Survival Analysis in Titanic

TENSORFLOW 2.0

12/16/2020

Section 1 Introduction and Data

Titanic, in full Royal Mail Ship (RMS) Titanic, was a British Passenger Luxury Liner that sank on April 14-15, 1912 during its maiden voyage, en route to New York City from Southampton, England killing about 1500 passengers and ship personnel. Being one of the most famous tragedies in modern history, it has inspired numerous stories, several films, and a musical and has been the subject of much scholarship and scientific speculation.

In this project, we aim to explore the presence of factors, if any, that could have increased the likelihood of any passenger to survive in the shipwreck. We hypothesize that besides luck, factors like gender, age and passenger class increased the chances of a person to survive the mishap.

The dataset consists of the information of 1310 passengers that traveled in the Titanic. The 14 columns included in the dataset are pclass, survived, name, sex, age, sibsp, parch, ticket, fare, cabin, embarked, boat, body, home.dest. While we retrieved the dataset from data.world, the principal source for data about Titanic passengers is the Encyclopedia Titanica (by the method of web-scraping) which consists of the facts and information of the real people that designed, built and sailed on the RMS Titanic.

Section 2 Methodologies

Our analysis methodologies can be divided into two subsections:

1. Data Analysis: In this subsection, we visualize the relationships between different variables of interest on our dataset and the survival of passengers and analyze them. First we begin by exploring the relationship between gender and survival. Then we explore other factors like age that could have contributed to the survival differences in the genders. Our second feature of interest is passenger class. After analysis of the fare and age distribution of passengers in each of the passenger classes to help us gain a deeper understanding of the composition of passengers in different classes, we explore the relationship between passenger class and the number of passengers who survived/died. Our analysis of passenger class is followed by the exploration of the relationship between port from which the passengers had embarked and their survival status. To further our quest to explore the variables that increased likelihood of survival and to make our models more efficient, we feature engineer variables like categorical fare, family size and title from the existing features such as fare, number of siblings aboard, number of parents or spouses aboard, and name of the passengers. Lastly, we visualize the relationship between our feature engineered variables and the survival of passengers and analyze them too.

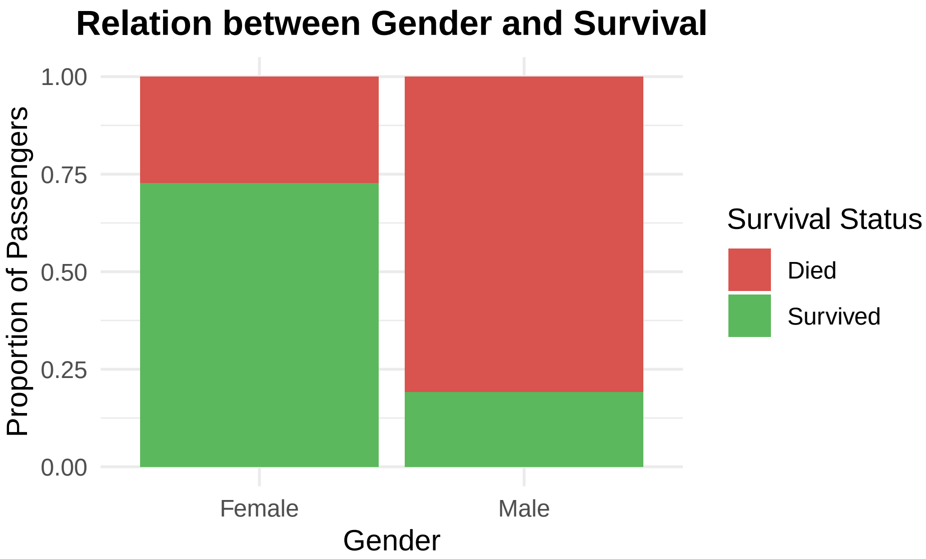
2. Predictive Modeling: We begin by splitting our data into train sets and test sets. Then we devise three logistic regression models by using the features that showed us signs of strong relationship with survival of the passengers based on our visualizations. Our first model consists of pre-existing factors like gender, passenger class, number of siblings, number of spouses/parents, fare, and port from which passengers had embarked as its predictor variables. Our second and third model consists of a mixture of pre-existing features like gender, passenger class and feature-engineered factors like categorical fare, family size, and title of the passengers. The difference between second and third model is that our third

