STA 4990 Project 1

Kyle Folbrecht, Anthony Mastrangelo, Michael Pena

March 21, 2023

Overview

Given a manifest of the titanic, our objective is to figure out who survived the famous catastrophe. We are given many factors such as the obvious factors of their ticket fare, their age, and their gender, since we figure according to the famous "women and children" first that they were the contributing factors. However, more obscure factors that were being considered such as their incomes and where they embarked on their journey from, were much more useful in order to sort the people who survived from those who didn't. The independent variables given were:

- **pclass** = Ticket Class, 1 = 1st Class, 2 = 2nd Class, 3 = 3rd Class
- **Sex** = Male or Female
- **Age** = Age in years, infants are decimals and unknown ages have the form xx.5
- **sibsp** = number of siblings/spouses the individual had on the titanic
- **parch** = number of parents/children the individual had on the titanic
- **ticket** = Ticket number
- **fare** = Passenger fare
- **cabin** = Cabin Number
- **embarked** = Point of Embarktion, where C = Cherbourg, Q = Queenstown, S = Southampton.

Initial Analysis

It is worth mentioning that a lot of the data had many holes in it, so we decided to separate the data with multitudes of holes. The one with many factors in it was cabin since most of the individuals were missing a cabin number. The cabin levels went from A-G, so we separated each level by its respective cabin level. If they belonged to a specific level, the value for the level was a 1, otherwise, it was a 0.

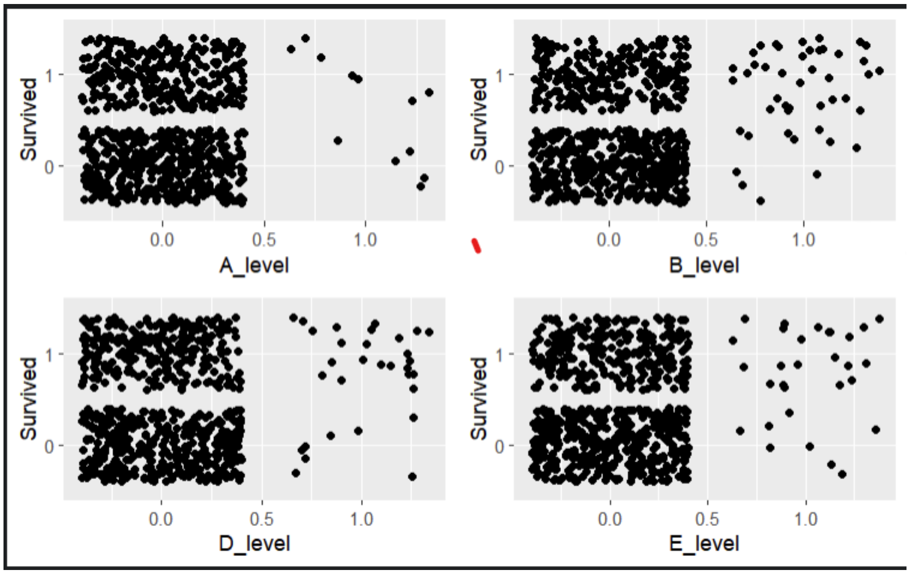


Figure 1: Survival based on Cabin number

Then we looked at how all the variables interacted with our dependent variable of survival. There were a couple of variables that stood out as having a clear interaction between survival. Objectively, the most important of these variables was the ticket fare, as it was one of two numerical variables in the dataset. The graph below shows that if a woman paid more than 200 dollars for their ticket, they survived the cinematic catastrophe. Furthermore, if the person spent more than 300 dollars in general, they survived.

