Important predictors for survival on Titanic

Thomas Selin

Introduction and data exploration

This is an analysis of passenger data from the Titanic ship maiden voyage in 1912 and whether a specific passenger survived the accident which caused the ship to sink.

I started with exploring the data. I first had a look at the age of the passengers, see Figure 1, and then had a look at the distribution of the interesting variables **Pclass**, **Sex** and **Survived**, see Figure 2.

Then analyzed which numerical variables in the dataset was related by calculating the correlations between the variables, see Figure 3.

As it is not specified if the variable **PassengerId** describes values on some sort of scale, it was not included in the model. Also, it had low correlation to the **Survived** outcome.

The **Ticket** and **Cabin** variables are very unclear as for the meaning of them, and I therefor considered these values are not valuable to use in the model. Also, the **Name** variable was not used in the model as I didn't expect it, without any other related data, to be a significant factor for predicting the survival chance and would lessen the explainability of the model.

Regression modelling and results

I performed a multiple logistic regression and backwards selection when creating the model. The variables that was statistically significant and was included in the final model was **Pclass**, **Sex**, **Age** and **SibSp**. Including **Parch** lead to only minimal increase in accuracy (but also a minimal increase in AIC value). It was therefore deemed as not adding sufficient value to the model to include it.

Removing any of the variables didn't decrease the AIC value of the model, which is 649.

The prediction accuracy when evaluating on the same data set is 80,2 %.

Discussion

This analysis aimed to discern the key variables that affected a Titanic passenger's chance of surviving the accident.

The data exploration and regression analyses revealed some important findings. Some limitations exist that should be taken in consideration when judging the accuracy of the model:

The findings might not very useful for predicting chance of survival on other ships that had a similar accident because of the limited variables that was included in the data. Variables such as data about the journey, ship construction and the safety of the ship and weather would likely be useful for making a model that is not so specific to Titanic.

Also, this happened long ago which can influence the relevance of the model in many unknown ways. This means, this model should be seen as interesting mainly from an historical perspective.

Perhaps one variable that would have been interesting if the data set contained would be the language the passenger was able to speak. If we also knew what language staff were able to speak, it could be interesting to see if whether a passenger could speak to the staff had a significant relation with the chance of survival. That could then be because of they had easier getting help from staff.

The result of the analysis indicates that the variables **Pclass**, **Sex**, **Age** and **SibSp** was most useful in modeling the chance of survival on Titanic.

