Machine Learning for Titanic on Kaggle

By: Huanjin Li, Cunyi Li, Zhenhua Ren, Yiyang Bo

Background

The sunk of the Titanic is one of history's most infamous event.

On April 15th, 1912, the Titanic hit an iceberg in its first voyage with 1502 passengers and crew death among 2224 in total. The horrible tragedy shocked the world and led to the better ship safety regulations.

The reason why the sunk of the ship and the large amount of death is the lack of lifeboats to save passengers and crew. Although surviving from the disaster can be little contributed to fortune, some people seemed to be more possible to be saved, such as women, children and superior levels.

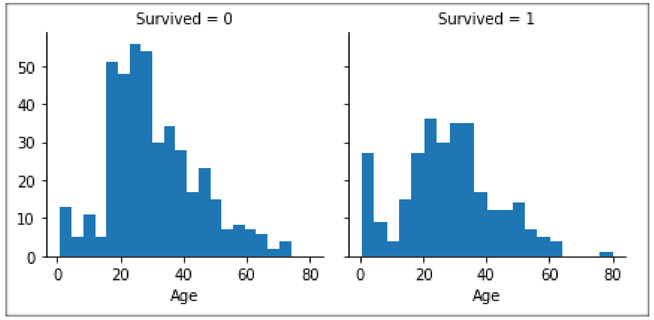
In this challenge, we need to analysis those who might be able to survive. We are supposed to use machine learning tools to predict which passengers may not died in the catastrophe.

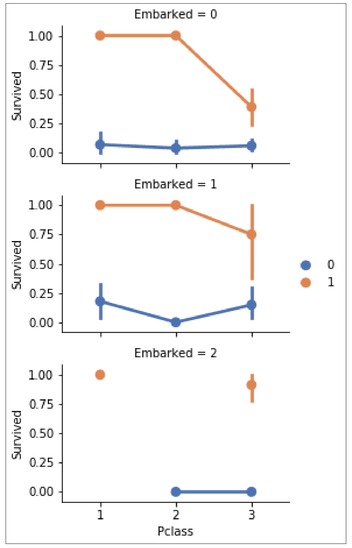
Simplification of the Dataset

- When the data has been imported into 'spyder', we can clearly discover the characteristics as follows:
- Categorical: 'Survived', 'Sex', and' Embarked'. Ordinal: 'Pclass'
- Continous: 'Age', 'Fare. Discrete: SibSp, Parch'
- With preliminary observation, we found that some of the data were 'NaN' which means the data has missed. Dealing with uncomplete data is essential in data analysis.
- 'Ticket' is combined with numbers and letters.
- Because there exists several ways to describe names, the characteristics may have mistakes.
- The number of the passengers of the sample is 891, accounting for 40% of the total(2224).
- 'Survived' is a Boolean value.
- 38% of the survival in sample means 32% of the people been saved in total.
- Most of the passengers(over 75%) traveled alone.
- About 30% had brothers, sisters or spouse.
- The prices of the tickets had massive differences, and only a little amount of the passengers paid the highest price(512 dollars).
- Elderly passengers(between 65-80) were less than 1%.
- Characteristics are all unique in the dataset.
- 'Sex' had two possible values and 65% of the passengers were male.
- The great lack of the values of 'Cabin' brought a huge trouble.
- Passengers were from three harbors and most of all got onboard at S.

	Sex	Survived
0	female	0.742038
1	male	0.188908

	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363





Preprocessing of data

1. Read data.
2. Drop 'Ticket' and 'Cabin' column.
3. Using regular expressions for the passenger's Name.
4. Set unusual 'Title' into 'Rare', and transform some 'Title' into ordinal forms.
5. Delete 'Name' and 'PassengerId' columns in 'train' dataset and 'Name' column in 'test' dataset.
6. Convert 'Sex' into numbers(0/1)
7. Complete the missing value of 'Age' and establish an empty array.
8. Combine 'Pclass', 'Sex' and 'Age' into a new feature.
9. Use median of 'Age' to complete the missing values of the 'Age' feature.
10. Establish a new feature 'FamilySize' to take the place of 'Parch' and 'SibSp'.
11. Create new features 'IsAlone' and 'Age*Class'.
12. The missing data of 'Embarked' and 'Fare' can be filled with its mode, then convert them into ordinal.

Algorithms and Results

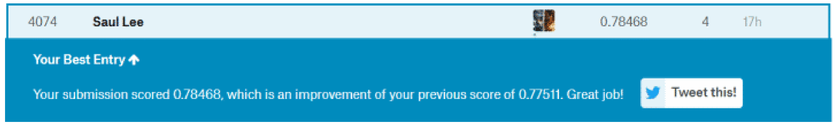
In order to solve the problem and to get better results, we have tried several algorithms and get different accuracies.

What we have used are as follows: SVM, Random Forest, KNN, Logistic Regression and Voting.

Here are the results:

```
Accuracy: 0.82 (+/- 0.03) [SVM]
Accuracy: 0.81 (+/- 0.03) [Random Forest]
Accuracy: 0.80 (+/- 0.02) [KNN]
Accuracy: 0.81 (+/- 0.01) [Logistic Regression]
Accuracy: 0.81 (+/- 0.02) [ensemble]
```

Here is the result of submission in Kaggle:



Conclusion

This is a less complex topic for data processing with only two forecasting results, but with many features, unrelated ones need to be excluded. After the exclusion, you need to preprocesses the data. As for the algorithms you might choose, there are multiple outstanding ones for you to solve the problem. After several attempts, we have better understanding of those algorithms we have used.