# A Statistical Analysis of the Titanic Dataset

Ryan Jackson

03/11/2021

## 1 Introduction

When it comes to the world of Data Science and Machine Learning, there are many datasets that are popular all over the world but probably none more so than the Titanic dataset which contains information on the fate of passengers aboard the infamous British ocean liner which tragically sank on $15^{th}$ April 1942 after striking an iceberg on its maiden voyage from Southampton to New York City. The data was obtained from the popular online data science community Kaggle and contains information on whether passengers survied or not, their sex, age, what class they were in, whether they had siblings/parents/children aboard, what fare they paid for their ticket and where they embarked from. In this analysis we begin by examining the raw data in Section 2 and doing some data cleaning to ensure it is appropriate for use. In Section 3 we then visualise the data to try and assess whether any of the variables in the dataset appear to have an effect on whether passengers survived or not. We then apply a Logistic Regression classifier to our data in Section 4 and assess its classification performance. Concluding remarks on the data and the analysis are then presented in Section 5.

## 2 Data Cleaning

As mentioned above, before we can begin visualising our data and fitting our classification method, it is important that we ensure the data is useable for our analysis and our first step is to make sure all our variables are of the appropriate data type. When examining the data, we see that the Survived and Pclass variables which correspond to whether a passenger survived or not (1 or 0) and the class they belonged to (either $1^{st}$, $2^{nd}$ or $3^{rd}$) are given as integers instead of factors so that will be our first change.

After that change, we now look to missing values in the dataset and in order to give a wholistic view of the number of missing values we can observe Table 1 below

Table 1: Count of Missing Values for each Variable.

| Survived | Pclass | Sex | Age | SibSp | Parch | Fare | Embarked |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 177 | 0 | 0 | 0 | 0 |

As we can see there are no missing values for 7 of the 8 variables but for the Age variable we have 177 missing values which is a pretty considerable number meaning that if we were to simply remove these rows with the missing values we would likely be dealing with a large loss of information thus, instead we will use a technique known as *imputation*. For this technique what we will do is look at the average and median age of passengers in each passenger class since we know there are no missing values for that variable and depending on what class each passenger with a missing age belongs to, we'll apply either the average or median age of their class to them. Table 2 below shows us the average ages by class

Table 2: Average and Median Age of Passengers in each Class.

| Class | Average Age | Median Age |
|---|---|---|
| 1 | 38.23 | 37 |
| 2 | 29.88 | 29 |
| 3 | 25.14 | 24 |

So from Table 2 we see firstly that the average and median ages don't appear to be hugely different suggesting that we aren't dealing with heavily skewed data and so it would be appropriate here to just use the mean age instead of the median. We can also observe that those in $1^{st}$ class are on average quite a bit older than those in $2^{nd}$ who are in turn slightly older than those in $3^{rd}$ class. So we can take these average ages and apply them to those passengers in the respective classes that have a missing age value.

Doing a final check on the missing values in Table 3 below we see that the steps outlined above to impute the values of the average age of each class to those passengers whose age was missing from the dataset have worked successfully.

Table 3: Count of Missing Values for each Variable.

| Survived | Pclass | Sex | Age | SibSp | Parch | Fare | Embarked |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 3 Exploratory Analysis

Now that our data has been cleaned and is in working order, we can begin to explore it in more detail with various visualisations. The first two visualisations are shown below in Figure 1,