Supplementary material

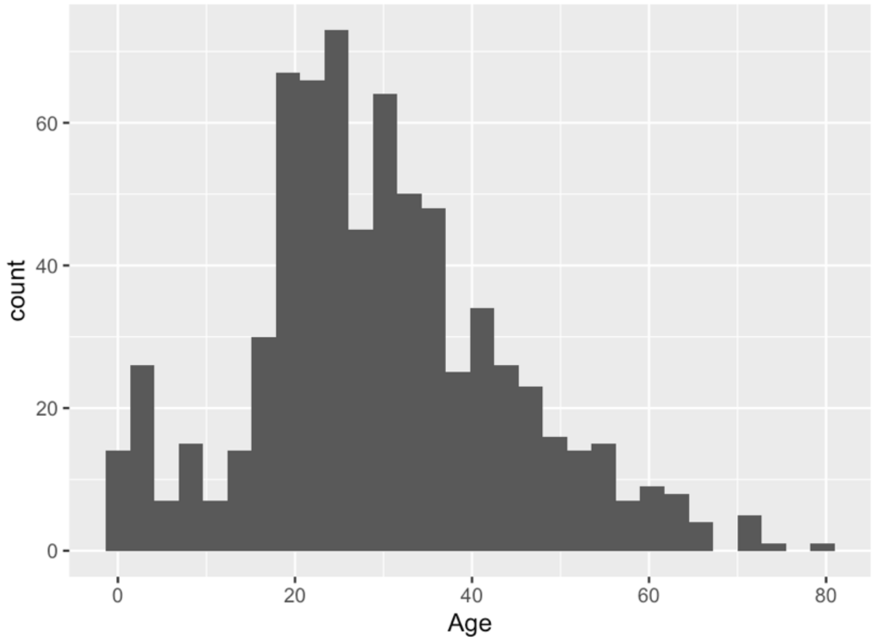

Figure 1. Histogram of age of passengers

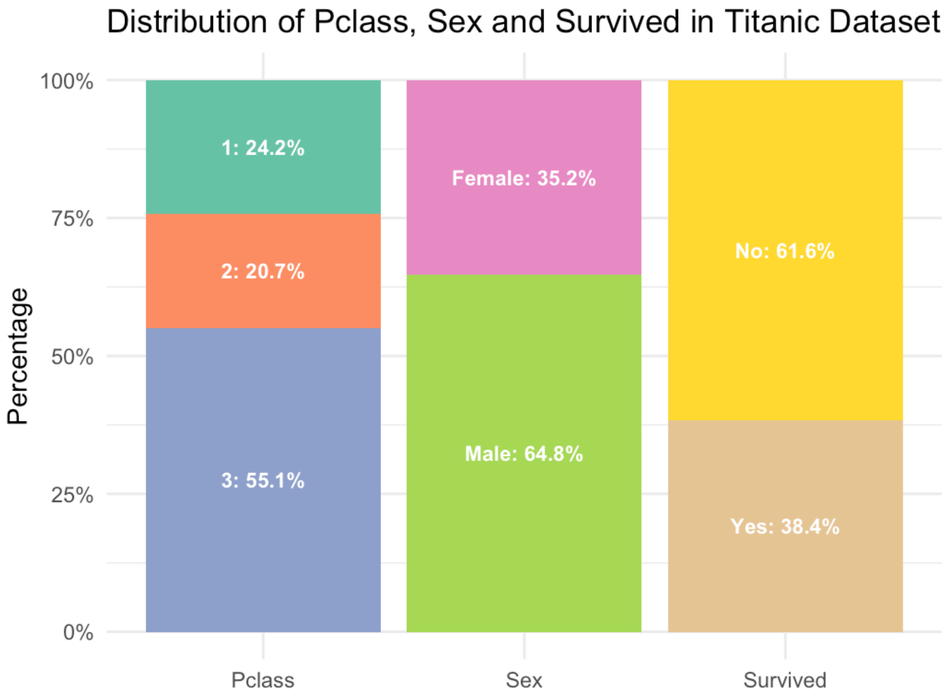

Figure 2. Distribution of Pclass, Sex and Survived in Titanic Dataset

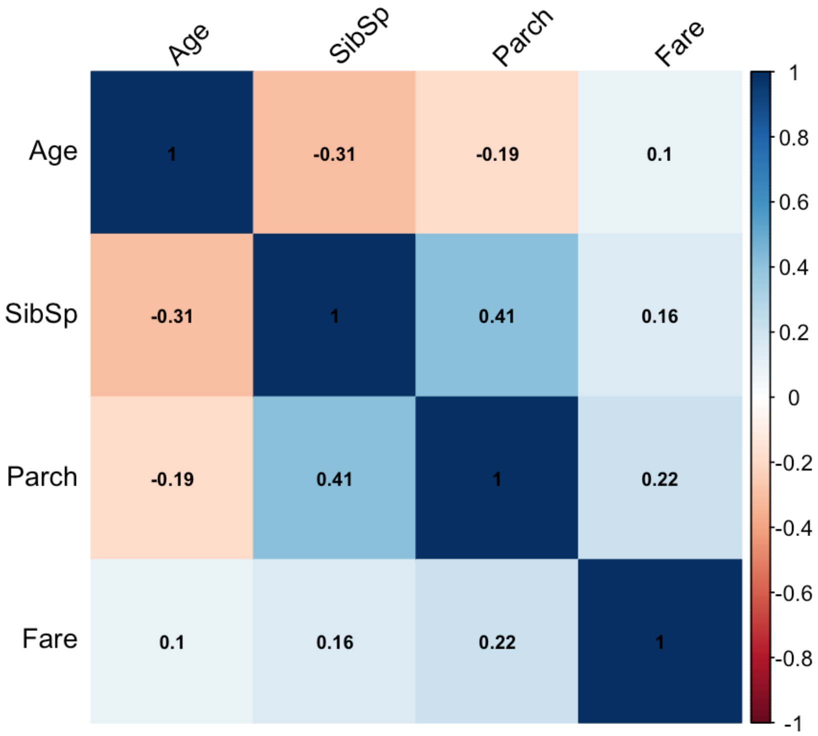

Figure 3. Correlations between variables