Data Mining

Table of Contents

Pre class preparation

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

Pre class preparation (exercise data and tool download address)

Exercise data

Titanic.csv

Houseprice.csv

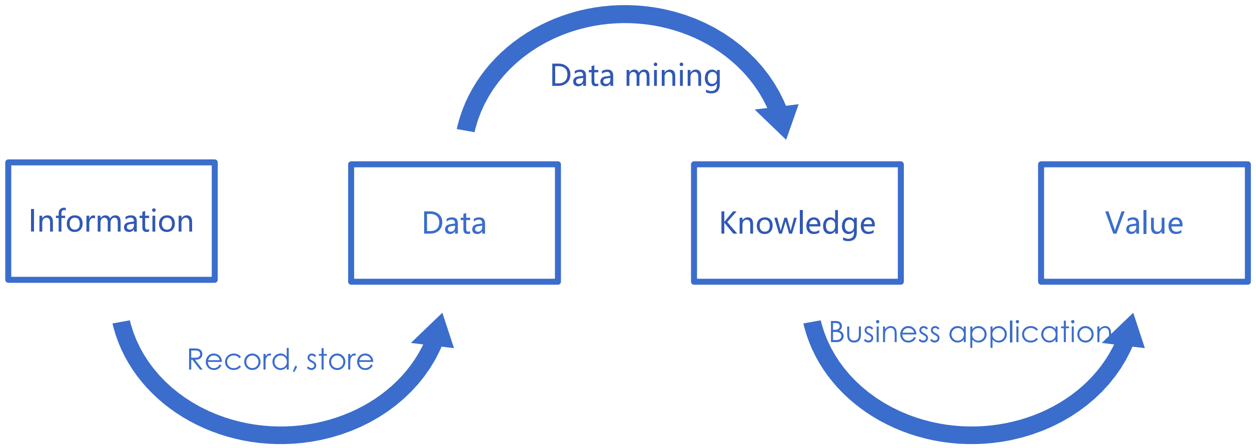
Download of YModel tool

http://c.raqsoft.com/tag/Download?t=%E6%98%93%E6%98%8E%E6%99%BA%E8%83%BD%E5%BB%BA%E6%A8%A1

Chapter 1 The concept of data mining

The concept of data mining

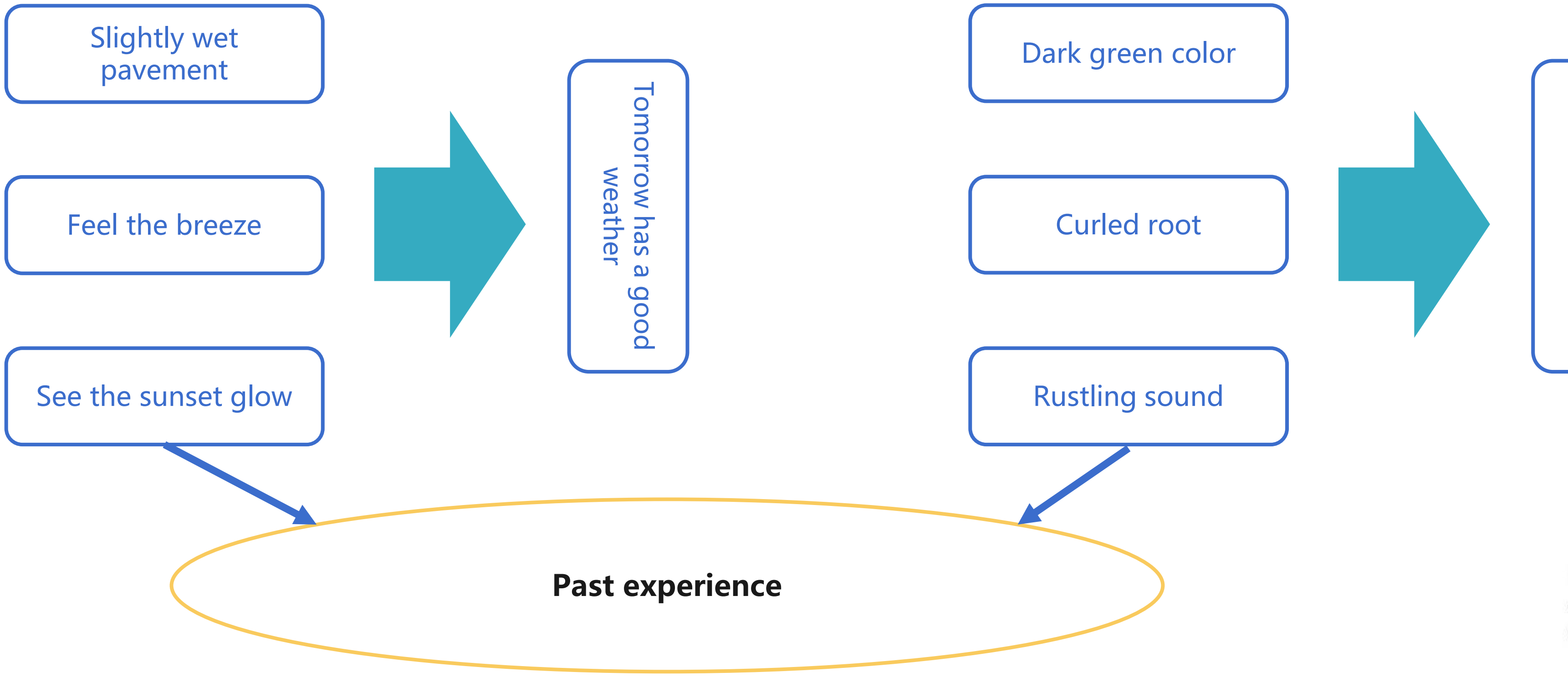
Data mining is a process of extracting hidden, unknown and potentially useful information and knowledge from a large number of incomplete, noisy, fuzzy and random practical application data.



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

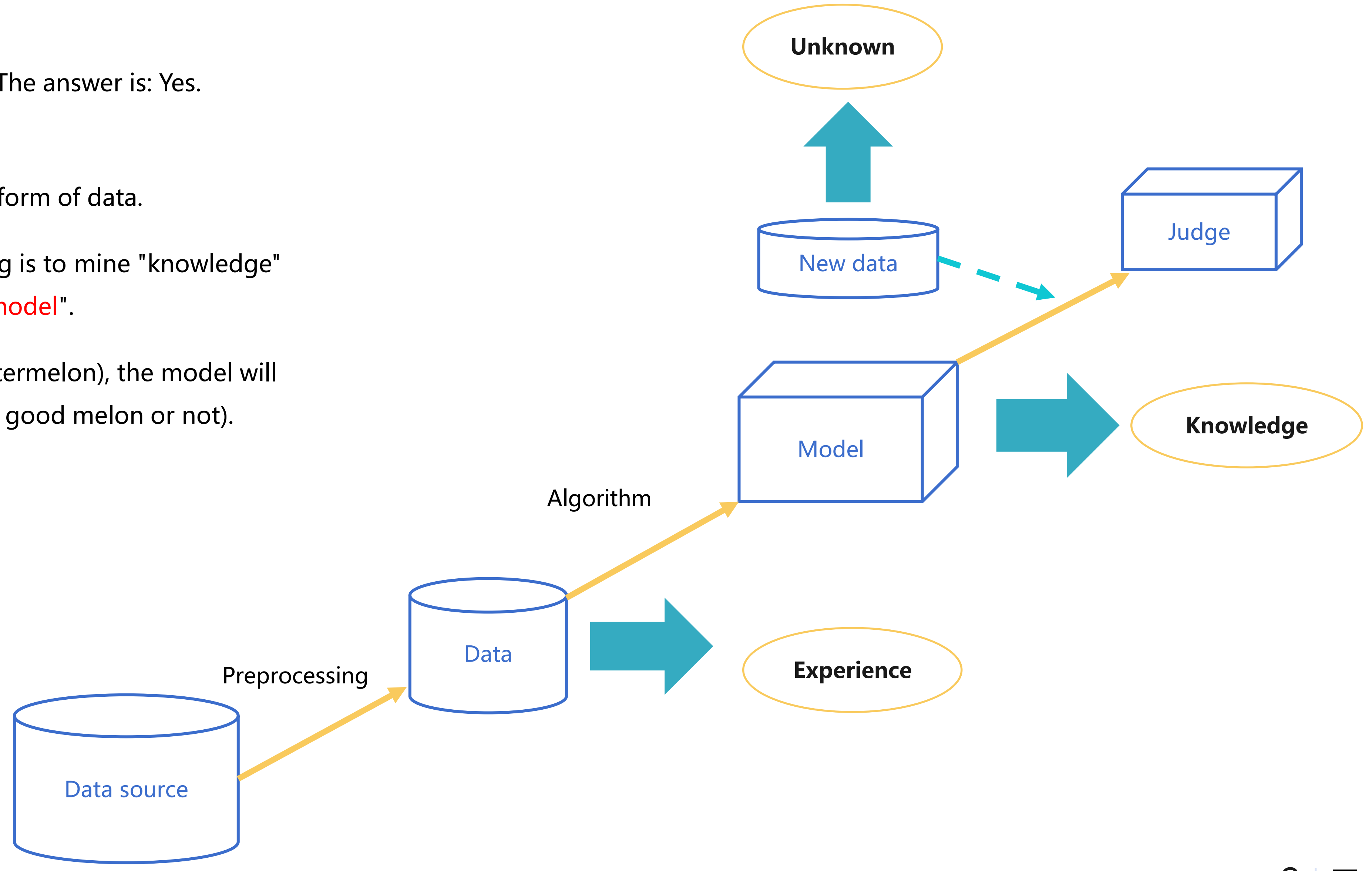
In the evening, the road surface of the street is wet after a little rain, and the gentle breeze blows. Look up at the sunset glow in the sky. Well, tomorrow has fine weather. Go to the fruit stand, pick up a dark green watermelon with curled root and rustling sound, and look forward to enjoying it.



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.

variable

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	…	…	…	…

Variable value

Label

Dataset

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 built, 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.

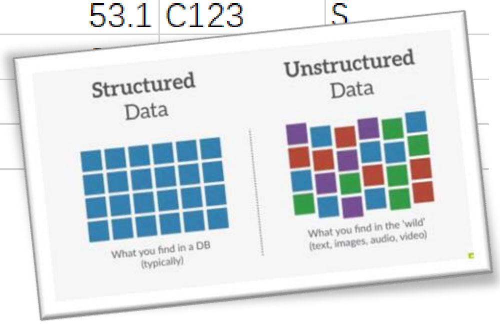
The concept of data mining

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			
6	0	3	Moran, Mr. James	male		0	0	330877			
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463			



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

Chapter 2 Data exploration

2.1 The significance of data exploration

View features

Using tools to view the characteristics of data

Perceive value

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

Understand data

Understand the statistical characteristics of variables and the correlation between variables

2.2 Data type identification

	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， ...]

2.2 Data type identification with YModel

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

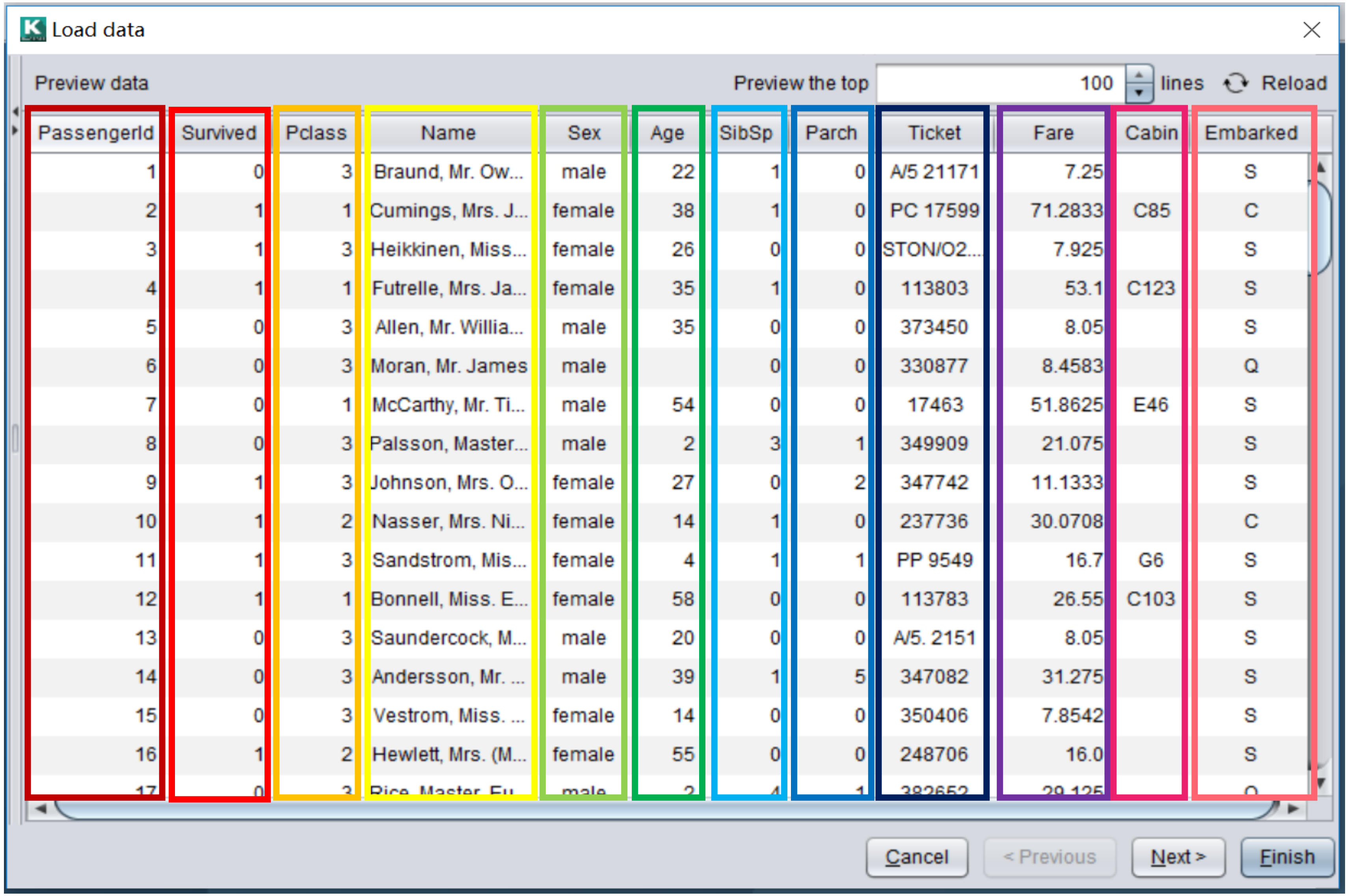
Data：Titanic.csv

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

2.2 Data type identification with YModel

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

1. Data preview



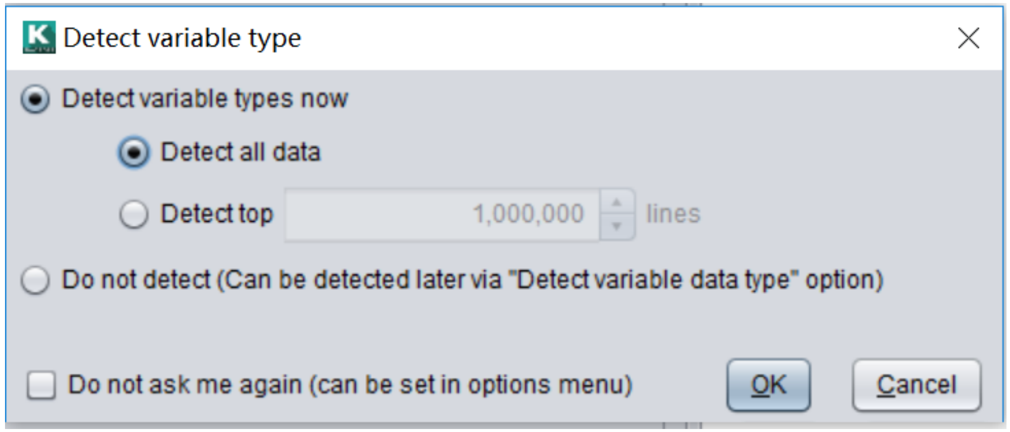
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Ow...	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. J...	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss...	female	26	0	0	STON/O2...	7.925		S
4	1	1	Futrelle, Mrs. Ja...	female	35	1	0	113803	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

2.2 Data type identification with YModel

Target variable: Survived | Set | Variable filter

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



2. Select detection data range

3. Automatic identification of data types by YModel tool

Target variable: Survived | Set | Variable filter

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		☑

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.

2.2 Data type identification with YModel

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

2.3 Quantitative data exploration with YModel

Data：Titanic.csv, row represents passenger sample, column represents different information of each passenger

There is a quantitative variable(column) named **“Fare”** in the Titanic dada, use YModel to explore “Fare”

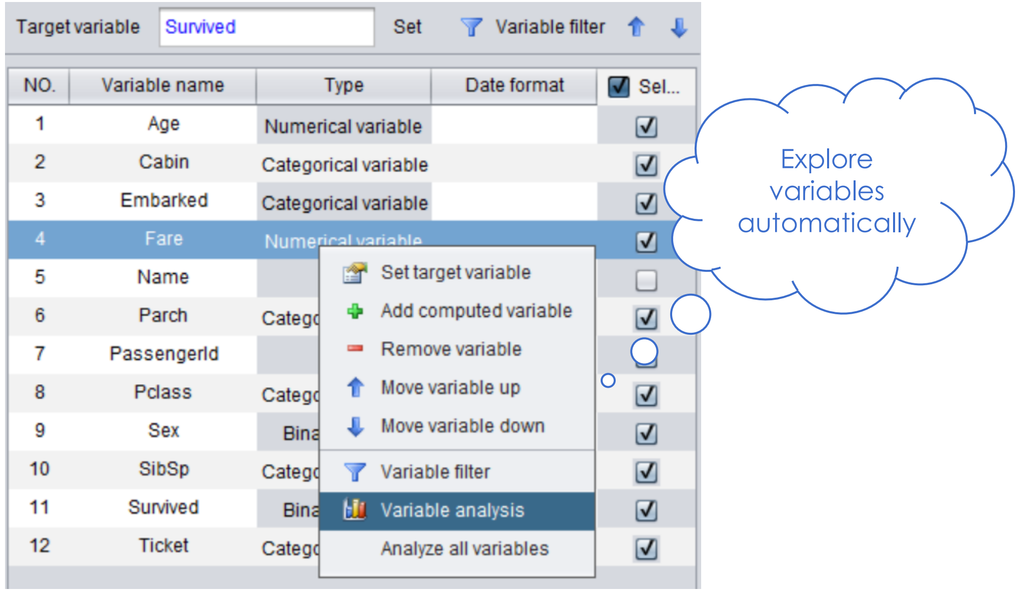
PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q
7	0	1	male	54	0	0	17463	51.8625	E46	S
8	0	3	male	2	3	1	349909	21.075		S

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

Analyze data distribution?

2.3 Quantitative data exploration with YModel

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.

