Titanic Disaster Analysis

By Deep Blue Team:

Gloria Sun

Jim Lu

Dennie Tan

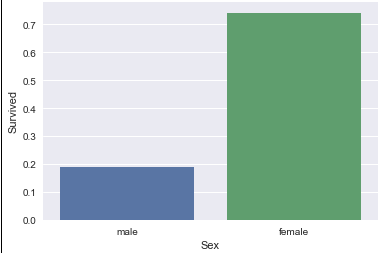
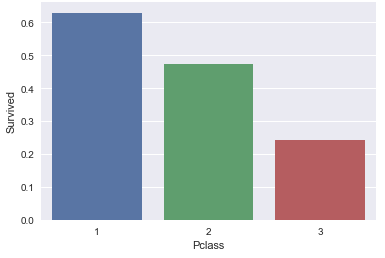
**Dataset of the Sad Story:**

We run the code to get the shape of both training and testing dataset on the raw data. The train dataset has 891 rows and 8 columns. The test dataset has 418 rows and 8 columns.

**Analyze the features:**

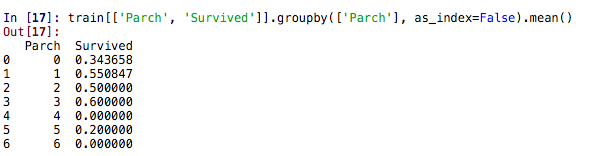
According to the raw data, the survival rate is 38.4%, and the non-survival rate is 61.6%.

Just based on guessing, there is a 61.6% non-survival rate without applying any algorithm. However, the non-survival rate and survival rate should be approximately 80% and 20%.

The two histogram show above are Gender and Ticket Class variables. We can use it to visualize how the sex and Pclass related to survival rate. Female has higher survival rate than male, and first class has higher survival rate than second and third class.

Parents and Children

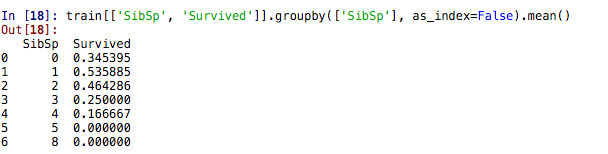


The above figure shows the relationship between survival rate and children or parents travel with each passenger.

The survival rate for 3 Parch is the highest (60%).

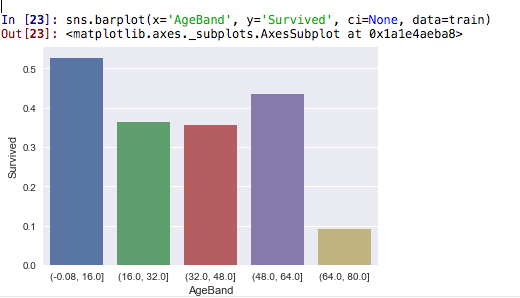
The survival rate for 6 Parch is the lowest (0%).

Siblings and Spouses:



The above figure shows the relationship between the survival rate and SibSp (spouse or siblings travel with each passenger). The survival rate is the highest (53.6%) if the passenger travels with one spouse. The survival rate is lowest (0%) if the passenger travels with 5 and 8 spouse.

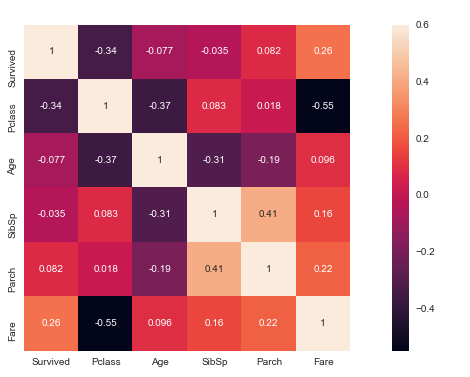
Does age make a difference:



Ageband: the age is segregated into five groups.

The first group (age 0 to 16), has the highest survival rate. Younger passengers have higher survival rate. The oldest passengers don’t have enough strengths to run and catch up the boat so that they have lowest survival rate.

