Titanic Survival Analysis

Cristobal Torres-Valderas

**Abstract**

In this paper we analyze the Titanic data set in attempt to identify the features and variables that are most indicative if a person would have survived the Titanic crash. The variables proven significant were: Gender, Age, and Ticket Class.

**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. Our objective is to determine who would survive the crash. We will be utilizing R Studio to analyze the dataset and create our predictive model. We will be using a classification algorithm to create a decision tree that will guide us to features that have more weight in predicting survival of a passenger.

To optimize our predictive model, we will be exploring the information of the data to help us gain insights about the variables included and which will provide us insightful meaning to predictive model. Once we explore the data overall, we will narrow down our analysis to eliminate categorical variables and then conduct test to see whether certain variables are statistically significant and apply this information to optimize our model.

**Information about the Data**

We are given two data sets, one is a test.csv and the other is a train.csv. This will be our initial data to train and test our predictive model. Our training data has 891 rows of data and our test data has 418 rows of data.

The data is concise and is real world data in regard to survival on the Titanic. We began by analyzing several categorical variables that will be eliminated such as ticket number, cabin number, and port of embarkation. *Ticket Number* is just a number with no significance. *Cabin Number* similar to ticket number did not provide key information on the passenger’s survival. *Port of Embarkation* is irrelevant as well with the assumption that survival was no indicated by this information.

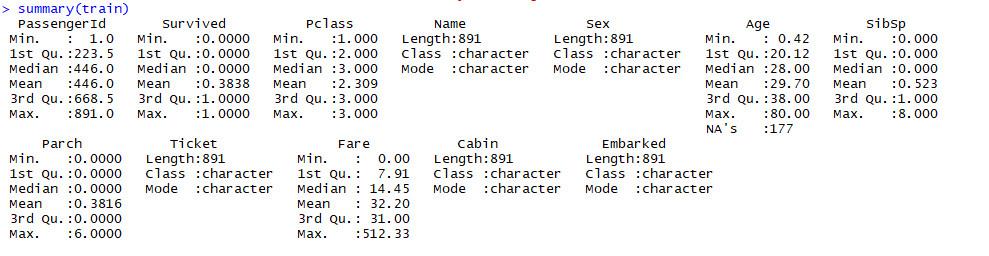
Our data dictionary:

|  |  |  |
| --- | --- | --- |
| 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.*

We can run a summary of the table to get a quick analysis of the variables.

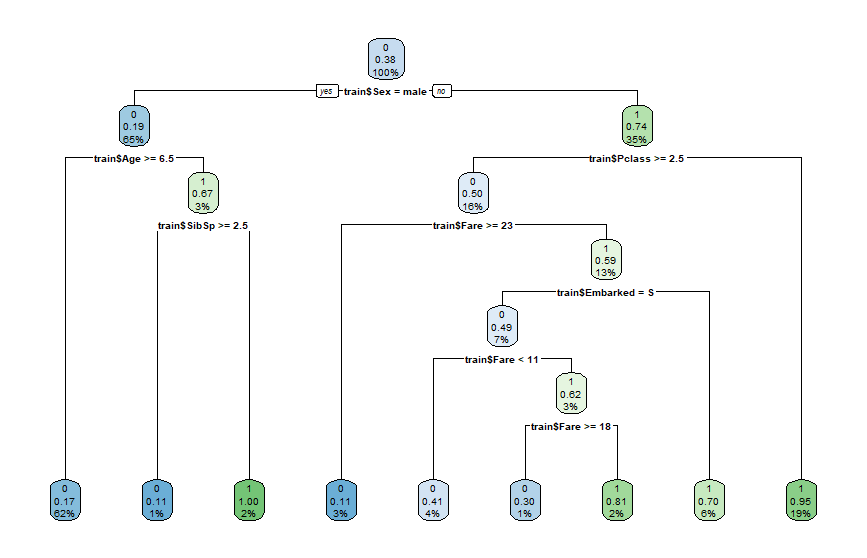


After we run the summary we can pull some basic information of the data. We learn the average class is 2 meaning majority middle class tickets. We can see a survival rate of about 38% from the mean of survived. We can see the average age was about 30 years old.

