Survival chances of Titanic passengers and predictors influencing their outcome - Assignment 3

Emelie Wärmlund

2ST065 Data analytics with R, Summer 2024

Introduction and exploratory data analysis

The data used for this analysis is Titanic data, to explore which attributes can explain the survival chances of a passenger. More information regarding the data set can be found here.

First, the variables were compared to each other using exploratory data analysis, to get a better picture of which variables could relate to each other, and specifically the survival attribute. Due to the variable being predicted being a factorial, barcharts and boxplots were used to explore relations between different attributes instead of a correlation plot.

When comparing two categorical variables, bar charts were used, see Figure 1. The charts showed that more men tend to not survive than women and more passengers in third class tended not to survive compared to the other two classes.When comparing numerical values to the survival variable, boxplots were used, see Figure 2. The plots showed a small tendency of more passengers with a lower fare price tended to not survive than passengers with a higher fare price. This boxplot did also show that the data had a lot of outliers, which can be data errors or extreme values. The boxplots also showed a small tendency of people at a lower age to have a larger chance of survival.

Regression modeling

Based on the results of the EDA, a logistic regression model was fitted using Survival as the response variable. Due to Survival being a categorical variable, a logistic regression model was used. The fit was explored in various models using a mixed selection strategy, continuously removing variables with insignificant p-values. The VIF values were also checked for each fit to assess the potential risk of multicollinearity and for all attributes used in the final fitted model are low, all below *1.2*.

Results

The attributes in the fitted model are *Pclass*, *Sex* and *Age*. The variable *Fare* was not significant as initially thought during the EDA and was therefore discarded. The coefficients of each attribute suggest that the probability of survival decreases if you're male, was a passenger in class two or three and if you're older.

The coefficients and their respective p-values for the attributes in the model considered the best fit is shown in Figure 4. The AIC of the fitted model is at *657.28* and the accuracy at *78.85%*.

Discussion

This analysis showed some important attributes that affect the probability of survival at the Titanic using the data given. The results show a negative correlation between survival and age, sex and passenger class. If you were an older male aboard the Titanic in class three, the chances of survival were not high. I had some prejudices before starting my analysis, based on watching documentaries and the 1997 movie "Titanic". I therefore thought about the saying

"Women and children first", thinking that the variable for gender, age and the variable describing the number of children per passenger would have a high significance. I was correct regarding the variables of age and sex, but not the variables describing the number of children per parent, *Parch*. I also had a prejudice about passengers at class one having a higher chance of survival than the passengers at other classes. Therefore I also thought that the ticket prices being higher for a seat in first class would be higher than in third class, and thus the variable for Fare could be a good predictor for survival rate. This however was not correct. The variable for class had a low p-value and low VIF value, but the fare variable had no significance and a high p-value. This may be explained by the high number of outliers in the data, giving an incorrect result. For future studies, a suggestion would be to remove the outlier data from this variable and see if the significance changes.

The analysis has some limitations, one being the number of observations compared to actual passengers. The number of observations in the dataset is at *891,* when the actual number of passengers aboard the Titanic were at *1316* excluding the crew members. (Geller, 1998). The data therefore doesn't cover all passengers and the fitted model can lose accuracy in its prediction due to missing data. Another limitation is that the prediction uses the same data as the fitted model. This can cause the model to be "overfitted". This model will work well for this data and perhaps other passenger data of the Titanic, but due to not being predicted on test data, the risk is high that it will not work as well for other passengers lists from other voyages. Another limitation is that this model is highly specific, and its usage may not be that wide. The model may only work for predicting survival rates on similar boats during the same time period with the same structure of passenger classes. This model would not work for more modern cruises due to the great development in technology and safety over the last hundred years.

Suggestions for future studies would be to use other test data from the same time period with similar circumstances, (for example time period), for the prediction to evaluate and find ways to make the model more accurate for wider sets of data. Furthermore, the model can perhaps later on be compared to models based on data from more modern voyages to explore if the same predictors are still the same many years later.

