Predicting the Survival on Titanic Disaster

This is a predictive machine learning project using R based on Kaggle competition: Titanic: Machine Learning from Disaster.

This is my first attempt at Kaggle. This work is influenced by Becky Wang's and Amber Thomas's analysis.

1. Introduction

The sinking of the RMS Titanic 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. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as **women**, **children**, and the **upper-class**.

In this challenge, we are required to complete to predict which passengers survived the tragedy, using the tools of machine learning.

2. Data Overview

The data for this analysis can be downloaded by clicking here. The data has been split into two groups:

- training set (train.csv)
- test set (test.csv) The training set is used to build machine learning models. For the training set, the outcome (also known as the "ground truth") for each passenger is also provided. The test set should be used to see how well our model performs on unseen data. For the test set, the outcome is not provided. For each passenger in the test set, we use the model we trained to predict whether or not they survived the sinking of the Titanic.

We see that there are 1309 observations (891 from train dataset and 418 from test dataset.) and 12 variables.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	

fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

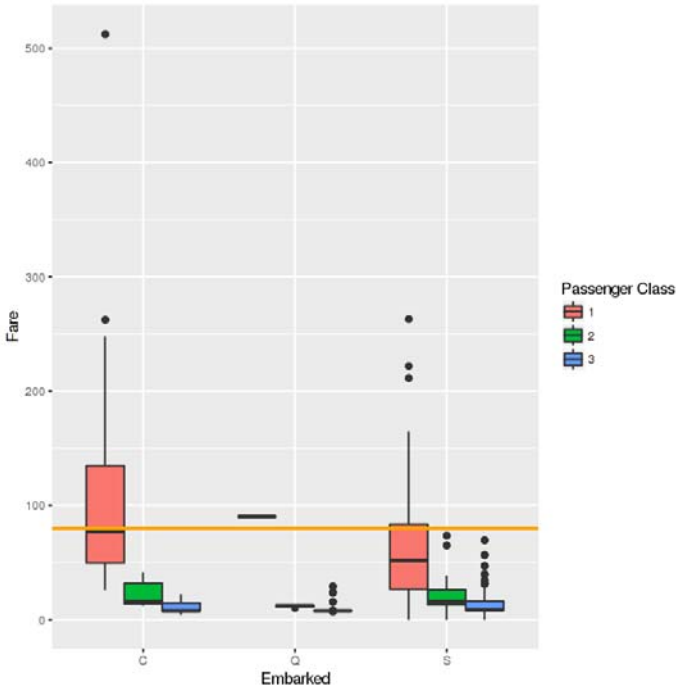
- pclass: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower
- age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- sibsp: The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)
- parch: The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

Missing Values

We see that Cabin has so many missing rows, but we think that this is not an important variable anyway so we ignore it. We then fill in the missing data in Age, Fare and Embarked columns.

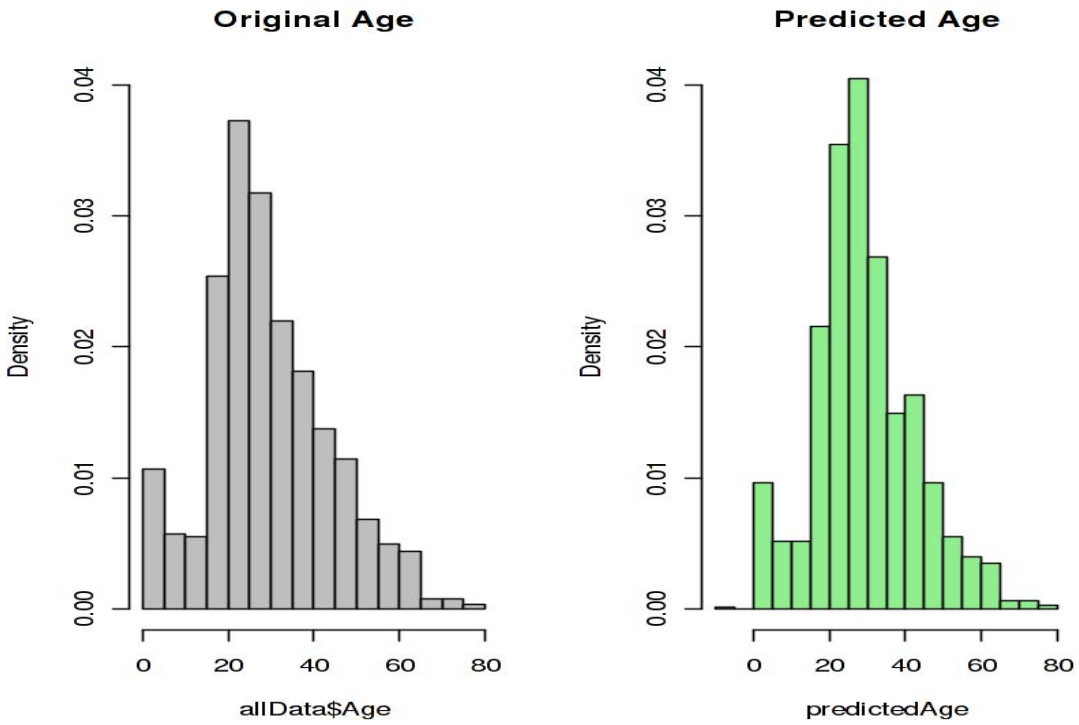
There was 1 missing fare and the passenger belonged to Pclass = 3. The median fare of PClass=3 passengers is: 8.05 which we used to fill in the missing fare.