Based on this we can grasp better the variables we are working with and how to explore them.

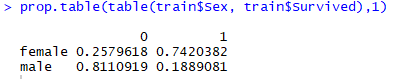
**Exploring the data**

Using RPart we can create a decision tree model in which we can explore the data to better grasp. Lets explore the dataset using a decision tree and move forward from there.

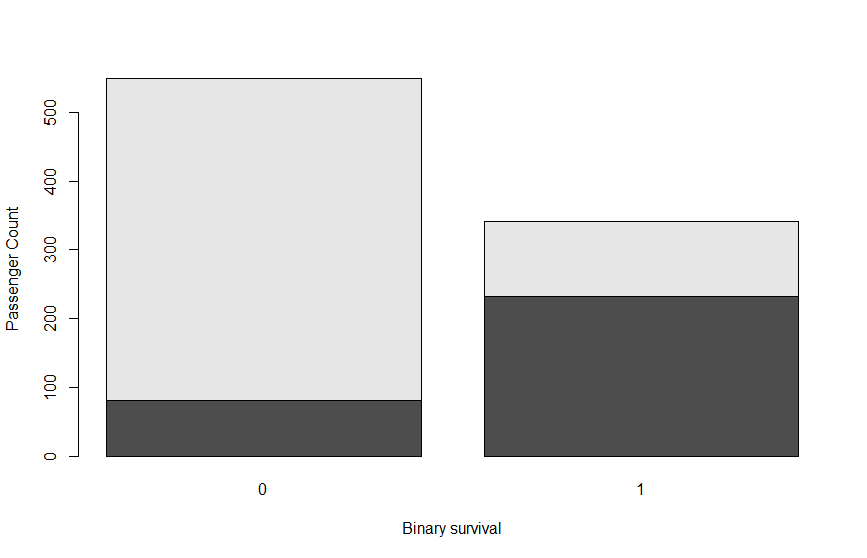


The root of the decision tree stems from initially separating genders. Depending on if it is male or female then the decision tree expands further to different branches. If the passenger is male, then age becomes the next significant factor. We can see that males kids younger than 6 years old have a better chance of survival. Now that we can see an overall gist of the data let’s explore the relationships.

**Explore Gender**



Here we run a proportion table to see the proportion on which survival was based on gender. We must view the proportions in the 1st dimension to view survival rate as in separate groups between male and female. What we discover here is that women are more likely to survive due to the relationship suggesting that about74% survived of all the women on board. Now when looking at the males you can quickly notice that about 81% of the males did not survive.



With this bar plot we can get a better visualization of the survival based on gender. We see that overall more people died than survived which also exemplifies the knowledge as discovered earlier. We can also see that a majority of those the did not survived are male(light) and a miniscule number of females(dark) died. We can look at those that survived and quickly realize that majority of survivors were woman reiterating our discovery from earlier.

The gender category gives us a strong basis on predictability of the variables that will help us predict who will survive. From this we can see alone that a female has far more likelihood of surviving.

**Explore age**



Quick variables about age we can see the youngest is practically a new born with the oldest being 80 years old and the average being about 30.

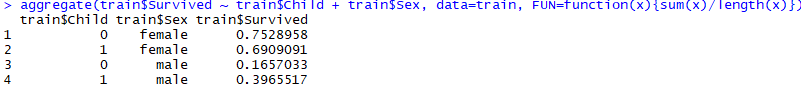
Lets add another category to the training data that will be binary to inform us if the person is less than 18 years old.

After adding the child category, we can now analyze the survival rate of a child by aggregating it. In order to specifically aggregate the proportion on survival we will create a function that returns the proportion of survivals in our aggregate function. Below is our result.