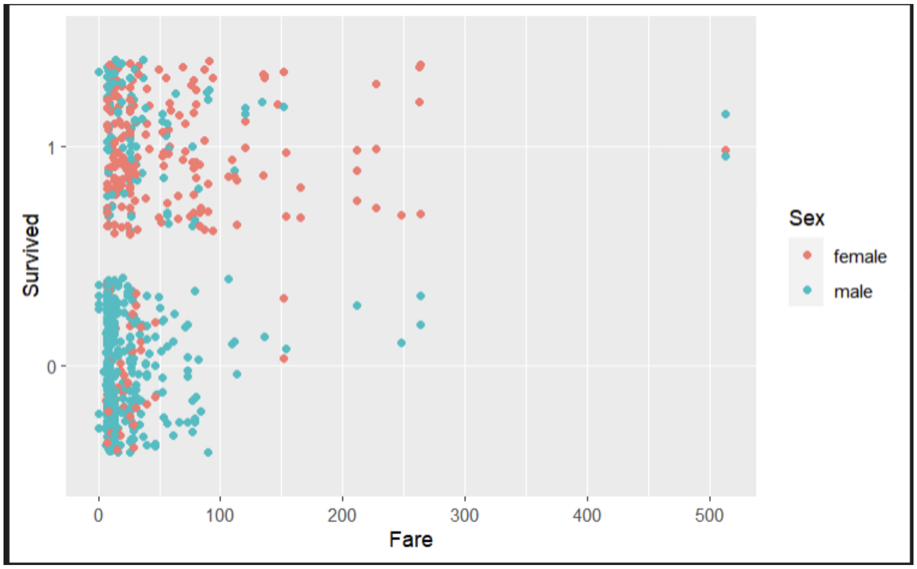


Figure 2: Survival based on Ticket Fare

The amount of siblings/spouses also seemed to have a very strong factor, but in the opposite sense. According to the data below, if an individual had more than a composite number of 4 siblings and spouses, then they were not going to survive the Titanic. If the individual was a woman, then they were more likely to survive if they had more than 2 siblings and spouses than if they were a man, as the proportion of women who survived was much higher in that group than the amount of men, as only 3 men survived if they had more than 2 siblings/spouses.

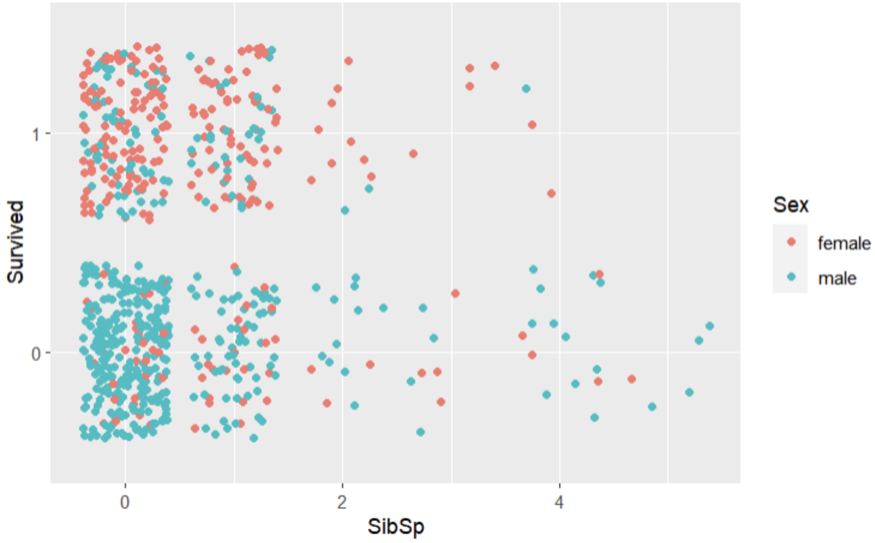


Figure 3: Survival based on Siblings and Spouses

The final variable that had a strong correlation between survival was the number of parents/children the individual had. If a person was to have more than 4 cumulative parents and children on the Titanic, then they were more than likely going to die. Furthermore, the likelihood of survival with more than 2 parents and children was much higher in women than it was in men, with only 4 men surviving in this demographic. In fact, if any men had no parents or children on the titanic with them, they were incredibly likely to die on the famous ship.

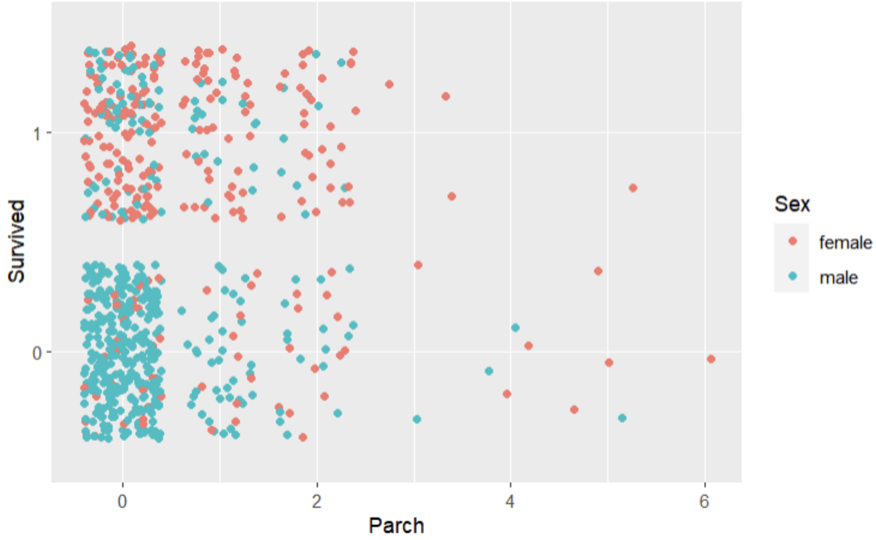


Figure 4: Survival based on Parents and Children

Modeling Methods

Naturally, our sample was taken from Kaggle. Kaggle offered two files consisting of train and test sets. We took Kaggle's train set and split that data, partitioning 70% into a new set of training data and 30% into a new test set. It was from this new split data that we would work to find an optimal model for this project and submit for a Kaggle score.

