Machine learning Report

Student Number: 19033698

Student Name: Xinquan Li

- Introduction

In 1912, the ship RMS Titanic struck an iceberg on its maiden voyage and sank, which is is one of the most infamous shipwrecks in history. In this project, it requires to explore how to build a model to find the predictors might have led to somebody survive or not. This means it is a classification problem.
The purpose of this project is on the process of learning machine learning, it means there is no need to implement some specific algorithms.Using the third-party libraries will give developers many tools.For example, numpy, pandas and scikit-learn are used in the project.

- Numpy, python scientific computing library
- Pandas, data analysis processing library
- Scikit-learn, Machine learning library(contain many algorithms)

Some important concepts:
- **Feature variables,** also called independent variables, are the features that can be observed by the sample, usually the input of the model.
- **Label**, also called target variable. The variable that needs to be predicted is usually the label or output of the model.(Survive, in this project)
- **Train data**, labeled data, provided by the organiser.
- **Test data,** the label is unknown, is the data used by the competition to evaluate the score.

- Method

1. Data preprocessing

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
<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)
```

From these graph, it could find many useful information.

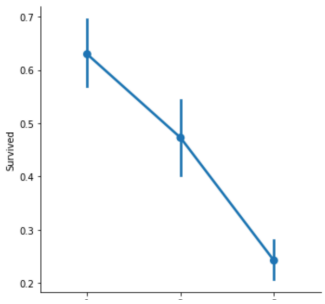
- Age is not missing too much, and Age is an important feature and should be retained.
- Name, PassengerId and Ticket may be useless In this project, they will be dropped.

2. Features preprocessing

- Pclass

Observe the relationship between Pclass and survive. This features are divided:1,2,3.
From the graph, it could say that the passengers in higher class
usually have higher survival rate.

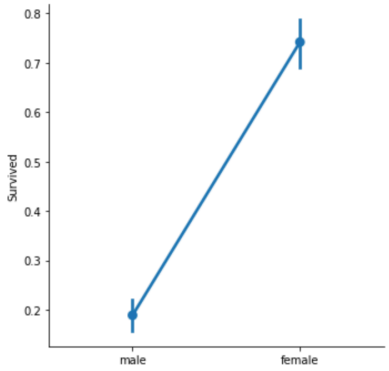
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```



- Sex

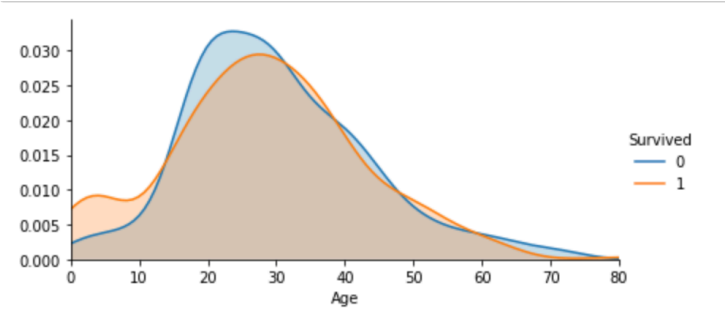
Observe the relationship between Sex and survive, the survival rate of female is significantly higher than that of male.Dividing Sex feature into male and female.

```
male      577
female    314
Name: Sex, dtype: int64
```



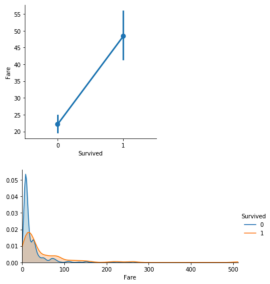
- Age

Calculating the mean and variance to fill the missing values. Observe the relationship between Age and survive, Age add two features, age under 15 (young) and age between 15 and 30.Because the range of 15 to 30 have significant influence on result.

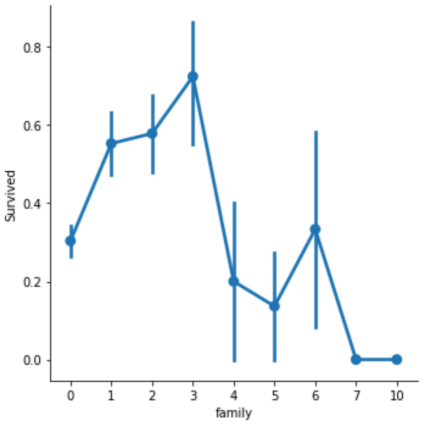


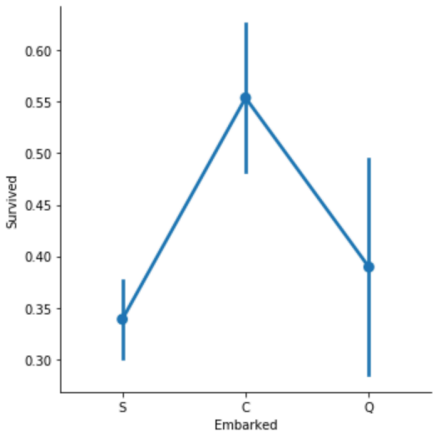
- Fare

Use average fill missing values. The higher fare will have higher survival rate.



• SibSp and Parch
It was found that the larger the two values, the lower survival rate.Delete the two two features, and combine the two features into a new feature(family)





• Embarked
Divide feature Embarked into S, C and Q. And C will be new feature.

- Results

• Use four models train the data.(Logistic regression, RandomForest, SVC and KNN)
• Use different features combination to train the model.
• test1:all features, test2:delete young, test3: delete young ,c test4:delete fare test5:delete C test6:delete Fare and young

	logistic	rf	svc	knn
test1	0.791246	0.814815	0.747475	0.740741
test2	0.791246	0.812570	0.764310	0.757576
test3	0.796857	0.818182	0.763187	0.757576
test4	0.793490	0.802469	0.830527	0.780022
test5	0.795735	0.808081	0.763187	0.738496
test6	0.790123	0.819304	0.831650	0.775533

In a conclusion, the test4 and SVC perform best.(the highest accuracy)

- Discussion

- Data processing has a significant influence on results. For example, when dealing with age, use median or mean to fill missing values will generate different accuracy on different models.

- Cross validation is a statistical method used to verify the performance of classifier. The basic idea to group the raw data in a certain sense, one is a training set and the other is a validation part.It is used to test the trained model as an evaluation index of the classifier.

- Random forest is composed of many decision trees, and there is no correlation between different trees. When perform a classification task, new input samples are entered and each decision tree in the forest is judged and classified separately. Each decision tree will get its own classification result, which one of the classification results of the decision tree is classified. At most, then the random forest will treat this result as the final result. For this project, when change the parameters of this model, it will get a different result.For example, when change the n_estimators(the number of decision trees) 10 to 50, the accuracy will have a significant rise.

- During the project, the importance of different features on final result is different.For example, age usually will be more important than Fare.This means it is a good idea to select features.

- The combination of different models may have better result, it may give weight to different models. The final result will generate from these models.

- Conclusion

This project will be a good start of machine learning.During the project, people will gain some practical experience. For example, when facing a real project, it need link it to a task of machine learning.Then data exploration, statical analysis, visualise, data processing, feature engineer, Feature Extraction, Feature Selection, Model and Validation. Luckily, these models have three-party library and there is no need to recreate wheels.But it is import to know the principles of models when using them, at least adjusting parameters.

reference

[Titanic: Machine Learning from Disaster]
[TitanicLearningQI]