model has port from which passengers had embarked removed from its list of predictor variables. We perform 5-fold cross validation on these models and we analyze their efficiency based on accuracy and area under the ROC curve. We then select the most efficient model to make predictions on our testing data and analyze its efficiency based on the area under the ROC curve again. We finally, analyze the confusion matrix based on different cutoff probabilities and choose the most suitable one for our model.

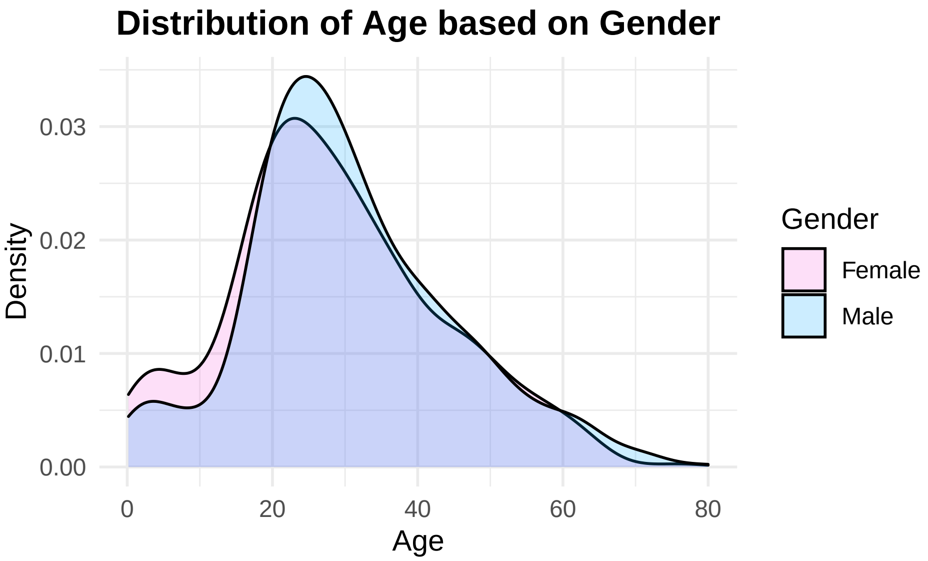
Section 3 Results

Section 3.1 Data Analysis

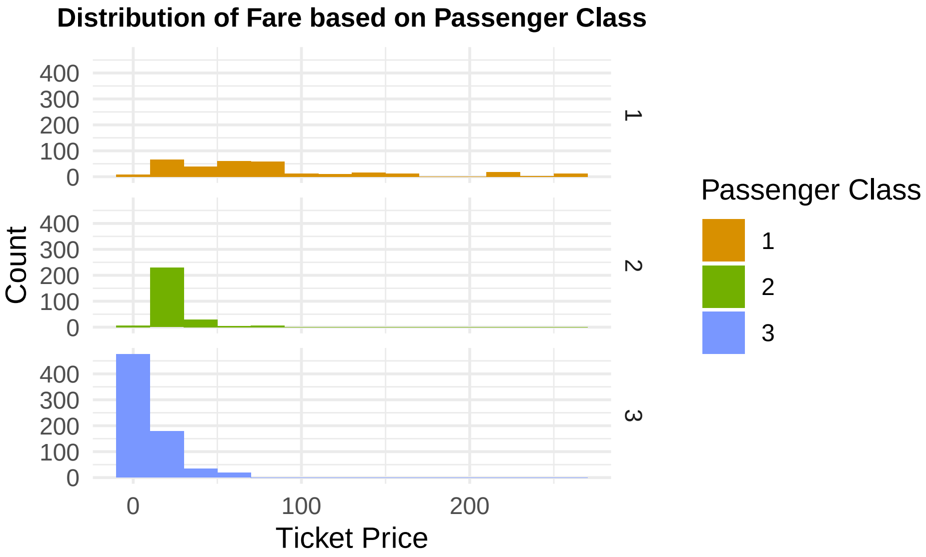
Lets begin our analysis by visualizing the relationship between the gender of the passengers on the ship and the proportion of passengers who survived and died. We believe that females had higher chances of survival than that for males.