There were two passengers with missing embarked. They are both first class passengers and paid $80. We see that the median of first class passengers is $80 for those who embarked from C. We can thus assume that both these persons embarked from C also.



Various methods can be used to fill age columns, for example, we can use mean value or we can use a random sample. But here we prefer to use a linear model for age prediction.

The distribution has changed a little bit, but it still looks fine. So we can go ahead with it.



3. Data Visualization

In this section, we shall try to visualize our data and try to find out the impact of various factors on the survival rate.

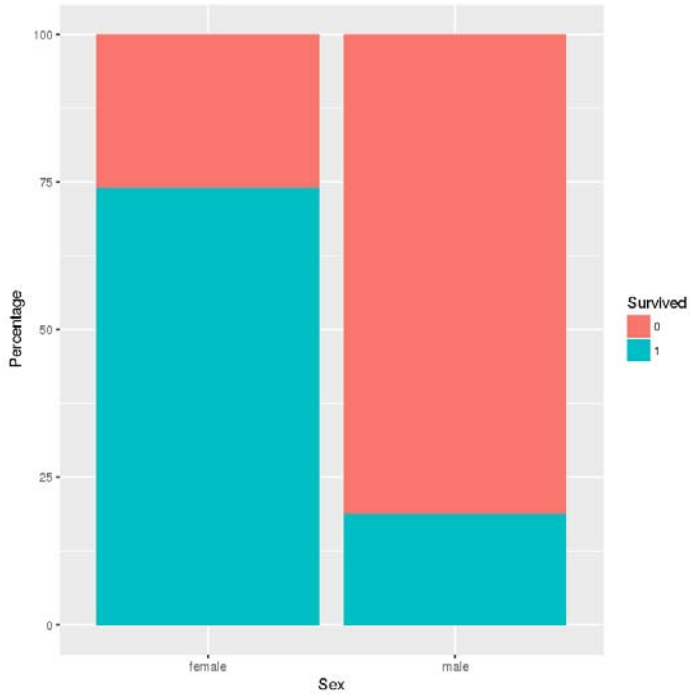
We shall not use absolute values for these graphs and use percentages instead. We think that it is easier to visualize the impact when the data is in percentages.

We first check the percentage of person surviving the disaster:

The percentage of people surviving the disaster: 38.38%.