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

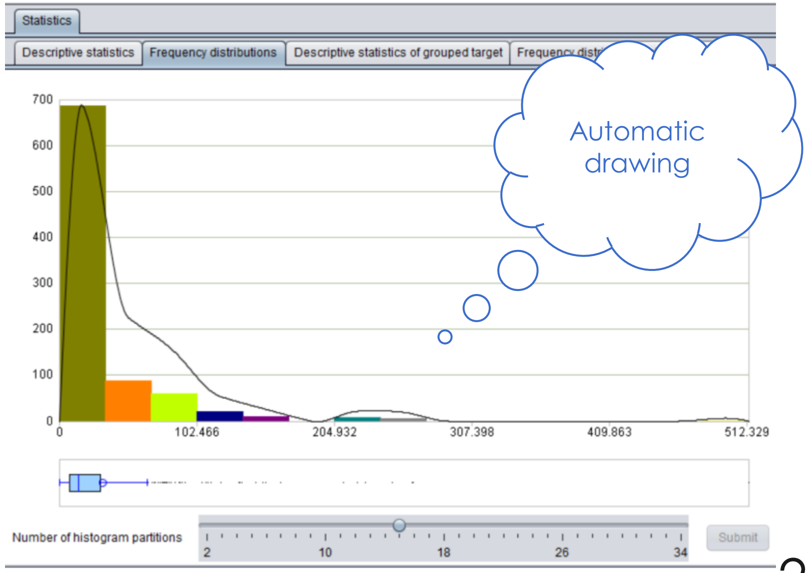
It can be seen that the maximum is 512, minimum is 0, the average value is greater than the median, and the data skewness is large, indicating that the rich are in the minority.

calculate various statistical values automatically

Statistics

Descriptive statistics | Frequency distributions | Descriptive statistics of grouped target | Frequency distributions of grouped target

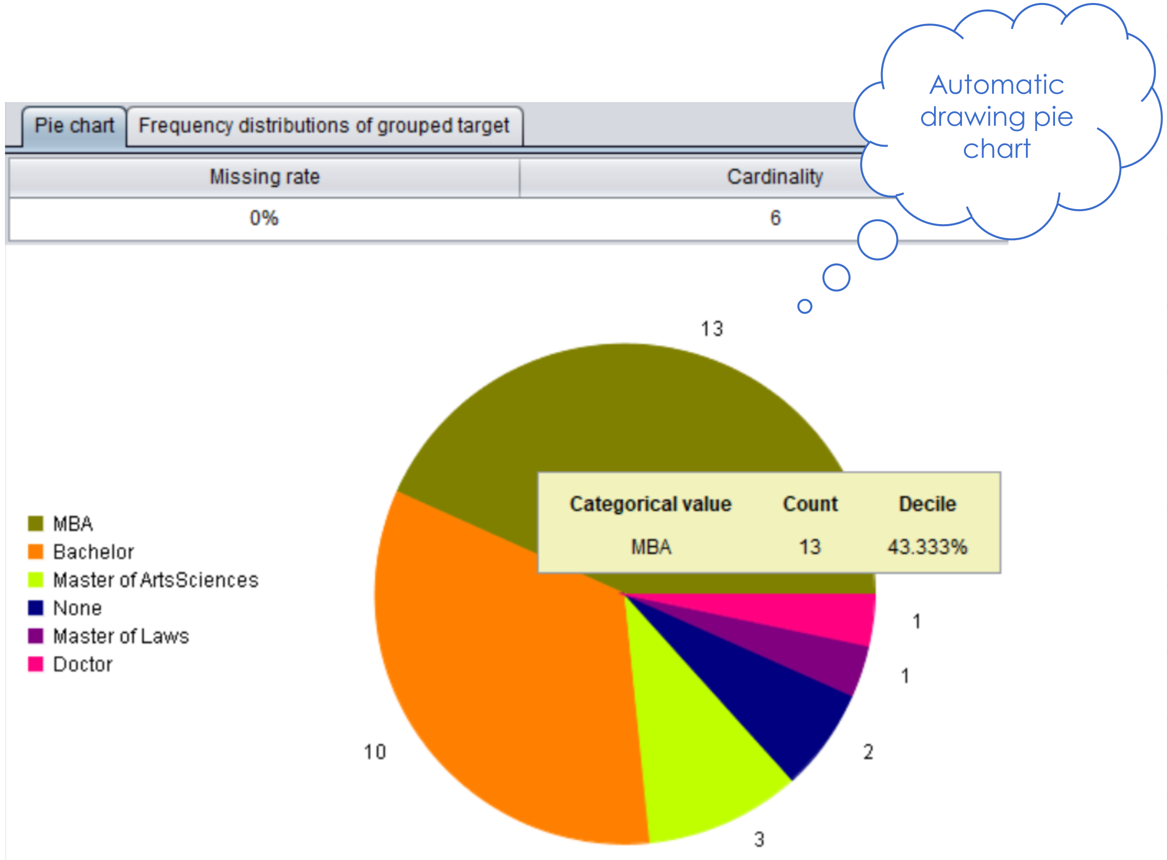
Missing rate	Minimum	Maximum	Average	Upper quartile	Median	Lower quartile	Standard deviati...	Skewness
0%	0.0	512.329	32.204	31.0	14.454	7.896	49.693	4.779



2.4 Qualitative data exploration with YModel

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

Please use YModel to explore the variable.



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

2.5 Correlation analysis with YModel - continuous variables

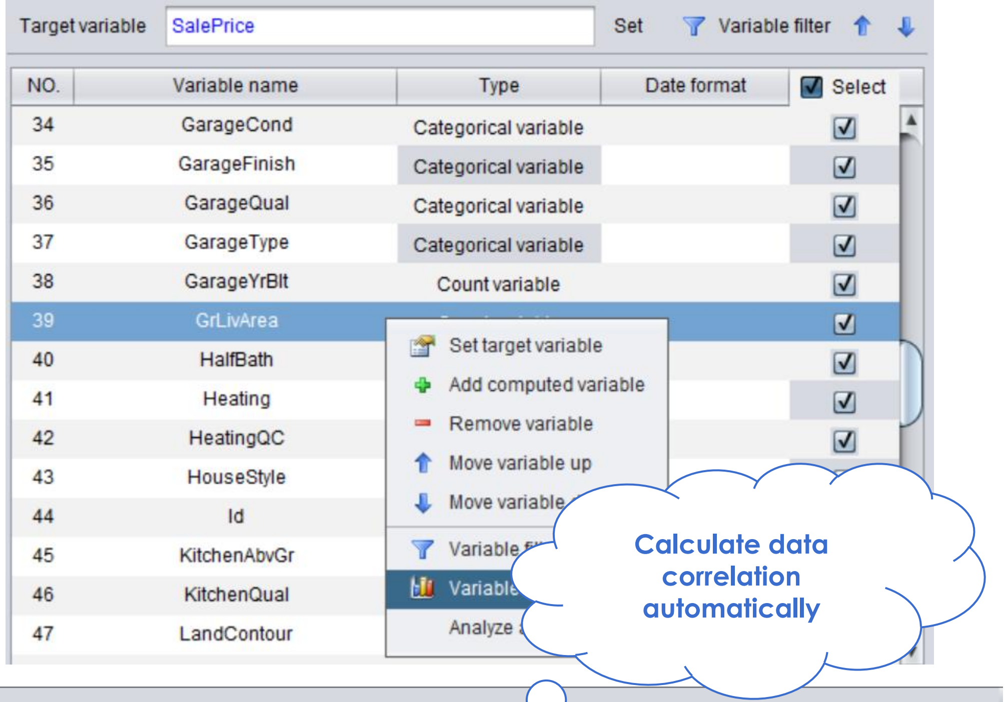
For example, in the case of housing price prediction on kaggle, practice using tools to analyze Pearson and Spearman correlation coefficients of GrLivArea” and " SalePrice ".

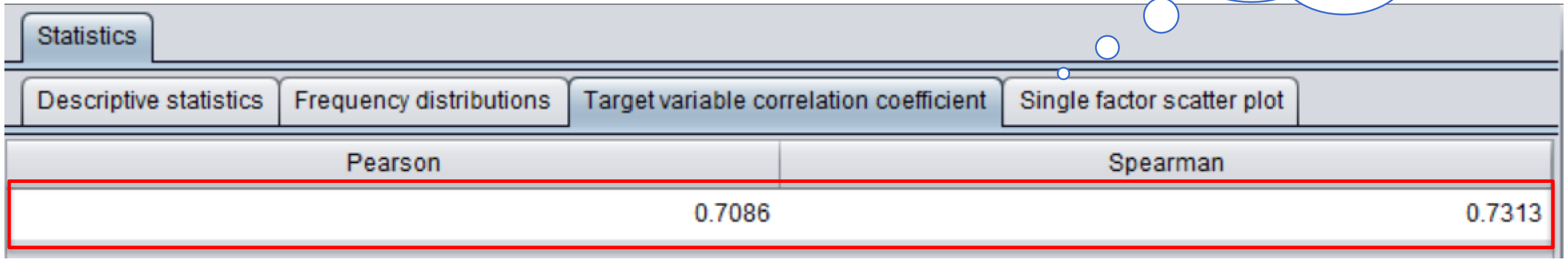
Id	GrLivArea	SalePrice
1	1710	208500
2	1262	181500
3	1786	223500
4	1717	140000
5	2198	250000
6	1362	143000
7	1694	307000
8	2090	200000
9	1774	129900
10	1077	118000

2.5 Correlation analysis with YModel - continuous variables

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

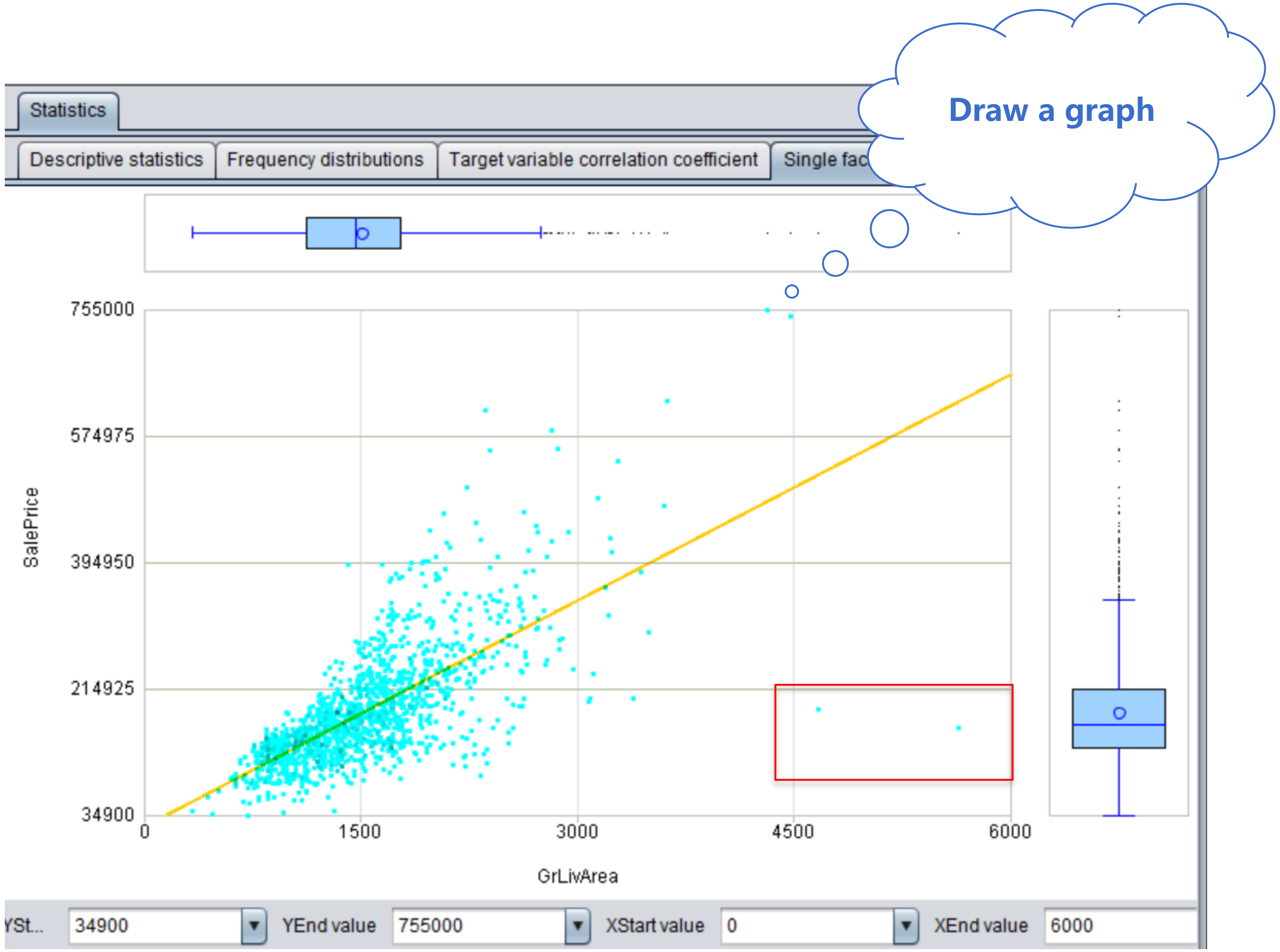
For example, in the case of house price prediction on kaggle, use YModel to view the correlation between " GrLivArea " residential area and "SalePrice".





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

2.5 Correlation analysis with YModel - continuous variables



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.

2.5 Correlation analysis with YModel - 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

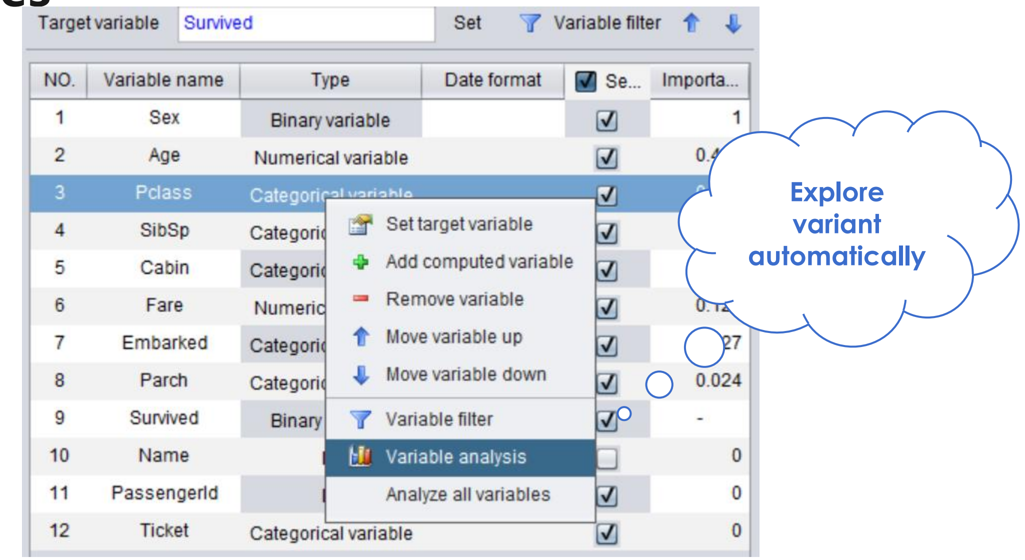
Explore the "Pclass" variable in Titanic data

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?

2.5 Correlation analysis with YModel - categorical variables

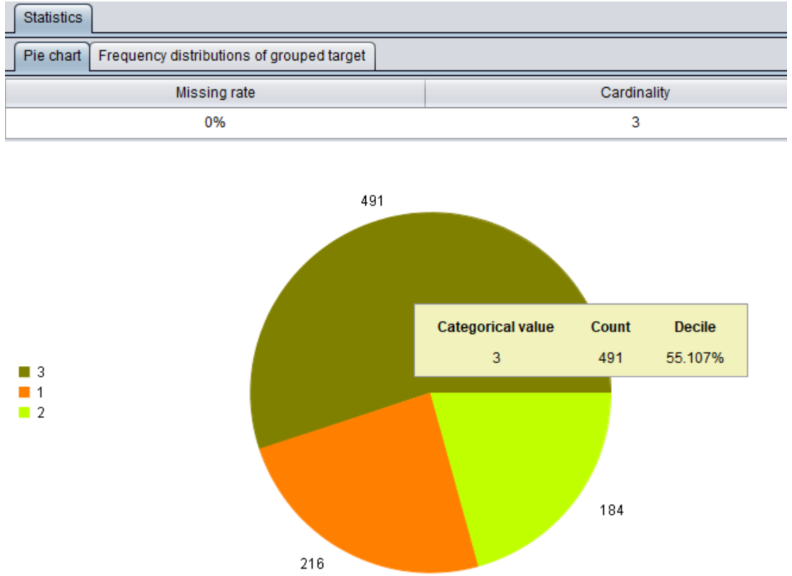


Survived	Pclass	Name	Sex
0	3	Braund, Mr. Owen Harris	male
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female
1	3	Heikkinen, Miss. Laina	female
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
0	3	Allen, Mr. William Henry	male

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.



Statistics

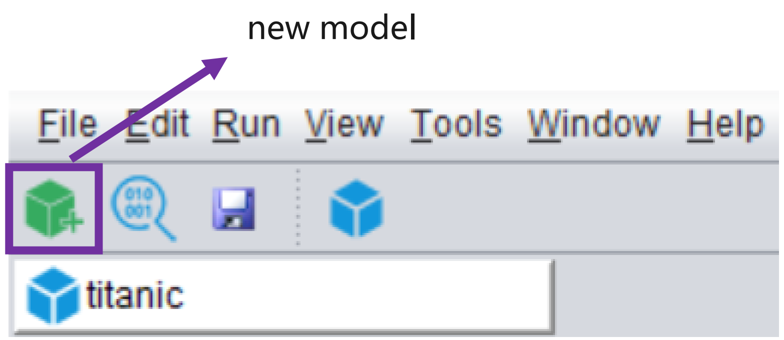
Pie chart | Frequency distributions of grouped target

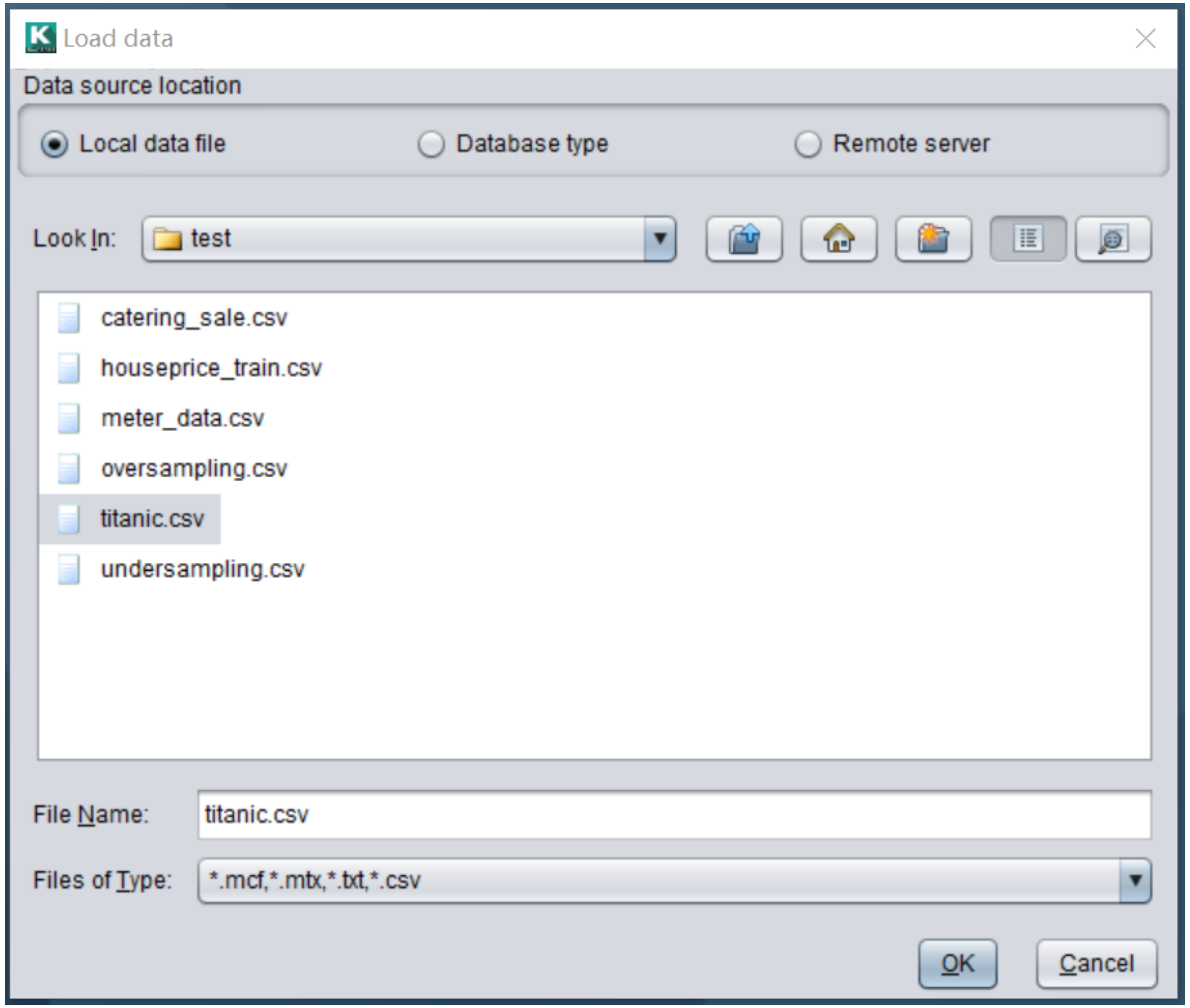
Categorical variable	Sample size	Positive cases size	Positive cases rate
3	491	119	24.236%
2	184	87	47.283%
1	216	136	62.963%