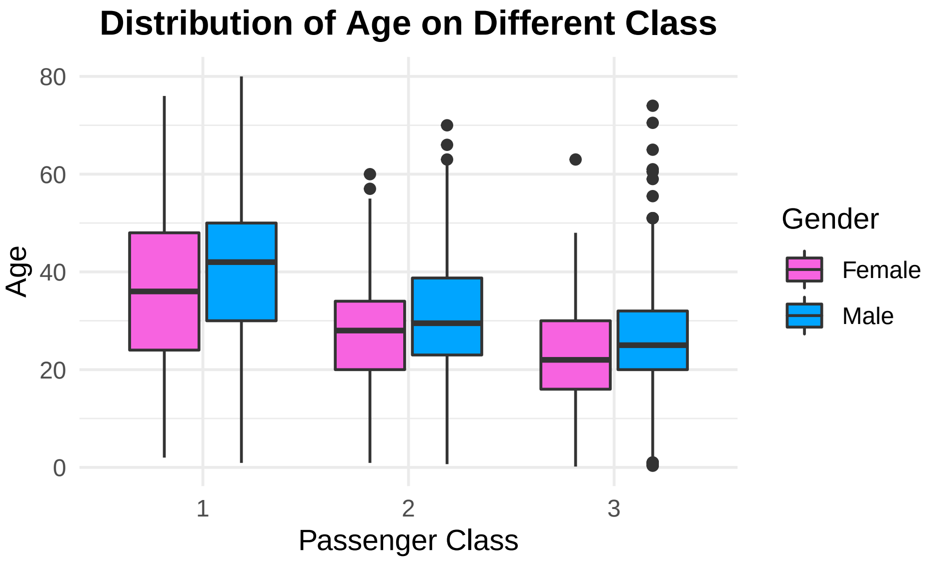
It is evident from the bar chart above that a large number of females survived compared to the males in the ship. The proportions of nearly 0.75 and 0.20 of survived females and males respectively suggest that every 3 out of 4 females survived the shipwreck while only 1 out of 5 males survived the shipwreck. Perhaps, this was because females were given priority and only once a large number of females were offered lifeboats, males were focused in the rescue efforts. But could this difference in proportion of people who survived in each of the genders be due to the difference in their ages? We believe any significant difference in age might contribute to the differences in the the survival chances of the genders as some age group might be given more priority over others. To explore this potential differences in age, lets visualize the distribution of age across males and females.



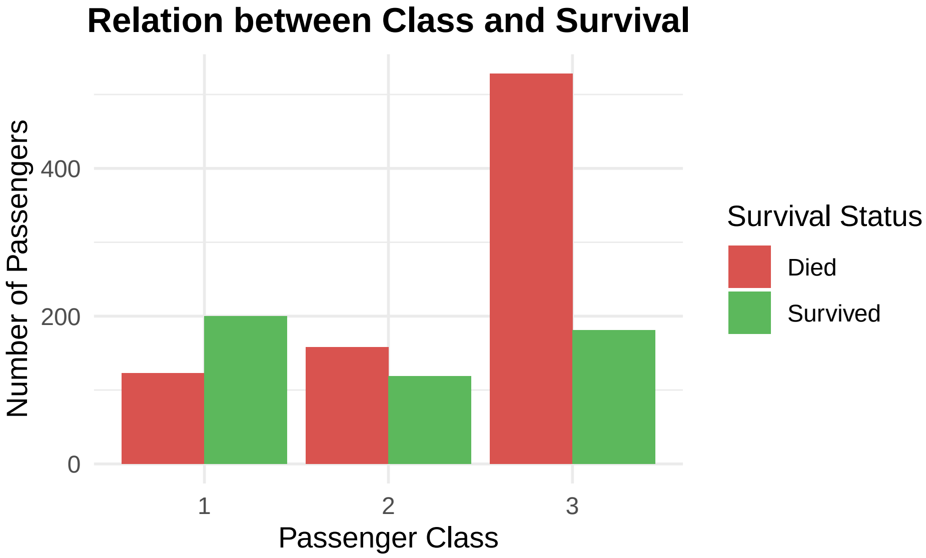
Both the genders have similar age distribution; majority of passengers were between the age group of 18-30 in both males and females. Since we do not see any significant differences in their distribution of age, we can fairly infer that the age did not have a significant effect on the survival difference in the gender. Now that we have analyzed the differences in survival chances based on gender, lets proceed to our second factor, passenger class, and check if passengers from certain classes had better chances of survival than the others. But before, lets deepen our understanding of the passenger class. To begin with, lets explore the fare distribution across each of the passenger classes.



