Titanic: Machine Learning from Disaster

Predicting Survival Based on Women and Children Groups

Kush Bhatt (10147050) – Michael Lee (10124458) – Matthew Vanderwey (00211874)

1.0 Introduction

WHITE STAR LINE
BOARDING PASS
PERMISSION GRANTED TO COME ABOARD
WHITE STAR LINE'S
R.M.S.
TITANIC
ISMAY, IMRIE & CO.,
34, LEADENHALL STREET, LONDON,
AND
10, WATER STREET, LIVERPOOL

The RMS Titanic was marketed by its owner's, the White Star Line, as unsinkable. However, at approximately 20 minutes before midnight on 14 April 1912 the ship struck an iceberg that would result in its sinking just a few hours later. More than 1,500 people perished in what was the greatest maritime disaster to date. This represented almost 2 out of every 3 people on board [1].

The goal of this machine learning competition, hosted by Kaggle, is to see if we can use what little information is known about each passenger to predict their survival. For this competition Kaggle provides a training set of 891 passengers and a test set of 418 passengers, roughly a 32% test-split. The data includes limited information about each passenger like their name, age, gender, passenger class, etc.

Before setting up our machine learning model, we first explored the available data to see what patterns of survival (if any) could be established. This process is known as Exploratory Data Analysis, or EDA, and it helped us also review the overall quality of the data. Including which data points were missing and which features were, or were not, useful in our modeling. We then investigated how to impute missing values in the features we wanted to use, re-factoring the training data, and engineering new features to make it easier for our models to predict survival.

A cursory review of the work done by past competitors showed promise in the use of Extreme Gradient Boosting (XGBoost) [2] and ensemble [3] based models. Therefore, we looked at the application of both forms of modelling in our investigation. We also looked at the application of the Receiver Operator Characteristic (ROC) metric [4] and precision recall curves in evaluating the performance of our models.

To condense this report to the required five-pages, all charts and figures have been relegated to Appendix A – Figures & Tables. For the reference of the reader, the python based Jupyter Notebook built for this project is also attached here as Appendix B.

2.0 Exploratory Data Analysis

Kaggle provides the data for this competition split into a training and testing set. However, to more accurately evaluate the characteristics of the passengers as whole population, we merged the training and testing data into one master data set.

Before we engaged in any feature engineering or machine learning we needed to explore both the patterns of survival and its correlations with the provided features. We started with looking at the split between those who survived and those who did not. The percentage of passengers in training data who survived is 38.4%, which is close to the 31.6% of total passengers and crew who survived [5].

One of the root causes that made this such a tragedy was that fact that there were simply not enough lifeboats on the ship for everyone to escape in. At the time the number of lifeboats that a ship was required to have was based on its gross tonnage and not the number of people it carried. Thus, the Titanic was only legally required to have 20 lifeboats when it needed at least 48 for the number of passengers aboard. Even more tragically it is estimated that the ship could have accommodated up to 64 lifeboats, but that this was cut back for esthetic and cost reasons [6].

The crew, knowing full well that there were not enough lifeboats, would have prioritized loading women and children first. So, for our analysis we will create a new feature called "Title" which will break the population into three categories, "woman" which will include both women and girls, "boy" for underage males and "man" for adult males. Figure 1 in Appendix A shows us the split between those who survived and did not amongst these three groups. Of the women and boys in the training data 72.3% survived. However, of the adult males only 16.0% survived. So, it appears that our earlier assumption stands and that there is substantial difference in the survival rates of women and children vs. adult males.

While 1912 may have been a chivalrous time, it was not an egalitarian one. Even more so than it is today, back then a person's importance was measured by their wealth and social standing. This leads us to the hypothesis that First-Class passengers were more likely to survive than Second-Class, and Second-Class passengers were more likely to survive than Third Class. In the training set the percentage of passengers who survived from First, Second and Third Class was 63%, 47% and 24% respectively. Thus, we can comfortably state that passenger class is a meaningful factor in predicting survival. A visual of this breakdown is provided in Figure 2 in Appendix A.

The dataset provided by Kaggle includes two categorical features that may be used to yield some insights into the composition of the groups traveling together on the ship. These features are 'SibSp' and 'Parch'. 'SibSp' represents the number of siblings and/or spouse that an individual was traveling with, and 'Parch' represents the number of parents and/or children that a passenger was traveling with. If we add these two features together (plus 1) we get the number of members in each family that were traveling together. Again, knowing that there were not enough lifeboats on the Titanic, and assuming that families would be reluctant to split up, we can reasonably hypothesize that large families would be less likely to survive. We can test this hypothesis by first calculating the size of each family and plotting that against survival. Figure 3 in Appendix A demonstrates that families with between 2 to 4 members were by far the most likely to survive. Those travelling on their own, or in families of 5 or more, were substantially less likely to survive.

The last categorical feature that we investigated is the point of Embarkation, 'Embarked'. Passengers onboard the Titanic that fateful night had boarded at either Cherbourg 'C', Queenstown 'Q', or Southampton 'S'. Looking at the data (see Figure 4 in Appendix A) it does appear that passengers who embarked at Southampton were more likely to perish. However, this is deceptive as vast majority of passengers embarked at this port. With such a grossly uneven distribution there is no way to make a meaningful conclusion regarding the relationship between embarkation and survival. Therefore, we left this feature out of our modeling.

After investigating the relationships between the categorical data and survival, it was time to look at the correlation between the continuous numerical features and survival, namely 'Age' and 'Fare'. However, before we could this, we needed to address the fact that the values for 'Fare' were per-ticket and not per-passenger. For example, The Taussig family were all traveling under Ticket No. 110413, but they did not each pay 79.650, but instead that was the cost for all three family members. Please see Table 1 in Appendix A for this example.

Figure 5 in Appendix A shows that there is some correlation between 'Age' and 'Survival', but an even stronger correlation between the per-person fare ('FareAdj') and 'Survival'. We investigated this a little deeper and reviewed the distributions of ages and adjusted fares for those who survived versus those who did not in Figures 6 and 7 in Appendix A. Figure 6 shows us that there is a substantial difference in survival for those under 15 years old. Figure 6 shows us that those who paid under 15 for their ticket were more likely to perish and those whose paid around 25-30 were more likely to survive.

So far, our EDA had helped us make the following determinations regarding survival:

- Primary Factors:
  - Adult males were the least likely group to survive
  - Women and children were the most likely groups to survive
- Secondary Factors:
  - Passenger class and the size a family travelling together were significant factors in survival
  - Age and the per-person fare paid were moderate factors in survival
- Features we can ignore:
  - There are simple not enough data-points for cabins to make any meaningful determination with them as over 77% of the values are missing
  - The poor distribution of embarkation points makes this feature also unsuitable for our analysis

These determinations were important because they indicated that if we could find adult males who survived, and conversely women and children who perished, we should then be able to make accurate predictions regarding survival in general. We had at this point also demonstrated that we should be able to use passenger class, family size, per-person fare and age to help us find these groups of rare individuals.

