

Titanic - Machine Learning for Disaster

Brief Description of the Data Set

In this project, we get two sets of data: training data and test data. The main objective of this project is to study the training data and make predictions based on our study, then fit it into the test set and submit it on Kaggle. The variables in both data sets are the same except in the training set we are given if the passenger survived or not. The table below describes all the variables and their meanings.

Variable Name	Variable Meaning
PassengerId	PassengerId's assigned to every passenger on board
Pclass	The socio-economic status of every passenger with 1 being the highest and 3 being the lowest.
Name	Names and titles of each passenger
Sex	The gender of each passenger
Age	The age of each passenger
SibSp	How many sibling(s) or spouse each passenger has with them
Parch	How many parent(s) or child(ren) each passenger has with them
Fare	How much each passenger paid for the ticket
Embarked	Where did each passenger abroad
Ticket	The ticket number of each passenger
Cabin	The cabin number of each passenger

We noticed that most of the variables can be used as factors. For example, the variable 'Sex' has the factor of male and female and 'Pclass' has the factors 1, 2, and 3. However, there are some variables that cannot be used as factors, for instance, the variable 'Name' has individual names for each passenger and 'Ticket' is a string of numbers assigned to each passenger on board. Based on this observation, we decided to 'clean up' the raw data given to us.

Initial Observations and Feature Engineering

Name

First we want to change the name of every passenger to just their titles (such as Miss, Mrs, etc.) We noticed that 'Master' is used to refer to males that are younger than 12, and 'Miss' is used for young females and females that are not married. Additionally, the title 'Mr' refers to males that are older than 12, and 'Mrs' refers to married females. There are also titles like 'Rev', 'Lady', 'Sir', and 'Countess'. We decided to only keep the title 'Rev' since there are more than 3 of them in the training set, and for titles like 'Lady', 'Sir', and 'Countess', we categorized them as 'Others' since there are not a lot of them with these titles.



Figure 1

From Figure 1, it appears that the majority of those who survive are 'Master', 'Miss', and 'Mrs' from Pclass 1, and 2. We want to further investigate if the number of siblings/spouse and parents/children play a role in their survival. Figure 2 and 3 below graphs Pclass with 'SibSp' and 'Parch', we can see that most passengers in Pclass 1,3 did not bring more than 2 siblings/spouse or 2 parents/children.

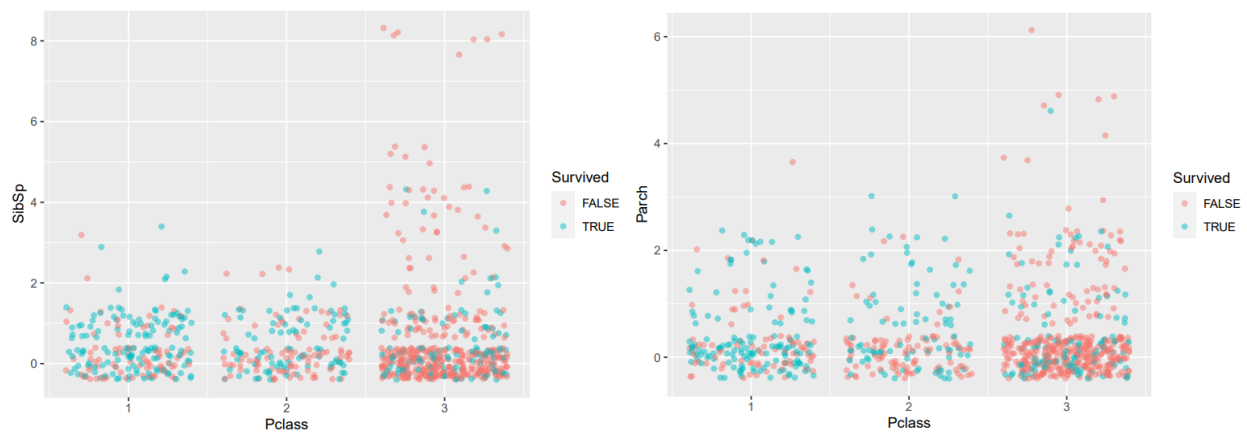


Figure 2

Figure 3

Age

From our observation, we found that 'Age' does play relatively significant role in passenger survival. The younger the passenger was, the more likely they have survived the ship wreck. For those that were missing an age, we took the average age of what was already in the dataset and added that to all the missing instances of "Age".

Fare

We also wanted to see how 'Fare' played into passenger's survival. From the figures below, we can see that passengers who are already in class 1 and paid more than 50 are more likely to survive.