Chapter 3 Data pre-processing and modeling

Modeling with YModel- Classification Model

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

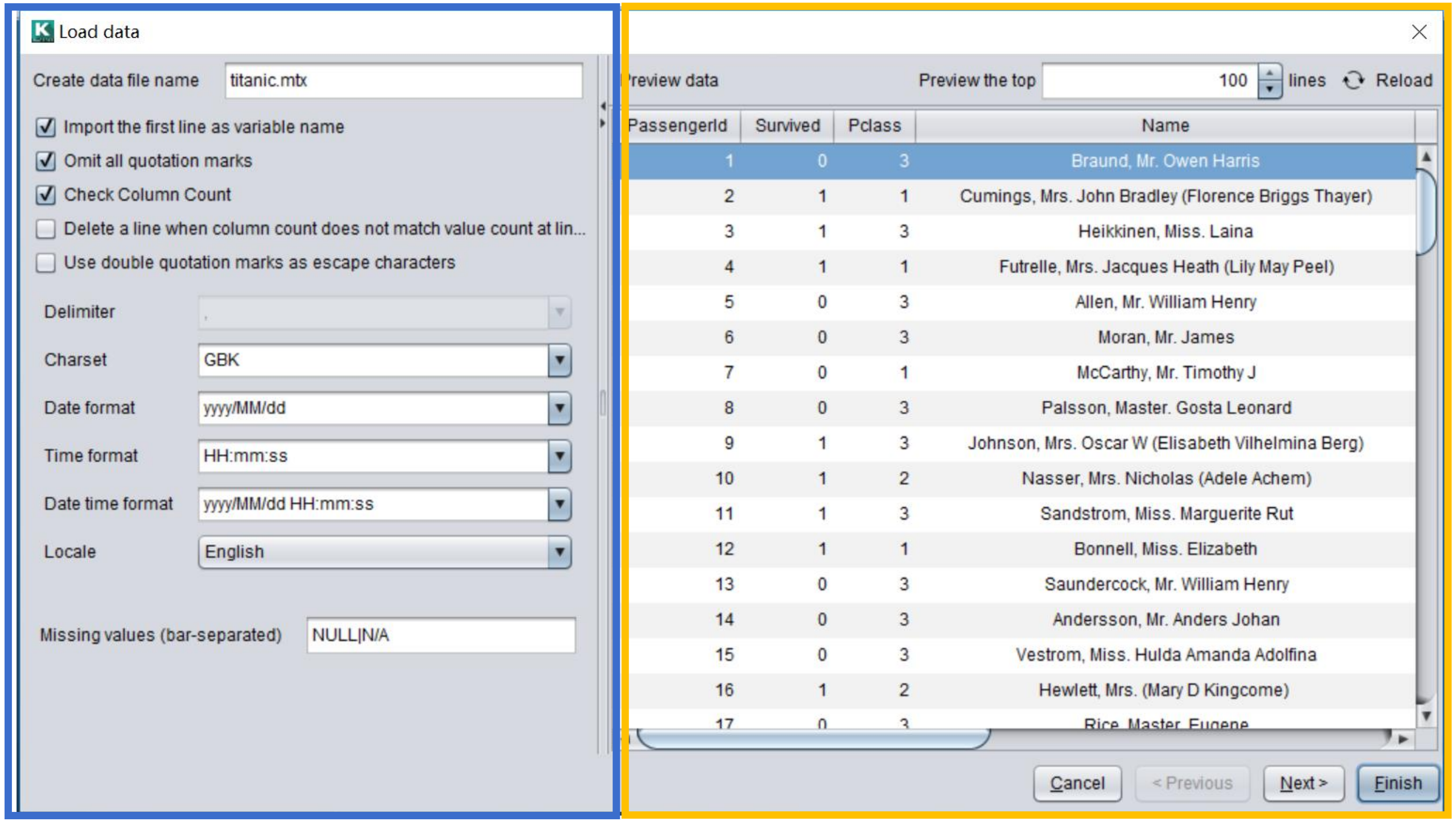




Modeling with YModel- Classification Model

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

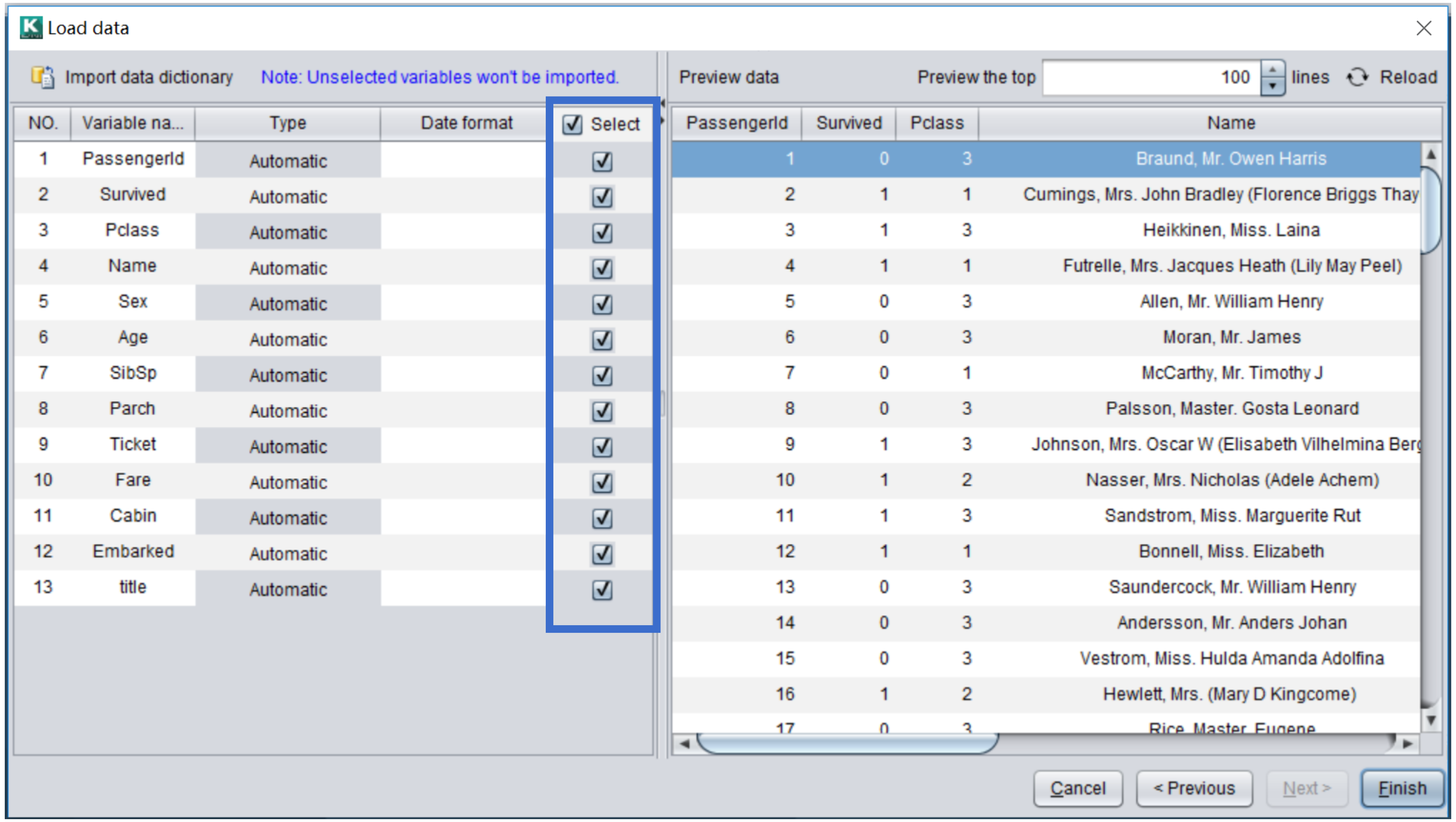
The left side of the page is configured with character set format, date time format and missing value format for automatic recognition by software.



Modeling with YModel- Classification Model

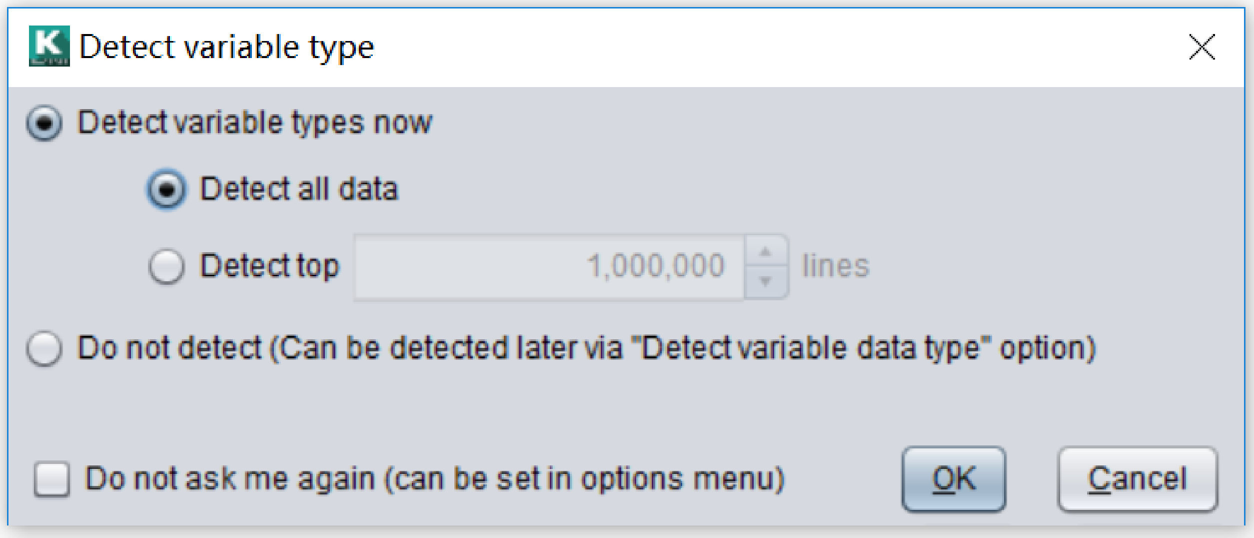
Select the variables involved in the modeling and click Finish.

Here we choose all variables.



Modeling with YModel- Classification Model

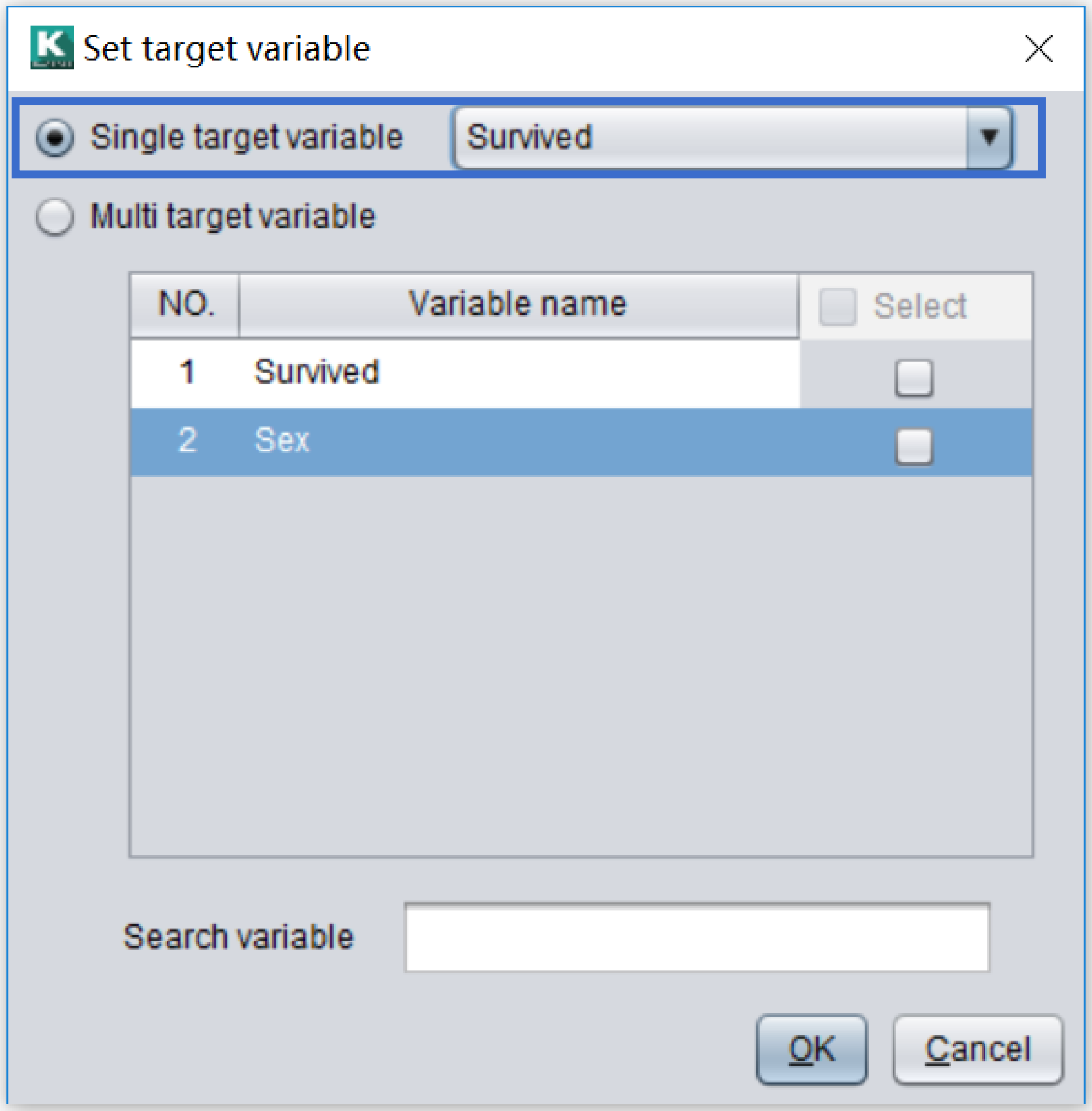
Select the amount of data to be detected. When the amount of data is small, all can be detected. When the amount of data is large, some can be detected, such as 50000 pieces, to improve efficiency.
Here we check all.



Modeling with YModel- 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.

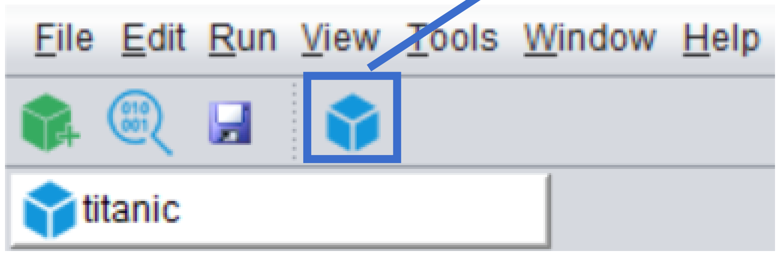


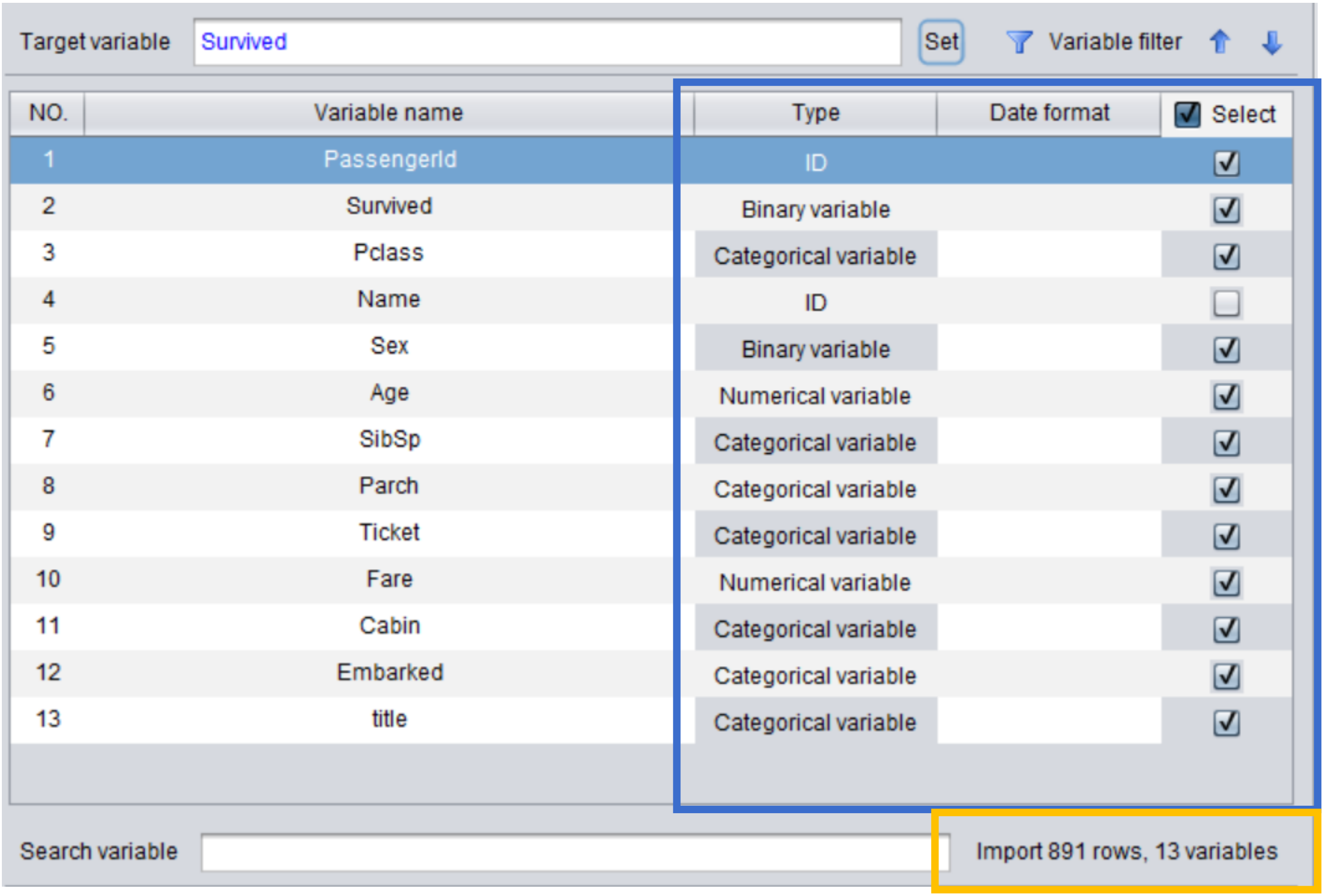
Modeling with YModel- Classification Model

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.

modeling





Target variable: Survived | Set | Variable filter

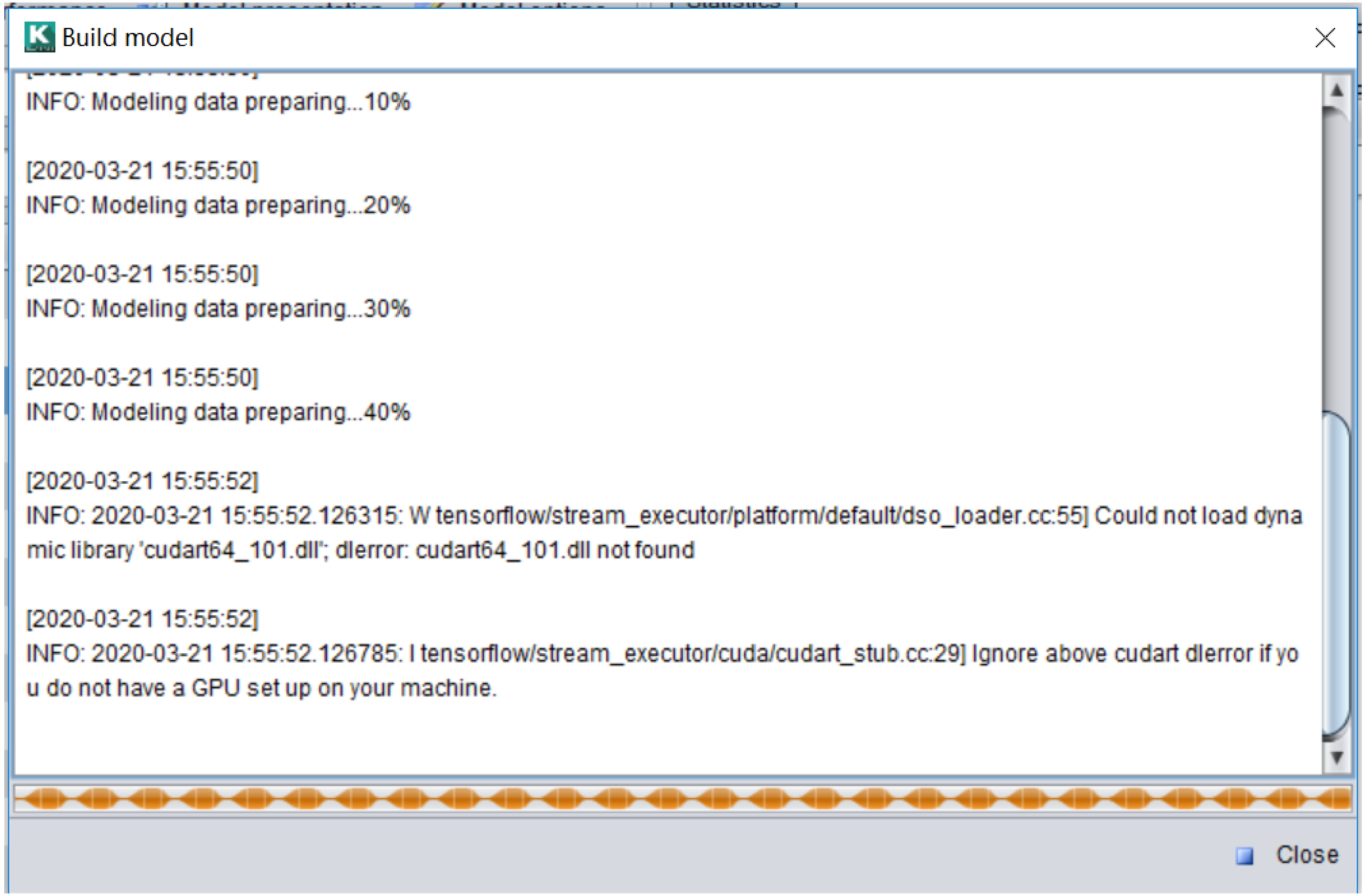
NO.	Variable name	Type	Date format	Select
1	PassengerId	ID		☑
2	Survived	Binary variable		☑
3	Pclass	Categorical variable		☑
4	Name	ID		☐
5	Sex	Binary variable		☑
6	Age	Numerical variable		☑
7	SibSp	Categorical variable		☑
8	Parch	Categorical variable		☑
9	Ticket	Categorical variable		☑
10	Fare	Numerical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

