David Nemirovsky & Jared Klug

BISTP8106 - Professor Sun

Final Report

5/13/21

Predicting Survival on the Sinking R.M.S. *Titanic*

Introduction

The R.M.S. *Titanic* is one of the most famous ships, and shipwrecks, known in popular culture. It was the largest luxury passenger ship of its days, equipped with the most advanced safety features of its time. The *Titanic* voyage departed with 2,240 passengers and crew on board.[1] What was thought to be an unsinkable ship was sunk after colliding with an iceberg, taking the lives of more than 1,500 passengers and crew.[2] The aim of this project is to predict which passengers were likely to survive the *Titanic* shipwreck based on their passenger data (sex, age, socio-economic class, etc).

Data was acquired from Kaggle.com's Titanic Machine Learning from Disaster competition.[3] For this model, data was already partitioned into a training and test set. The training and test set consisted of unique passenger data (passenger ID, sex, age, socio-economic status, number of siblings/spouse on board, number of parents/children on board, ticket number, fare, cabin number, and port of embarkation) and their survival status (coded as "1" for survived and "0" for died). Bagging imputation was used to fill in missing data from both the training and test set. The training data was tidied by removing noninformative data (ticket number, cabin number, name, and passenger ID). Data tidying was done using the 'tidyverse' package in R and the code can be viewed in this report's additional documents.