These determinations formed the basis of the feature engineering and machine learning which we applied in the following sections. However, before we could move forwards, we needed to first look at imputing missing values in the provided dataset. Details of how this was done are outlined in Appendix B, but in short, we used a Decision Tree Regressor (DTR) to impute the missing values in Age and Fare. We choose to do this as a more accurate method as opposed to simply taking the mean or median values. This should also avoid introducing any skewing into our data.

3.0 Finding the Surviving Men

After isolating adult males from the training set, we generated the Kernel Density Estimation (KDE) plots found in Figure 8 in Appendix A. These plots show the distributions of Age, Per-Person Fare ('FareAdj'), Family Size and Passenger Class between survivors and non-survivors. These plots show us that we are most likely to find surviving adult males at Age 30-40, a per-person fare of ~20, Family Size of 2 and in First Class.

To help us find patterns where all these features intersected, we first reduced them down to two dimensions. This made it possible to plot their relationships in 2D. For our analysis we chose to create two new features, 'x1' which is the Per-Person-Fare/10 (to scale the value) and 'x2' being the Family Size + Age/70. We did not include the passenger class in either 'x1' or 'x2' because there is already a strong correlation between the Per-Person-Fare and passenger class, as shown in Figure 9 in Appendix A.

By creating feature 'x2' we could view the 'SibSp', 'Parch' and 'Age' features all in one dimension. Creating 'x1' allowed us to see the impact of individual ticket costs (which are also a function of passenger class) on a similar scale as 'x2'. Looking at Figure 10 in Appendix A, we saw a cluster of surviving men around x2<=2.5 (Family Size <= 2 + Age < 35) and 2.5 < x1 < 3.5 (per person fare of 25 to 35) which would put those passengers in First Class. This aligned with what we found in Figure 8 earlier.

Before we dove into making predictions on the survival of adult males using XGBoost, we first tested if such a model could in fact find the patterns that we had already established. We did this by training a XGBoost model on our engineered 'x1' and 'x2' features and then made predictions, a lot of them. But not on the test data, instead on a large linear series points that coincided with the ranges of values for 'x1' and 'x2' where we expected to find survivors.

The exact implementation is explained in Appendix B, but in summary we made 10,000 predictions on survival using the 'x1' and 'x2' features and our XGBoost model. Since we had just two dimensions, we could test the performance of our model by superimposing our training values over our 10,000 predictions to see if they aligned. This exactly what is shown in Figure 11 in Appendix A, with the predictions in light blue/orange and the training values in dark blue/orange.

Looking at Figure 11, it appeared that our model was struggling to identify Second or Third-Class survivors but had done a fairly good job of finding patterns of survival for the First-Class passengers (round orange dots). To begin tuning our model, we applied the Receiver Operator Characteristic (ROC) metric [4] to optimize both the maximum depth of the boosted trees and the probability threshold for positive results. We tried max-depth values of 3, 4, 5, 6 and 7. Finding that a max depth of 5 provided the greatest Area Under Curve (AUC), that is, the highest ratio of True Positives to False Positives.

A plot of this curve is shown in Figure 12 in Appendix A. This curve also helped us visualize at what probability threshold we are likely to obtain the highest level of accuracy with. We did this by looking at where the slope of the curve is steepest and then taking the inverse of the True Positive rate. Looking at Figure 12, the slope is steepest between the True Positive Rates of 0.1 to 2.5, so we then checked how tuning our probability thresholds between 75% to 90% (1 - 0.25 to 1 - 0.10) impacted our accuracy.

The exploration of this is detailed in Appendix B, but in summary using a 90% threshold for positive predictions yielded the highest accuracy score on our training data. It was safe to assume that using such a high threshold would have a dramatic negative impact on our recall score, and that was supported by Figure 13 in Appendix A, which shows that our model has huge trade-off between precision and recall. This means that we were less likely to find surviving males, but if we did, they were very likely to be correct predictions.

We also applied a 10-fold cross-validation to our model and received an average score of 82.51%, which gave us the confidence to move forward. So, after all this work, how many adult males in the test set does our super-model predict will survive? Zero…. While this was at first highly frustrating, it does agree with the findings of some of the highest rated public models on Kaggle, including the much-referenced Titanic Mega Model [7].

4.0 Women and Children Groups

With the spectacular disappointment that was trying to find surviving adult males, we moved on to feature engineering our Women and Children Groups (WCGs). During our EDA in Section 2.0 we demonstrated that, by a wide margin, women and children were the most likely to survive. This also tells us that the survival of both women travelling with their families and the survival of children are highly correlated. Therefore, we will build our WCGs around the following hypothesis:

- All boys survive in families where all the females and other boys survive; and
- All females perish in families where all the boys and other females perish.

The procedures involved in performing this feature engineering are all outlined Section 4 of Appendix B. In summary, identifying the family groups required generating unique keys based on Surname, Passenger Class, Ticket No., Fare and point of Embarkation.

A good example of how forming these WCGs will help our survival predictions is demonstrated with the Wilkes-Richards-Hocking family [8]. There were seven members in this family including four women, two boys and one adult male. Two of the women are in the test set and the rest of the family are in the training set provided by Kaggle. From the training set we know that the two women and boys survived, even though the adult male did not. If our hypothesis is correct, then the two women in the test set should also survive and looking at the list of survivors [9], they do.

The “ultimate” test of our hypothesis was to submit our predictions generated by our feature engineering to Kaggle for scoring. The file submitted was the attached *2020-03-30_WCG.csv* and the resulting score was **0.83732**, which indicated to us that we were on the right track. The challenge was then to identify the next rare group, single women who perished.

5.0 All the Single Ladies

After all the work that we put into building and tuning an XGBoost model which found no surviving adult males, we decided to take a different approach for trying to predict single females who perished. For this we built and tuned an ensemble consisting of a Logistical Regressor, a Random Forest Classifier (RFC), a Support Vector Machine Classifier (SVC), a Decision Tree Classifier (DTC) and a K-Nearest Neighbours (KNN) Classifier. We used 5-fold cross-validation to test the accuracies of each individual model and to tune the weights of both hard and soft-voting ensembles consisting of all five models. We also used the cross-validation scores to tune the max depth of the DTC and the n-neighbours of the KNN Classifier.

In the end we chose to use the soft-voting ensemble because this allowed us to also tune the probability threshold for our predictions. We tuned this threshold by directly submitting multiple sets of predictions to Kaggle for evaluation. In the end we received our best score using a 70% probability threshold for negative results, that is, the predicted probability for survival had to be below 30% for us to accept that any single female perished.

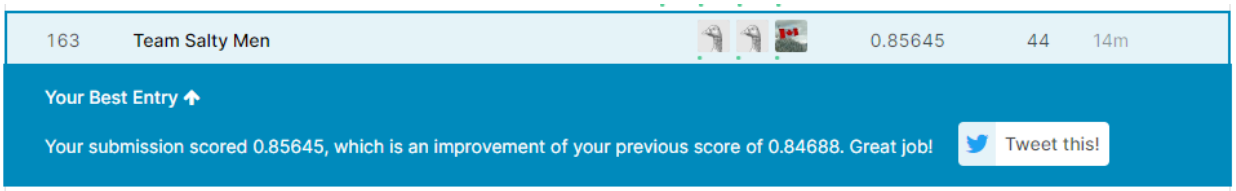
Our final predictions for this ensemble are included in the attached *2020-03-30_Ensemble.csv* which scored **0.84210** on Kaggle. As this was an improvement over our previous score, we can conclude that our ensemble did indeed find at least some single women who perished in the tragedy.