The optimal model would then be retrained using the entire training set and be tested with the test set given by Kaggle.

Our approaches consisted of logistical, K-nearest neighbor, LDA, Naive Bayes, Lasso, Ridge, and Forward Selection methods to then go about building $\hat{y}$ and probability columns into the data frames for all models in the data frame, allowing us to test our predictions (QDA was not processing, forcing it to be left out of the chosen methods). We used LOOCV as training control on each model but forward selection due to its long processing time to actually create a model.

Because in this situation neither a false negative nor a false positive is worse than the other one, the accuracy metric would be the best way to evaluate our data.

Feature Engineering

We wanted to see if there was any relevancy basing passengers off their Ticket Class and embarkation. For this, we were looking for a correlation between two predictors. We found that there was a correlation between Ticket Class, "Embarked" and "Survived." In addition, the correlation between peak class and fare showed that there was some correlation, and we created a variable for it that would represent whether a passenger paid above average or below for their Ticket Class.

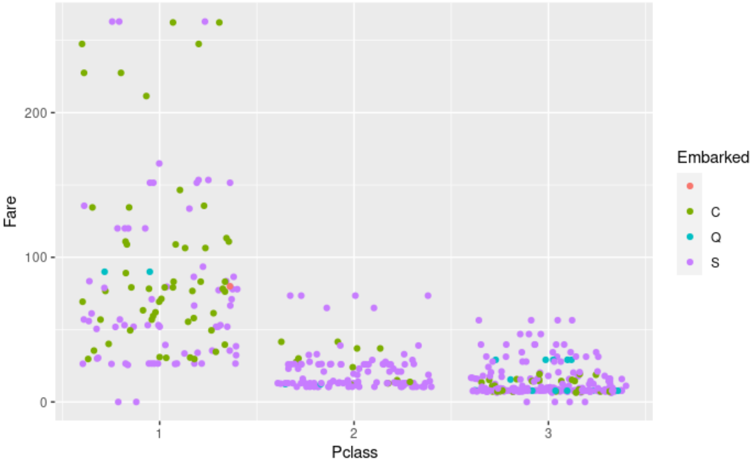


Figure 5: Correlation between Fare and Ticket Class

Using figures 5 and 6, we were able to include a variable that tells us if a passenger pays more than the average price for their ticket. As seen in figure 5, Ticket Class 1 seems to have wider range of fare prices while also being the most expensive. In figure 6, the right graphic seems to show that higher fares indicate more likely survival. We also see that Ticket Class 3, which seems to span the cheapest fares, is associated with more deaths.