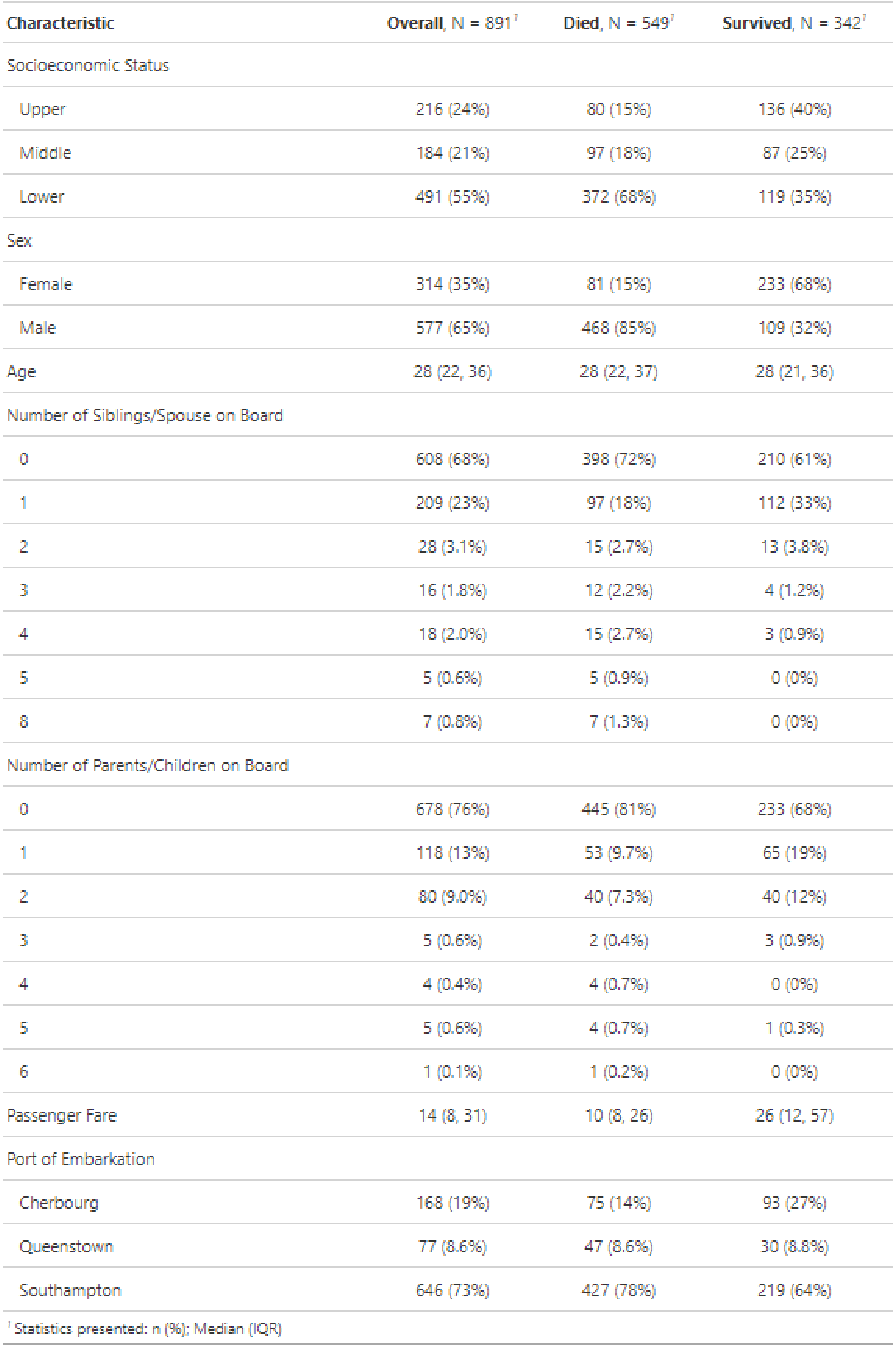
Methods

I. Exploratory Data Analysis

The variables used to predict survival were age (in years), sex (coded as "male" and "female"), socioeconomic status (denoted as 'pclass' and coded as "1" for upper class, "2" for middle class, and "3" for lower class), number of siblings/spouse on board (denoted by 'sib_sp'), number of parents/children on board (denoted by 'parch'), passenger fare (in USD, denoted by 'fare'), and port of embarkation (denoted by 'embarked' and coded as "C" for Cherbourg, "Q" for Queenstown, and "S" for Southampton).

In the training data, 19.87% of the passenger data was missing age information. Of the passengers with missing age data, the survival rate of these passengers was similar to the overall survival rate. This indicates that the missingness of data is at random (MAR), and that missingness should not be considered as a feature. Bagging imputation was used to fill in the missing age data. There were also 2 embarked observations missing. A large majority of passengers had embarked from Southampton, so that value was used for the two missing observations. Table 1, below, shows the descriptive statistics, grouped by survival, after missing data was imputed.

Table 1 – Descriptive statistics of passengers on the *Titanic*, grouped by survival (training set)