6.0 Conclusions

At the end of our journey, while we were able to obtain a top 200 score (out of 17,000+ teams) the actual use of machine learning itself to get there was limited. In fact, we could have likely achieved a score over 0.80 using purely feature engineering and basic statistics. As noted earlier, we are confident that the key to this competition is identifying the Women and Children Groups. Knowing that there are likely more groups such as the Wilkes-Richards-Hocking family in the dataset, further review of the data is likely to yield even more accurate WCGs than what we did here. From there, if we had more time, we would experiment with various ensembles for revising our predictions for both surviving adult males and single women who perished. Possibly using an ensemble of XGBoost, KNN and DTC as these seemed to perform best. This makes sense because most of the provide features are either categorical or of discrete numerical values.

7.0 Epilogue

After completing our predictions and taking machine learning as far as we could conceive, we conducted a further experiment in MS Excel. In a spreadsheet we compared our results with those of both the *Titanic Mega Model* [8] and *Titanic WCG+XGBoost* [2]. We treated the three sets of predictions as an ensemble itself for predicting which single females perished. From this we derived the attached *2020-03-30_Excel.csv*, which helped us achieve our best score on Kaggle of **0.85645**.



163 Team Salty Men 0.85645 44 14m

Your Best Entry ↑

Your submission scored 0.85645, which is an improvement of your previous score of 0.84688. Great job!

Whether or not this should really be counted as our best score, since it is based on not just our results but also those of others, is for the reader to decide.

8.0 Contributions

We all worked together on testing numerous models and methods of feature engineering before landing on our final selections. Each member of this team is comfortable with taking equal credit for the delivery of this final report.

References

[1] D. Fowler, "The History in Numbers", *titanicfacts.net.* [Online]. Available: https://titanicfacts.net/. [Accessed: 30 March 2020].

[2] C. Deotte, "Titanic WCG+XGBoost[0.84688]". *Kaggle.com*. [Online]. Available: https://www.kaggle.com/cdeotte/titanic-wcg-xgboost-0-84688. [Accessed: 30 March 2020].

[3] Y. Ghouzam, "Titanic Top 4% with ensemble modeling". *Kaggle.com.* [Online]. Available: https://www.kaggle.com/yassineghouzam/titanic-top-4-with-ensemble-modeling. [Accessed: 30 March 2020].

[4] Scikit-learn. "Receiver Operator Characteristic (ROC)". *Scikit-learn.org.* [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html. [Accessed: 30 March 2020].

[5] D. Fowler. "Titanic Survivors". *Titanicfacts.net*. [Online]. Available: https://titanicfacts.net/titanic-survivors/. [Accessed: 30 March 2020].

[6] D. Fowler. "Titanic Lifeboats". *Titanicfacts.net*. [Online]. Available: https://titanicfacts.net/titanic-survivors/. [Accessed: 30 March 2020].

[7] C. Deotte. "Titanic Mega Model – [0.84210]", *Kaggle.com*. Sec. B.4. [Online]. Available: https://www.kaggle.com/cdeotte/titantic-mega-model-0-84210#B4.-All-Adult-Males-Die. [Accessed: 30 March 2020].

[8] C. Deotte. "Titanic Mega Model – [0.84210]", *Kaggle.com*. Sec. B.6. [Online]. Available: https://www.kaggle.com/cdeotte/titantic-mega-model-0-84210#B6.-Meet-the-surviving-Wilkes,-Hocking,-Richards-family. [Accessed: 30 March 2020].

[9] D. Fowler, "Titanic Survivor List", *titanicfacts.net.* [Online]. Available: https://titanicfacts.net/titanic-survivors-list/. [Accessed: 30 March 2020].