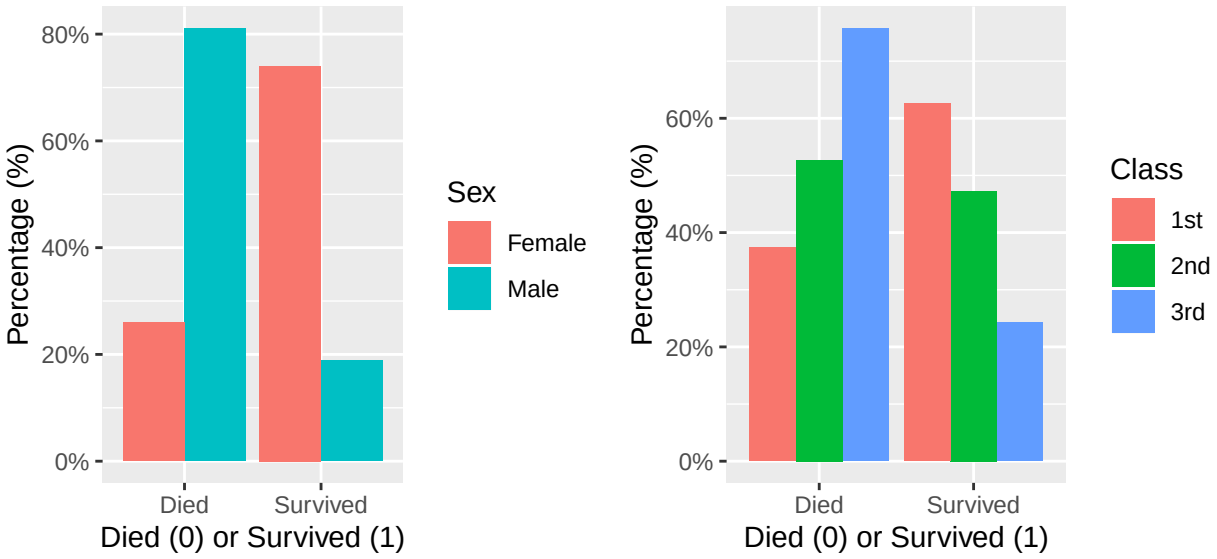


Figure 1: Bar Plots of the Percentage of Survivors by Gender (L) and Passenger Ticket Class (R).

To the left of Figure 1 we have a bar chart of the percentage of passengers that died and those that survived grouped by their gender. We see that of those passengers who were female, roughly 25% of them died and about 75% survived whereas for males aboard the Titanic, just over 80% of them died and slightly under 20% survived. This represents a huge discrepancy between the two genders and could likely be due to women being given priority for spots on the limited number of lifeboats. On the right of Figure 1 we see the same plot as that on the left however this time we have a breakdown of the percentage of passengers who died or survived based on their passenger class. The plot shows us that the passengers who survived the most came from $1^{st}$ class where just over 60% survived whereas only about 48% of those in $2^{nd}$ class survived and about 24% of those in $3^{rd}$ class survived.

Another point of interest for our analysis is to see how the ages differ of those passengers who died and those who survived, we can see that to the left of Figure 2 below that the ages between those that died and those that survived don't differ by much. We can observe that theres a slight right skew for the passengers that died, that is, there may be more older passengers who died compared to those that survived although we do see some outliers at the upper whisker of the survived class. The plot on the right hand side of Figure 2 gives us slightly more detail with a breakdown by passenger class where we can see that the passengers in $1^{st}$ class do appear to be older than those in the other two classes but those in $1^{st}$ class that died on the Titanic were older on average compared to those that survived. For passengers in $2^{nd}$ class, there it seems to be a bit more even when it comes to the ages of those that died and survied with those having died being ever so slightly older on average. Lastly for those in $3^{rd}$ class there is a much greater spread when it comes to the ages of those who died and survived - especially for the passengers that died where we see a huge right-skew suggesting those that died were on average older passengers - but when it comes to those that survived we see a huge left-skew suggesting those that survived from $3^{rd}$ class were on average younger passengers.