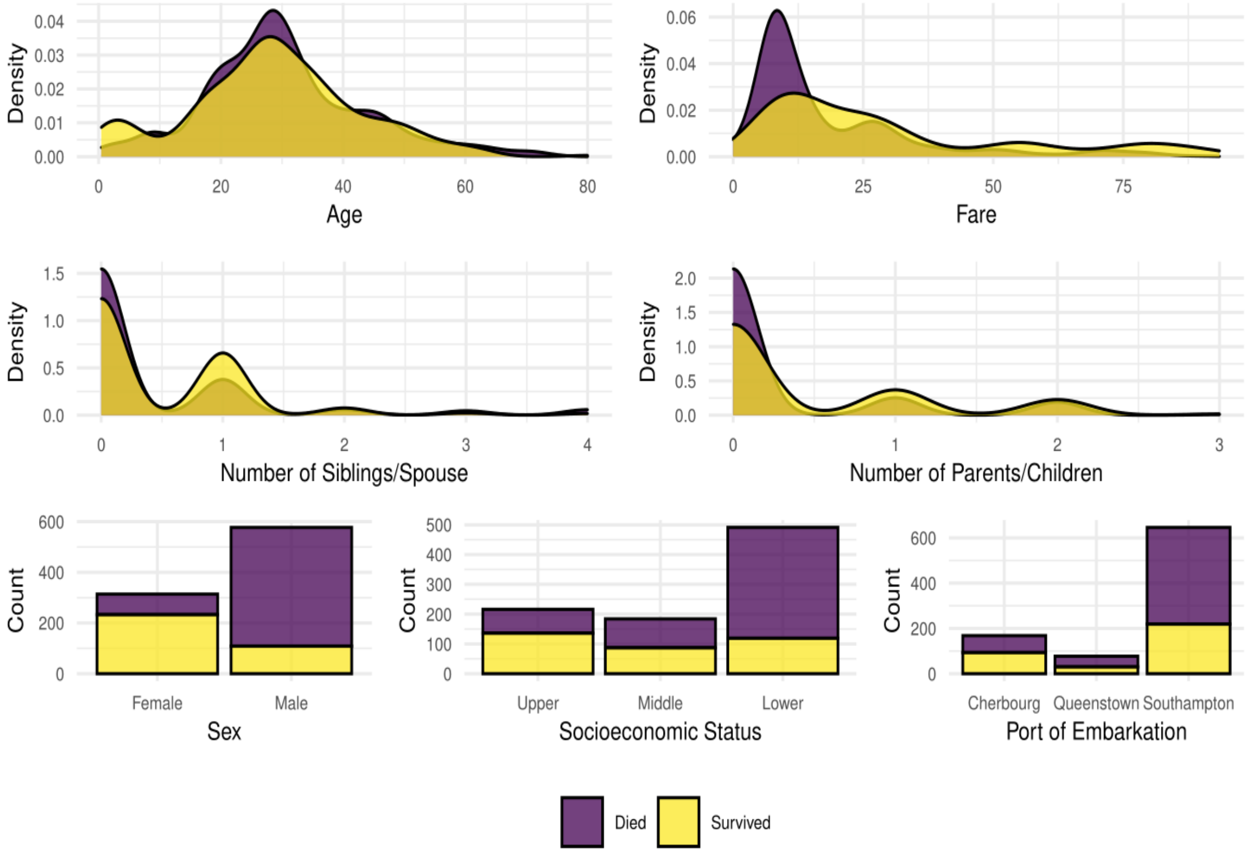


Characteristic	**Overall**, N = 891[1]	**Died**, N = 549[1]	**Survived**, N = 342[1]
Socioeconomic Status			
Upper	216 (24%)	80 (15%)	136 (40%)
Middle	184 (21%)	97 (18%)	87 (25%)
Lower	491 (55%)	372 (68%)	119 (35%)
Sex			
Female	314 (35%)	81 (15%)	233 (68%)
Male	577 (65%)	468 (85%)	109 (32%)
Age	28 (22, 36)	28 (22, 37)	28 (21, 36)
Number of Siblings/Spouse on Board			
0	608 (68%)	398 (72%)	210 (61%)
1	209 (23%)	97 (18%)	112 (33%)
2	28 (3.1%)	15 (2.7%)	13 (3.8%)
3	16 (1.8%)	12 (2.2%)	4 (1.2%)
4	18 (2.0%)	15 (2.7%)	3 (0.9%)
5	5 (0.6%)	5 (0.9%)	0 (0%)
8	7 (0.8%)	7 (1.3%)	0 (0%)
Number of Parents/Children on Board			
0	678 (76%)	445 (81%)	233 (68%)
1	118 (13%)	53 (9.7%)	65 (19%)
2	80 (9.0%)	40 (7.3%)	40 (12%)
3	5 (0.6%)	2 (0.4%)	3 (0.9%)
4	4 (0.4%)	4 (0.7%)	0 (0%)
5	5 (0.6%)	4 (0.7%)	1 (0.3%)
6	1 (0.1%)	1 (0.2%)	0 (0%)
Passenger Fare	14 (8, 31)	10 (8, 26)	26 (12, 57)
Port of Embarkation			
Cherbourg	168 (19%)	75 (14%)	93 (27%)
Queenstown	77 (8.6%)	47 (8.6%)	30 (8.8%)
Southampton	646 (73%)	427 (78%)	219 (64%)

[1] Statistics presented: n (%); Median (IQR)

Exploratory data analysis (EDA) showed differences in survival among the various predictors used. Looking at the continuous variables, age seemed to have a similar distribution among survivors and non-survivors, except when looking at ages ranging from 0 to 10 years old. Of the survivors, there were more passengers in that age range than in the non-survivors, indicating that age plays a role in survival for that age range. Passenger fare appeared to also play a role in survival, where the non-survivors had a much higher percentage in the $1 to $20 range than those who survived. Survivors also had a larger proportion in the passenger fare range greater than $50. For the discrete variables, the number of siblings/spouse on board was distributed relatively evenly among survivors and non-survivors, with slight differences for those having 0 and those having 1 sibling/spouse on board. Of the survivors, there were more passengers with 1 sibling/spouse on board, and of the non-survivors there were more passengers with 0 siblings/spouse on board, indicating that having 0 versus 1 sibling/spouse on board could impact survival. A similar trend can be seen with the number of parents/children on board. Of the survivors, there were more passengers with 1 parent/child on board, and of the non-survivors there were more passengers with 0 parent/child on board, indicating that having 0 versus 1 parent/child on board impacts survival. Looking at the categorical variables, sex seemed to play a significant role in determining survival. It can be seen that most of the female passengers survived, whereas an overwhelming majority of the males did not survive. For socioeconomic status, it appears that most passengers in the lower class did not survive, whereas those in the upper class did survive, indicating that socioeconomic status impacts survival. For port of embarkation, it appeared that embarking from Port Cherbourg saw the highest survival among passengers, although not by much. Graphs depicting all of these associations are shown in Figure 1 below.