Search variable

Import 891 rows, 13 variables

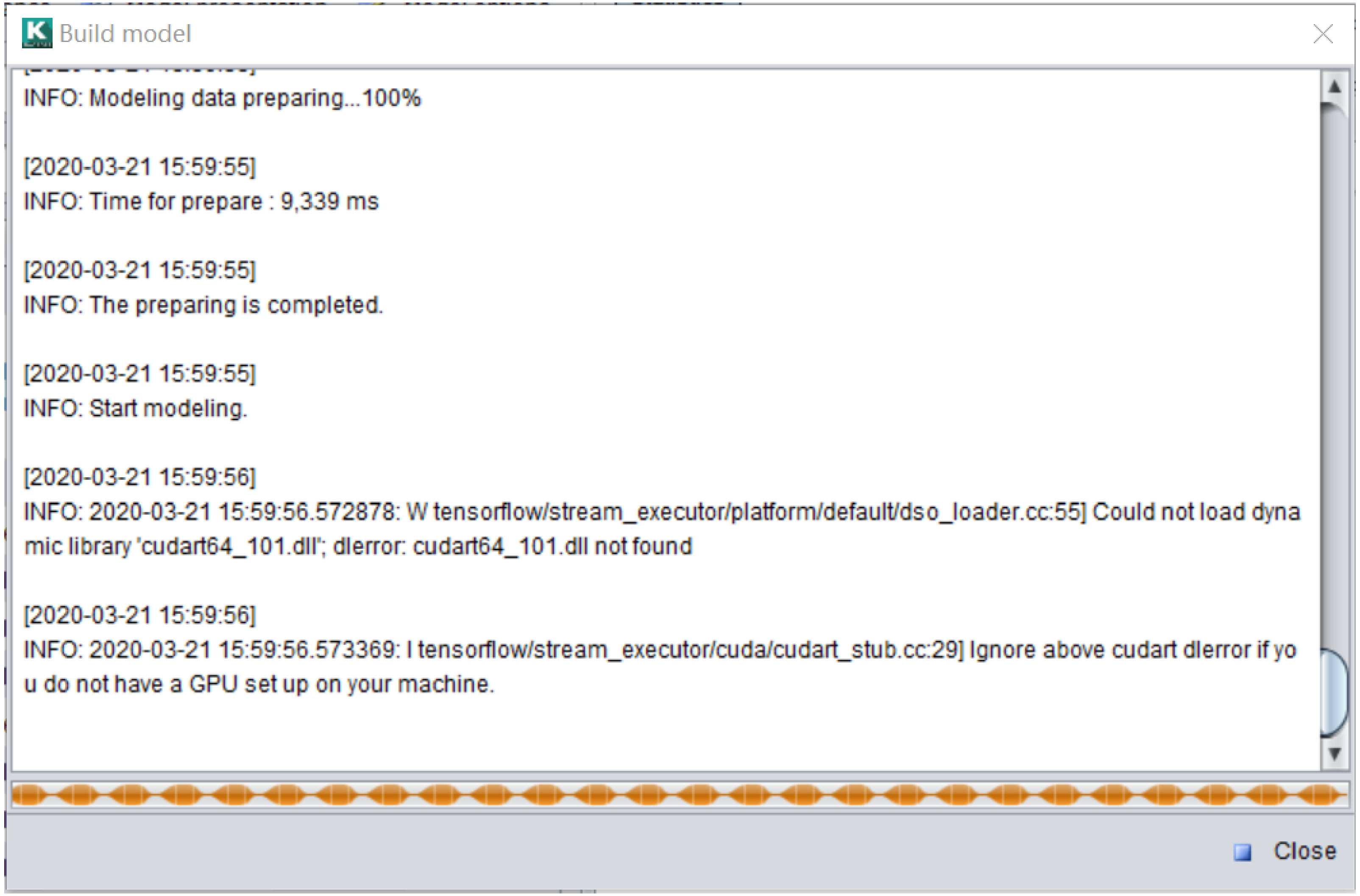
Modeling with YModel- Classification Model

Automatic data preparation and display the preparing progress.



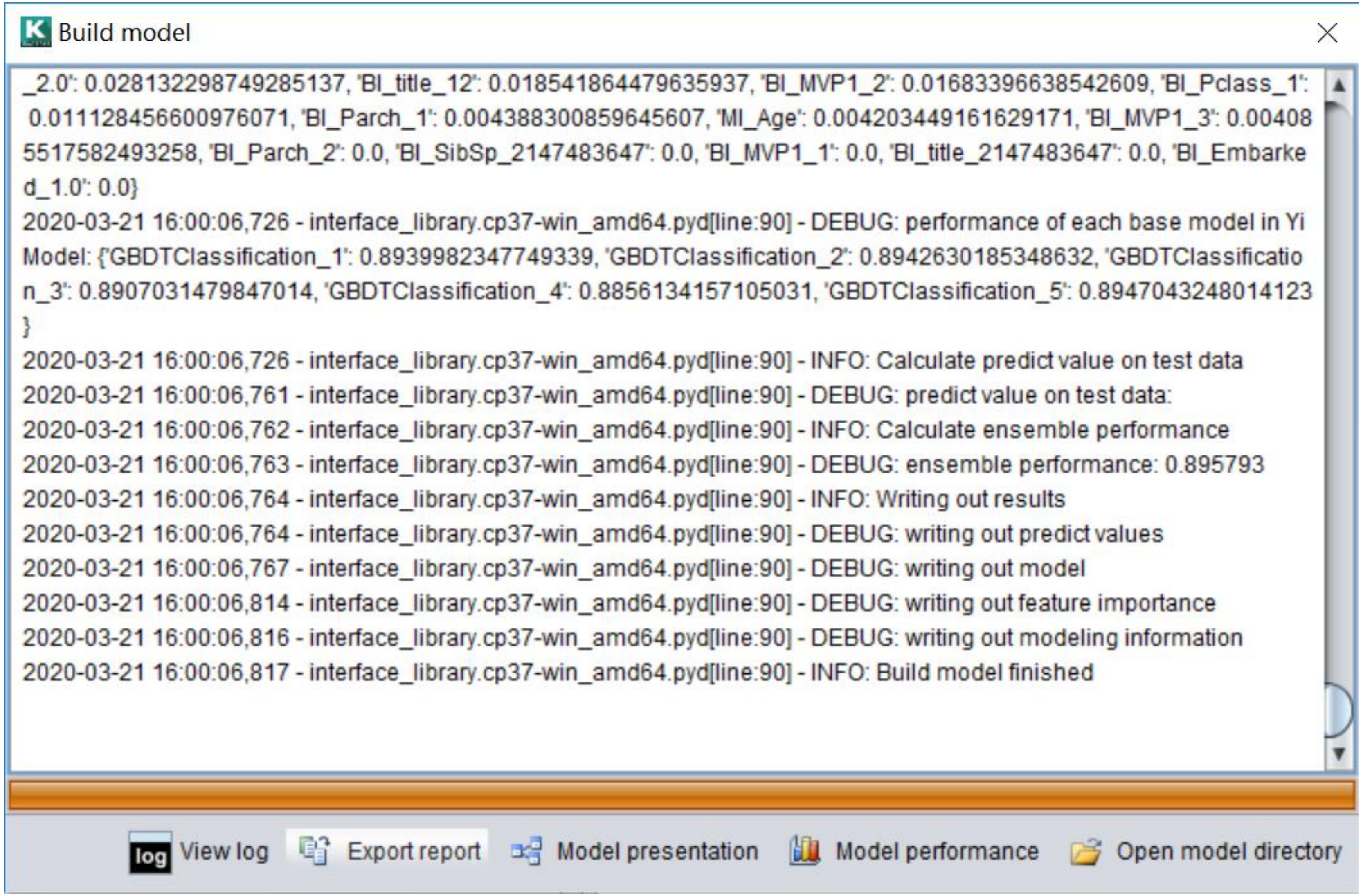
Modeling with YModel- Classification Model

Preparation is completed, start modeling.



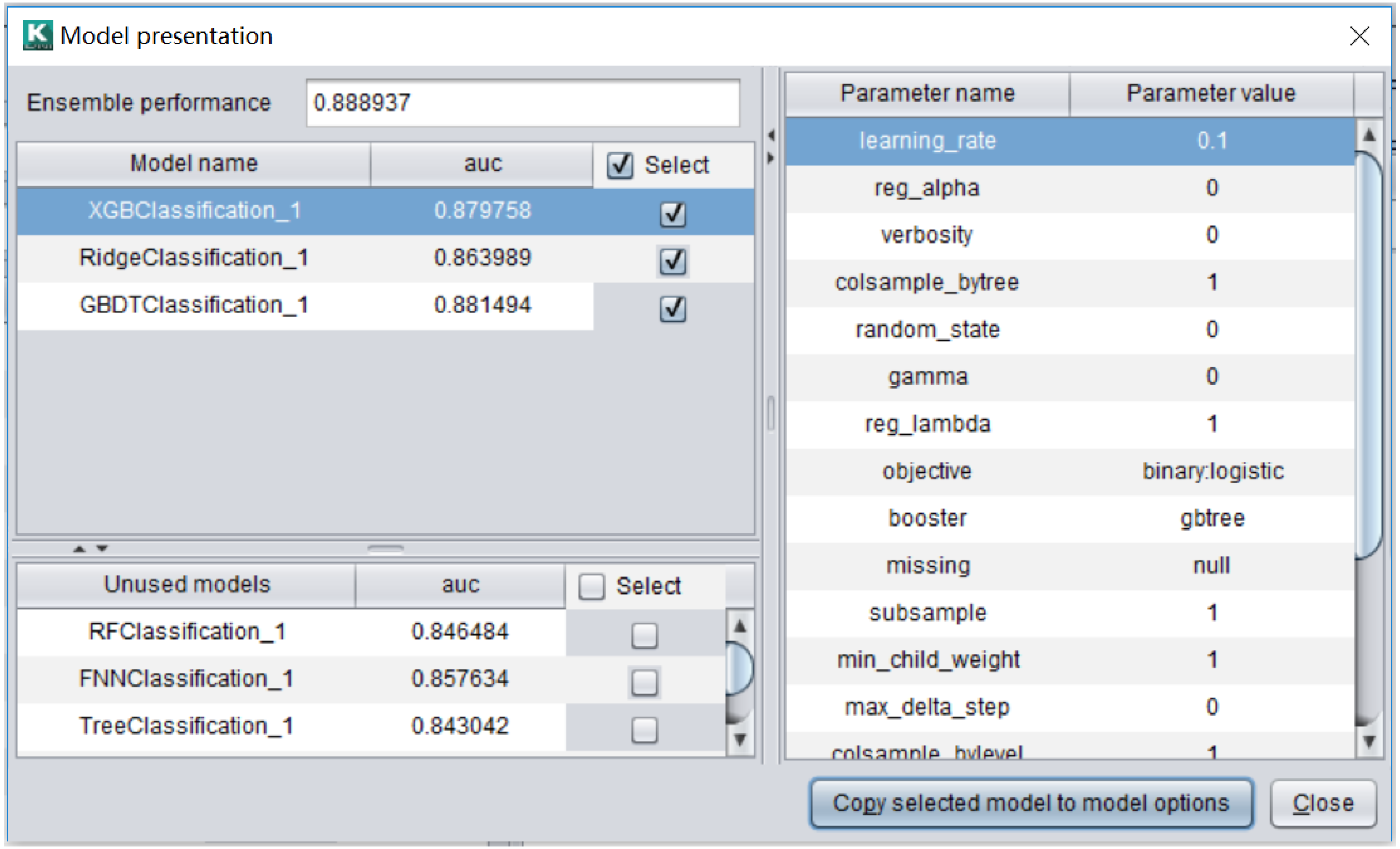
Modeling with YModel- Classification Model

Automatic modeling completed,
time consumed 10s



Modeling with YModel- Classification Model

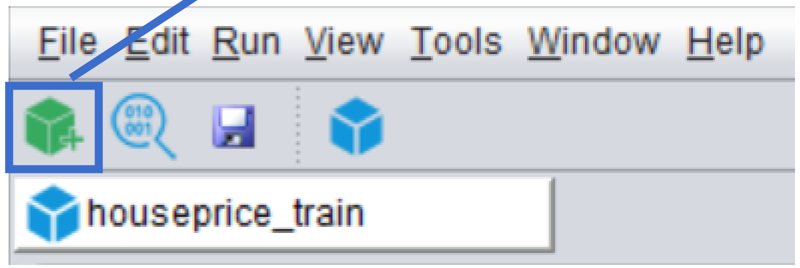
The models and parameters with better effect are selected automatically by using various algorithms to model respectively.