Appendix A – Figures & Tables

Figure 1 - Survival by Title

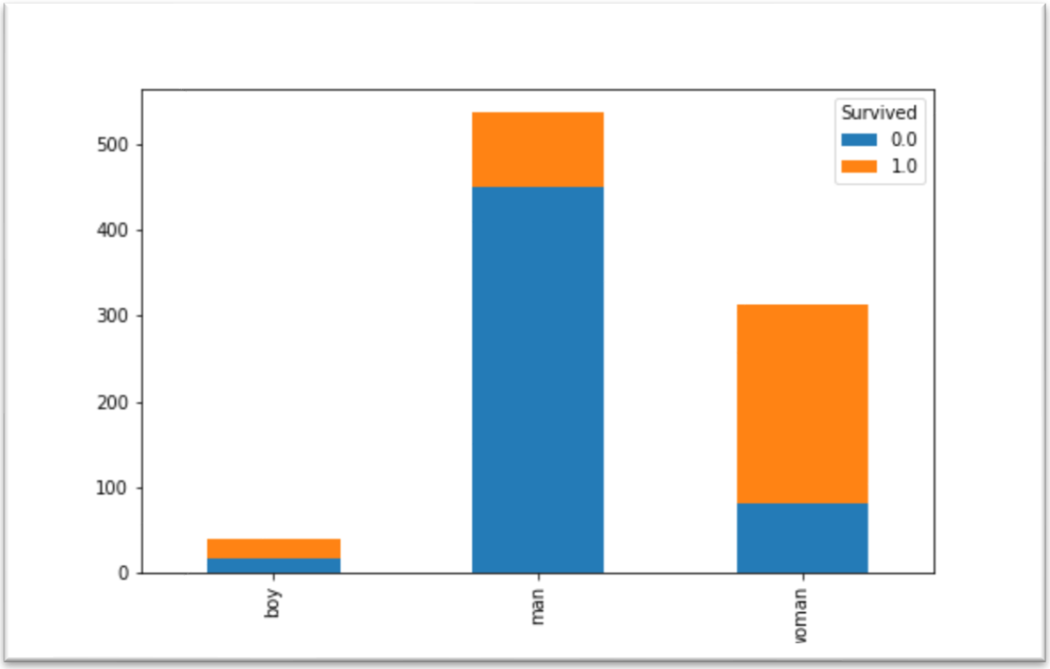


Figure 2 – Survival by Passenger Class

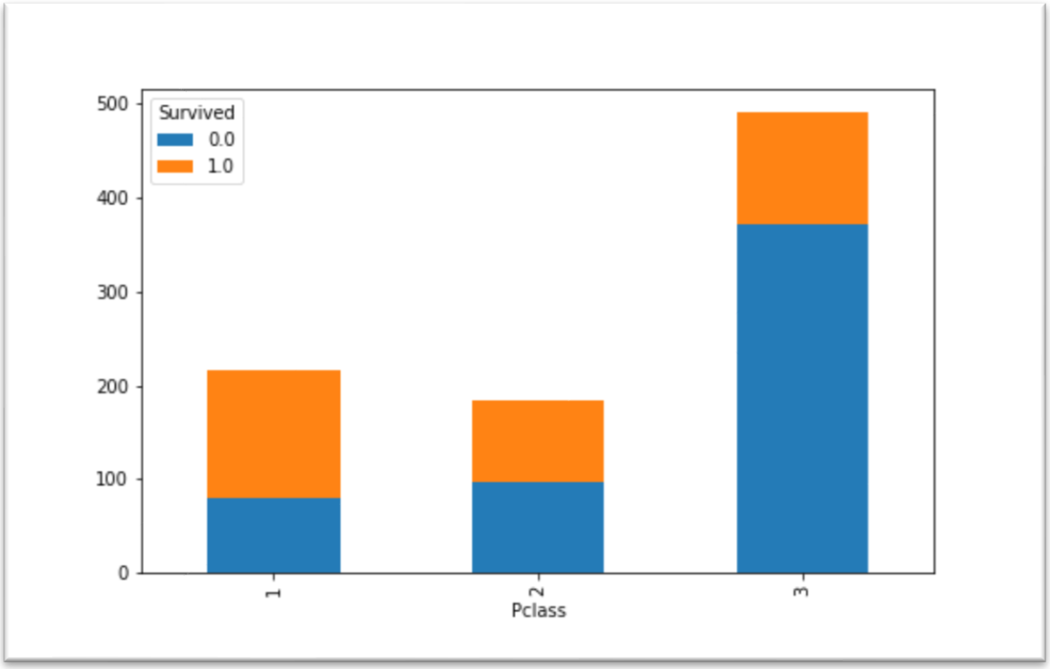


Figure 3 – Survival by Family Size

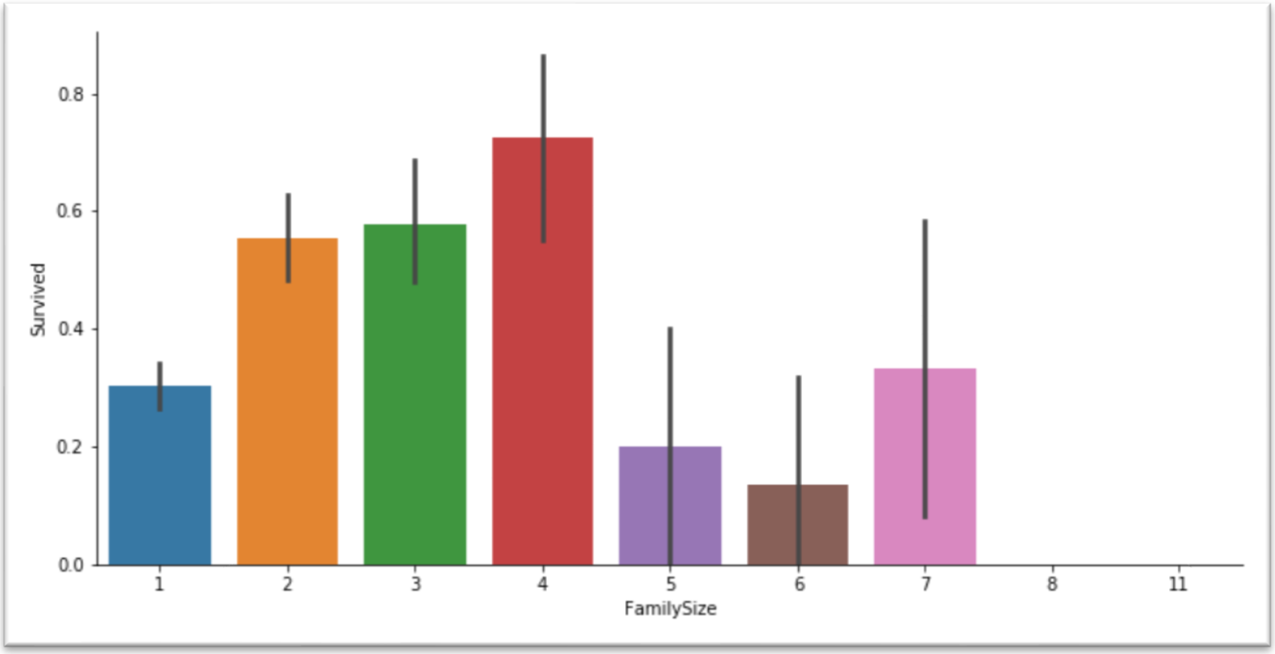


Figure 4 – Survival by Port of Embarkation

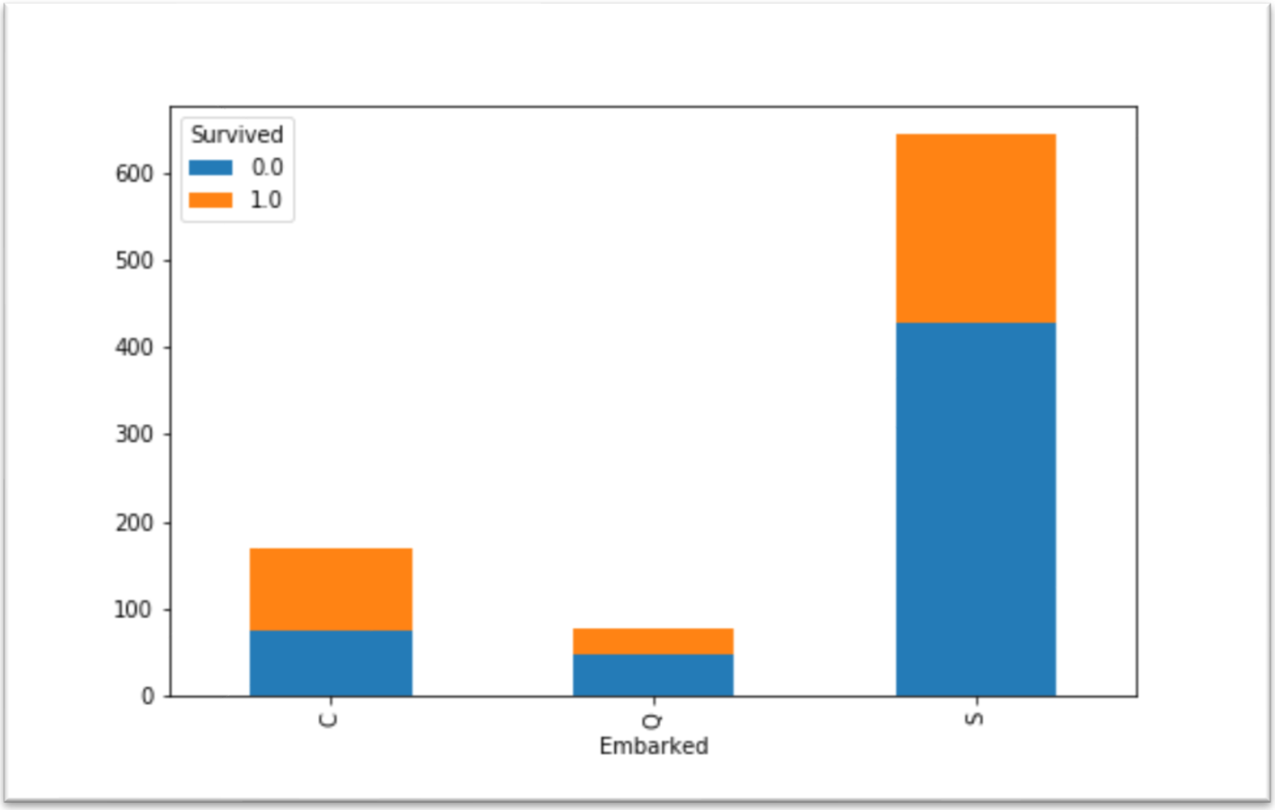
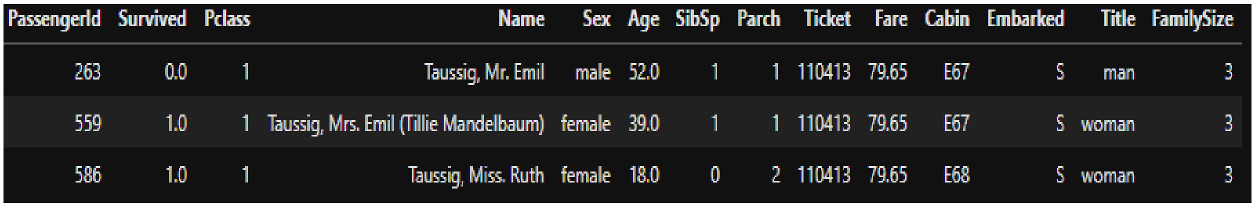