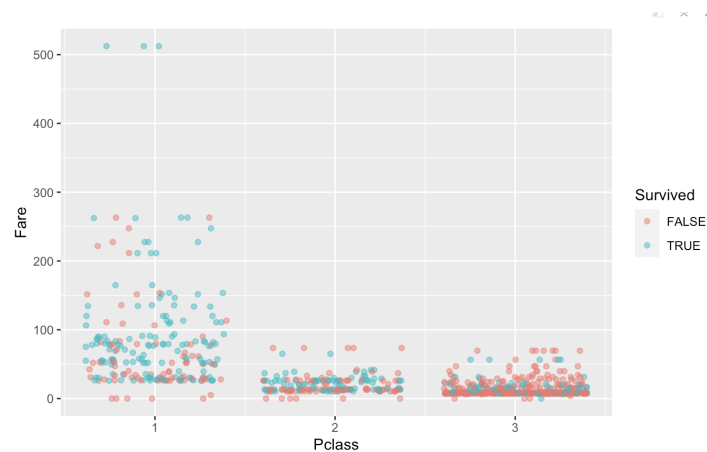


Figure 4

Removed Variables

We decided to remove variable 'Ticket', 'Cabin', 'PassengerId' because we believe those variables do not play a significant role for our model. In retrospect, we should have run cross-validation to ensure assumptions. That is something we need to improve on for our next project.

Additional Feature Engineering

To better predict the result, we decided to add three new variables: 'Companion', 'Rich', 'Prime'.

- 'Companion' combine every passenger's "SibSp" and "Parch" count.
- 'Rich' classifies passengers who are in Class 1 and paid more than \$50 for their tickets. (Shown in figure 5,6)

- ‘Prime’ classifies people with the titles “Master”, “Miss”, and “Mrs.” that are in “Class” 1 and 2. (Shown in figure 7, note that we converted the name titles to numeric values: Master = 0, Miss = 0, Mrs = 0, Mr = 1, Rev = 1, Other = 1)

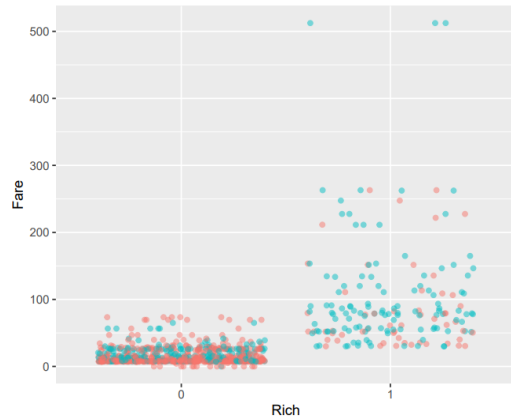


Figure 5

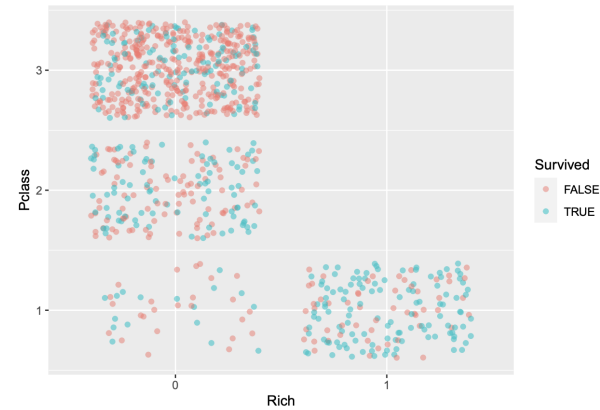


Figure 6

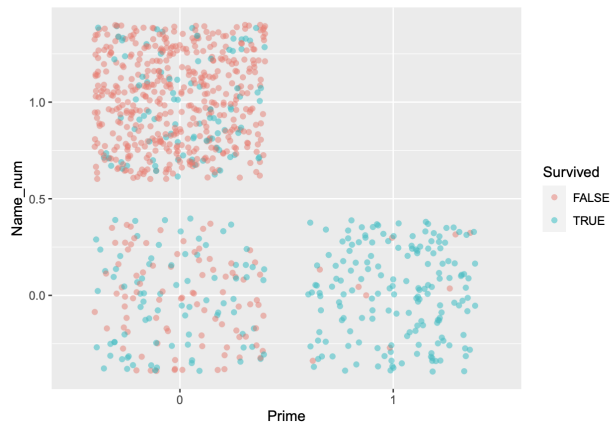


Figure 7

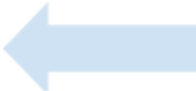
Fitting the Models into the Training Set, then the Test Set

To check our model accuracy, we split the training data into training set and validation set. We decided to split 70% of the training data into the training set(ship_train) and 30% of the training data in the validation set(ship_test_1).

