Report

Introduction

The sinking of the Titanic was one of the most notorious shipwrecks in history. On April 15, 1912, when the Titanic collided with an iceberg on its maiden voyage, 2,224 passengers and crew members died. This sensational tragedy shocked the international community and led to better safety rules for ships.

One of the reasons for the loss of life was that there were not enough lifeboats for passengers and crew. Although there are some elements of luck in surviving the shipwreck, some are more likely to survive than others, such as women, children and the upper class.

Based on the above description, we can analyze which people are more likely to survive.

Missing data processing

From the train data, I can see not all the information is intact. There 891 train data and 418 test data, and the data of 'Age' have 714, the data of 'Cabin' only have 204 and the data of 'Cabin' have 889. In test data, in addition to the above data kinds, the data set still lost one 'Fare' data. At first, I need to consider about how to deal with the lost data. Generally, there are two ways to deal with the lost data one way is completing data and the other is remove this kind of data. About the 'Cabin' and 'Fare' the lost data is only one or two, I add them with the most data to complete it. About the 'Cabin' both in train data and test data lost too much data. So, I delete all 'Cabin' data. About the 'Age', the age is an important feature, so I cannot delete it. For the 'Age' I try two ways to deal with it. The first method is to take the mean completion missing data. The second way is to take the random value between the mean minus the standard deviation and the mean plus the standard deviation. This two functions the first one is easy to operate, and the second have a low influence on the distribute of the value. However, the end accuracies have no big difference. So, I final choose the mean value to complete the lost data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId     891 non-null int64
Survived        891 non-null int64
Pclass          891 non-null int64
Name            891 non-null object
Sex             891 non-null object
Age             714 non-null float64
SibSp           891 non-null int64
Parch           891 non-null int64
Ticket          891 non-null object
Fare            891 non-null float64
Cabin           204 non-null object
Embarked        889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

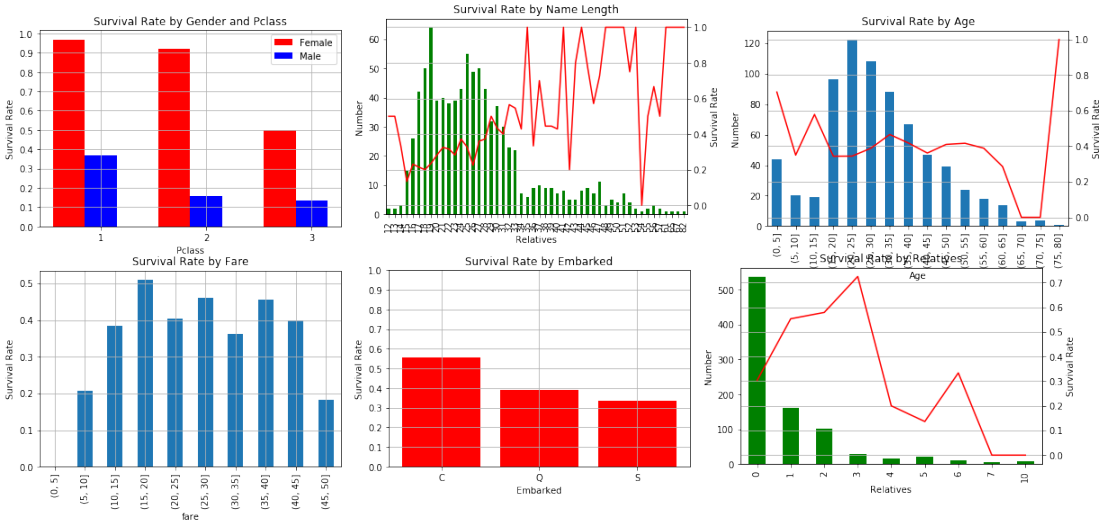
train data information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId     891 non-null int64
Survived        891 non-null int64
Pclass          891 non-null int64
Name            891 non-null object
Sex             891 non-null object
Age             714 non-null float64
SibSp           891 non-null int64
Parch           891 non-null int64
Ticket          891 non-null object
Fare            891 non-null float64
Cabin           204 non-null object
Embarked        889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

test data information

Data Analysis

After missing data processing, I analysis the data to find a link between data and survival rates. In the first picture, that is the survival rate by gender and public class. In this, the female survival rate is higher than the male. The high public class survival is higher than the low public class (1 > 2 > 3). The second picture is the survival rate by name length, from the chart I really cannot get some useful information. So, I delete it. The third is about the age. The child survival rate is higher than the adult. The forth picture is about fare, to some extent it reflects public class. The fifth picture is about embarked. From this the survival rate is C > Q > S. The last picture is about the how many relatives in Titanic. The survival rate of relative number is 2 or 3 is higher than others.



Methods

About the machine learning model, I choose the SVM and random forest. In SVM

model, each training instance is marked as one or another of the two categories. In the Kaggle Titanic, the label of the data is this kind 1 and 0. So, I think this model is suitable to solve problem. The advantage of random forest is that, for a variety of data, it can produce highly accurate classifiers. In the Kaggle Titanic, we have 12 kinds of data. So, I think random forest is also a suitable one to solve this problem.

Results

```
=================this for SVM
the accuracy of the cross validation score :  0.832796150972
the accuracy of the test set :  0.782296650718
the mean f1 score of the test set :  0.778362640264
=================this for RF
the accuracy of the cross validation score :  0.826815642458
the accuracy of the test set :  0.777511961722
the mean f1 score of the test set :  0.775452384776
```

Discussion

Accuracy is the proportion of true positive in the prediction positive (TP/(TP+FP)).

mean F1 score is the harmonic mean of precision and recall (F1=2PR/(P+R), P=TP/(TP+FP), R=TP/(TP+FN)).

		fact	
		1	0
prediction	1	True Positive(TP)	False Positive(FP)
	0	False Negative(FN)	True Negative(TN)

In those two models, the accuracy and mean F1 score is nearly. Both, the accuracy and mean F1 score are nearly. In SVM model I adjust the parameters to get the result the kernel is rbf, the C value is 100 and the gamma value is 0.01.

About the name feature, I add the name feature into my model. But I found it is useless. The change of accuracy and mean F1 score is little. After getting this result, I think the

function of name reflected in the appellation, such as Miss, Mrs., Mr. However, the Sex feature is also can reflected this. So, the role of name is very small.

Conclusion

In this coursework, I have learned how to process machine learning data and the simple machine learning project development process. And find the fun of machine learning. So, I will to learn more about machine learning.

Reference List

[1] : https://www.kaggle.com/c/titanic