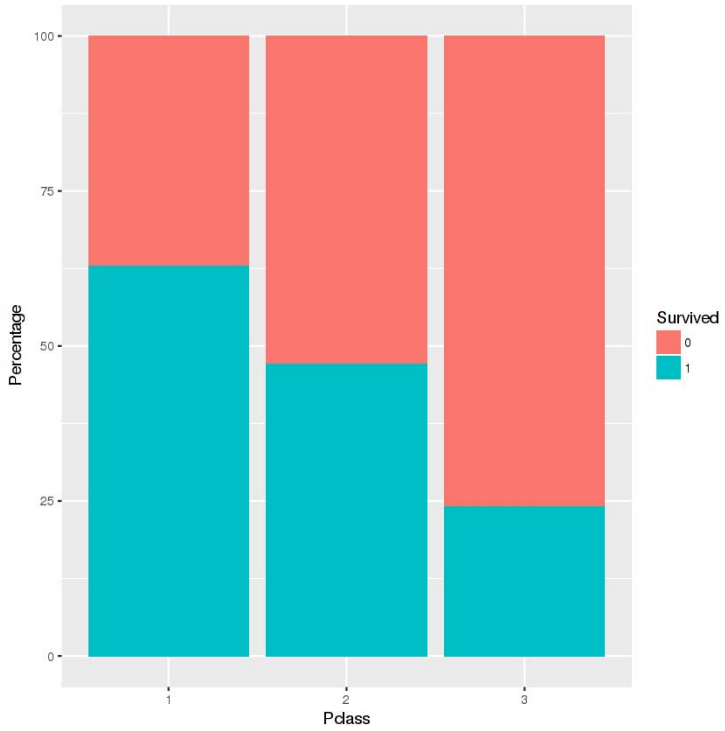
3.1. Gender vs. Survival

We can clearly see from the histogram above that female's survival rate is greater than the average survival rate (38.38%).

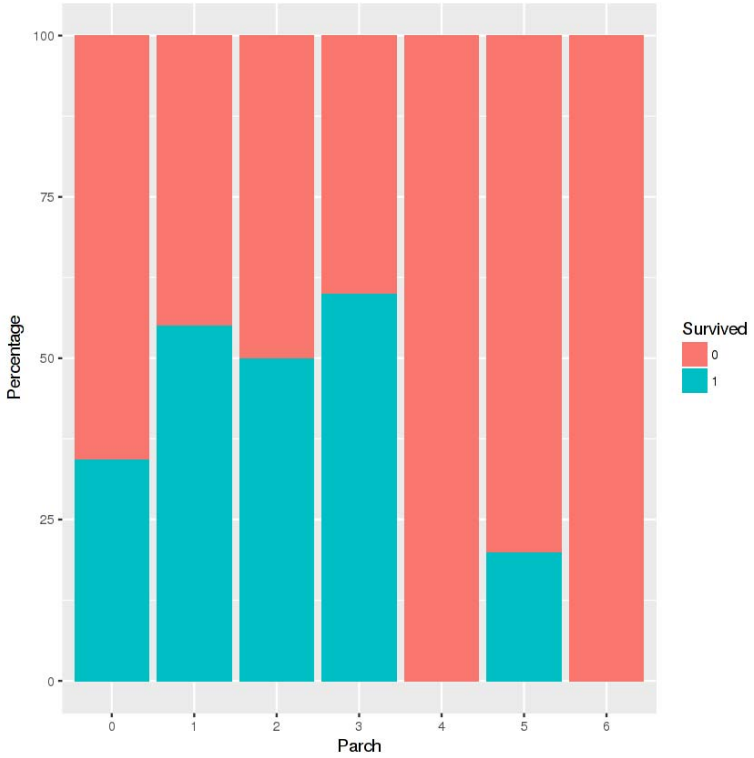


3.2. Passenger class vs. Survival

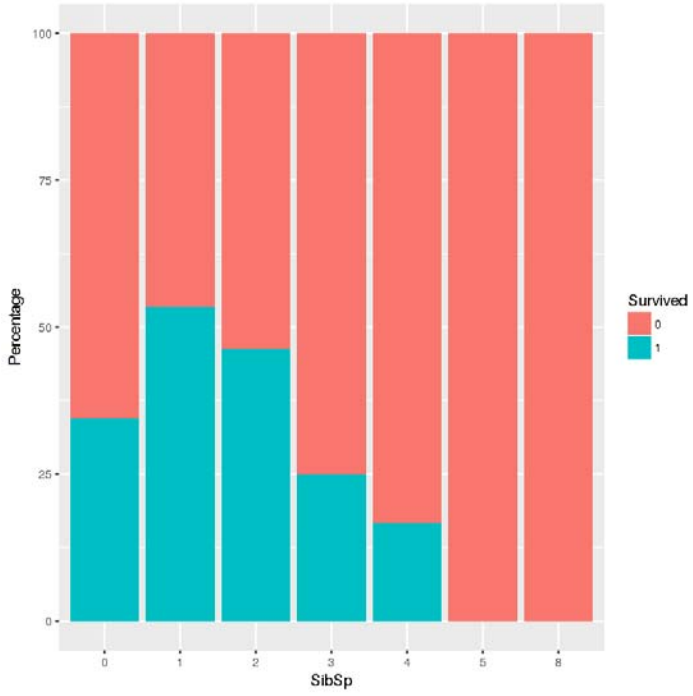
From the histogram, we notice that Pclass = 1 group has the highest survival rate. Pclass = 3 group has the lowest survival rate within these three groups, and it is even lower than the average survival rate of 38.38%.