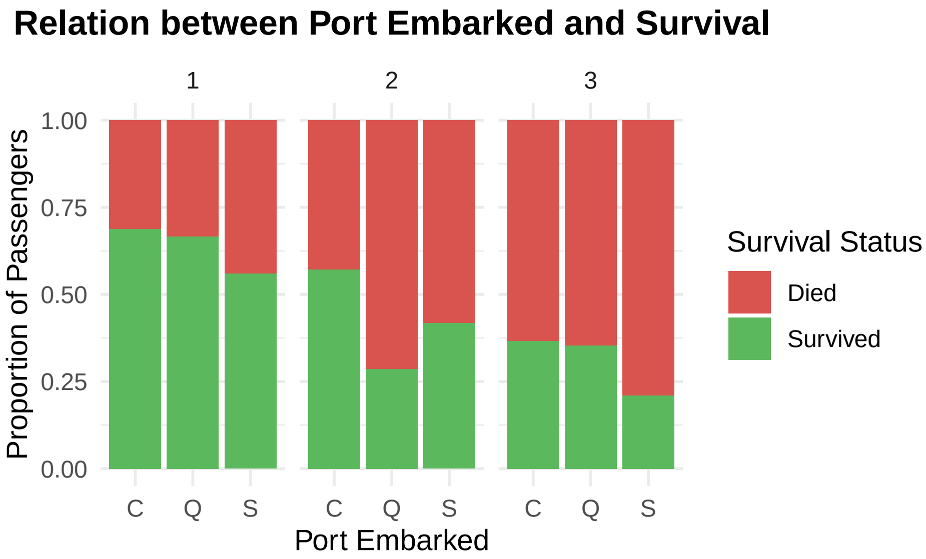
Overall, the passengers from the class 1 had the highest median fare and the passengers from class 3 had the lowest median fare, with the fare from passengers from class 2 in between them. The fare for passengers of class 1 was spread on a wider range with some as high as greater than USD 250. Passenger class 1 had rich and affluent businessmen, class 2 had luxury tourists and class 3 had young migrants who hoped to made a new living in the US and Canada. The distribution of fares accurately reflects the economic status of passengers in each of these classes. Now, lets get one step further to deepen our understanding of passenger class. Lets explore the gender distribution of passengers in each of these classes.