The data suggest that the age function indicates statistically significant survival rate if the passenger is child. The data indicates that the survival rate for an adult is 36% and for a child is 53%. This alone allows us to interpret that age is an important variable.

We can even go a step further and aggregate this with gender to get a better feel of the data.



What we can ultimately conclude that age is a significant increase in survival rate if the passenger is male. However, we can also see that if the passenger is female they will still most likely survive and no major difference if the female passenger is a child or not.

Interesting as we progress and analyze our model…

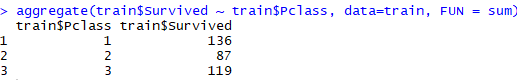
**Explore PClass**

We can see from the basic exploration as done in the beginning that PClass indicates in which social class the passenger is affiliated. With this in mind we can create a hypothesis to assume that higher on the social class could indicate a higher survival rate. Lets run a basic summary to remind us of a basic analysis.

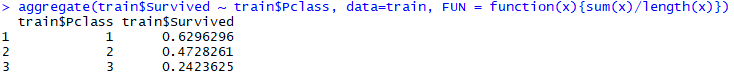


From this summary we can conclude that the average passenger is middle class. With a median of lower class, we can also assume that there is a significant amount of lower class patrons aboard. This can enable our hypothesis that upper class was given preferential treatment of survival

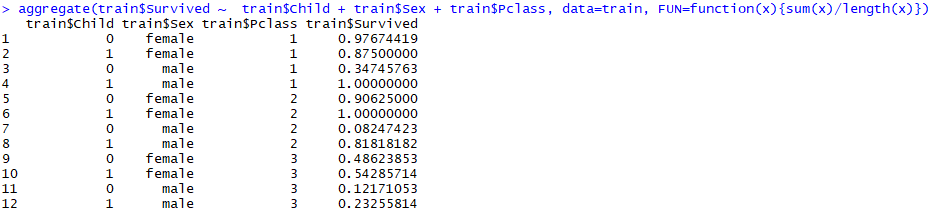
We can now aggregate the variable with survival to see specifically how many survived of each social class.



We see that a majority of those that survived our upper class affirming our hypothesis. However, we can also notice that more lower class survived than middle class. This aggregate is based purely on the sum lets analyze the ratio of each social classes survival rate to gain deeper insight.



Now we when we look at proportion it gives us more affirmation to our hypothesis. From this aggregate we now see the ratio of survival based on social class. As we assumed, if you are in the upper class you have a 62% survival rate. Middle class 47% survival rate and lower class has a 24% survival rate. Interesting we can clearly see that being upper class is significant in determining survival lets aggregate this with the other variables we have explored (gender and age) to see how this all fits together in determining predictability and to see if our other feature analysis collude in a higher predictability.