3.3. Family vs. Survival

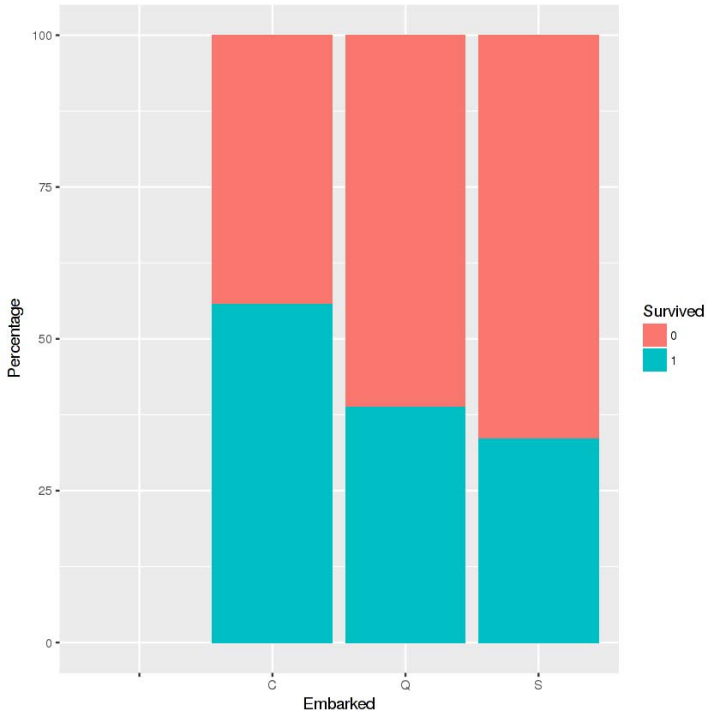


In [22]:



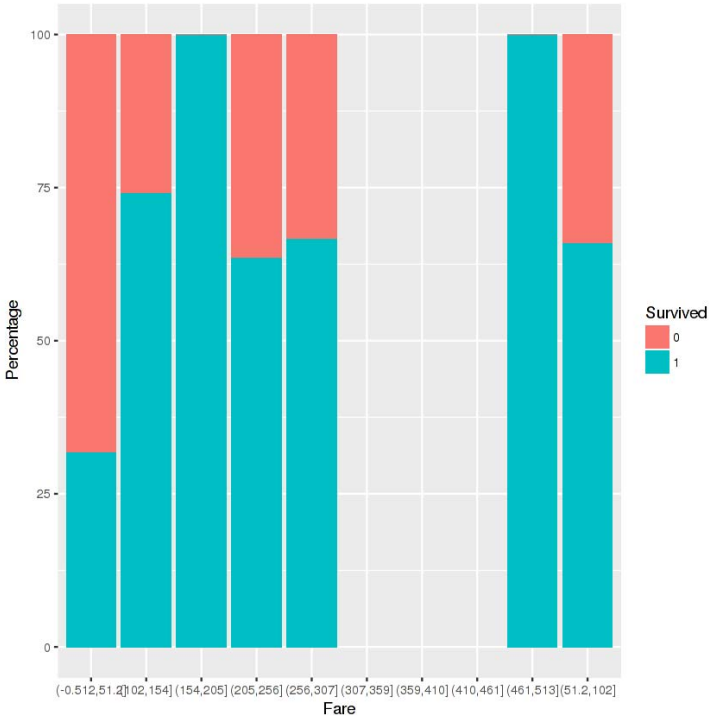
3.4. Place of Embarkment vs. Survival

We see that people who embarked from Cherbourg have a greater chance of survival.