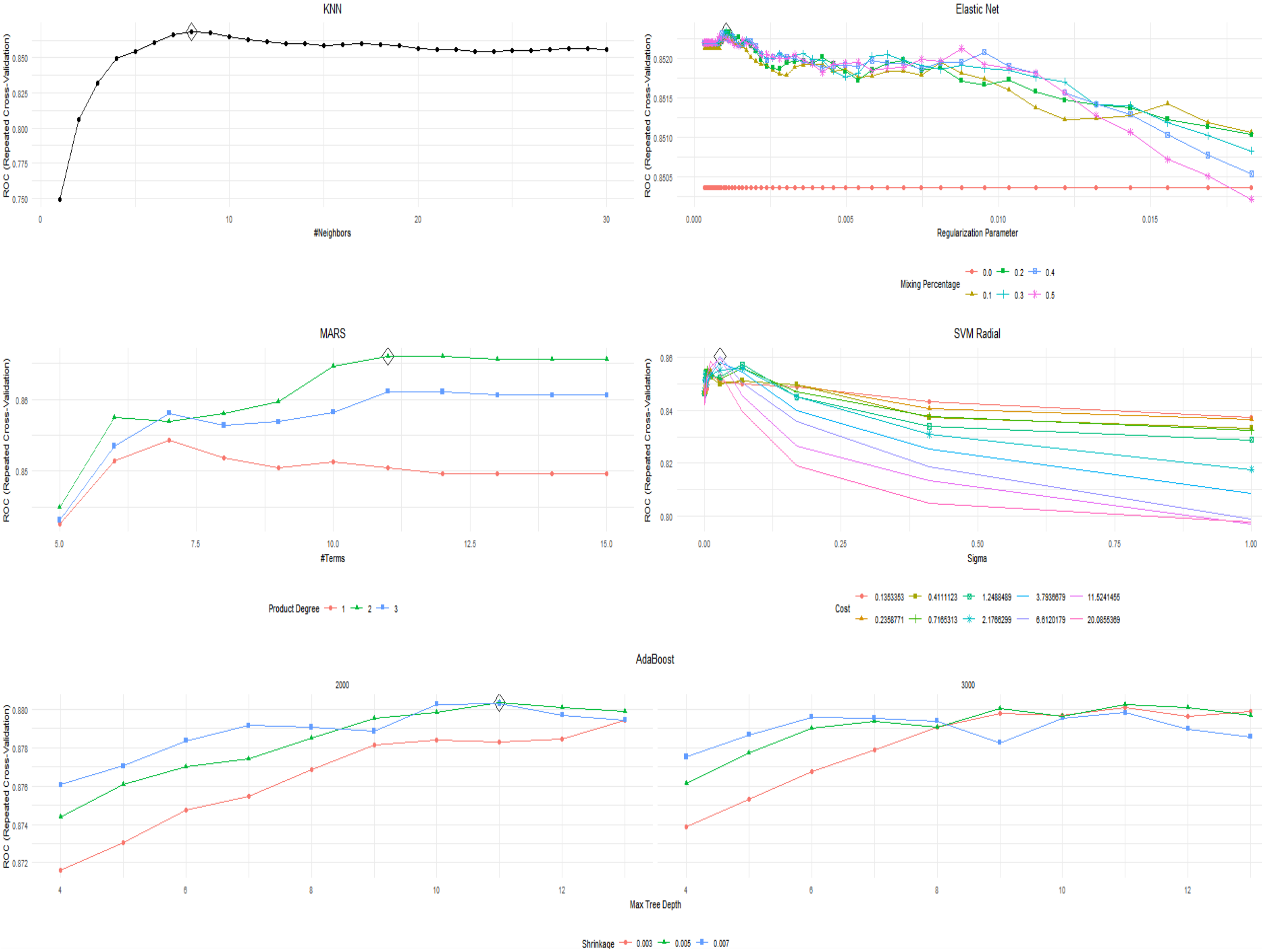
Figure 1 – Distributions of predictor variables, grouped by survival