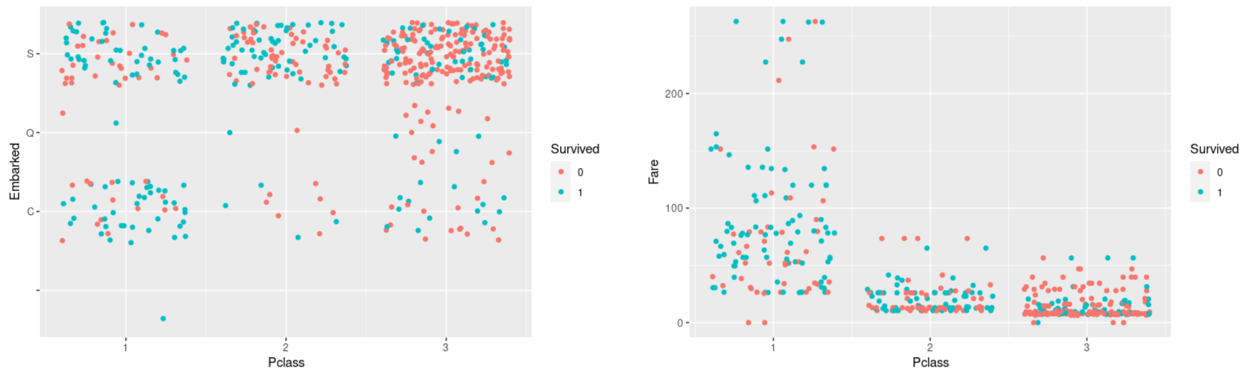


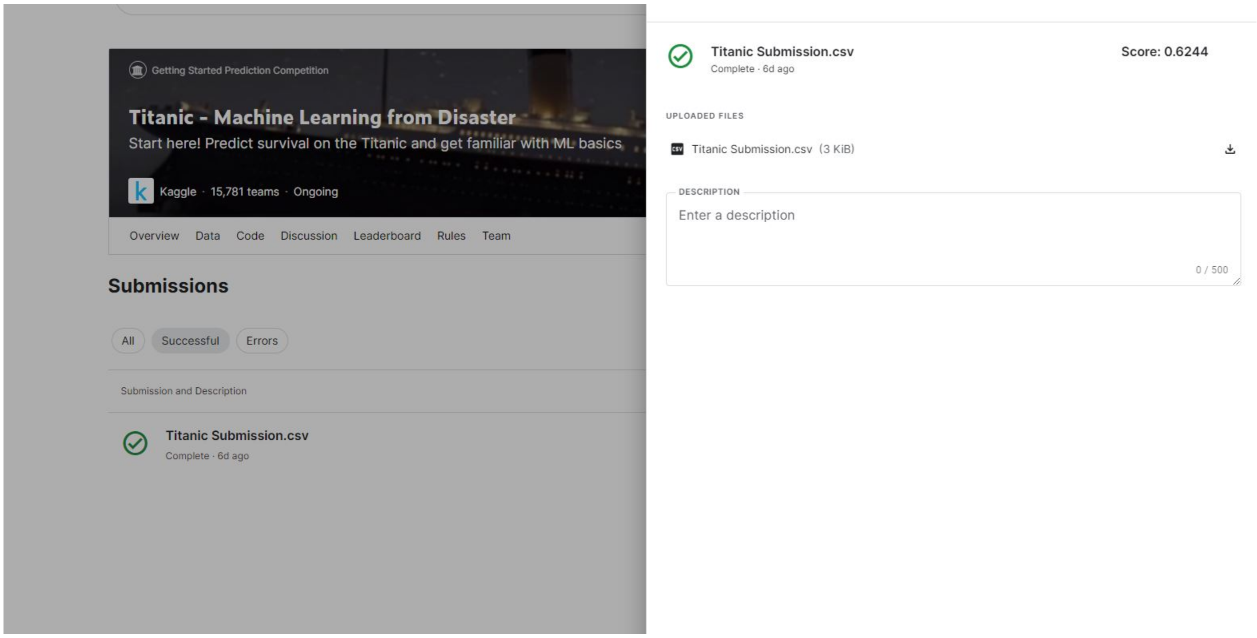
Figure 6: **(Left)** Correlation between Point of Embarkation and Ticket Class with Survival color coded. **(Right)** Correlation of Fare amount to Ticket Class with Survival color coded.

Results

The models with the best accuracy score in the training set were also the same three models that had the best accuracy score in the test set. Notice that LDA's performance was comparable to forward selection, even though forward selection did not having training control. If forward selection performed significantly worse than the other models we would have tried to run it again with training control but based on the results below we did not find it worthwhile.

Model	Train Accuracy	Test Accuracy
Ridge	0.828	0.832
Logistic	0.818	0.804
Forward Selection	0.816	0.808
LDA	0.814	0.808
Lasso	0.778	0.785
Naive Bayes	0.734	0.738
k-Nearest Neighbor	0.732	0.696

The model that we decided to use to make our prediction with was our Ridge model because it performed the best on both the training and test set. Also with the large amount of variables we were working with ridge's penalty would help the model not overfit. This model would be tested against the Kaggle test set and would be submitted to Kaggle.


Getting Started Prediction Competition
Titanic - Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Kaggle · 15,781 teams · Ongoing
Overview
Data
Code
Discussion
Leaderboard
Rules
Team
Submissions
All
Successful
Errors
Submission and Description
UPLOADED FILES
DESCRIPTION
Enter a description
0 / 500

Titanic Submission.csv
Complete · 6d ago

Titanic Submission.csv Score: 0.6244
Complete · 6d ago

Titanic Submission.csv (3 KiB)

Figure 7: Kaggle submission results

Conclusion

The score we received on our final model was not as accurate as we were hoping, in fact it is only marginally better than assuming everyone perished. The greatest contributor to this was missing data, each time a passenger that we were predicting was missing some data on the prediction would automatically be NA. Each prediction with the value NA would automatically be wrong because it would not match any possible values of the binary variable Survived.

One of the largest improvements to our model would be some way to fill in the gaps of data so that we could use our model on each passenger rather than only on the passengers who were not missing any data. If we filled in these gaps we could also train our model on more passengers, which could potentially change the model, and would also be able to predict on each variable rather than for some passengers having no way to attain the correct prediction other than random guessing.