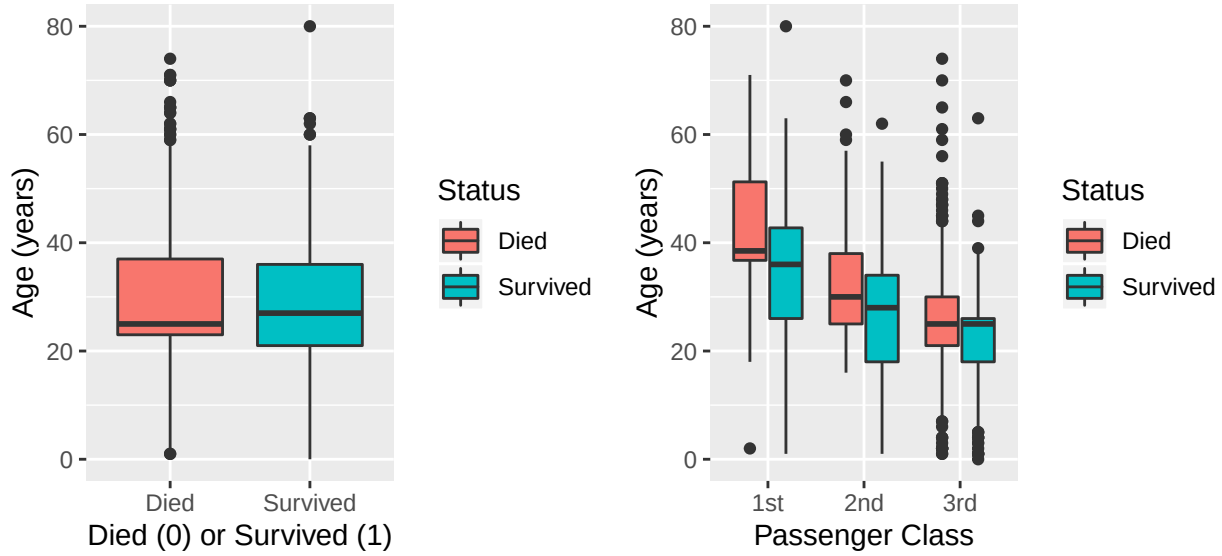


Figure 2: Boxplots of the Ages of Passengers who Died and Survived (L) and by Passenger Class (R).

Our final visualisation for this section is displayed in Figure 3 where we have boxplots showing the Fare prices paid depending on where passengers embarked the Titanic and whether they died or survived.

