Titanic Case: Machine Learning Report

Titanic Data Science Solutions Titanic sinking is one of the most infamous shipwrecks in history. "On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew." After this disaster, someone collected the data about almost all of the passengers' personal detail information and whether they survive in the end. The following codes are about some analysis of what sorts of people were likely to survive. And I applied some tools of machine learning to predict which passengers survived the tragedy.

In order to analyze the data, the first step is to take a look at the data dictionary. There are 12 variables in train.csv. 1. Survived (0 = Not survived, 1 = Survived) is the output that we need to predict about. 2. Pclass: (Ticket class (1 = 1st class, 2 = 2nd class, 3 = 3rd class) means passenger's ticket class. 3. Name: Passenger Name. 4. Sex: passenger gender (male / female). 5. Age: Age in years. 6. SibSp: number of siblings / spouses aboard the Titanic. 7. Parch: number of parents / children aboard the Titanic. 8. Ticket: Ticket number. 9. Fare: Passenger fare. 10. Cabin: Cabin number. 11. Embarked: Port of Embarkation C = Cherbourg, Q = Queenstown, S = Southampton. 12. Passenger ID：is not useful variable for prediction of survival.

Secondly, I dig into data to figure out the relationships between survival and each variable. For the numerical variables, I utilized "train_df[['variable', 'Survived']].groupby(['variable'], as_index=False).mean().sort_values(by='Survived', ascending=False)" to figure out the relationship between them. Pclass, Sex, Sibsp, Parch, and Embark are numerical variables. For instance, the output of the Pclass looks like the following table:

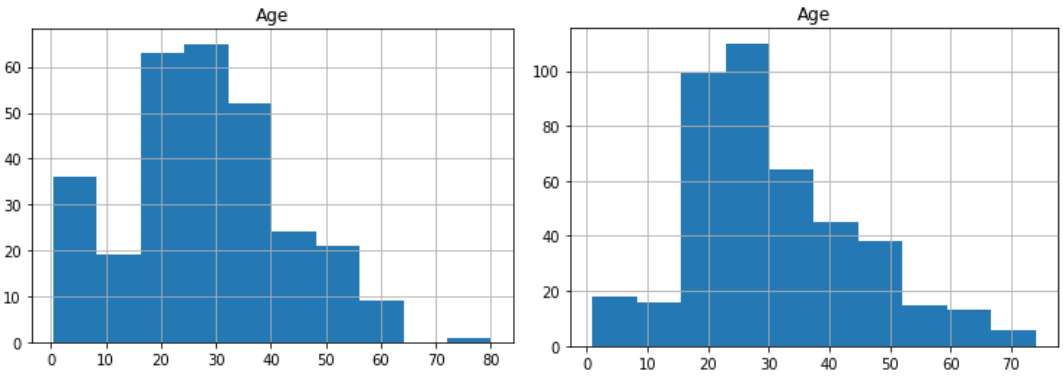
	Pclass	Survived
0	1	0.629630
1	2	0.472826
2	3	0.242363

As you can see, the survived column number represents the correlation between survival and Pclass. With the increase of the Pclass number, the correlation becomes less and less, which means there is some relationship between the Pclass and Survived.

Hence, after figuring out the correlation, we decide to keep these variables, which exist some relationships between it and survived, including Pclass, Sex, and Embarked.

Then, I utilized histogram to figure out the correlation between the categorical variables and survived. For instance, here is the output of the Age: According to compare two histograms of age, we can figure out the large part of the child (0 to 10) and old people (older than 75) are survived. Hence, the

age variable is relevant to the survived, which means I will keep this variable in the last model building.



(left one is the histogram about the name of people who survived in the end, right one is one of people who did not survived). Meanwhile, the histograms of fare also show some relationship with survived. So, I also keep it for the model building.

The last step is the model choosing and comparing. There are only 5 variables I kept for the last step, including Pclass, Sex, Age, Fare, and Embark. For model choosing, because this case is a classification and regression problem, we want to identify some relationship between the survived and other variables or features. Hence, for narrow down the choice of models to a few I tried these three models: Stochastic Gradient Descent, Decision Tree, and Random Forest. Then, after train our model, I build up a table about the ranking of evaluations of these three models. We can see the ***Random Forest*** is the highest one, and I chose it for the final model. And submit my answer. Finally, I gained the final score as 0.78947.

	Model	Score
1	Random Forest	97.98
2	Decision Tree	97.98
0	Stochastic Gradient Decent	75.08

Here is the link of my Kernel page: https://www.kaggle.com/mqzhong/titanic-data-science-solutions

Here is the link of my Colab page:

https://colab.research.google.com/drive/1CGpyzRtQ2pDgb4ieKEsagMceTV84dueN