We fit the model $\text{Survived} \sim \text{Prime} + \text{Rich} + \text{Age} + \text{Companion}$ into the validation set using 5 different methods: QDA, LDA, KNN, Naive Bayes, and Logistic Regression.

The accuracy of the models are shown in the graphs below:

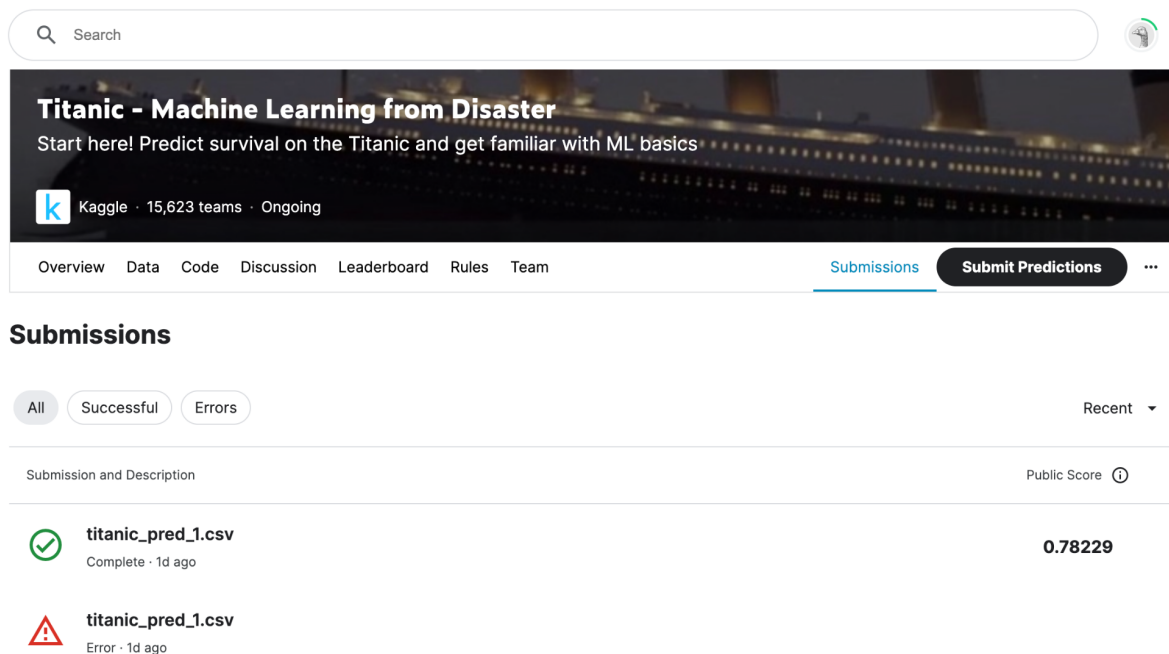
```
ship_test_1 %>% accuracy(truth = as.factor(Survived),  
  estimate = Survived_qda)  
  
## # A tibble: 1 x 3  
##   .metric .estimator .estimate  
##   <chr>   <chr>      <dbl>  
## 1 accuracy binary    0.786  
  
ship_test_1 %>% accuracy(truth = as.factor(Survived),  
  estimate = Survived_knn)  
  
## # A tibble: 1 x 3  
##   .metric .estimator .estimate  
##   <chr>   <chr>      <dbl>  
## 1 accuracy binary    0.729  
  
ship_test_1 %>% accuracy(truth = as.factor(Survived),  
  estimate = Survived_log)  
  
## # A tibble: 1 x 3  
##   .metric .estimator .estimate  
##   <chr>   <chr>      <dbl>  
## 1 accuracy binary    0.808
```



The LDA and Logistic Regression models were the most accurate.
We chose to use the logistic regression model to predict on the test set.

The Results

We submitted our final prediction on Kaggle and the score is 0.78229.





Titanic - Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 15,623 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules Team Submissions Submit Predictions

Submissions

All Successful Errors Recent

Submission and Description	Public Score
 titanic_pred_1.csv Complete · 1d ago	0.78229
 titanic_pred_1.csv Error · 1d ago	