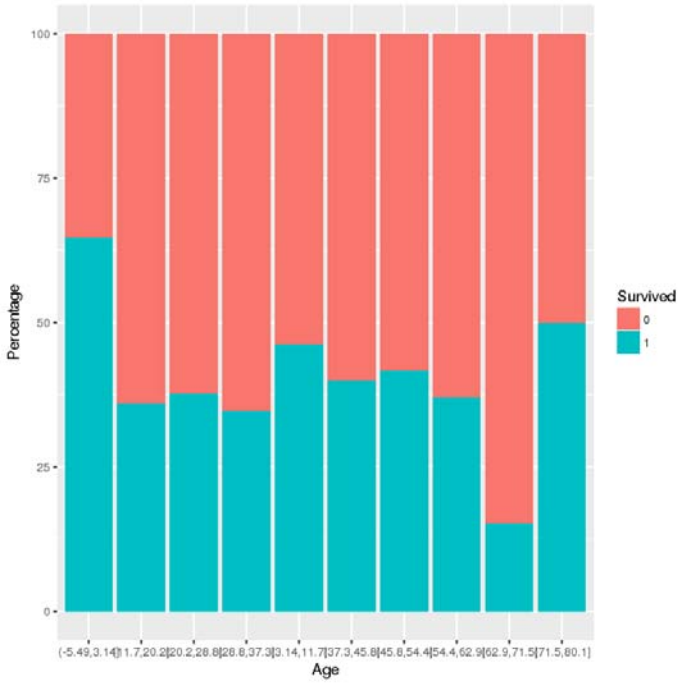
3.5. Fare vs. Survival

We see that people with low fare have a lower chance of survival.

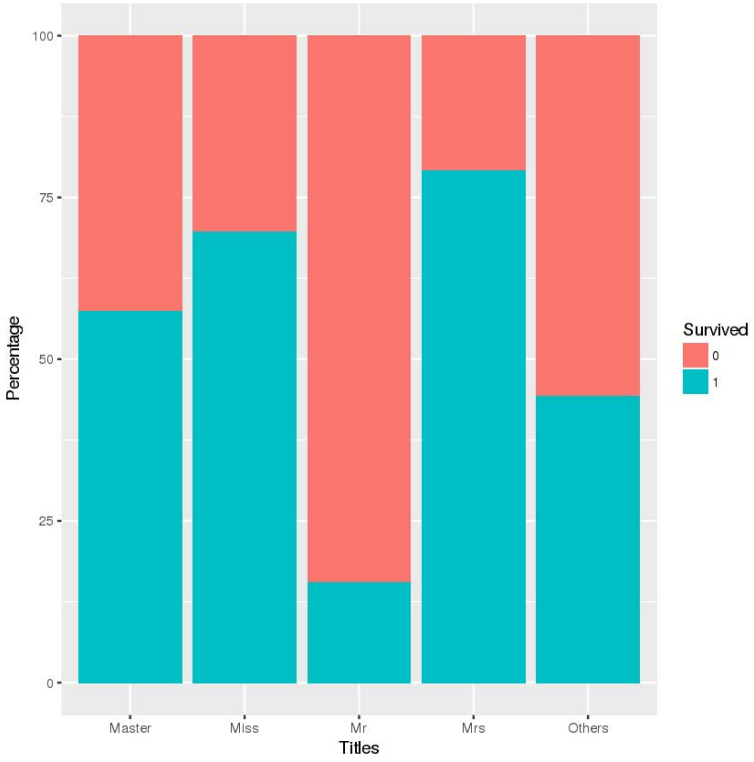


3.6. Age vs. Survival

We see that younger children have a higher rate of survival.



3.7. Titles vs. Survival



4. Data Modelling

We first choose our features that have significant effect on survival according to the visualization above.

Survived is our dependent or response variable, whereas Age, Titles, Pclass, Sex, Parch, Sibsp, Fare, and Embarked are our independent variables.

We consider common machine learning model such as Logistic Regression, Decision Tree, Random Forest and SVM.

We measure the model accuracy using the formula given below:

Accuracy = (True Positive + True Negative)/(All outcomes)

Model	Accuracy (%)
Logistic Regression	83.73
Random Forest	83.05
Decision Tree	83.95
SVM	83.84

5. Prediction of Survival

Since all models have similar prediction accuracy, we can use all of them to predict the survival from Titanic disaster:

Model	0	1
Logistic Regression	251	167
Random Forest	277	141
Decision Tree	262	156
SVM	257	161