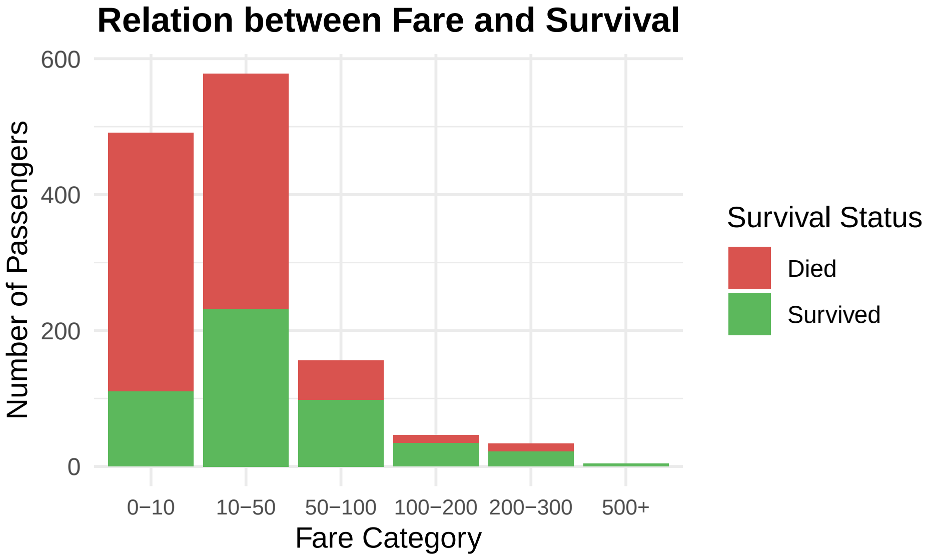
Passenger class 1 had passengers(both male and female) with highest median age group and passenger class 3 had passengers with the lowest median age group. Gaining status and wealth in society takes time and perhaps this might be the reason that class 1 passengers were relatively older in age. Similarly, class 3 passengers consisted of the most number of young people who were moving for a better future and starting their new life and this might be the reason for their lowest age among all the passenger classes. Now, that we have had sufficient understanding of passenger class, lets explore its relationship with the survival of passengers.



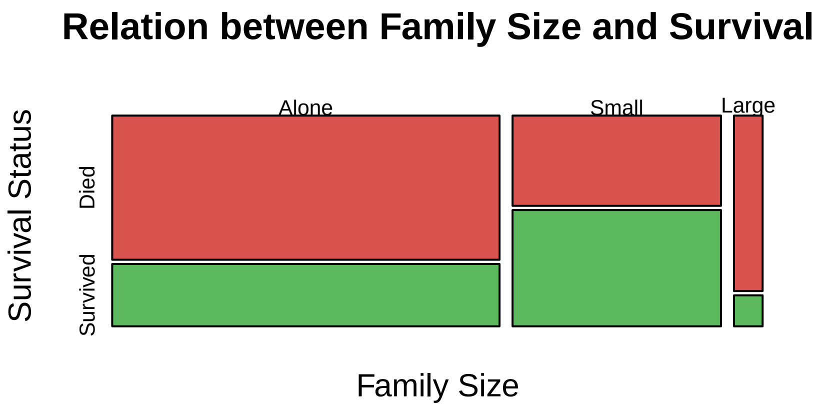
Only Passenger Class 1 had more number of passengers who survived than the number of passengers who died. Passenger Class 2 and 3, in stark contrast, had more number of passengers who died than the passengers who survived. Passengers from class 2 appear to have better survival chances than the passengers from class 3 as the number of passengers who died and those who survived are similar in class 2 while as in class 3, the number of passengers who died is far greater than the number of those who survived. Hence we conclude, that the survival rate of passengers decreased as we move from passenger class 1 to 3. The rich and the powerful appear to have been given priority in the rescue effort. Some reports also claim that a large number of class 2 and 3 passengers were not even unlocked from their compartments and abandoned to die because of insufficient amount of rescue boats and this is aptly reflected in our chart above. Now, lets visualize how the survival varied based on the port from which the passengers had embarked in each of these passenger classes.



