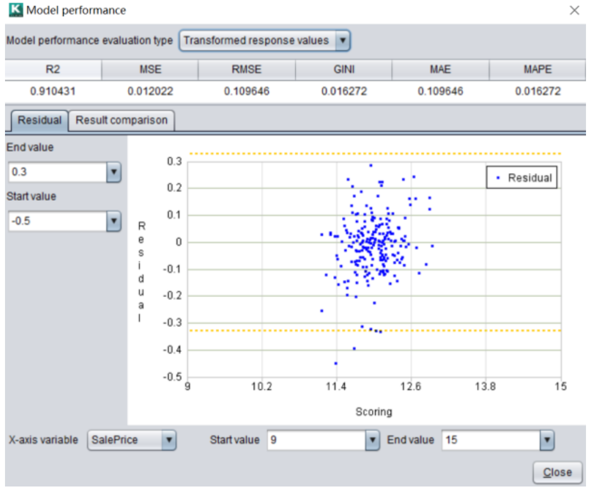
End value: 90000

Start value: -60000

X-axis variable: SalePrice

Start value: 50000

End value: 450000



Model performance evaluation type: Transformed response values

R2	MSE	RMSE	GINI	MAE	MAPE
0.910431	0.012022	0.109646	0.016272	0.109646	0.016272

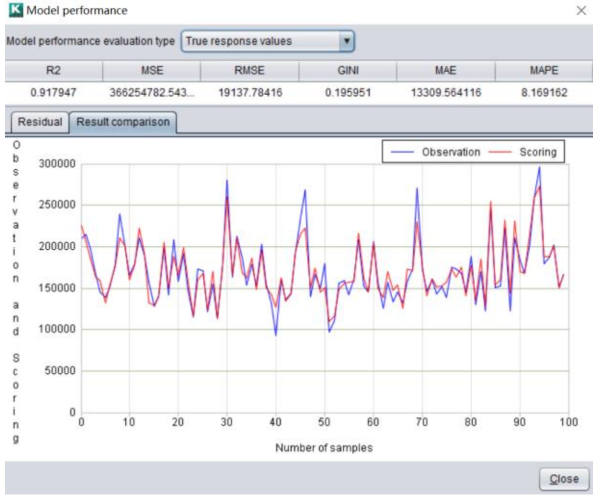
End value: 0.3

Start value: -0.5

X-axis variable: SalePrice

Start value: 9

End value: 15



Model performance evaluation type: True response values

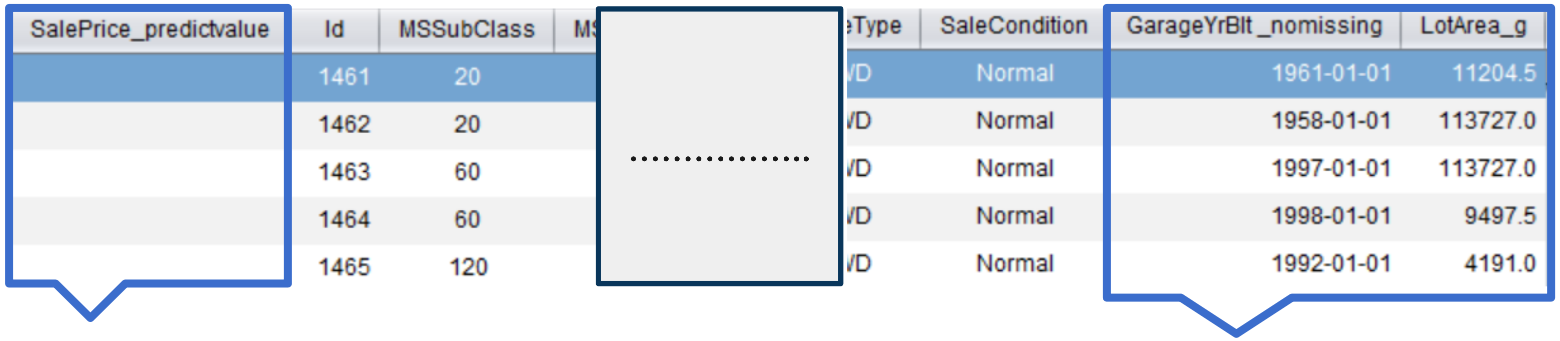
R2	MSE	RMSE	GINI	MAE	MAPE
0.917947	366254782.543...	19137.78416	0.195951	13309.564116	8.169162

Here we remove the variables whose importance is less than 0.2. From the evaluation index to the residual chart to the result comparison chart, the model has not much difference. Therefore, we can selectively remove some non important variables when modeling, which will not affect the model effect, but the modeling efficiency will be much higher. (this suggestion is not suitable for competition and can be used in actual production.)

Model application

After the model is built, it is used to predict the test set.