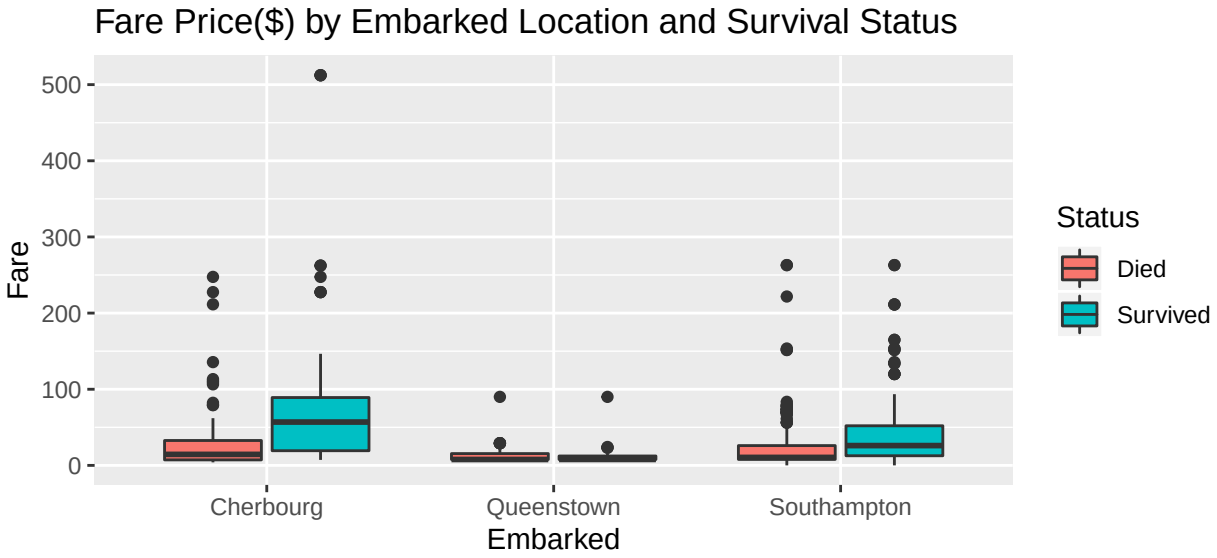


Figure 3: Boxplots of the Fare Price by Embarked Location and Survival Status

Looking to Figure 3 we observe that of the three embarked locations, it appears as though Queenstown had the cheapest fare prices overall, followed by Southampton and then Cherbourg. When we examine the prices paid by passenger survival status, we begin to see a similar trend to that in Figure 1 which showed the higher the class of ticket you had, the greater the chance of survival. For passengers that embarked at Cherbourg we see that for those that survived, they paid a much greater ticket fare than those that died and this is broken down into a numerical summary in Table 4 which shows the mean fare price for those that died were\$35.44 whereas for those that survived, they paid \$79.72 on average. We do also see a huge difference between the median prices paid aswell. Looking at passengers who boarded in Queenstown, there isn't actually much of a difference between the groups with the mean prices for the died and survived groups being \$13.34 & \$13.18 respectively. Similarly, the median fare prices are also inline with one another. We do see a slight difference for the two groups for those passengers that embarked in Southampton with the mean price paid for those who died being \$20.74 and \$39.55 for those that survived. Figure 3 does show quite a significant number of outliers which is what we would expect to see given passengers could pay for either $1^{st}$, $2^{nd}$ or $3^{rd}$ class tickets with more people likely being able to afford the lower two classes which would cost substantially less than $1^{st}$ class.

Table 4: Mean and Median Fare Prices for Embarked Location and Survival Status.

| | Embarked Location | | |
|---|---|---|---|
| | C | Q | S |
| **Died** | | | |
| Mean | 35.44 | 13.34 | 20.74 |
| Median | 14.46 | 7.75 | 10.50 |
| **Survived** | | | |
| Mean | 79.72 | 13.18 | 39.55 |
| Median | 56.93 | 7.81 | 26.00 |

# 4 Logistic Regression Classifier

The next stage of our analysis is the application and assessment of the Logistic Regression classifier but before we begin, we start by splitting our dataset into training and test sets in a 60%-40% split with 60% going to the training data and 40% going to the test data. The reasoning behind this is that we use the training data to fit our classification method and this is then used to predict the responses for the observations in the test set where we assess our models predictive performance.

We use a Logistic Regression model for our data here because the outcome variable- whether passengers survived or not- is binary and as such this model with a logit link function is appropriate. Our starting point for the model is to consider the fully saturated model, ie. the model containing all covariates in the dataset and we will then apply stepwise regression using both backwards elimination and forward selection to determine the optimal model. The final model deemed to be optimal will be the one which minimises the Akaike Information Criterion (AIC) score.

After fitting the model through the process described above, the optimal model returned an AIC score of 489 and contained the covariates Passenger Class, Sex, Age and the number of siblings/spouse onboard. This therefore means that the fare price paid by passengers, where they boarded the Titanic and the number of parents/children they had aboard the Titanic were not deemed to have a statistically significant effect on whether a passenger died or survived. We can now move to visualise the effects of the covariates in the final model on passenger survival below in Figure 4