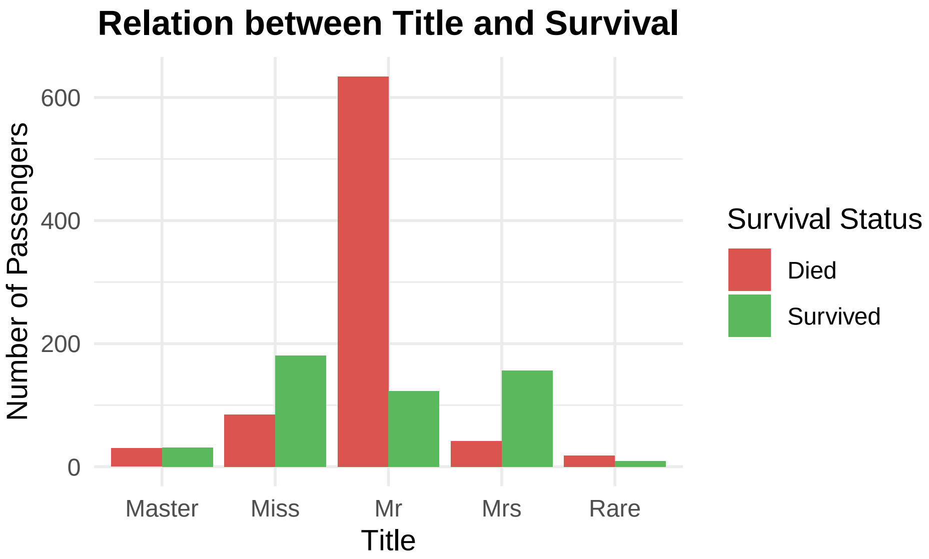
The survival rates in all the ports decreased with increasing passenger class with an exception of 2nd class passengers from port Q(Queenstown). Passengers who embarked from C(Cherbourg) have the highest survival rate in all the classes but it is not significantly higher than passengers who embarked from Q(Queenstown). To be specific, both class 1 and 3 have similar survival rate for the passengers who embarked from C(Cherbourg) and Q(Queenstown) with the aforementioned exception in class 2. In general, passengers who embarked from port S(Southampton) appear to have the lowest survival rate in all the classes. Perhaps, this was because a large number of male migrants had embarked from Southampton as this was the starting port. The relationship between port from which passengers embarked and their survival chances is not as significant as we had seen for in gender and passenger class. We will take this result into consideration on our modeling stage and devise our models accordingly. Now that we have seen the relationship between port of embarkation based on different classes and the survival chances, lets explore how fares affected survival chances. We feature a new variable by classifying the fare of the passengers into categories. We classifying the fares unequally in such a way so that this new feature will not only help us visualize the relationship between fare and survival chances but also make our models more efficient by efficiently capturing the patterns in our observations.



It is evident from the graph above that as the fare group increases, the survival chances also increases among the passengers. This helps to further our explanation for the difference in survival across passenger class. Based on the figure above and our analysis of passenger class and survival, we can infer that along with their social status, the economic status also might have helped passengers from class 1 in gaining priority in the rescue efforts. We will definitely use this feature in our models. After a successful feature engineering, lets create another variable to see how family size affects the chances of survival. First we begin by creating a variable called family by adding the number of siblings, parents, spouses and counting the passengers themselves. We then classify this family into a new variable in which people with no family aboard are classified as Alone, people with number of family member in between 2 and 4 inclusive are classified as having small family and people with number of family members greater than or equal to 5 aboard are classified to have a large family. Now, lets visualize the relationship between family size and survival of passengers.



It is evident from the size of containers in the mosaic plot above, that a large number of people had no family members aboard and only a few passengers had a large family in the ship. Passengers with small family appear to have the best survival chances, which is evident from the large proportion of people who survived than who died in people having small family. People who had a large family had the least chances of survival and the chances of survival of people who were alone fell in between that for small and large family. A large number of passengers who were alone belonged to the third class and this might be the reason for their low survival chances. Moreover, people from large family couldn't easily be placed in a single boat which must have decreased their survival chances. On the contrary, majority of passengers with small family were either first or second class passengers who were given priority and could easily fit in a boat and hence had high survival chances. Now, finally, lets explore how the title of a person affected his/her chance of survival. We first extract the title of the passengers from their names and then group them based on their similarity and classify the remaining titles with less number of occurrences as "Rare".



It is evident from above, that females(Mrs) and minors(Master and Miss) had better chances of survival. Males(Mr.), on the contrary, had the lowest survival chances as there were more number of male passengers that died than those who survived. This further supports our previous observation of females having better survival chances than males. We will definitely add this feature in our models.