II. Cross-Validation

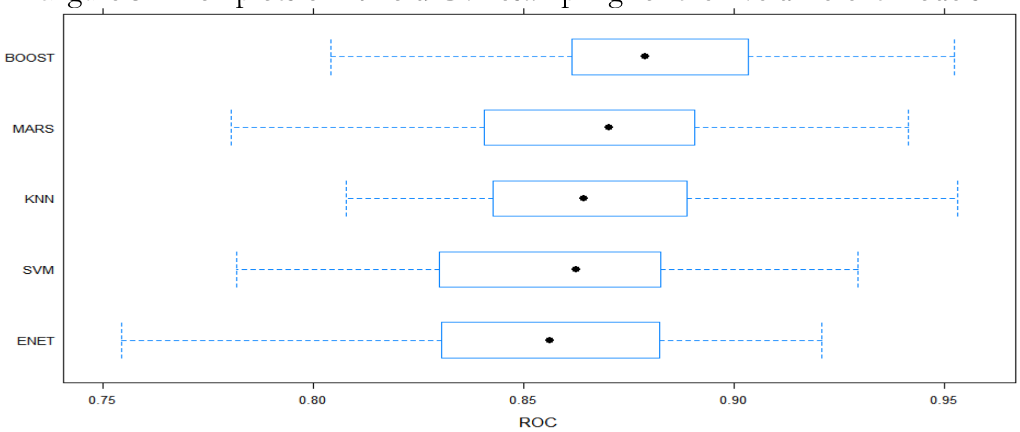
Since a training data set was already given, partitioning was not necessary for model building. Five different models were fitted on the training data using 10-fold cross-validation (CV), repeated five times: k-nearest neighbor (KNN), elastic net regularized regression (ENET), multivariate adaptive regression splines (MARS), AdaBoost regression tree (BOOST), and a radial kernel support vector machine (SVM). The optimal tuning parameter, k, for the KNN model was 8. The optimized tuning parameters for the elastic net model were 0.3 for alpha and 0.0011 for sigma. The best tuning parameters for the MARS model came from 2 degrees and 11 model terms. The optimal tuning parameters for the AdaBoost model was 2,000 trees with 11 splits in each tree (interaction depth) and a shrinkage (learning rate of boosting) of 0.005. For the radial SVM model, the optimized tuning parameters were 0.029 for sigma and 6.61 for cost. Figure 2, below, shows the optimized tuning parameters for each model during repeated 10-fold CV.

Figure 2 – Optimal tuning parameters for each model



For model selection, resampling was done using each model and the area under their receiver operating characteristic (ROC) curve was used as the metric for comparing models. Figure 3 shows how well the five models do, with a higher area under the curve (AUC) indicating a better fitting model. It can be seen that the AdaBoost regression tree model generates the best model, yielding the highest AUC values. CV and resampling was done using R's `caret` package.

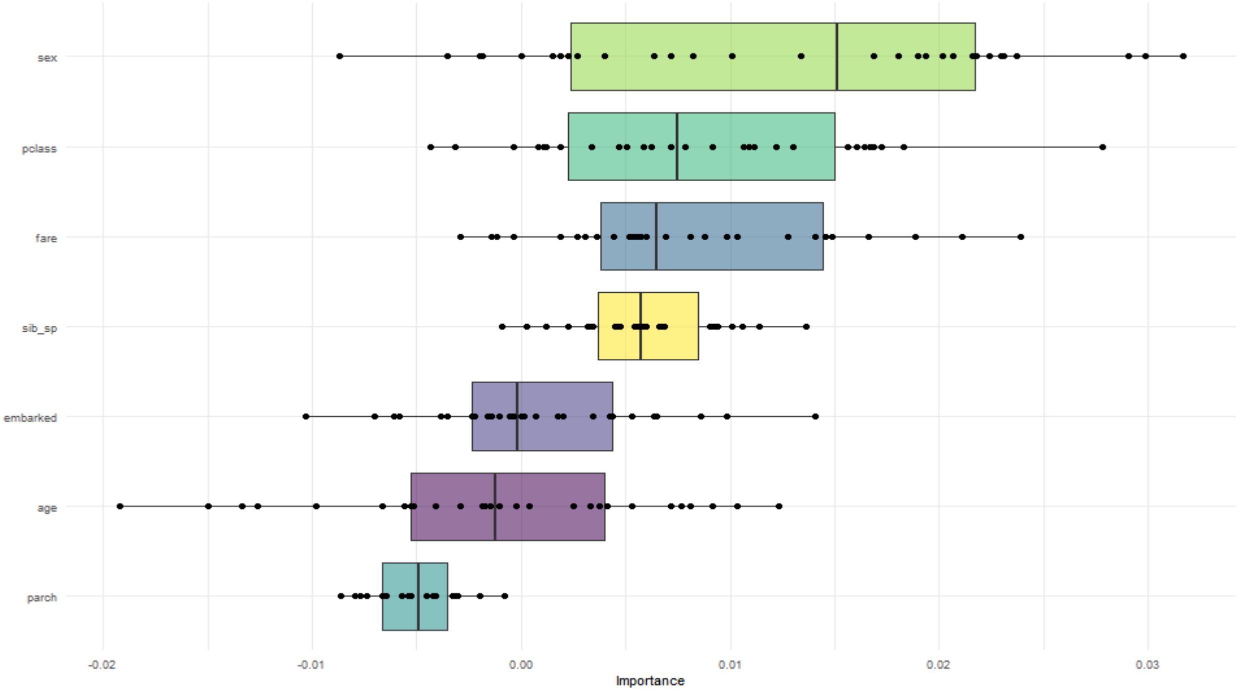
Figure 3 – Box plots of 10-fold CV resampling for the five different models