Table 1 – Example of Fare per Ticket



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	FamilySize
263	0.0	1	Taussig, Mr. Emil	male	52.0	1	1	110413	79.65	E67	S	man	3
559	1.0	1	Taussig, Mrs. Emil (Tillie Mandelbaum)	female	39.0	1	1	110413	79.65	E67	S	woman	3
586	1.0	1	Taussig, Miss. Ruth	female	18.0	0	2	110413	79.65	E68	S	woman	3

Figure 5 – Correlations Between Survival, Age and Fare

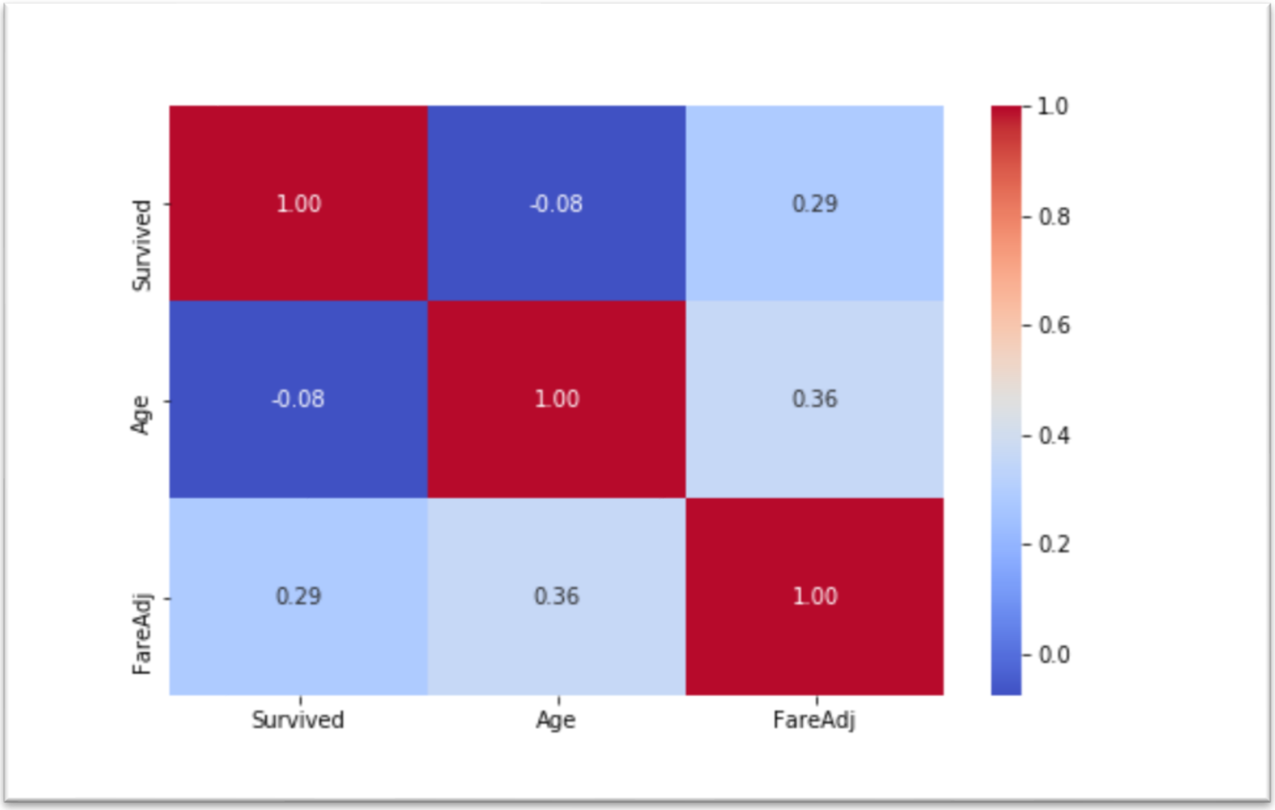


Figure 6 – Distributions of Age by Survival

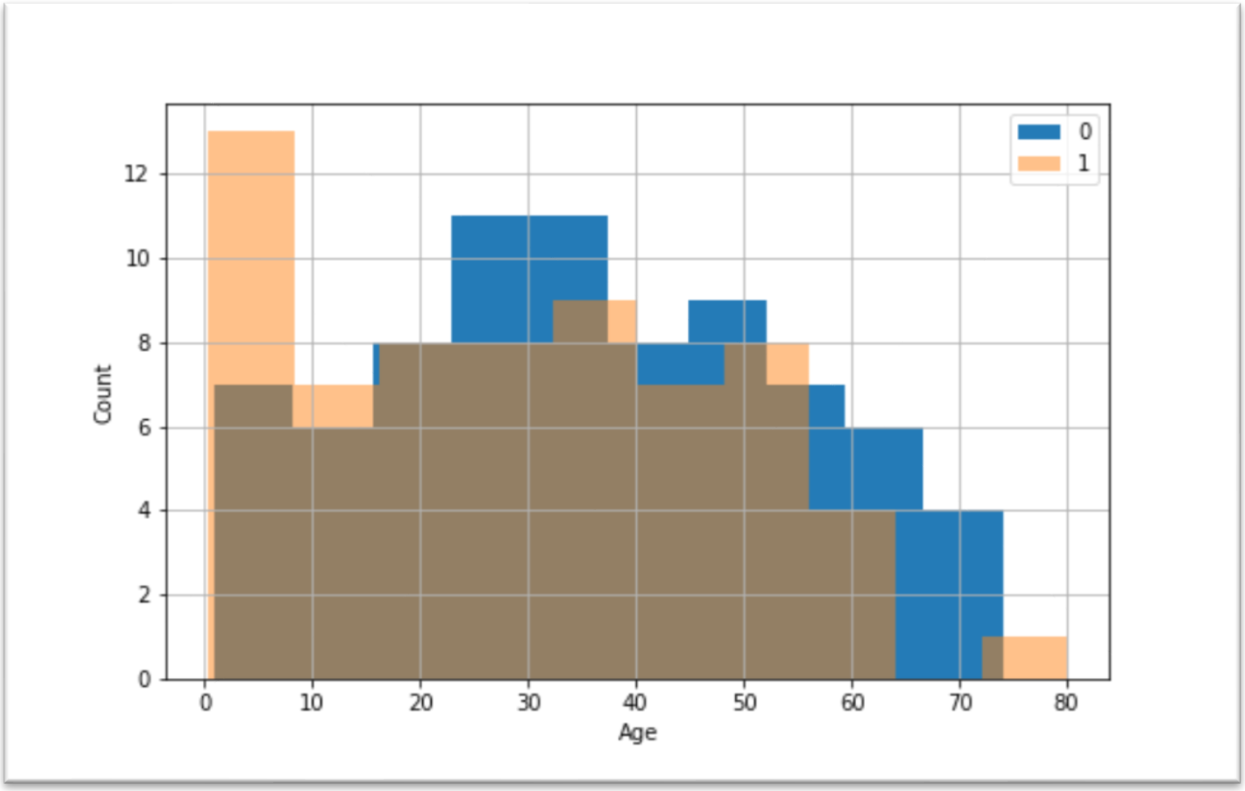


Figure 7 – Distributions of Adjusted Fares by Survival

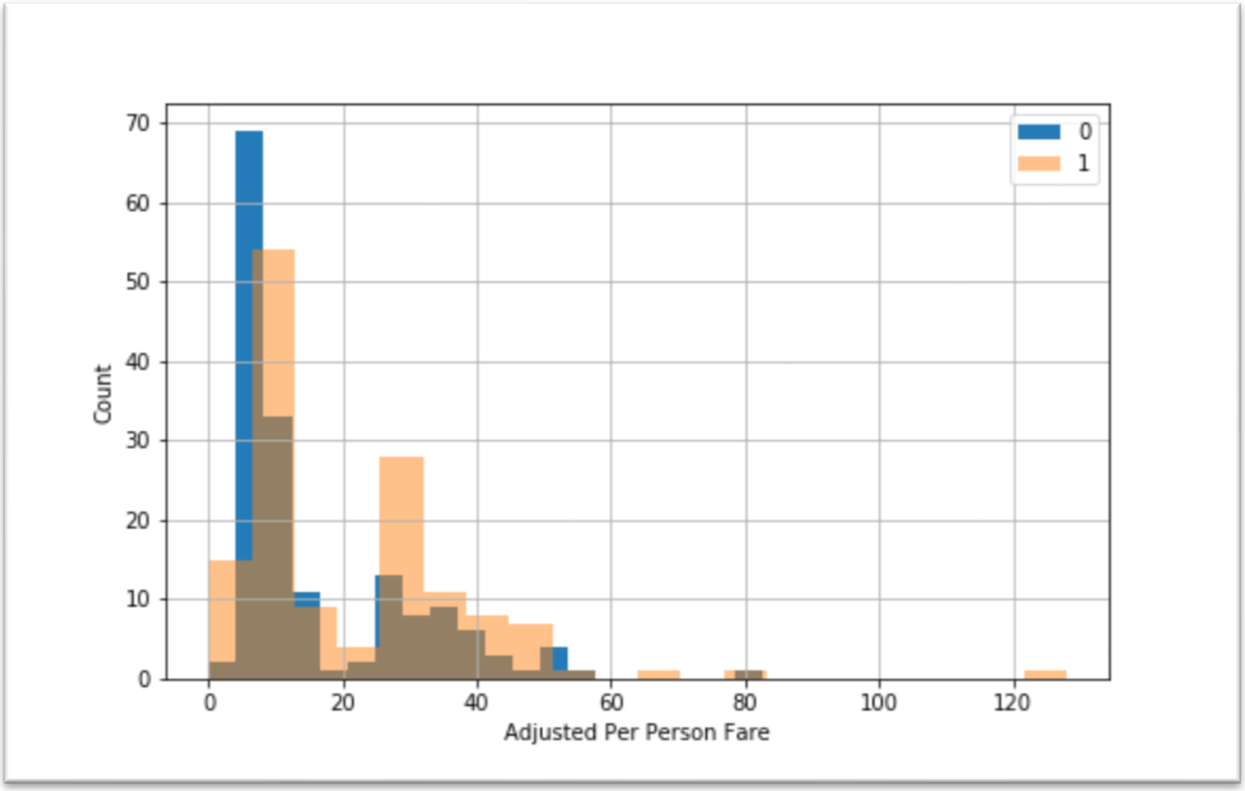