Section 3.2 Predictive Modeling

Based on the observations from the data analysis subsection, we found out that features like gender, passenger class, and fare showed a strong relationship with the survival of the passengers. We devise our first model to predict survival from the aforementioned features including the port from which the passengers had embarked(a feature which showed relatively weak relation with survival of the passengers). While our first model consists of only pre-existing variables in the dataset as their predictor variables, our second and third model consist of both the pre-existing variables and the feature engineered variables as their predictor variables. To be more specific, our second model uses gender, passenger class, categorical fare, family size, title, and port from which passengers had embarked to predict the survival status of the passengers. Similarly, our third model uses all the features that the second model employs excluding the port from which passengers had embarked to predict the survival status of the passengers. Now that we have given a brief description of our models, lets view their efficiency on predicting survival on our training data. The results of the average accuracy and average area under the ROC curve across different folds in our training data for model 1, 2 ,and 3 are as follows:

Table 1: Performance metrics for Model 1 on train set

.metric	.estimator	mean	n	std_err
accuracy	binary	0.782	5	0.008
roc_auc	binary	0.829	5	0.013

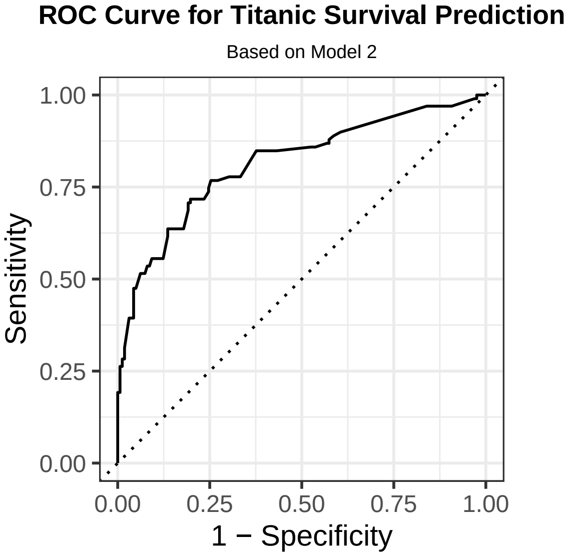
Table 2: Performance metrics for Model 2 on train set

.metric	.estimator	mean	n	std_err
accuracy	binary	0.800	5	0.006
roc_auc	binary	0.848	5	0.015

Table 3: Performance metrics for Model 3 on train set

.metric	.estimator	mean	n	std_err
accuracy	binary	0.801	5	0.008
roc_auc	binary	0.845	5	0.015

Both the models 2 and 3 performed better in terms of average accuracy and average area under the ROC curve than the model 1. Perhaps, our feature engineered variables more efficiently captured the patterns in our training set than the pre-existing "raw" variables alone. The average area under the ROC curve for model 2 is slightly higher than that for model 3. On the contrary, the average accuracy of model 2 is slightly lower than that for model 3. Since, majority of variables like gender, family size, title, etc in our dataset have a skewed distribution, we prefer area under the curve over accuracy to prevent overfitting of classes which have majority of observations. Hence, we choose model 2 for making predictions on our testing data. Now, lets view the ROC curve for our model 2 on the testing data.



The ROC curve above plots the True Positive Rate (Sensitivity) against False Positive Rate (1 - Specificity) for the different possible cut-points for our prediction. Our curve is significantly away from the 45-degree diagonal and more closer to the left hand top border of the ROC space, which shows that our model was significantly more efficient than random prediction of survival status of passengers. Its efficiency can be quantified by the value of the area under the curve for model 2 on our testing data which is given below:

Table 4: Performance metrics for Model 2 on test set

.metric	.estimator	.estimate
roc_auc	binary	0.813

The area under the ROC curve for model 2 on test data is 0.813. This value alone does not have significant meaning. However, it being more closer to 1 suggests that our model had captured some pattern from our training data and made predictions that was significantly better than predicting survival status at random. Now, lets explore the confusion matrices for the prediction of the survival of passengers with model 2 for different threshold values. The confusion matrices for threshold values of 0.3, 0.5, and 0.8 are as follows:

Table 5: Confusion matrix for threshold of 0.3

	Passenger died	Passenger survived
Passenger predicted to die	113	22
Passenger predicted to survive	49	77