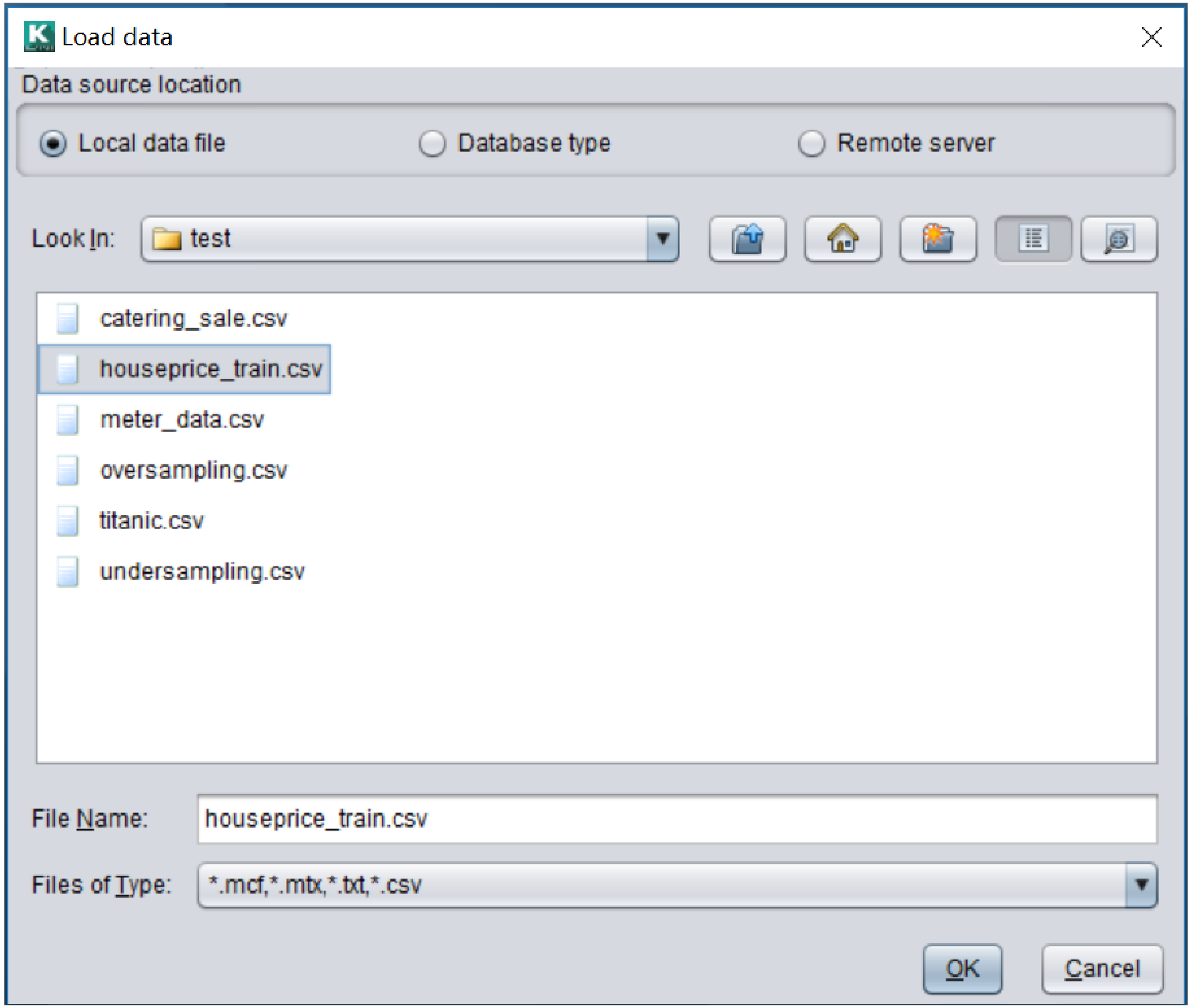


Modeling with YModel- Regression Model

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

new model

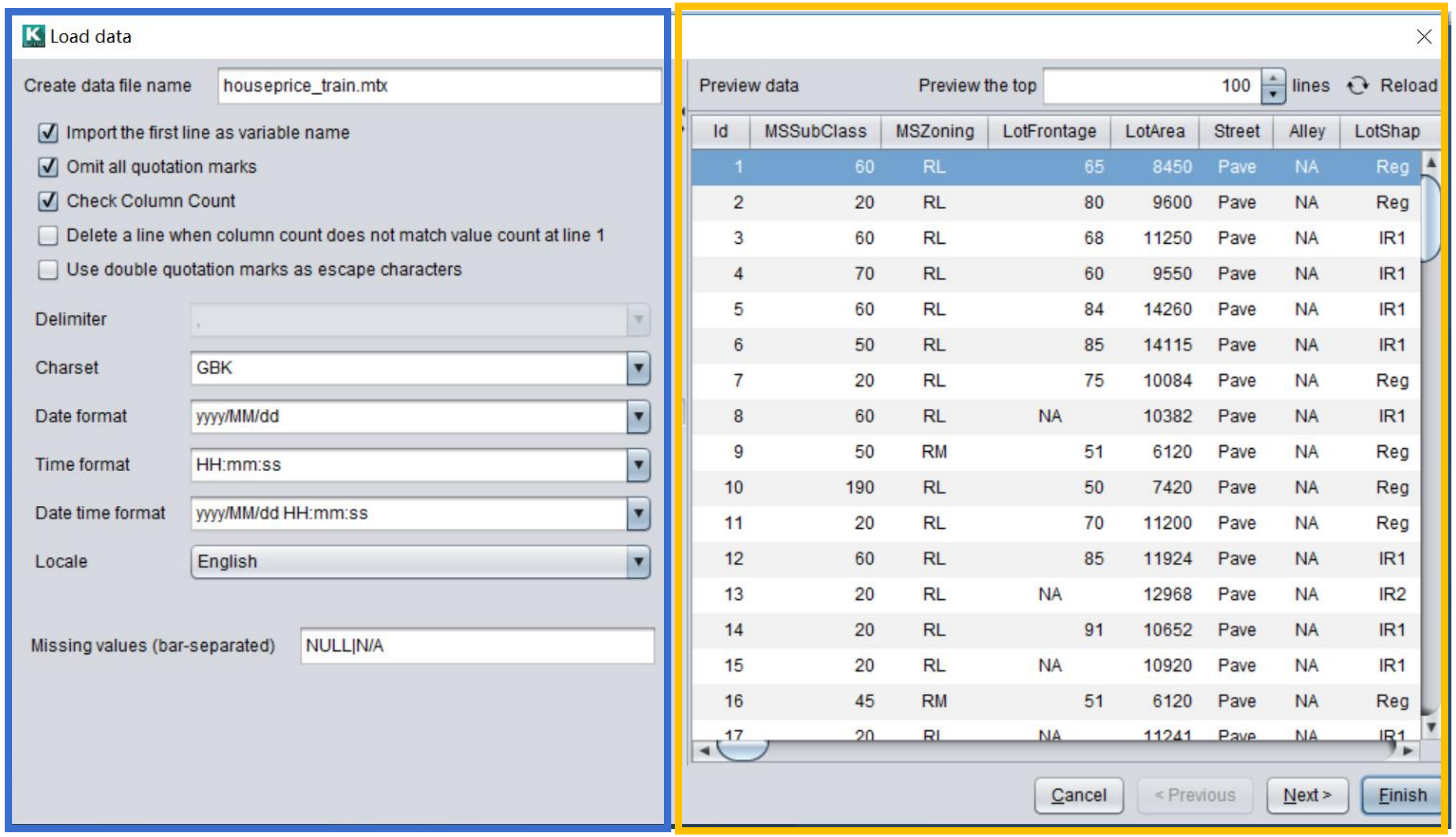




Modeling with YModel- Regression Model

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

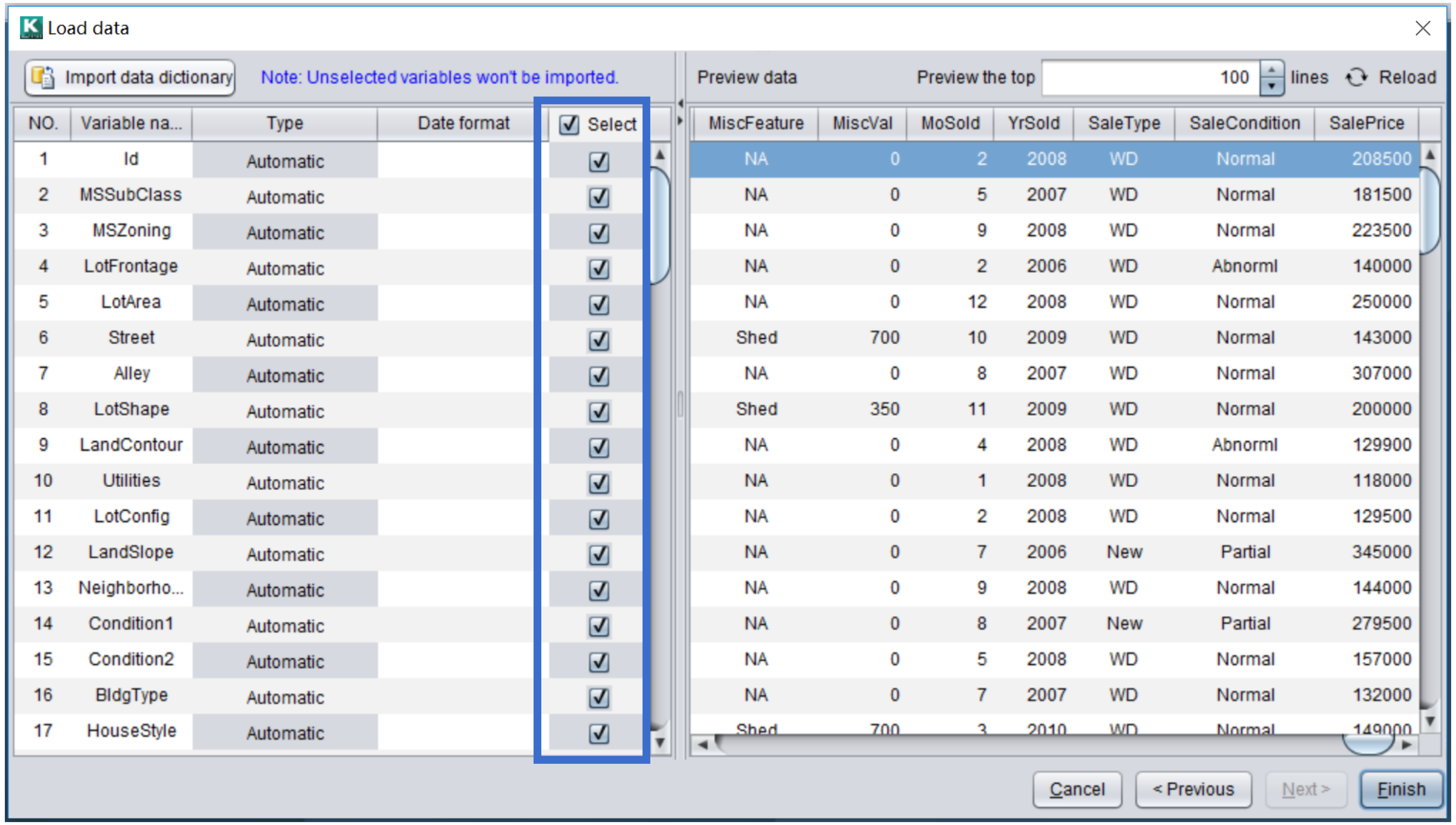
The left side of the page is configured with character set format, date time format and missing value format for automatic recognition by software.



Modeling with YModel- Regression Model

Select the variables involved in the modeling and click Finish.

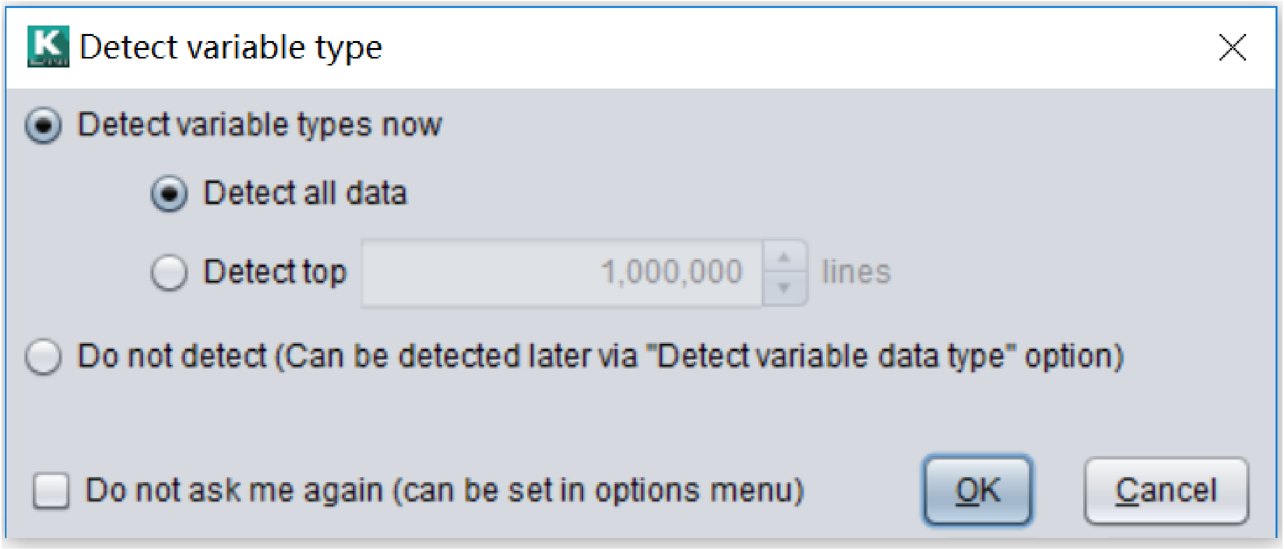
Here we choose all variables.



Modeling with YModel- 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.

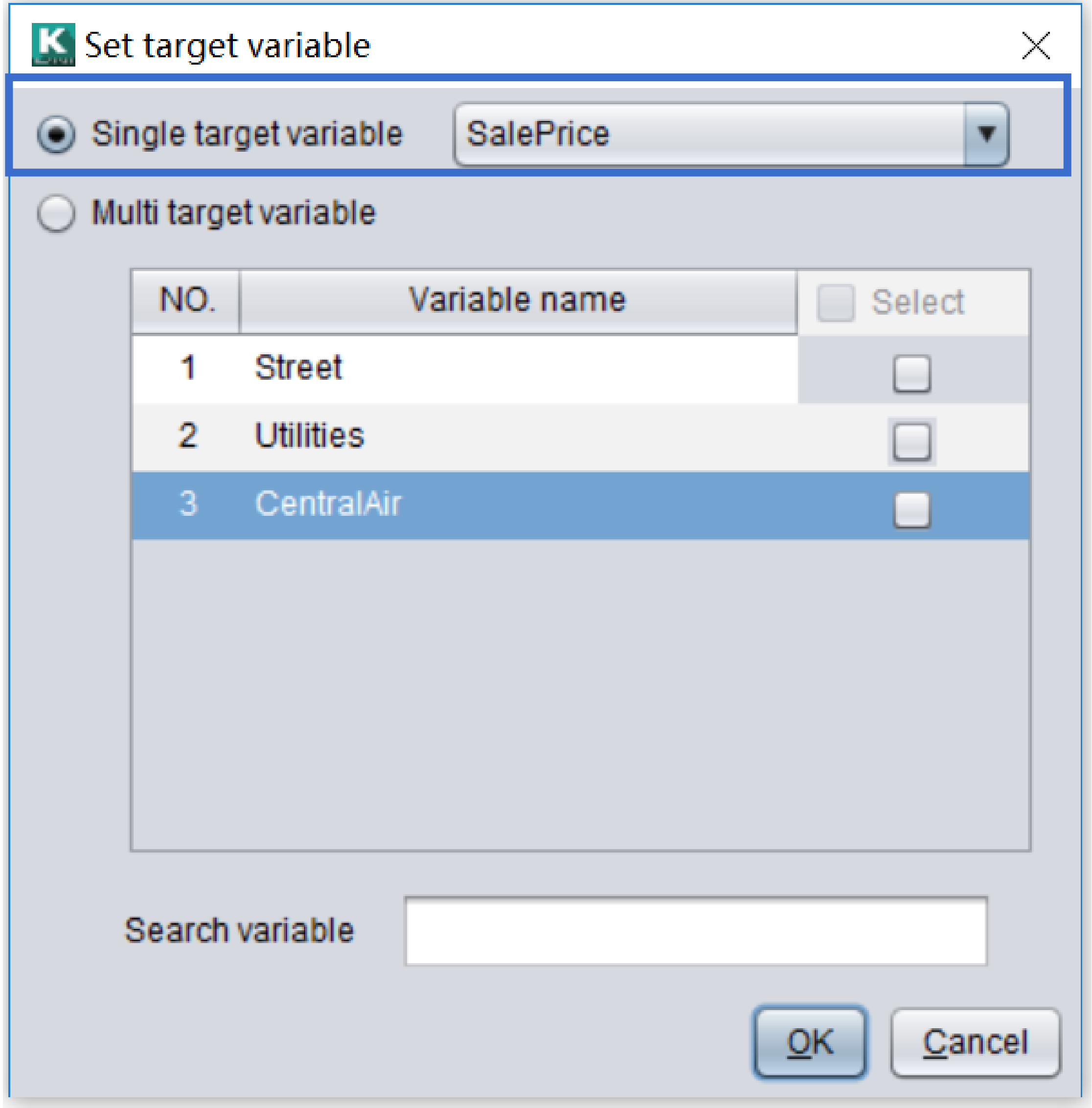


Modeling with YModel- Regression Model

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.

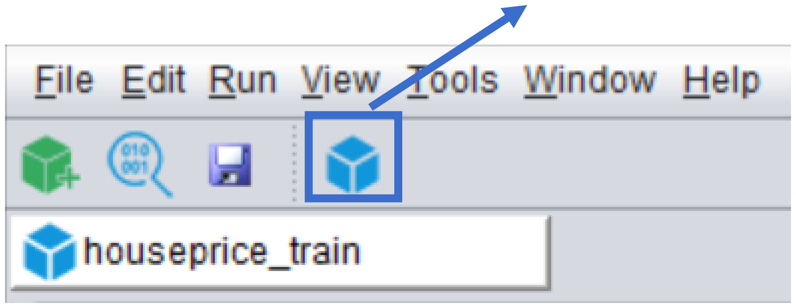


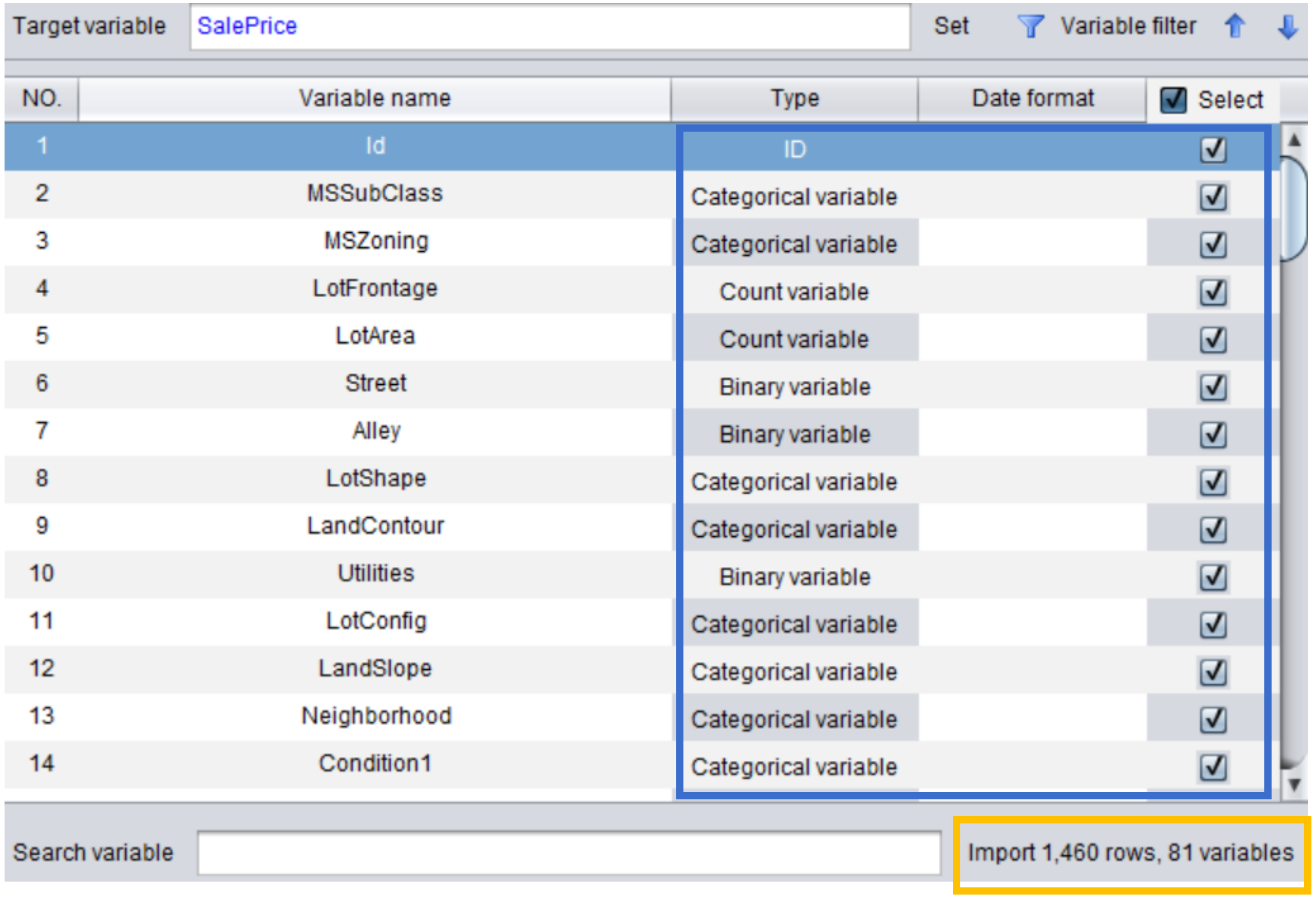
Modeling with YModel- Regression Model

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





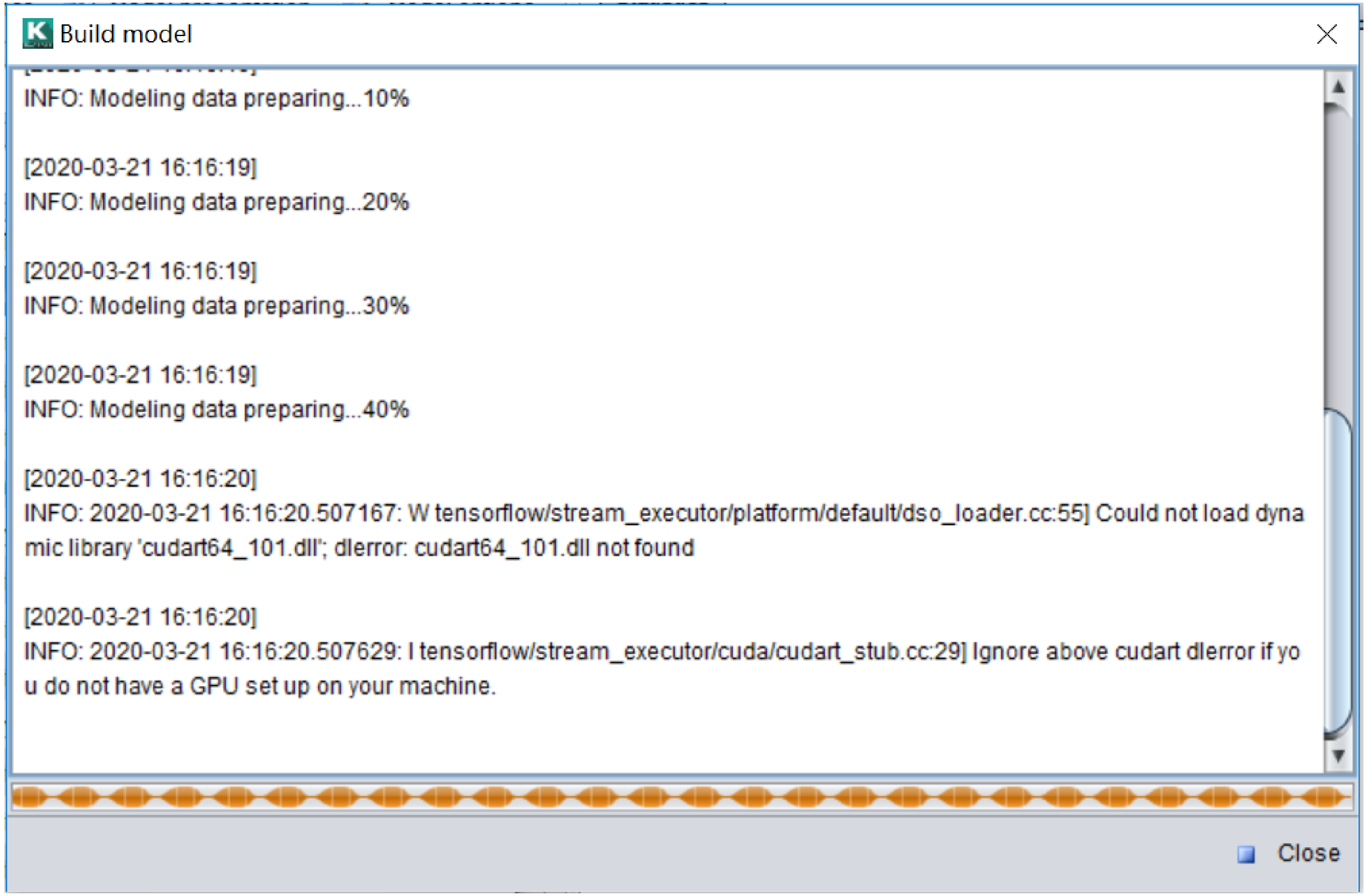
Target variable: SalePrice | Set | Variable filter