Figure 8 – KDE Plots of Survival Rates for Adult Males

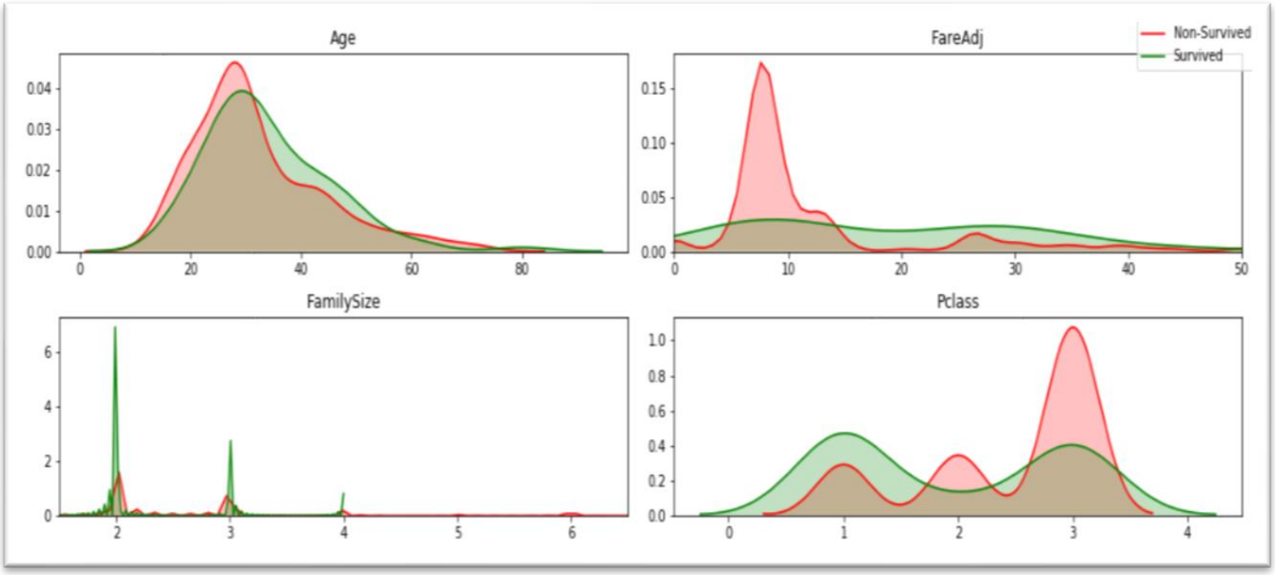


Figure 9 – Distribution of Per-Person Fares by Passenger Class

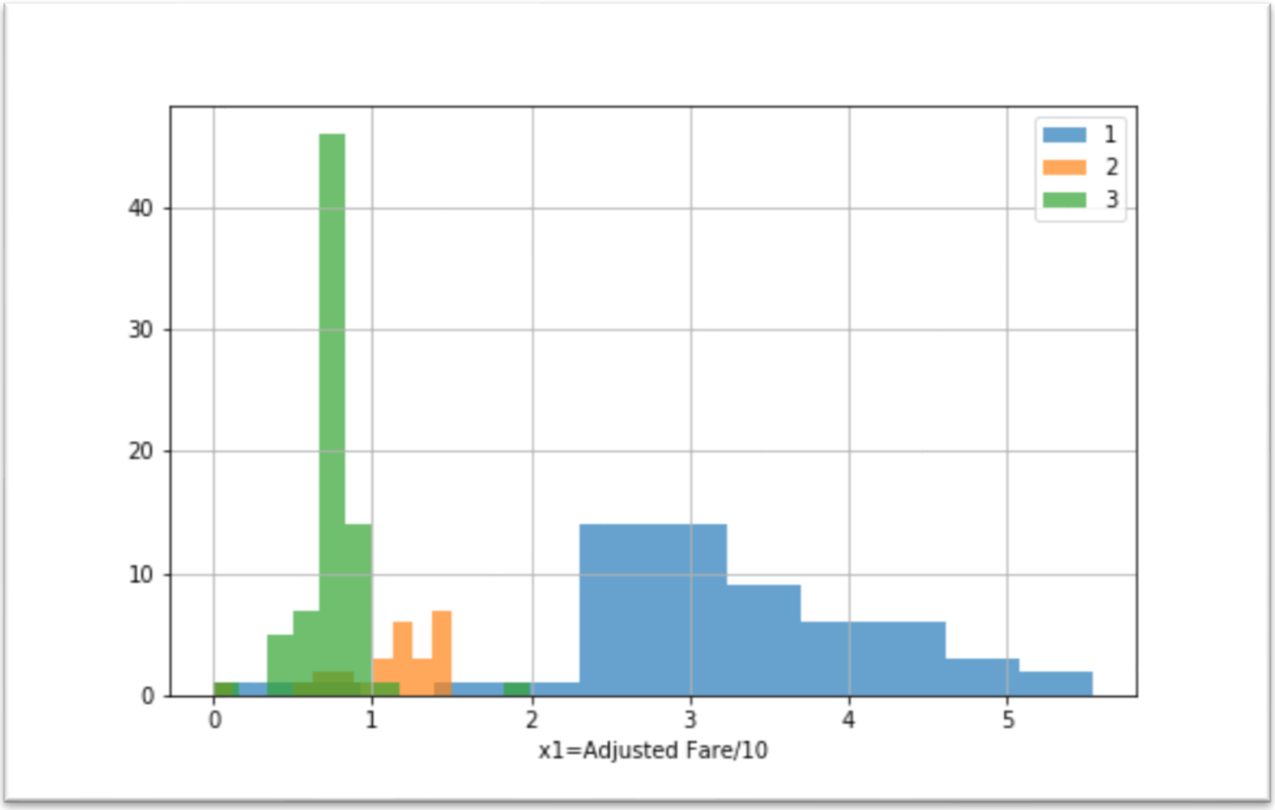


Figure 10 – Survival Patterns of Adult Males

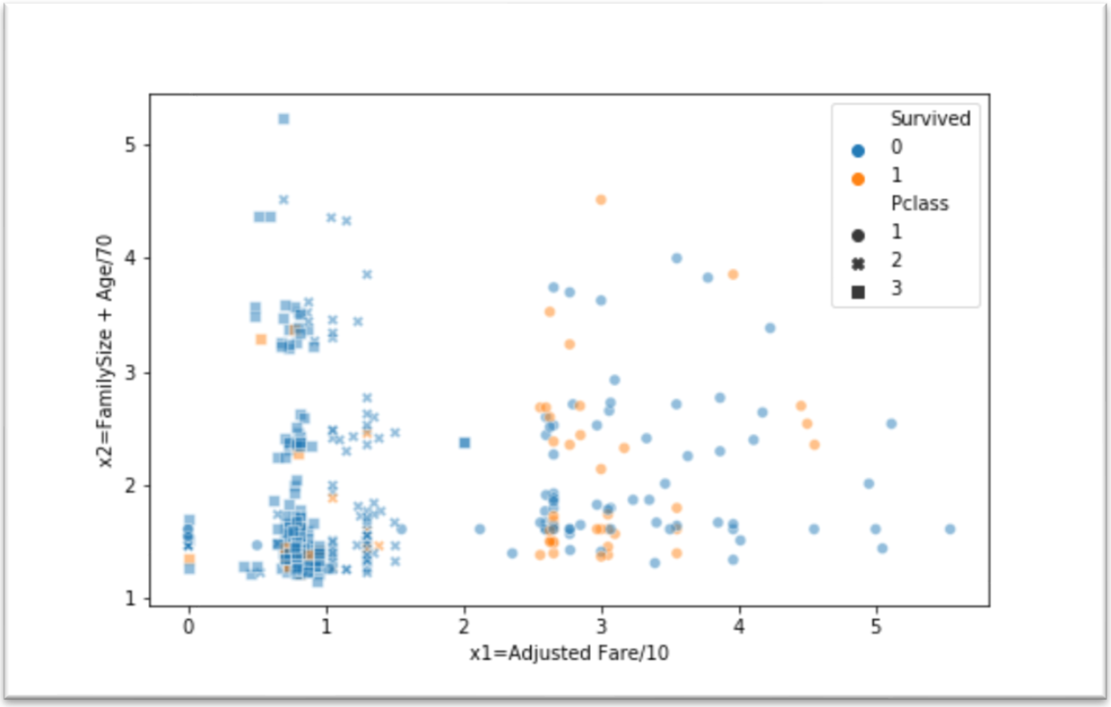