Table 6: Confusion matrix for threshold of 0.5

	Passenger died	Passenger survived
Passenger predicted to die	130	29
Passenger predicted to survive	32	70

Table 7: Confusion matrix for threshold of 0.8

	Passenger died	Passenger survived
Passenger predicted to die	155	60
Passenger predicted to survive	7	39

With the cutoff probability of 0.3, our model had significantly higher number of False Positives than False Negatives. It means that our model classified more number of passengers who were dead as survived than it classified the number of passengers who survived as dead. On the contrary, with the cutoff probability of 0.8, our model had significantly more number of False Negatives than False Positives. It means that our model classified more number of passengers who survived as dead than it classified the number of passengers who died as survived. Our model had a balanced number of False Positives and False Negatives with a cutoff probability of 0.5, signifying that our model had roughly similar number of passengers who were dead but classified as survived and who survived but were classified as dead. A cutoff of 0.3 should be used if one wants to minimize the number of survived passengers being mis-classified as dead and a cutoff of 0.8 should be used if one wants to minimize the number of dead passengers being mis-classified as survived. In our project, as much as we are interested in finding if a passenger survived, we are also concerned in exploring if the passenger died. Hence, we choose the cutoff of 0.5 in making prediction with our model if we are to conduct further analysis based on our predicted data.

Section 4 Discussion

It is evident from our analysis that our hypothesis (the existence of factors besides luck which increased the chances of survival of the passengers) was correct. Females, children, class 1 passengers, high-fared passengers, and passengers who had embarked from Cherbourg had the highest chances of survival while as adult males, class 2 and 3 passengers, low-fared passengers, and passengers who had embarked from Southampton had the lowest chances of survival. The logistic regression models that we devised from the mixture of pre-existing and feature engineered variables were comparatively good at predicting survival status than the model that was devised from just the pre-existing variables. Our model predicted the survival status with an area under the ROC curve of 0.813 and the cutoff probability of 0.5 was found to be the most suitable for our prediction as it puts equal emphasis on predicting if the passenger died or survived.

While we have tried our best to get the most accurate results for analysis and prediction, there are still rooms for improvements. There were other features like age, home/destination of the passengers, cabin, etc that we had completely ignored in our analysis due to them having large percentage (some as high as 60%) of missing values. These features could have been included to make our models more efficient in predicting the survival status of the passengers. After analysis of survival status of passengers based on gender, we had tried to find if any other factors had contributed to the difference in survival based on gender. We relied solely on visualization and analyzed the distribution of age based on gender to check if it had caused the differences. Similarly, throughout our analysis subsection we relied on visualization alone to check the relationship between different factors with survival. While visualizations can be a good predictor if two features have relationship, the relationship needs to be quantified some way or the other before the variables are used in our models. Similarly, the area under the ROC curve that we obtained on our final model for prediction on test set has little meaning on its own and it can still be improved because the maximum value the area under the curve can take is 1.

The missing values in the age could have been filled with median age of the passengers that fell under the same class, gender, and port and analysed. Similarly, the letters present in the cabin represented the compartments of the passengers and this could have been extracted to conduct further analysis. If we were to start over this project all over again, we would analyze all the existing features and conduct chi-square test to quantify the significance of different features with the survival status of the passengers(we do not calculate the correlation in our case because most of our variables, including our response variable, are categorical) and we would only use those features with significant statistical relationship in our models. We would also deploy other classification algorithms like Random Forest Classifiers, Extra Trees Classifiers, and Support Vector Machines in our training and testing phase and compare them with each other to predict the survival status with the highest efficiency.

References

Rippner, Noah, 2016, Titanic Disaster Dataset, https://data.world/nrippner/titanic-disaster-dataset
Scale Manual, https://ggplot2.tidyverse.org/reference/scale_manual.html
String Replace, http://rfunction.com/archives/2354
Theme Manipulation, https://ggplot2.tidyverse.org/reference/theme.html
Theme Minimal, https://ggplot2.tidyverse.org/reference/ggtheme.html