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		☑

Search variable | Import 1,460 rows, 81 variables

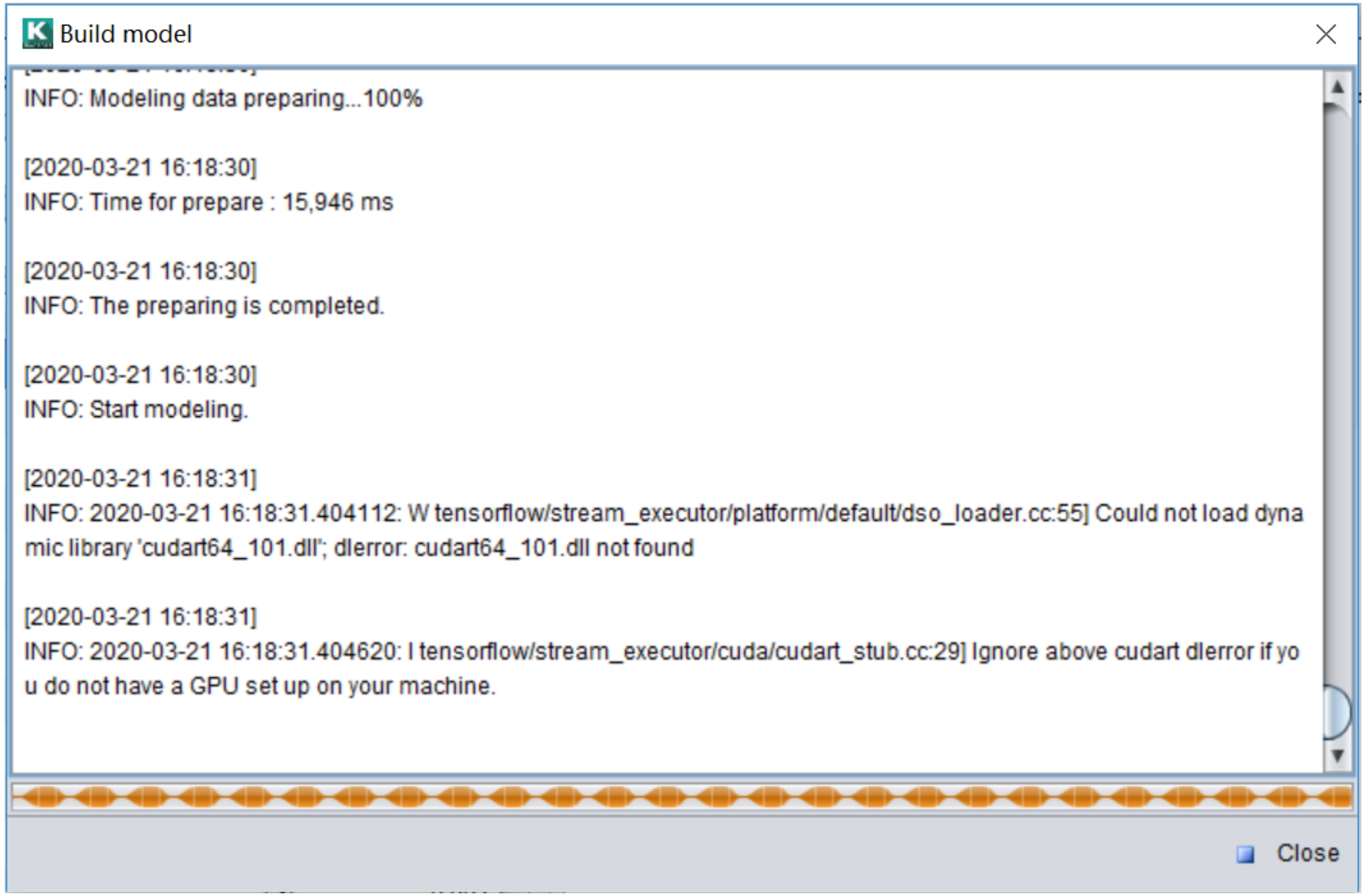
Modeling with YModel- Regression Model

Automatic data preparation and display the preparing progress.



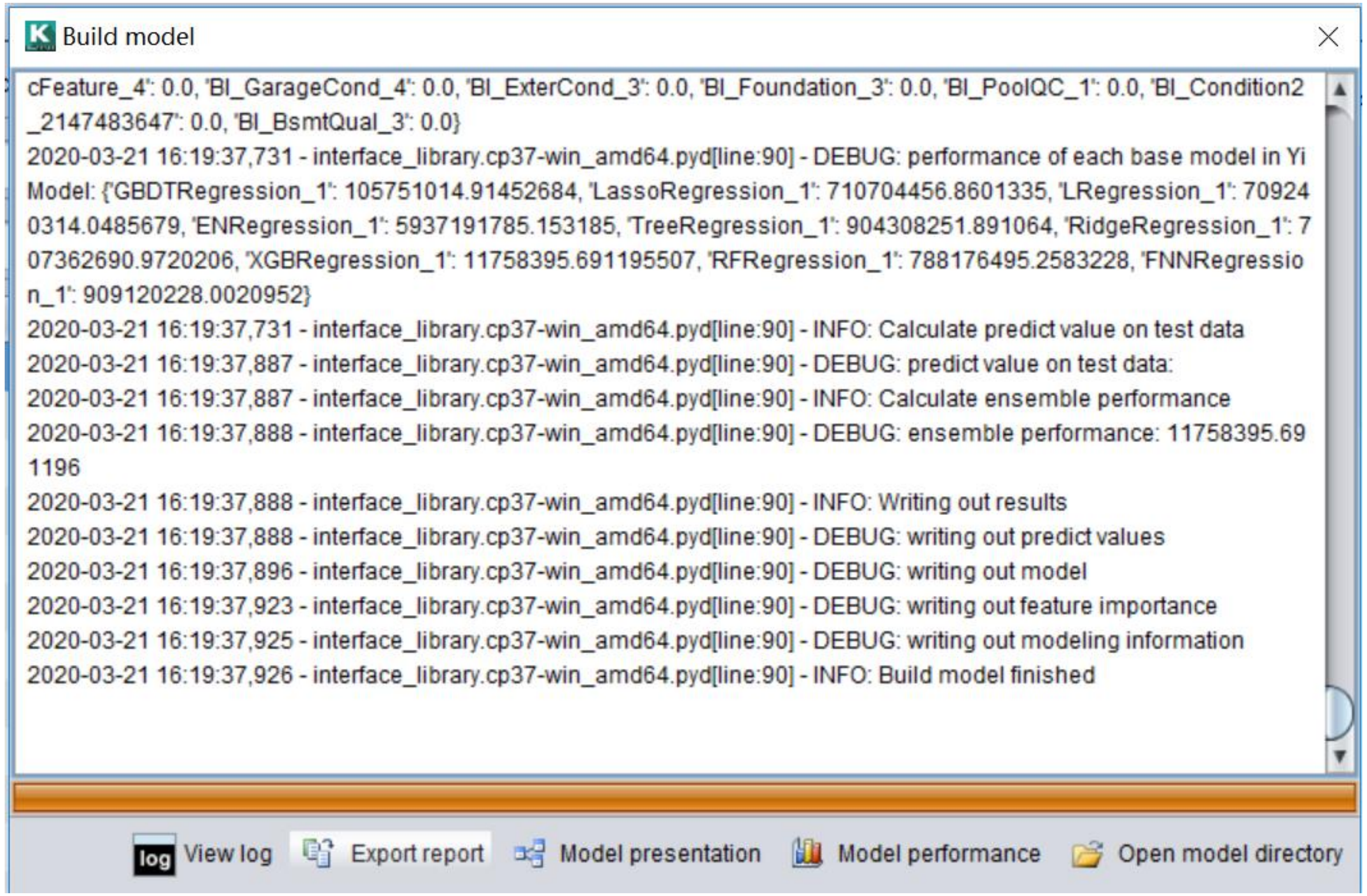
Modeling with YModel- Regression Model

Preparation is completed, start modeling.



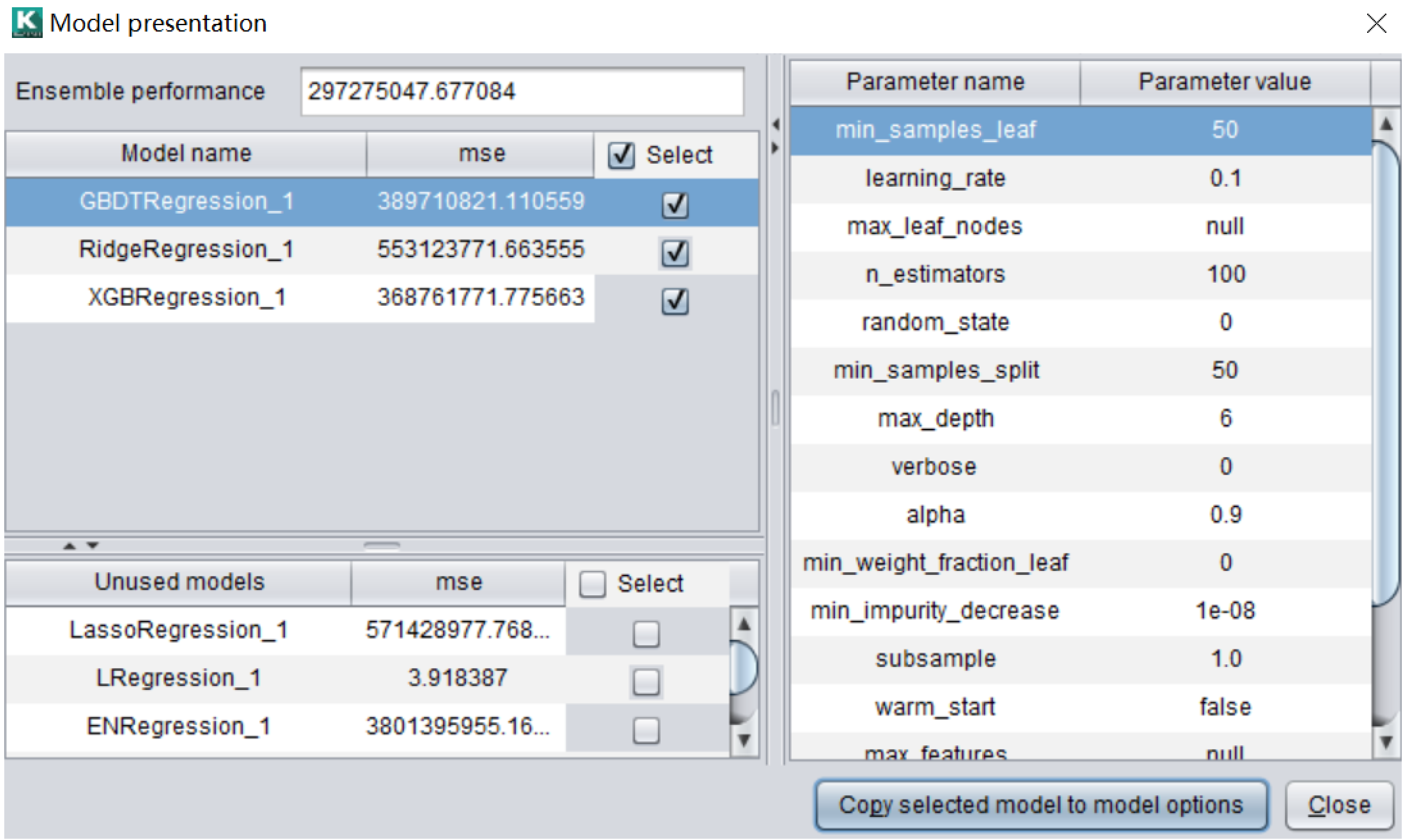
Modeling with YModel- Regression Model

Automatic modeling completed, time consumed 31s



Modeling with YModel- Regression Model

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

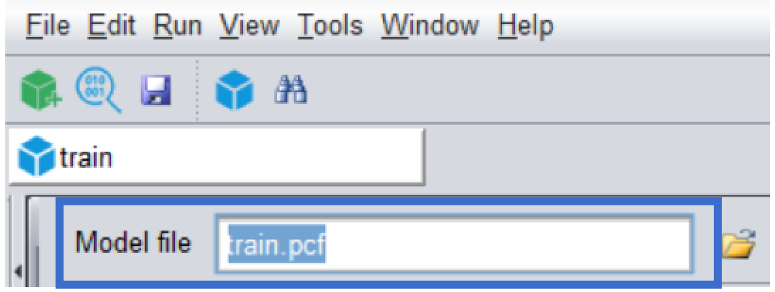


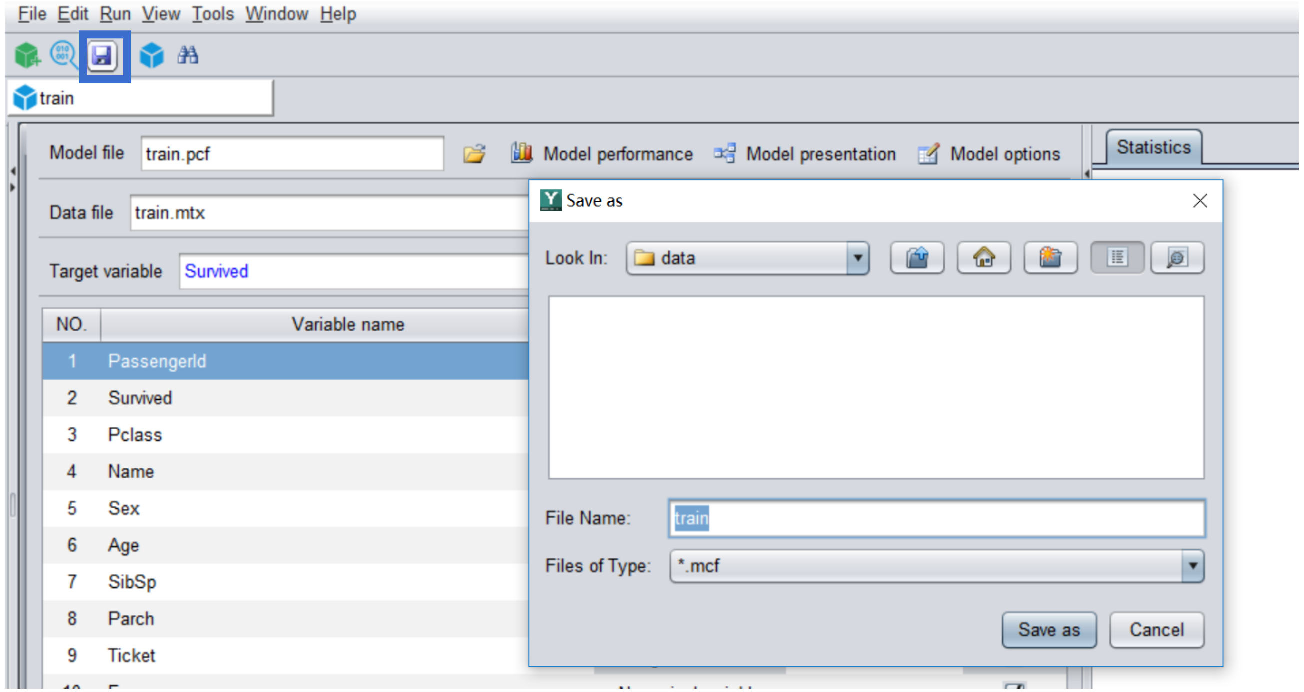
Model file

After the model is built, a model file with the .pcf suffix is written out to make predictions.

If you still need to Save the modeling process, click the "Save" button  to generate a modeling file with the .mcf suffix.

The PCF model file contains only the model information without data included, while the MCF file contains both data and modeling configuration information.

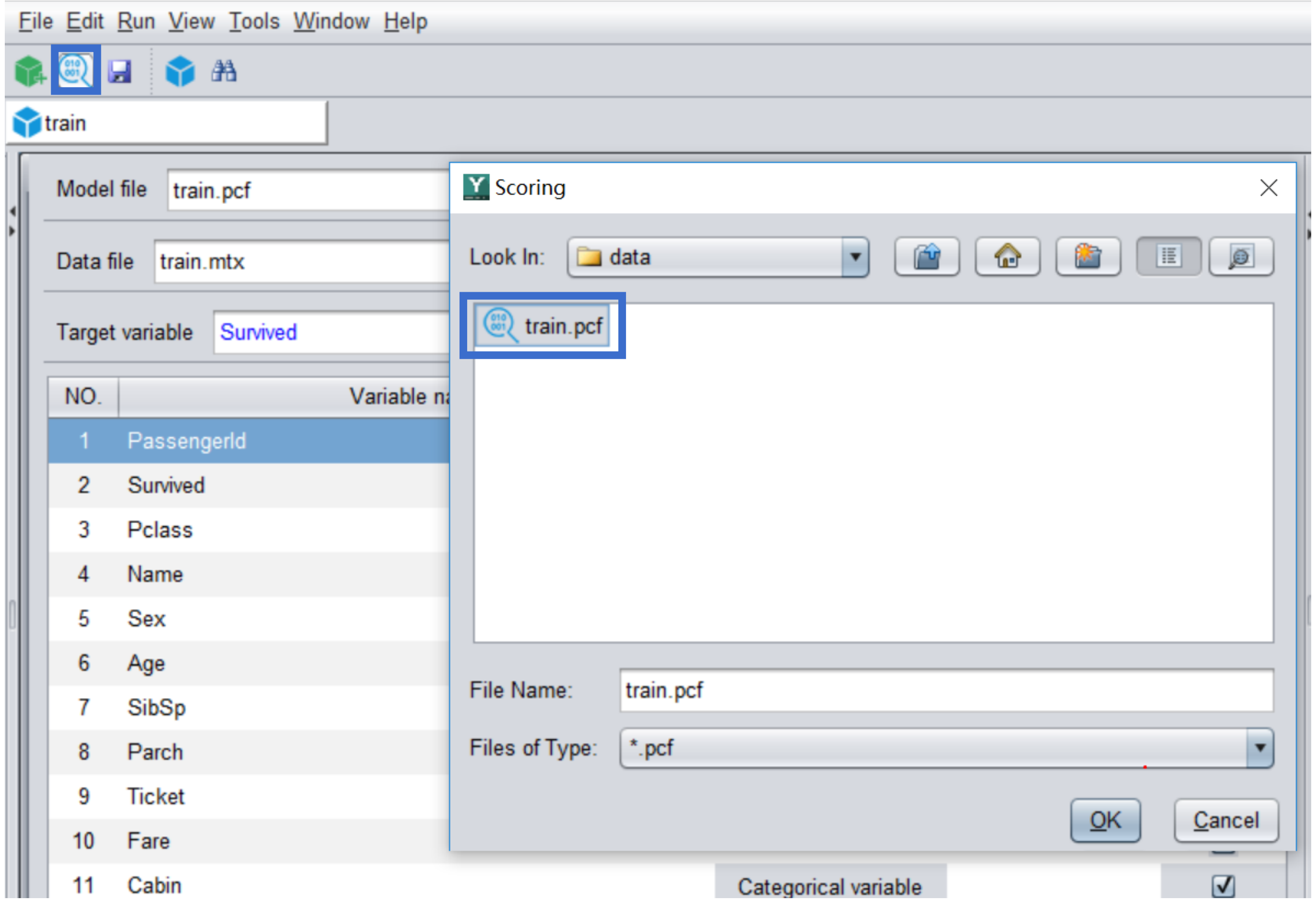




Chapter 4 Prediction

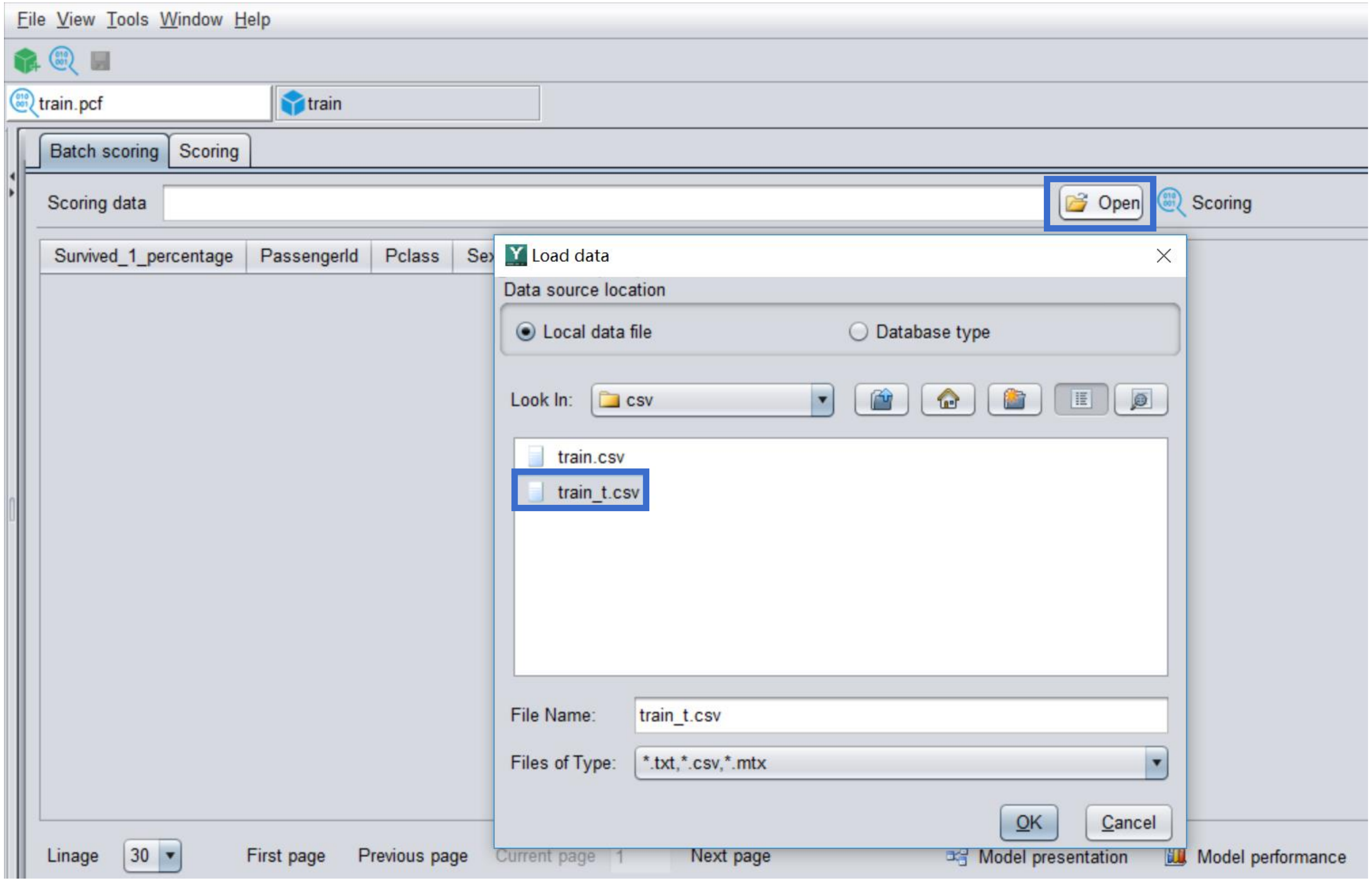
Prediction

Click the "Scoring" button on the top left of YModel, open the PCF model file which was generated in the previous chapter.