III. Model Analysis

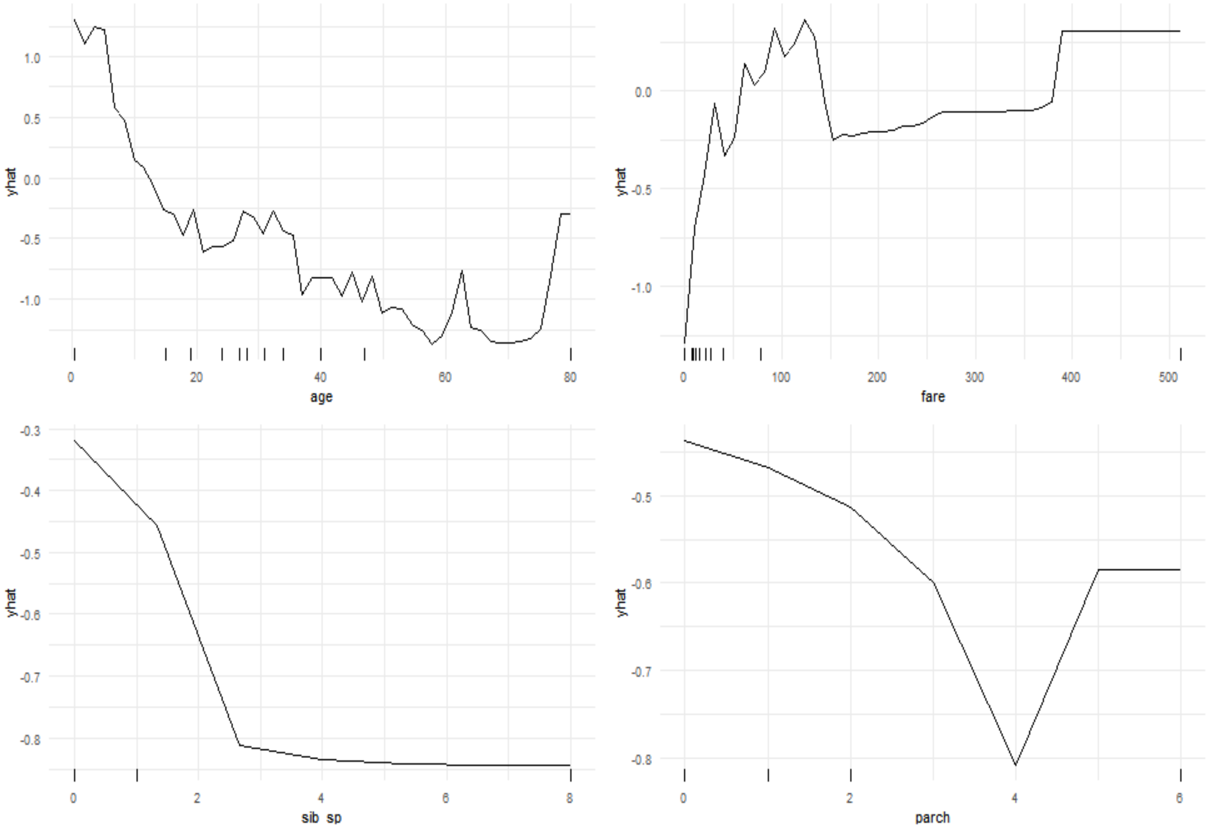
The important variables for the AdaBoost model were then examined using a variable importance plot (VIP). Figure 4 depicts a plot of this model's variable importance. It can be seen that the top three important predictors are sex, socioeconomic status, and passenger fare, whereas port of embarkment, age, and number of parents/children on board were found to be not important, using the AdaBoosted model.

Figure 4 – Variable importance plot for AdaBoost model



In order to understand how the four quantitative variables affect the survival probability, the partial dependence plots (PDPs) of each were created. Figure 5 shows how the four quantitative variables of age, passenger fare, number of siblings/spouse on board, and number of parents/children on board affect survival probability, while holding all other variables constant. Younger ages had a much higher chance of survival and the probability continues to decrease until around age 60, where survival probabilities increase for the oldest of passengers. Looking at the PDP for passenger fare, an increasing fare tends to increase survival probabilities. The PDP for the number of siblings and spouses on board shows that survival chances also significantly decrease as the number of spouses and siblings increase. A similar pattern appears for the number of parents and children on board, according to the PDP for the number of parents/children on board.