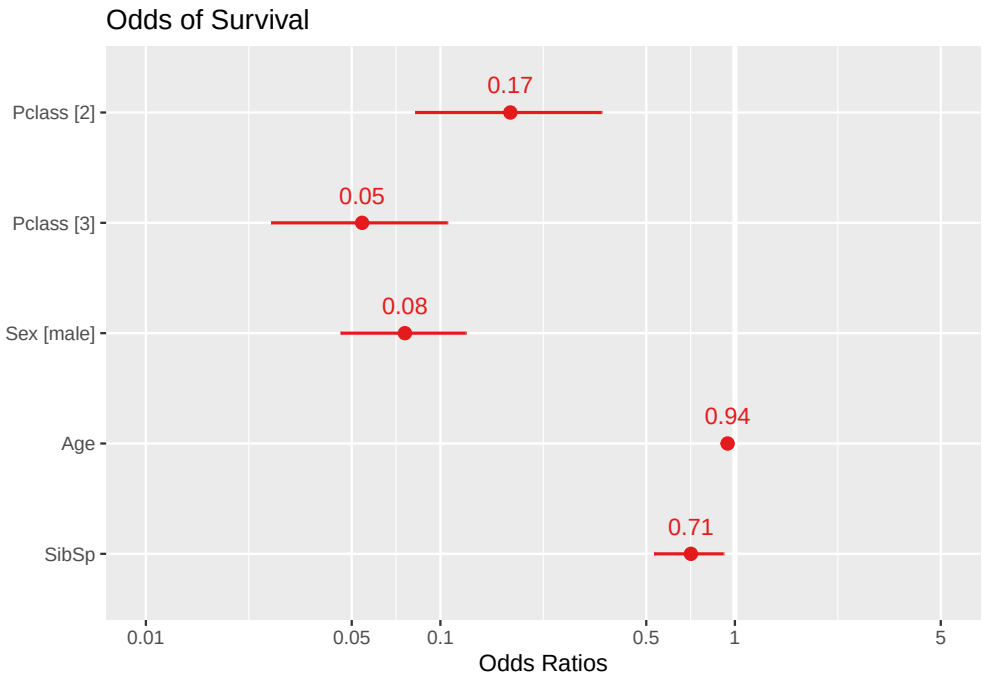


Figure 4: Effect of each Covariate on Odds of Survival.

From Figure 4 we see that the odds of survival for those passengers in $2^{nd}$ class and $3^{rd}$ class are on average 79% & 94% less than those passengers in $1^{st}$ class. This represents a huge decrease in the odds of survival and perhaps may be due to the $1^{st}$ class lounges being higher up on the ship so it took longer for the water to reach them or they may have perhaps been given preferential treatment when it came to boarding the life boats. We also observe that the odds of survival for those passengers who were male are 92% lower than female passengers and that for every year increase in the age of passengers, the odds of survival are about 5% lower on average. Similarly, for every extra spouse/sibling a passenger had aboard the Titanic their odds

of survival decreased by roughly 26% which could be likely due to them wanting to make sure they don't leave a family member behind before trying to disembark which could take valuable time.

We can now use our model to predict the probability and class that each passenger in the test set belongs to. In order to do so, we need to set a classification threshold/cut-off probability which in this case we make 0.5 so that if for example the probability of a passenger being assigned to the survived group is less than 0.5 then they are assigned to the died group and if greater than 0.5 then they are assigned to the survived group.

Table 5: Confusion Matrix of Ture and Predicted Class of Observations.

| | 0 | 1 | Sum |
|---|---|---|---|
| 0 | 182 | 33 | 215 |
| 1 | 38 | 103 | 141 |
| Sum | 220 | 136 | 356 |

Above in Table 5, we have a confusion matrix where we have the true classes as the columns and predicted classes as the rows and as we can see, it appears as though this Logistic Regression model with a cut-off of 0.5 has performed pretty well by correctly classifying 182 observations to the died group and correctly classifying 103 observations as belonging to the survived class. This therefore gives the classifier a Correct Classification Rate (CCR) of 80% and a misclassification rate (MCR) of 20%.

# 5 Conclusion

In conclusion, it appears as though all the covariates considered in this analysis played somewhat of a significant role in the survival status of passengers aboard the Titanic. Arguably the biggest factors that determined the survival status were the passenger ticket class and the sex of the passenger as the plots in Figures 1 show and also as suggested by the Logistic Regression model. The Logistic Regression model used for this analysis overall performed very well by correctly classifying 80% of the observations. To further improve this analysis in the future, we could consider the use of other classification methods such as K-Nearest Neighbours, Support Vector Machines and Random Forests, this would then allow us to form a comparison between the classifiers and understand what would be the best for these data.