Prediction

Import the data set to be predicted.



Prediction

The prediction data is still in CSV format, and the variables must be the same as in the modeling data (columns in CSV), but there is no target variable.

For example, the difference between the two tables is that the modeling data has a target variable and the prediction data has no target variable.

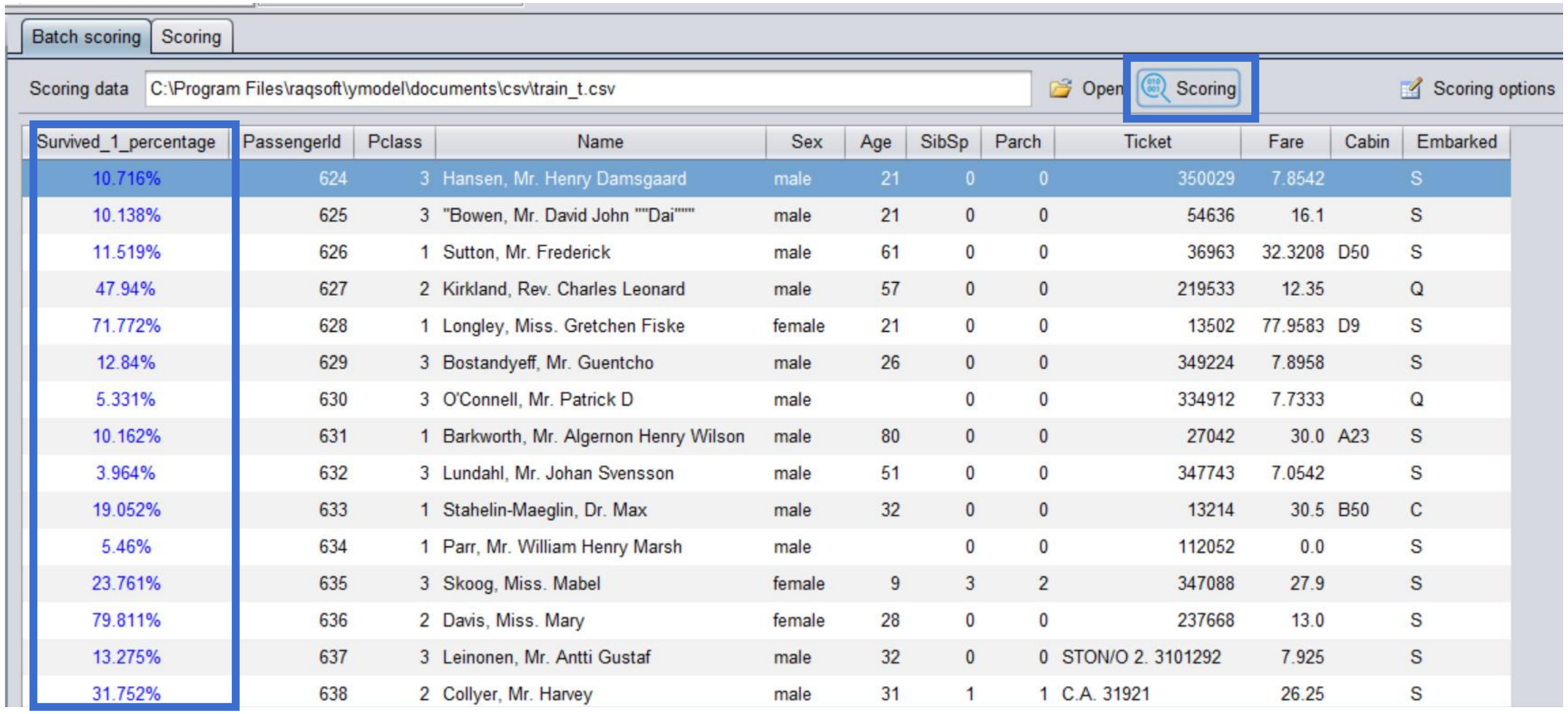
	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerI	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mı	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, I	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, Mı	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. \	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, N	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mr	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunderco	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson	male	39	1	5	347082	31.275		S

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerI	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	
2	624	3	Hansen, M	male	21	0	0	350029	7.8542		S	
3	625	3	Bowen, Mr	male	21	0	0	54636	16.1		S	
4	626	1	Sutton, Mr	male	61	0	0	36963	32.3208	D50	S	
5	627	2	Kirkland, R	male	57	0	0	219533	12.35		Q	
6	628	1	Longley, M	female	21	0	0	13502	77.9583	D9	S	
7	629	3	Bostandye	male	26	0	0	349224	7.8958		S	
8	630	3	O'Connell,	male		0	0	334912	7.7333		Q	
9	631	1	Barkworth,	male	80	0	0	27042	30	A23	S	
10	632	3	Lundahl, M	male	51	0	0	347743	7.0542		S	
11	633	1	Stahelin-M	male	32	0	0	13214	30.5	B50	C	
12	634	1	Parr, Mr. W	male		0	0	112052	0		S	
13	635	3	Skoog, Mis	female	9	3	2	347088	27.9		S	
14	636	2	Davis, Miss	female	28	0	0	237668	13		S	
15	637	3	Leinonen, I	male	32	0	0	STON/O 2	7.925		S	

Prediction

Click the "Scoring" button on the top right of the interface to make the prediction, and the following interface can be obtained after completion. The column on the far left is the prediction results.

In this example, percentage represents the probability of survival of passengers, and passengers with higher probability have a greater chance of survival.

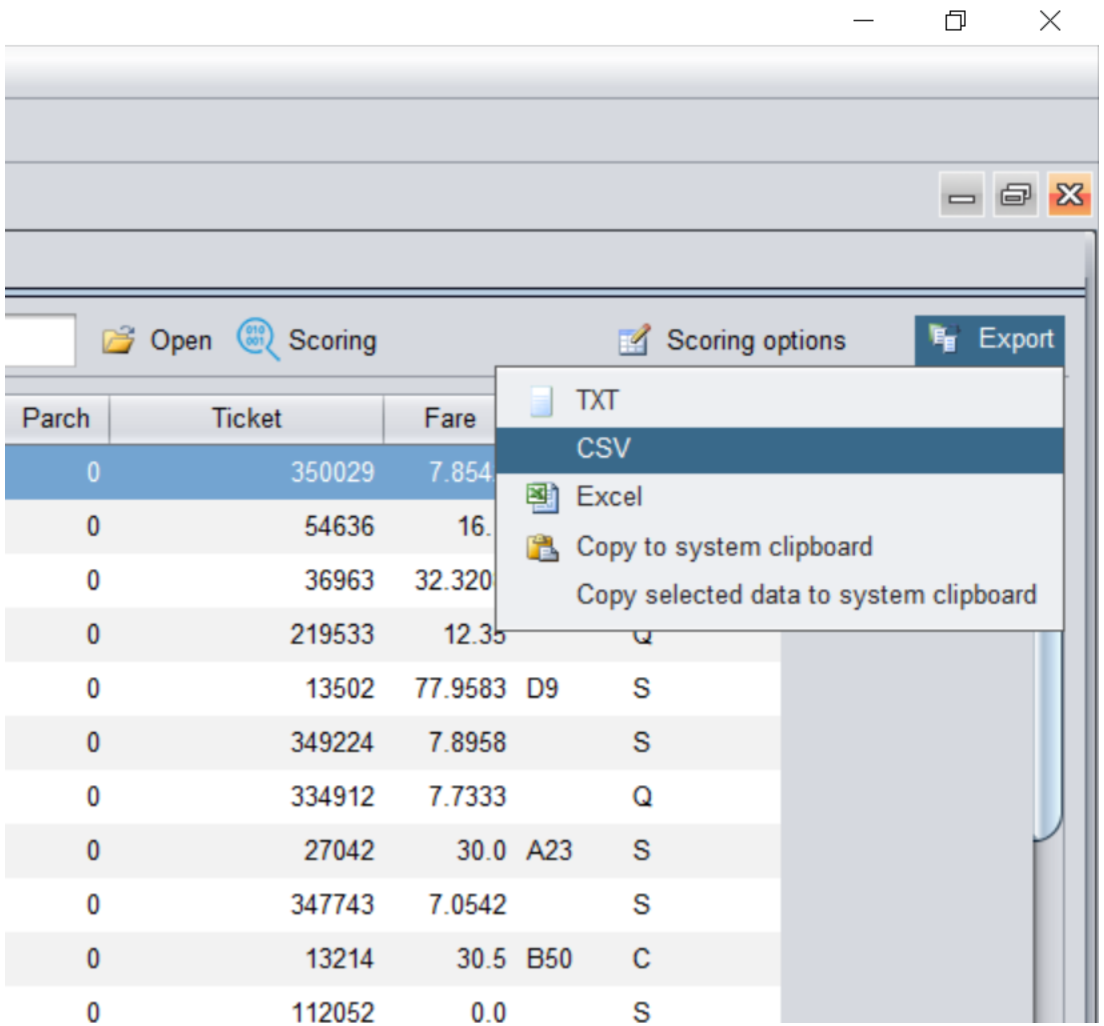


Survived_1_percentage	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
10.716%	624	3	Hansen, Mr. Henry Damsgaard	male	21	0	0	350029	7.8542		S
10.138%	625	3	"Bowen, Mr. David John ""Dai"""	male	21	0	0	54636	16.1		S
11.519%	626	1	Sutton, Mr. Frederick	male	61	0	0	36963	32.3208	D50	S
47.94%	627	2	Kirkland, Rev. Charles Leonard	male	57	0	0	219533	12.35		Q
71.772%	628	1	Longley, Miss. Gretchen Fiske	female	21	0	0	13502	77.9583	D9	S
12.84%	629	3	Bostandyeff, Mr. Guentcho	male	26	0	0	349224	7.8958		S
5.331%	630	3	O'Connell, Mr. Patrick D	male		0	0	334912	7.7333		Q
10.162%	631	1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30.0	A23	S
3.964%	632	3	Lundahl, Mr. Johan Svensson	male	51	0	0	347743	7.0542		S
19.052%	633	1	Stahelin-Maeglin, Dr. Max	male	32	0	0	13214	30.5	B50	C
5.46%	634	1	Parr, Mr. William Henry Marsh	male		0	0	112052	0.0		S
23.761%	635	3	Skoog, Miss. Mabel	female	9	3	2	347088	27.9		S
79.811%	636	2	Davis, Miss. Mary	female	28	0	0	237668	13.0		S
13.275%	637	3	Leinonen, Mr. Antti Gustaf	male	32	0	0	STON/O 2. 3101292	7.925		S
31.752%	638	2	Collyer, Mr. Harvey	male	31	1	1	C.A. 31921	26.25		S

Prediction

This result can also be exported to CSV, XLS and other formats.

At this point, the prediction is done, and the process is fairly straightforward.

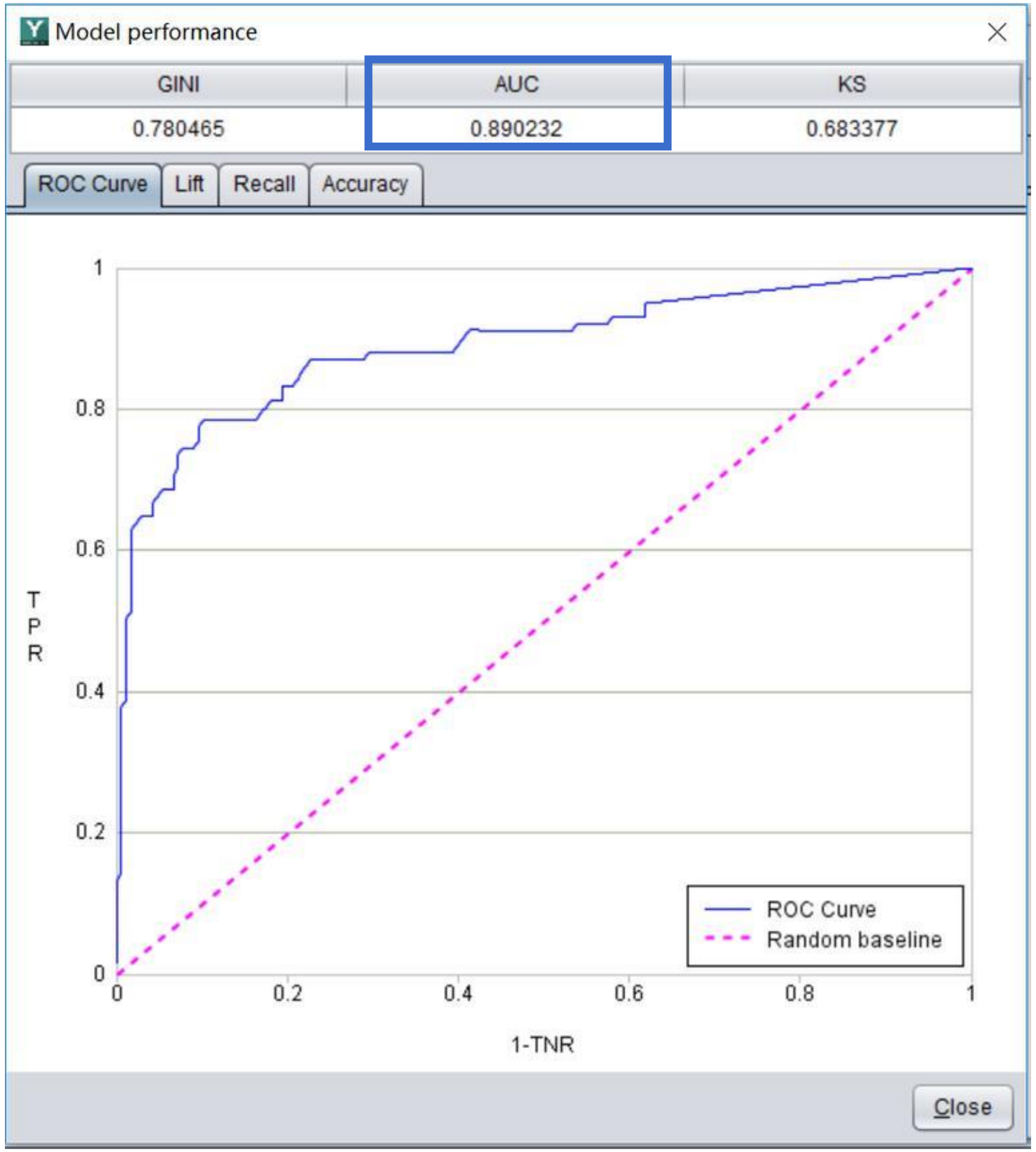


Chapter 5 Model evaluation and business application

5.1 How to evaluate a general prediction

How to evaluate a classification model?

Usually we check an index called AUC, which is greater than 0.5 but less than 1. The larger AUC value the better model performance.



General prediction

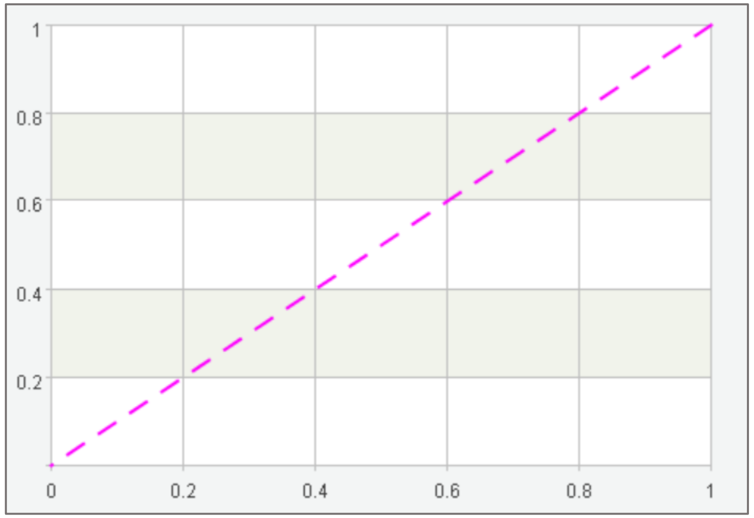
AUC = 0.5: the prediction model is the same as the random model, i.e. the discrimination between positive and negative samples is not better than the random model.

0.50 < AUC ≤ 0.65： poor

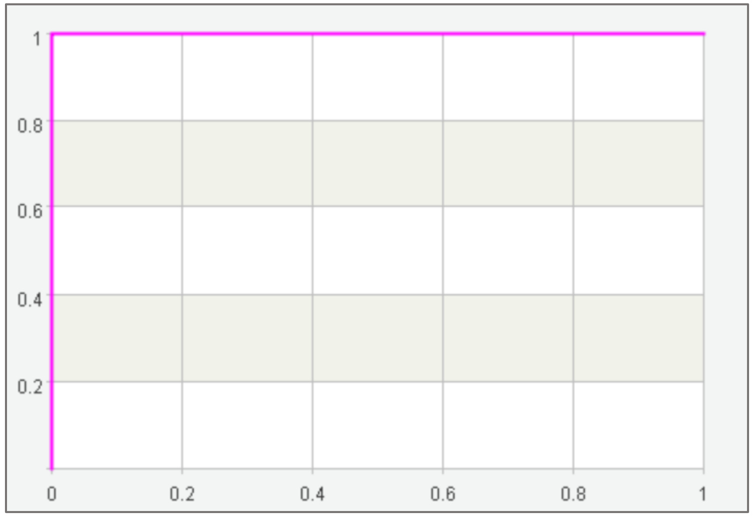
0.65 < AUC ≤ 0.80： medium

0.80 < AUC ≤ 0.90： good

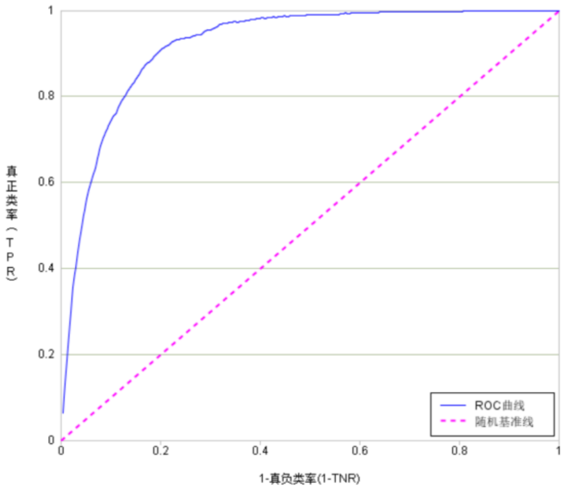
0.90 < AUC ≤ 1.00： excellent



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

5.2 How to improve the marketing success rate

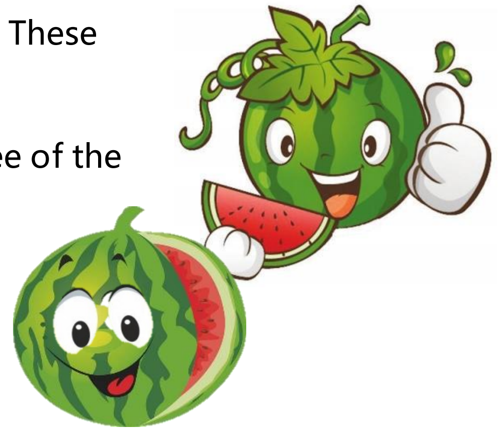
In a marketing scenario, in addition to check AUC value, also there is another useful index called Lift.