Figure 11 – XGBoost Predicted Patterns of Adult Male Survival

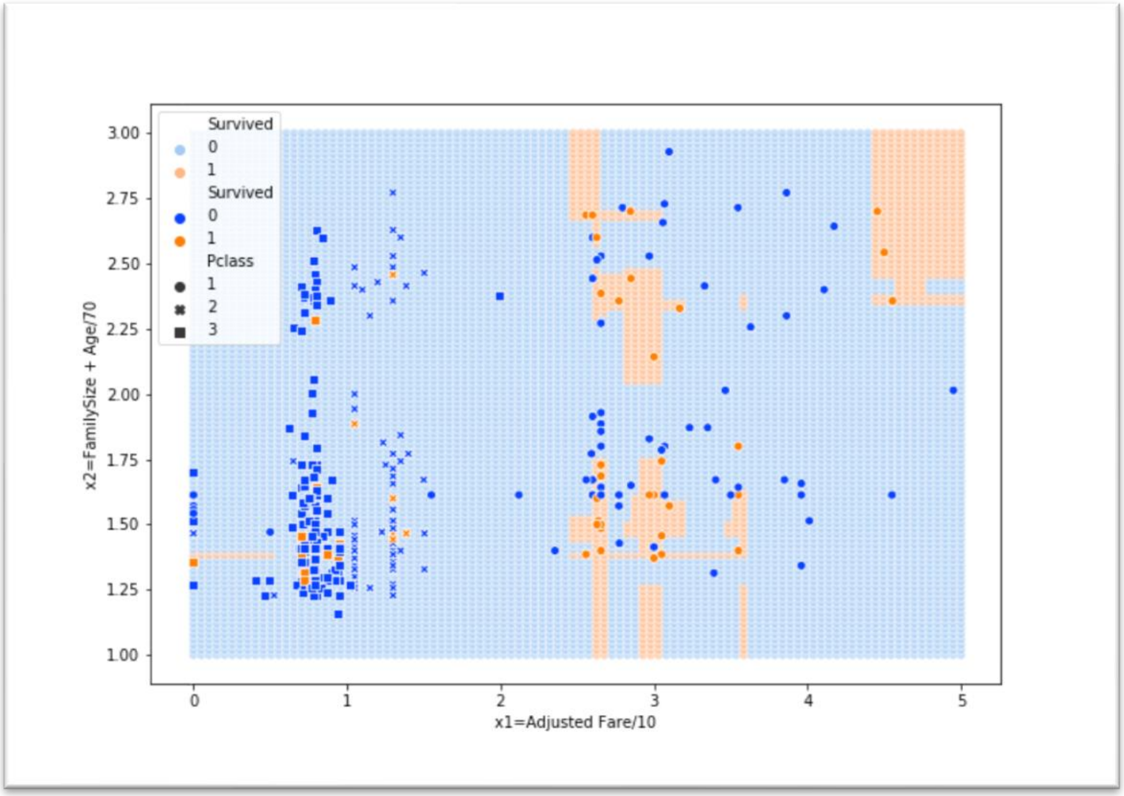
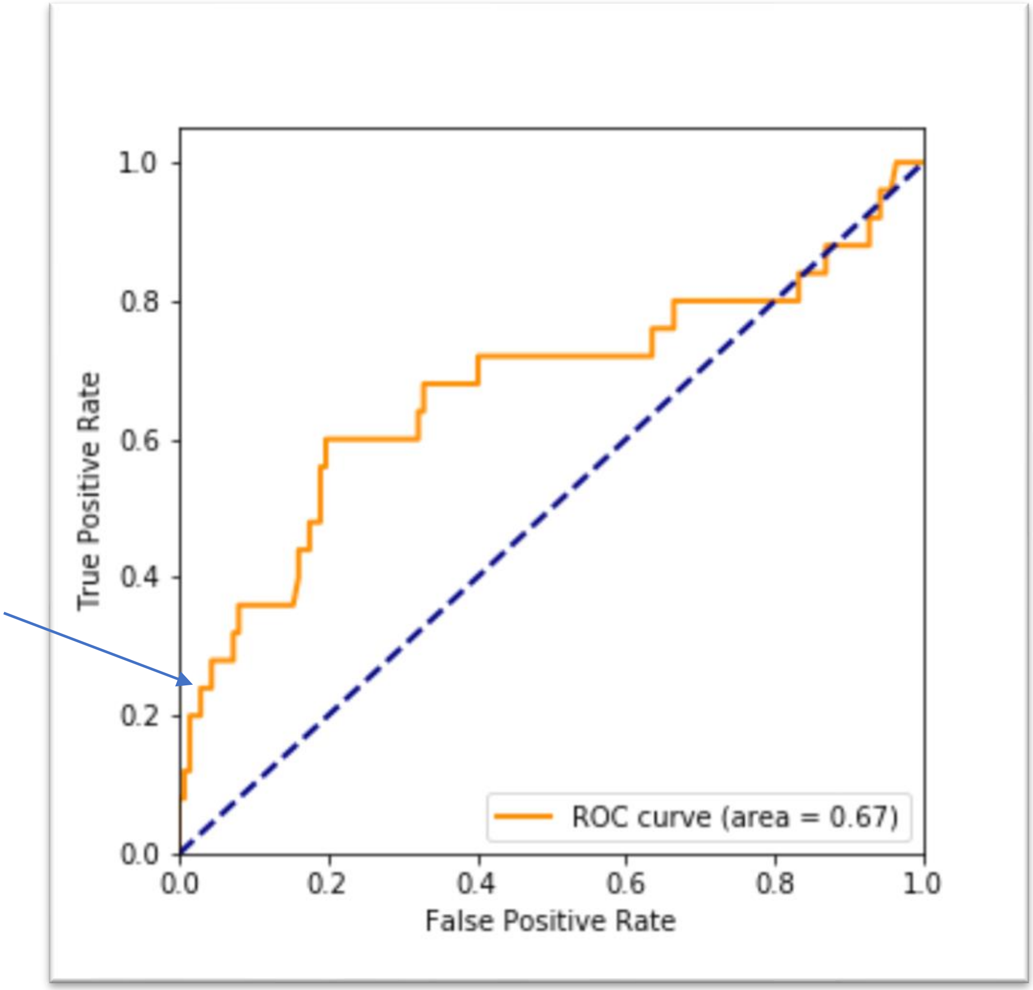
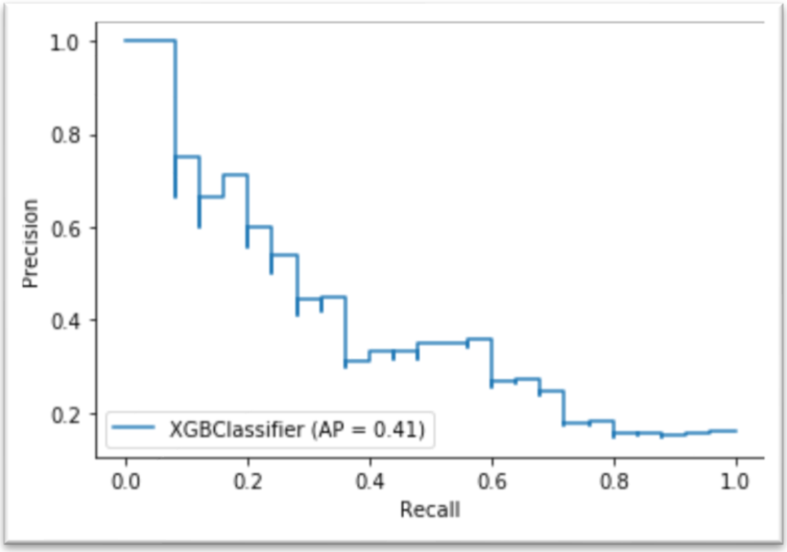


Figure 12 – ROC Curve



75% Probability Threshold

Figure 13 – 2-Class Precision Recall Curve



Appendix B – Project Jupyter Notebook (Python Code)