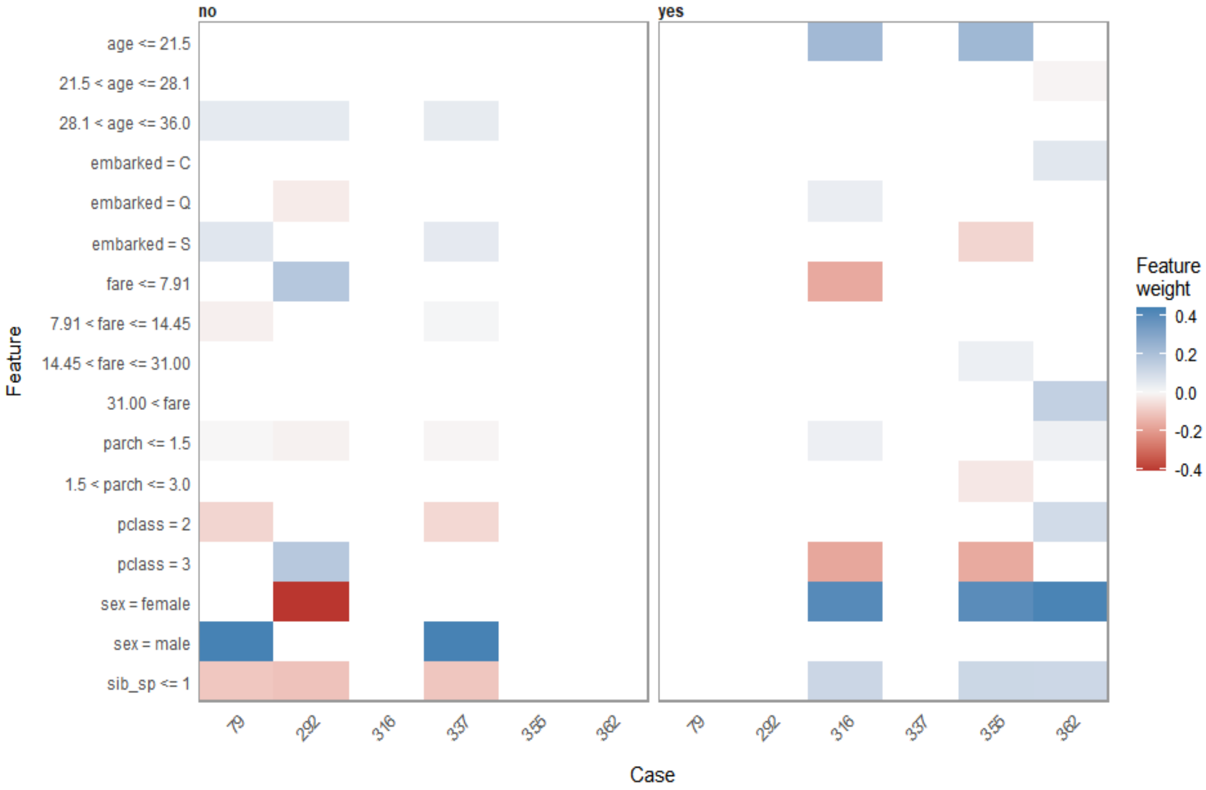
Figure 5 – Partial dependence plots for age, passenger fare, number of siblings/spouse, and number of parents/children on board, respectively



IV. Predictions & Explanations

Before assessing model predictability, it is important to understand this complex "black box" model. A test data set was already supplied, so no partitioning was necessary, just like for the training data. The test data was missing a few points, so bagging imputation was again used to impute the missing values. Taking a look at a random sample of six cases from the testing data set, it can be seen how each parameter impacted the passenger's survival probability. Figure 6, below, depicts the predictions and feature weights (weights of predictor variables) for this random sample of six *Titanic* passengers. It can be seen that sex was the predictor that carried the most weight in determining survival for the passengers. Among those who were predicted to survive, all of them were female, so despite their low socioeconomic status and low passenger fare, they were able to survive due to their sex. Looking at the two male cases (Case #79 and #337), they were of the middle class who had less than or equal to one sibling/spouse on board. However, despite their socioeconomic status and number of sibling/spouse on board, being male pulls their prediction towards death. A particularly interesting passenger was Case #292, who had identical features to Case #316. They were both females of the lower class with less than or equal to one sibling/spouse and less than or equal to one parent/child on board, and both of their passenger fares were less than or equal to $7.91. Where they differed was their age, with Case #316 being less than or equal to 21.5 years old and Case #292 being between the ages of 28 and 36 years old. This was the reason why Case #316 was predicted to survive and Case #292 was predicted to not survive. This goes to show that survival prediction does not solely depend on one factor and that age can heavily influence survival prediction, especially among those aged younger than 20 years old.