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
	……								

5.2 How to improve the marketing success rate

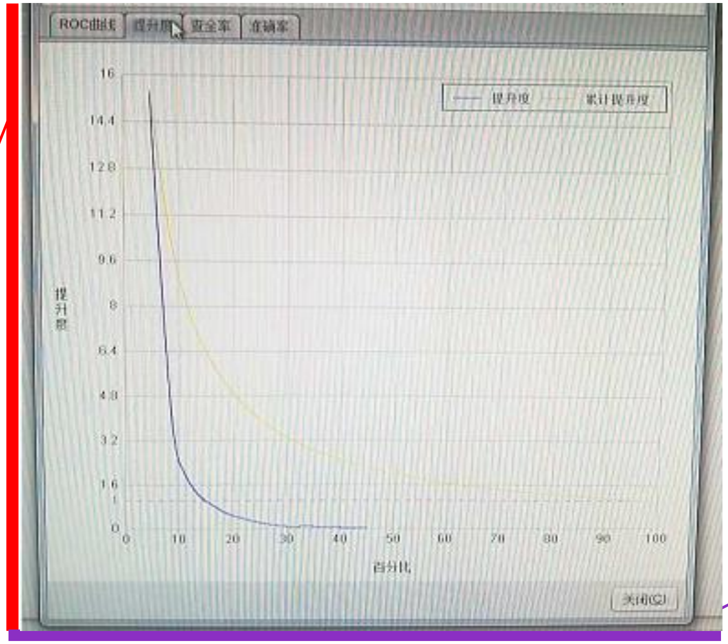
Lift is particularly suitable for targeted marketing scenarios

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** in randomly selected **100** customers will buy the product.

After using the model, the lift of the top 5% of the predicted probability is 14.4, that is to say, **21.6** (1.5 * 14.4) people in 100 people 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



Horizontal axis: grouping

Lift diagram

5.3 How to do multi-product portfolio marketing

If there are many kinds of products to be sold, such as a dozen or even hundreds of products, we can further improve the success rate and marketing value by exploring customers' interests and recommending product combinations to them. For example, Banks may have dozens of financial products to market, home appliance companies may have a variety of home appliance products to sell, supermarkets or e-commerce companies may have a variety of products to sell, and insurance companies may have various types of insurance products to market....

The classic case of beer and diaper in history is to increase the sales of both diapers and beer by mining data rules and selling two seemingly unrelated product combinations. For another example, there are many kinds of financial products in the bank, so we can combine several products with high purchase probability to sell by mining users' purchase preferences.

5.3 How to do multi-product portfolio marketing

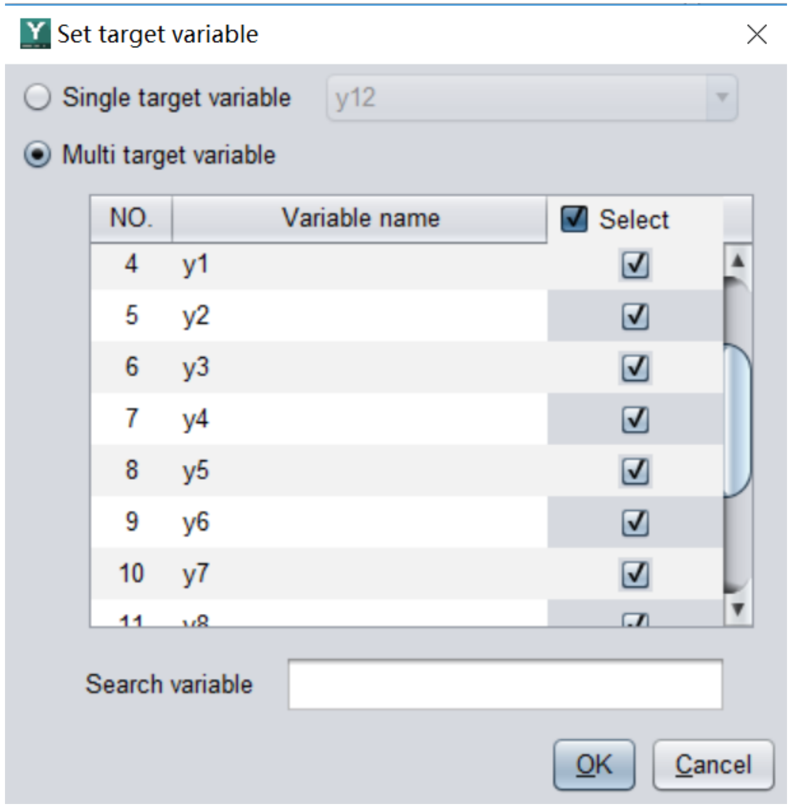
The prediction of a multi-product portfolio purchase list is also very simple, with off-the-shelf functional modules in the YModel. Specific operations are as follows:

(1) Modeling data set: prepare a tabular table of multiple objectives. Make a tabular table of historical information and all target variables needed to predict the products, as shown in Figure y1, y2, y3...Represents historical data on whether or not each product was purchased, these to be called multiple targets.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	poutcome	y1	y2	y3	y4	y5
58	manageme	married	tertiary	no	2143	yes	no	unknown	5	may	261	unknown	no	0	no	0	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	unknown	no	0	no	0	no
33	entreprene	married	secondary	no	2	yes	yes	unknown	5	may	76	unknown	no	0	no	0	no
47	blue-collaı	married	unknown	no	1506	yes	no	unknown	5	may	92	unknown	no	0	no	0	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	unknown	no	1	no	0	yes
35	manageme	married	tertiary	no	231	yes	no	unknown	5	may	139	unknown	no	0	no	0	no
28	manageme	single	tertiary	no	447	yes	yes	unknown	5	may	217	unknown	no	0	no	0	no
42	entreprene	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	unknown	yes	0	no	0	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	unknown	no	0	yes	0	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	unknown	no	0	no	0	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	unknown	no	0	no	1	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	unknown	no	0	no	0	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	unknown	yes	1	no	0	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	unknown	no	0	no	0	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	unknown	no	0	no	0	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	unknown	no	0	no	1	no
45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	unknown	no	0	no	0	no
57	blue-collaı	married	primary	no	52	yes	no	unknown	5	may	38	unknown	no	0	no	0	no
60	retired	married	primary	no	60	yes	no	unknown	5	may	219	unknown	no	0	no	0	no

5.3 How to do multi-product portfolio marketing

(2) When configuring the target variable, change the single target variable to multiple target variables, as shown in the figure. YModel will automatically combines products based on user preferences.



5.3 How to do multi-product portfolio marketing

Other operation steps are the same as single product purchase prediction. After the prediction is done, the results will appear as follows:

Batch scoring | Scoring

Scoring data C:\Program Files\raqsoft\ymodel\data\bank-full1.mtx Open Scoring Scoring options Export

Recommended combination	Combination_percentage	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.339%	55	management	divorced	tertiary	no	7010	no	no	cellular	22	sep	155	1	195
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.331%	34	management	single	tertiary	no	340	yes	no	unknown	3	jun	633	28	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.329%	48	unemployed	married	primary	no	1296	no	no	cellular	5	mar	119	1	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.321%	26	technician	single	secondary	no	354	no	no	cellular	16	jul	586	1	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.319%	35	management	married	secondary	no	724	yes	no	cellular	18	oct	102	3	96
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.316%	46	management	married	tertiary	no	699	no	no	cellular	2	feb	241	1	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.312%	27	blue-collar	single	secondary	no	294	yes	yes	unknown	28	may	694	2	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.311%	24	technician	single	tertiary	no	368	no	no	cellular	7	oct	197	2	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.298%	70	retired	married	secondary	no	616	no	no	cellular	27	oct	89	1	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.296%	61	technician	divorced	secondary	no	6481	no	no	cellular	19	oct	131	1	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.294%	28	management	single	unknown	no	560	no	no	cellular	2	jun	135	1	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.291%	45	admin.	married	secondary	no	362	no	no	cellular	12	aug	536	4	
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.286%	33	management	single	tertiary	no	718	yes	no	cellular	18	oct	104	1	111
y2,y4,y6,y8,y1,y3,y5,y7,y9	90.279%	39	management	married	tertiary	no	1373	no	no	cellular	19	aug	386	2	

The first column on the left is the content of the product portfolio, and the second column is the probability that the user will buy the portfolio. Once the results are exported, the product portfolio purchase list is generated, and the top customers with higher probability can be marketed.
It should be noted that for the combination probability without Lift curve, the number of top customers will depend on the situation (usually this number will be more than the number of customers for a single product).

5.4 How to predict rare cases

In many business scenarios, there is a data imbalance phenomenon, such as bank loan defaults, only a small number of people default ; Insurance fraud, fraud is also an individual phenomenon; There are also the proportion of defective products in product quality, non-planned parking phenomenon in industrial production....
The occurrence rate of these rare phenomena is very low, but once happen, there will be a large loss, so it’s better to predict and avoid them.

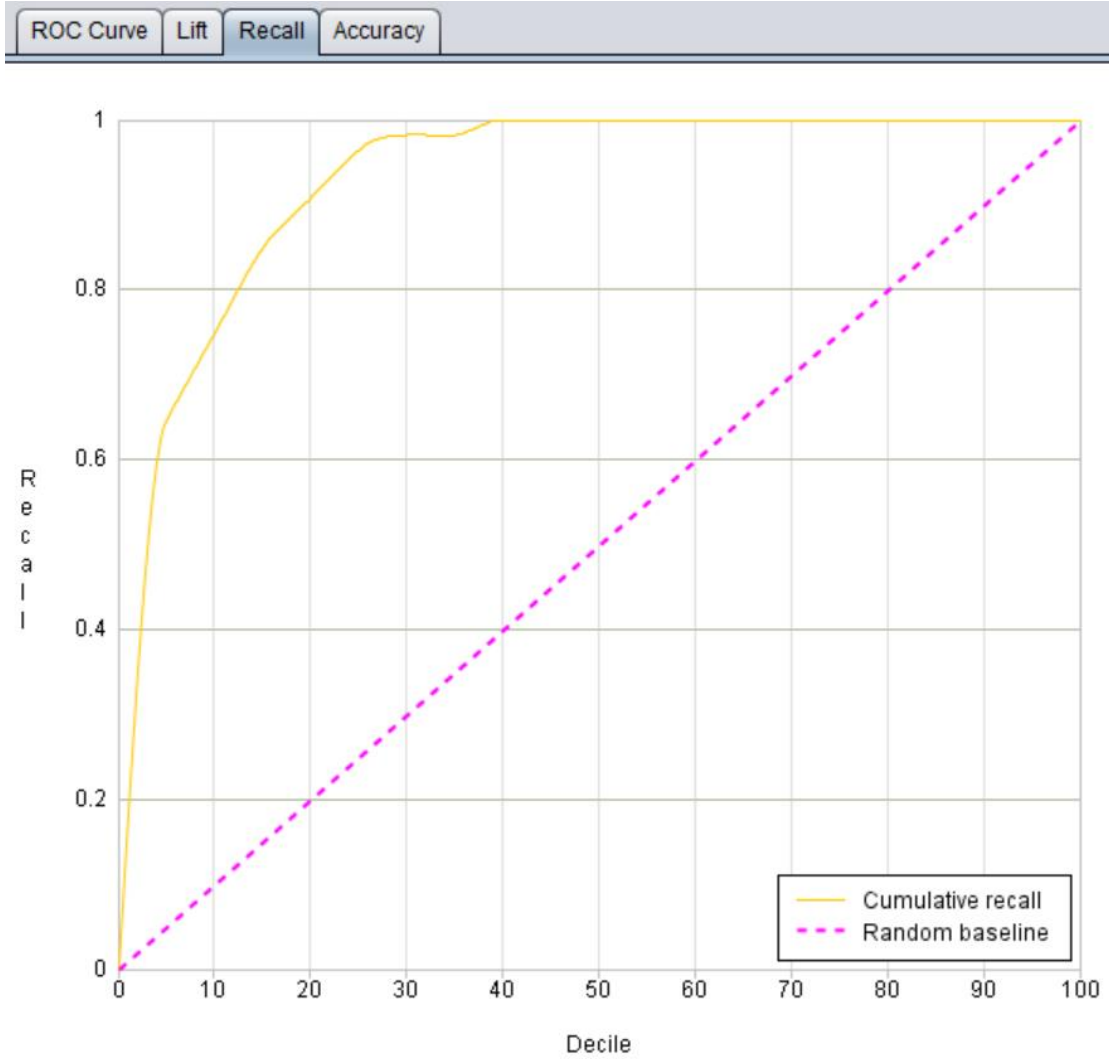
In the scenario with unbalanced data distribution, it is meaningless to just check the accuracy. What is more significant to us is Recall. Recall is how many of the positive samples were correctly predicted.

For an exaggerated example, to identify terrorists in airport, only five terrorists in 1 million passengers, because the terrorists are rare, if accuracy is used to assess the model, as long as all people are recognized as normal people, its accuracy can reach 99.9995%, but apparently this doesn't make sense, no terrorist would be caught, that is to say, although the model accuracy is very high but the recall rate is 0/5 = 0. On the contrary, the other model predicted that 100 people would be high risk group, and that all five terrorists would be included in that group. The accuracy was down to 99.9905% (95 people were wrong), but the recall rate was 5/5=1, and the terrorists were caught. Such a model would make more sense.

5.4 How to predict rare cases

In YModel, Recall curve is used to judge the Recall rate. As shown in the figure, the abscissa represents the number of the prediction probability of rare occurrence in order from high to low, 10,20...Represents the top 10%, 20%..., and the ordinate represents the recall value corresponding to each ranking stage.

The recall rate corresponding to the abscissa 10 in the figure is about 0.75, indicating that 75% of the rare phenomena can be captured in the top 10% of the predicted probability. That is to say, compared with all the screening, we can find 75% of the rare (abnormal) cases with 10% of the workload. The closer the Recall curve is to the upper left corner, the better the ability of the model to capture rare phenomena(default, fraud, defective products, abnormal equipment...).



5.5 Other evaluation indexes-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> $Gini$ ≥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.

5.5 Other evaluation indexes-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.
$0.2 > ks \geq 0$: 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.

5.6 Regression model evaluation

How to evaluate the model?

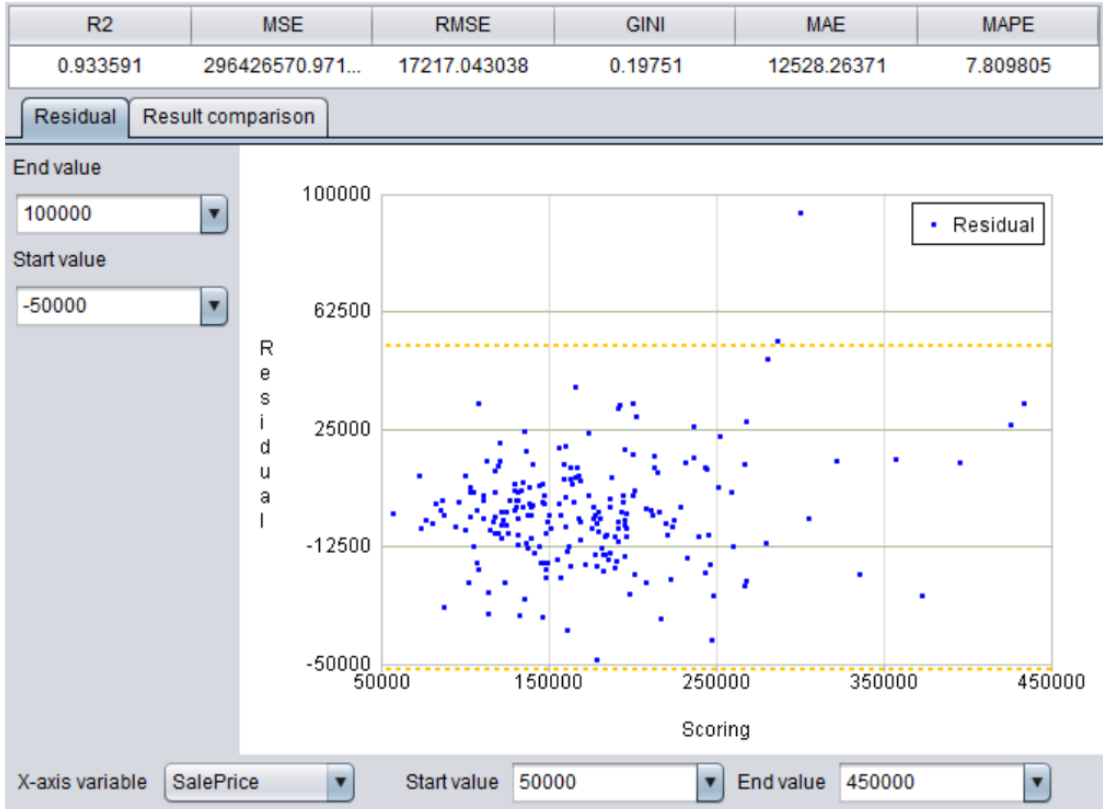
Regression model evaluation

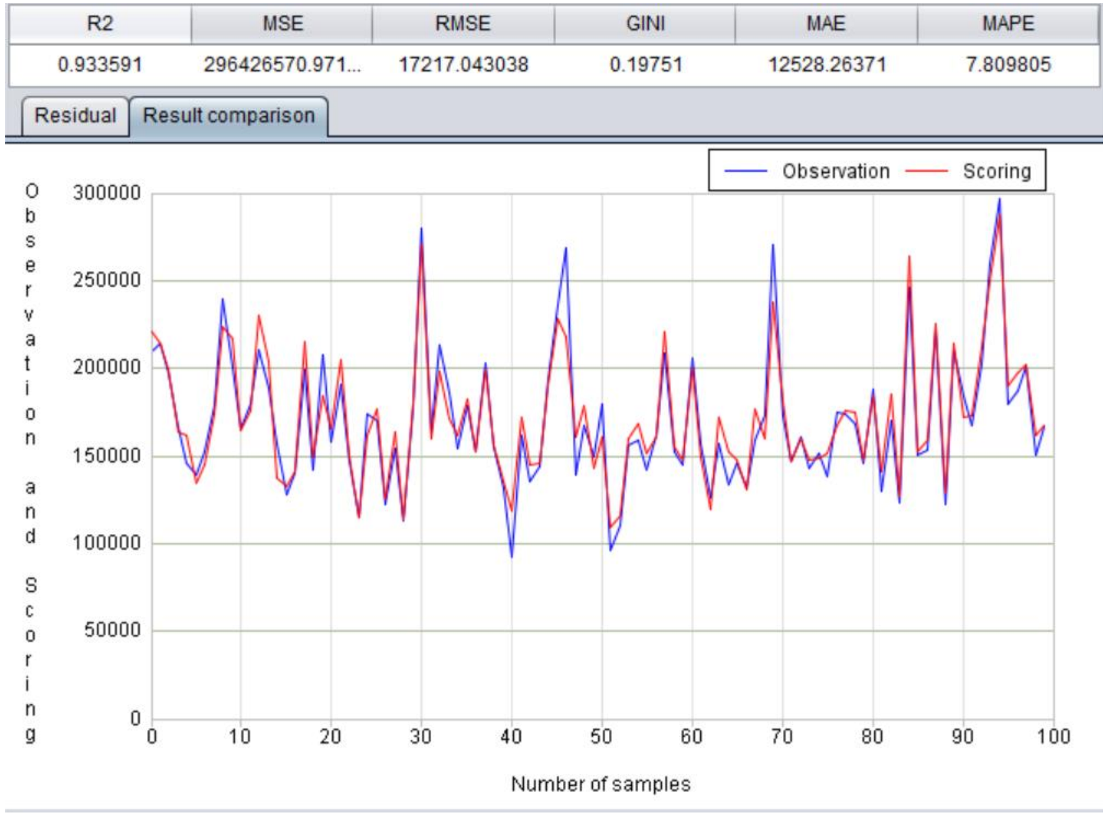
Common indexes:
R^2, MSE, RMSE, GINI, MAE, MAPE, etc.

Graphs:
Residual diagram, result comparison diagram

5.6 Regression model evaluation with YModel

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





THANKS