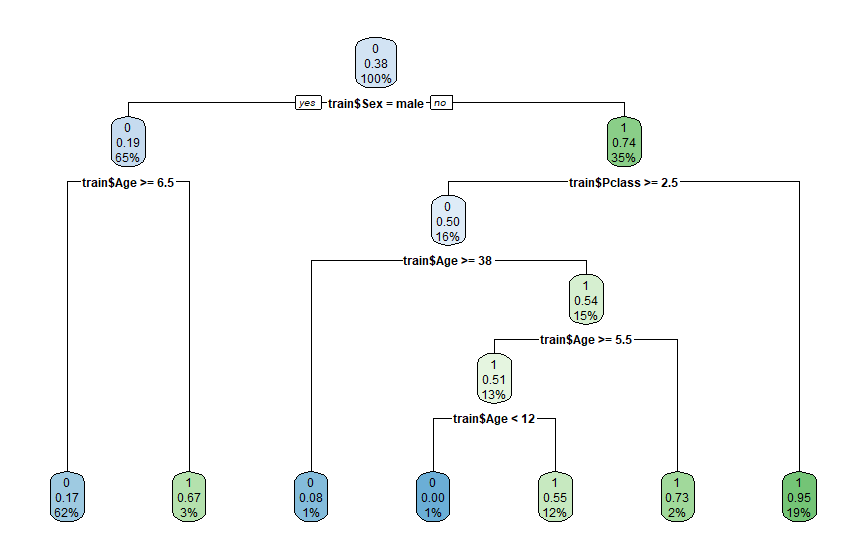


What we want to mainly analyze is the entourage effect of the variables and how the correlate in survivability. Based off the numbers we can quickly see that a female and upper class has a significantly high survival rate disregarding whether the passenger is an adult or not. We can also see that being a child, male and upper class had a 100% survival rate. We also see female and middle-class children and adults have a >90% survival rate and also a significantly high survival rate from middleclass male children.

With these analyses of our data we begin to gain a better picture on who survived and what were the determining factors in whom would survive. We have built a strong idea of what variables will be dominant in creating a successful predictive model.

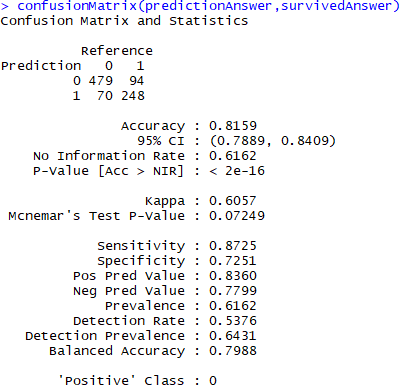
**Building the prediction model**

We have successfully explored three variables in which we have discovered will have the strongest indication and weight in survival. These variables are gender, age, and pclass. From this we can visualize the prediction model, we see that first the question is the gender of the passenger after we would want to know the age of the passenger and then last know the social class of the passenger. As our analyses showed previously, an optimal passenger for survival would be a female, child, and upper class. From this we can build a prediction model that will classify the likelihood of survival based on these values.

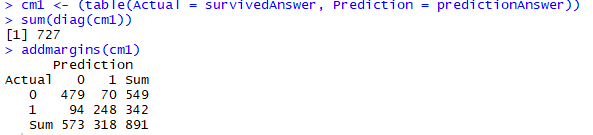


We have created a fit with our specified variables and narrowed down the variables that were initially used when we explored the data.

Now we must validate our model with a confusion matrix. This is were the major issues and troubleshooting came for us and we were unable to get the code running. After a couple of days of troubleshooting we were able to get it working. Here is the confusion matrix from our prediction model.



We can quickly see that our prediction accuracy is 81%. We can verify this calculation with our true positives and true negatives. We will now add the sums to get a better confusion matrix model.



We can now manually calculate the values and analyze the significance to validate our model.

Prediction accuracy (479+248)/891= .81 \* 100 = 81% as shown in the initial confusion matrix.

Sensitivity rate is 87% from this we can conclude that the model was correct almost 90% of the time that a positive outcome was predicted. This means the model is competent at predicting a survivor.

Specificity rate is 72% from this we can see that about ~3/4th of the time the model was correct with negative outcomes. The means the model is decent at predicting a passenger who will not survive.

**Conclusion**

After exploring significant variables individually and together, we found that passengers who were female, upper class, and above the age of 18 had an extremely high survival rate on the Titanic at 98%. The most significant variable that had an impact on survival was gender as about 74% of women on board survived. In addition to gender, age factored in heavily. The percentage rate of a child surviving almost doubles the percentage rate of an adult. The tests we ran to predict survival rate, based on these variables, proved to be helpful in predicting whether a passenger would have survived the tragic sinking of the Titanic in 1912. We believe our results would allow you to accurately predict your own chances of survival without running any more tests and only using the data collected.

After creating our model, we created prediction model with 81% prediction accuracy by using a classification algorithm and construct a decision tree. We can now think about how the model could be optimized further. We could’ve done some feature engineering to engineer our variables to provide more statistical significance. The task to create a predictive model truly brings all the information we have learned together and apply them as if we were a data scientist. The experience provides great insight into how innovative technologies such as AI and ML operate and how the mathematics and concepts are applied with R programming.