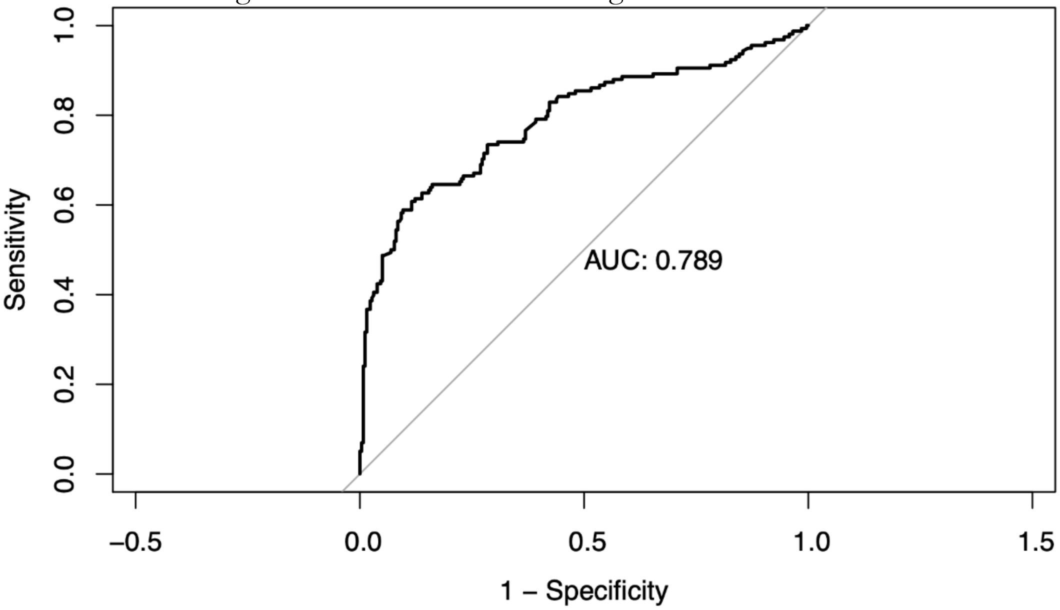
Figure 6 – Feature explanation plot for random sample of six *Titanic* passengers



Results

The fitted AdaBoost model was tested against the test data provided by Kaggle. R's `predict( )` function was used to get the probability of survival for each *Titanic* passenger in the test data set. The model achieved an accuracy of 76.56%, a sensitivity of 85.00%, and a specificity of 62.66%. Having a lower specificity indicates a higher false positive rate, meaning that the model had a tendency to classify those who didn't survive the *Titanic* shipwreck as survivors. Additionally, having a large sensitivity, therefore a larger true positive rate, meant that the model did a good job in classifying survivors of the shipwreck for those who actually survived. Figure 7, below, shows the ROC of the boosted model on the test set. The model gave an AUC value of 0.789, indicating a decent fit.

Figure 7 – ROC of AdaBoost regression tree model



Conclusion

Although the AdaBoost model only achieved about 76.56% accuracy, this model outperformed the other model choices during repeated 10-fold CV. It was found that variables such as sex, socioeconomic status, and passenger fare were the most important in the final model. However, key values of other predictor variables, such as age being lower than 20 years and the number of siblings/spouse on board being less than 2 could impact prediction survival. In order to further improve the model, feature engineering must be done. Generating more information such as "family relationship" (i.e. husband, wife, child, single) could help improve the accuracy of classification. From the EDA, it is shown that younger passengers, as well as female passengers were more likely to survive compared to their counterparts. Therefore, having new variables indicating family relationships may help develop a more accurate model. Additionally, different imputation methods for missing values should be considered, leading to a more representational effect of age on survival.

References

[1] US Department of Commerce NOAA. National Oceanic and Atmospheric Administration (NOAA) Home Page. NOAA Office of General Counsel. Published December 6, 1998. Visited 10 May 2021. https://www.gc.noaa.gov/gcil_titanic-history.html

[2] National Geographic Society. Sinking of the Titanic. National Geographic Society. Published November 9, 2012. Visited May 13, 2021. https://www.nationalgeographic.org/media/sinking-of-the-titanic/

[3] Kaggle.com. "March Machine Learning Mania 2021 – NCAAM competition". Visited 10 May 2021. https://www.kaggle.com/c/titanic/