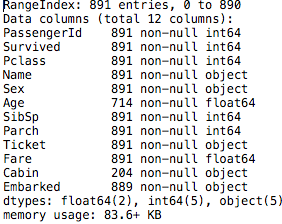
Correlation:



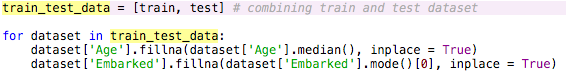
All 5 variables data type are integers, can’t over 0.7. If is over 0.7, it is redundant.

**Complete the dataset:**

We will start the cleaning process.

****

Using the code: train.info(), we can get the figure above including the data type for each features. Age, Cabin, and Embarked have missing data.



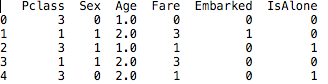
The above figure shows the code we used to fill the missing data for Age and Embarked. Since Cabin is not useful, it is removed and we didn’t fill the missing data for it.

**Convert the dataset:**



The figure show above is the code we used to convert Sex and Embarked to numeric value. In Sex, female is converted to 1 and male is converted to 0. Similarly, in Embarked, S = 0, C = 1, Q = 2. The object data type can’t be converted so all object data needs to be dropped. In this case, we dropped PassengerId, Name, and Ticket columns. We also dropped Cabin because it does not have enough information.

**Prepare the dataset:**

****

These are the features we choose for the algorithms, total 6 features. All features have either integer or float data type so that they can be converted to numeric value.

IsAlone means whether a passenger is with family or not.

**Models**

We pick the following models: logistic regression, decision tree, and random forest.



DT = Decision Tree

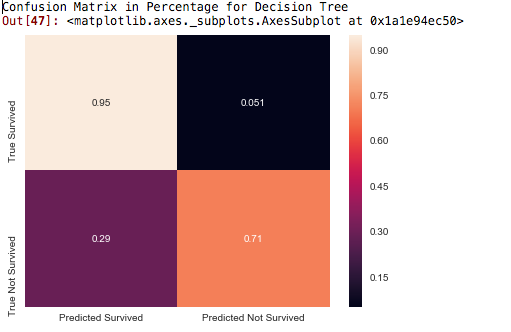
RF = Random Forest

According to the model evaluation: DT has 85.63 accuracy, RF has the same accuracy score as the DT, and LR has 79 accuracy score.

**Confusion Matrix**

We used the training and testing datasets to build all of the three contribution matrix.

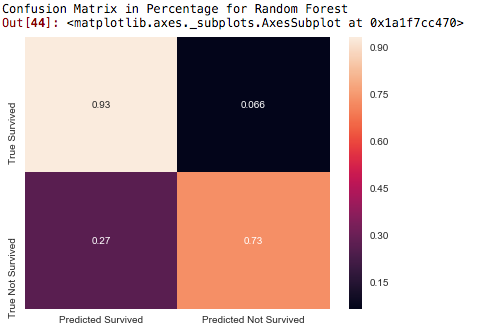
Decision Tree



Accuracy = = = 0.830

The predicted positive survival rate vs. the actual positive survival rate is very close to 100%. The true negative not survived rate is 0.71, which is a slightly lower than the true negative from random forest algorithm.

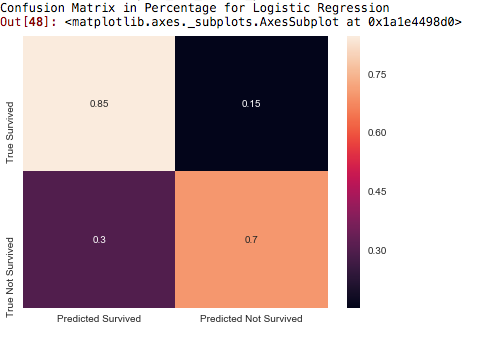
Random Forest



Accuracy = = 0.832

The true predicted survival rate is 0.93. The true predicted non-survival rate is 0.73.

Logistic Regression



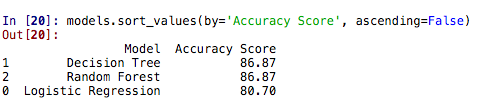
Accuracy = = 0.775

The logistic regression algorithm has the lowest accuracy score compare to decision tree and random forest. Although the predicted survival and non-survival rate are 0.85 and 0.70 respectively, they aren’t as good as the rate from decision tree and random forest.

**Improve the models:**

The way to improve the models:

1. Add or drop features



1. Try different algorithm. For example, KNN.
2. Collect more training data if possible.