The YModel tool will automatically process the set to be tested and generate derived variables without manual processing, as shown in the following figure:



SalePrice_predictvalue	Id	MSSubClass	MS	eType	SaleCondition	GarageYrBlt _nomissing	LotArea_g
	1461	20		D	Normal	1961-01-01	11204.5
	1462	20		D	Normal	1958-01-01	113727.0
	1463	60		D	Normal	1997-01-01	113727.0
	1464	60		D	Normal	1998-01-01	9497.5
	1465	120		D	Normal	1992-01-01	4191.0

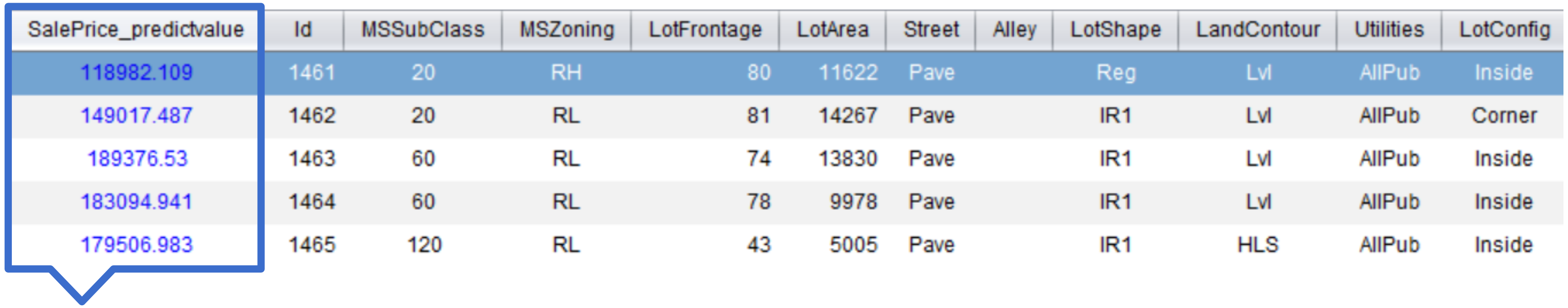
....

Prediction result column

Derived variables

Because we only use two derived variables with high importance in the end, only two variables are derived from the test set.

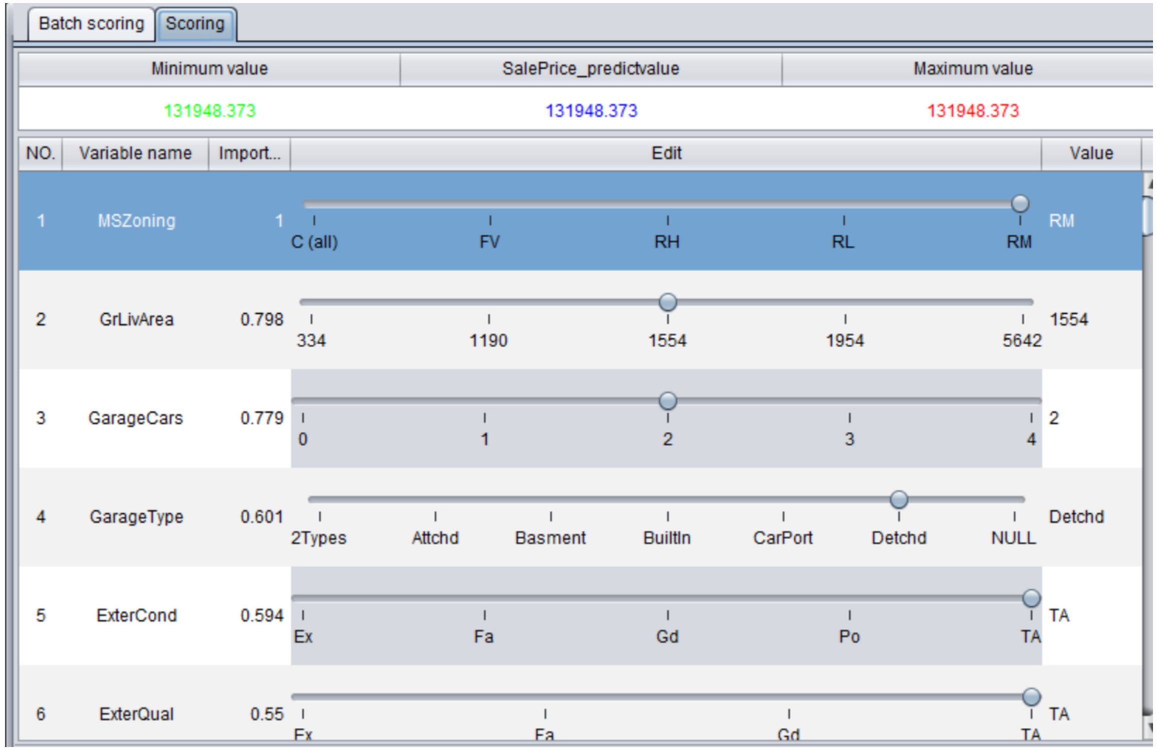
Model prediction results



SalePrice_predictvalue	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig
118982.109	1461	20	RH	80	11622	Pave		Reg	Lvl	AllPub	Inside
149017.487	1462	20	RL	81	14267	Pave		IR1	Lvl	AllPub	Corner
189376.53	1463	60	RL	74	13830	Pave		IR1	Lvl	AllPub	Inside
183094.941	1464	60	RL	78	9978	Pave		IR1	Lvl	AllPub	Inside
179506.983	1465	120	RL	43	5005	Pave		IR1	HLS	AllPub	Inside

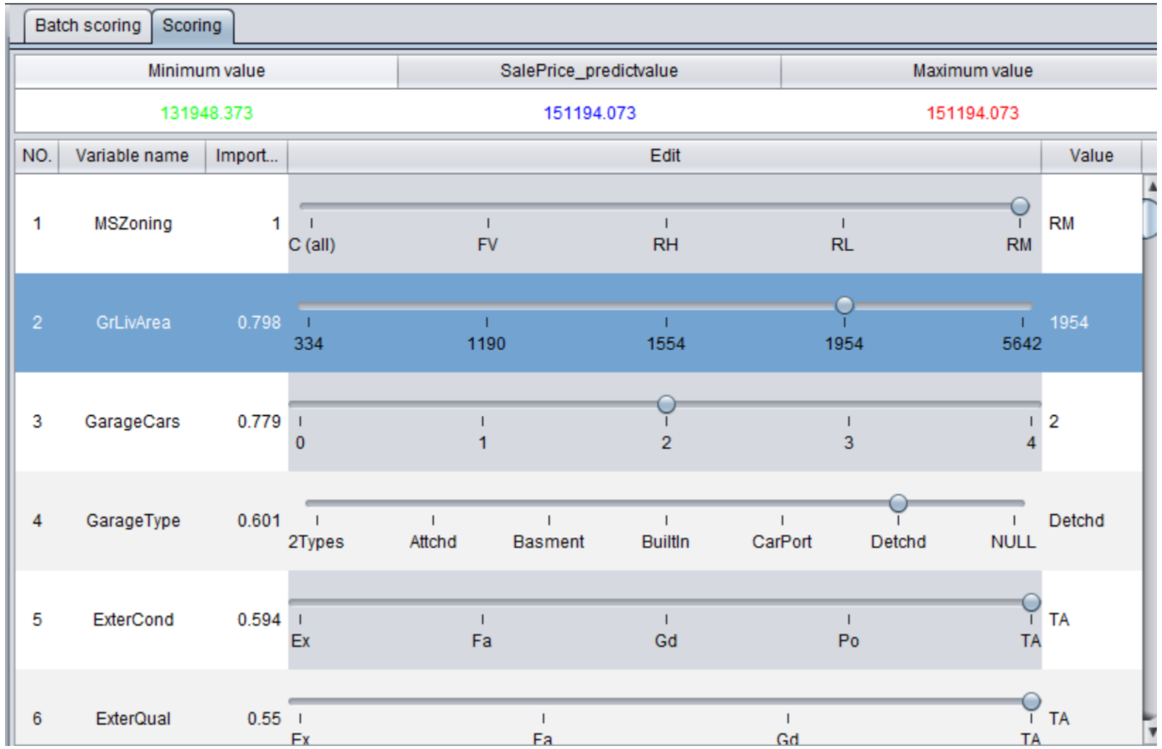
According to the prediction results, the predicted value of the sale price, for example, the first prediction result is 118982.109, which means the price of the house is about 118982.109.

We can also make a single prediction



Minimum value	SalePrice_predictvalue	Maximum value
131948.373	131948.373	131948.373

NO.	Variable name	Import...	Edit	Value
1	MSZoning	1	C (all) FV RH RL RM	RM
2	GrLivArea	0.798	334 1190 1554 1954 5642	1554
3	GarageCars	0.779	0 1 2 3 4	2
4	GarageType	0.601	2Types Attchd Basment BuiltIn CarPort Detchd NULL	Detchd
5	ExterCond	0.594	Ex Fa Gd Po TA	TA
6	ExterQual	0.55	Ex Fa Gd TA	TA



Minimum value	SalePrice_predictvalue	Maximum value
131948.373	151194.073	151194.073

NO.	Variable name	Import...	Edit	Value
1	MSZoning	1	C (all) FV RH RL RM	RM
2	GrLivArea	0.798	334 1190 1554 1954 5642	1954
3	GarageCars	0.779	0 1 2 3 4	2
4	GarageType	0.601	2Types Attchd Basment BuiltIn CarPort Detchd NULL	Detchd
5	ExterCond	0.594	Ex Fa Gd Po TA	TA
6	ExterQual	0.55	Ex Fa Gd TA	TA

If someone wants to buy a house and other variables remain unchanged, when the living area GrLivArea is 1554, the predicted value is 131948.373; when the living area GrLivArea is 1954, the predicted value is 151194.073. According to this, it can give the customer reference and facilitate the customer decision-making.

THANKS