Figures

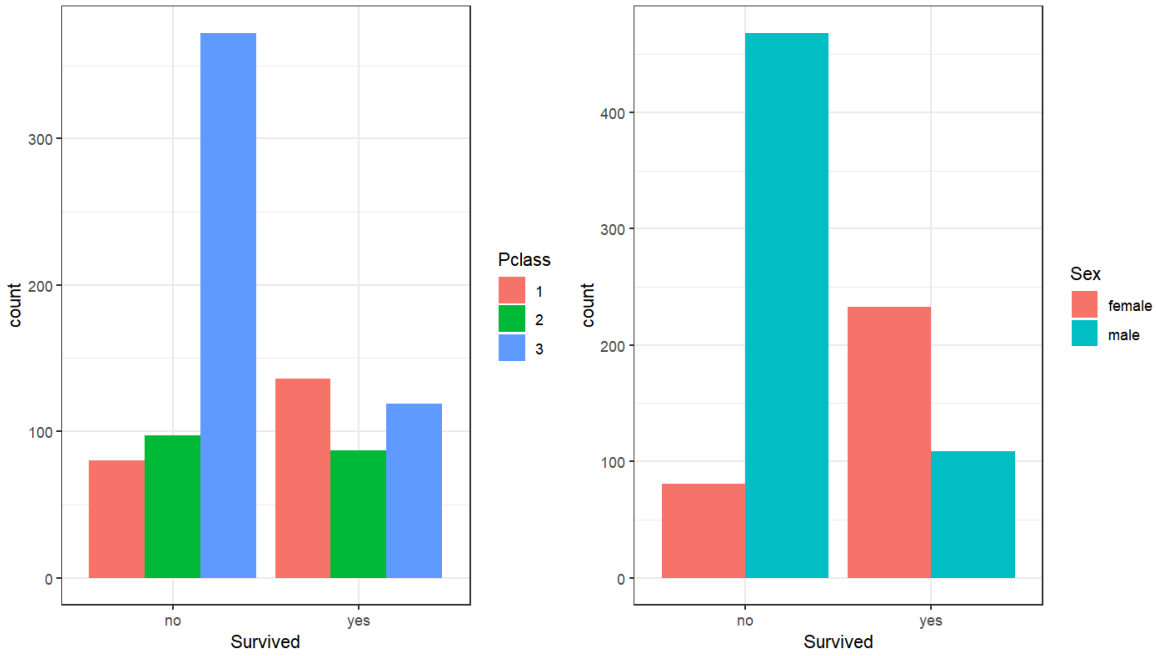

Figure 1: Barcharts showing the number of survivors in relation to gender and passenger class.

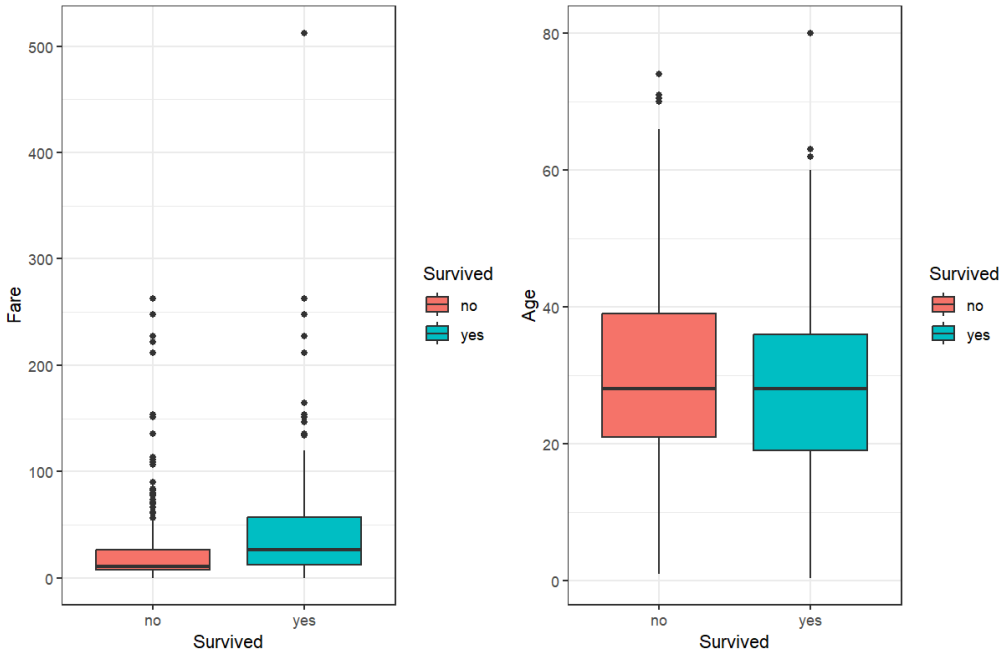

Figure 2: Boxplot showing the distribution of fare prices in relation to survival rate

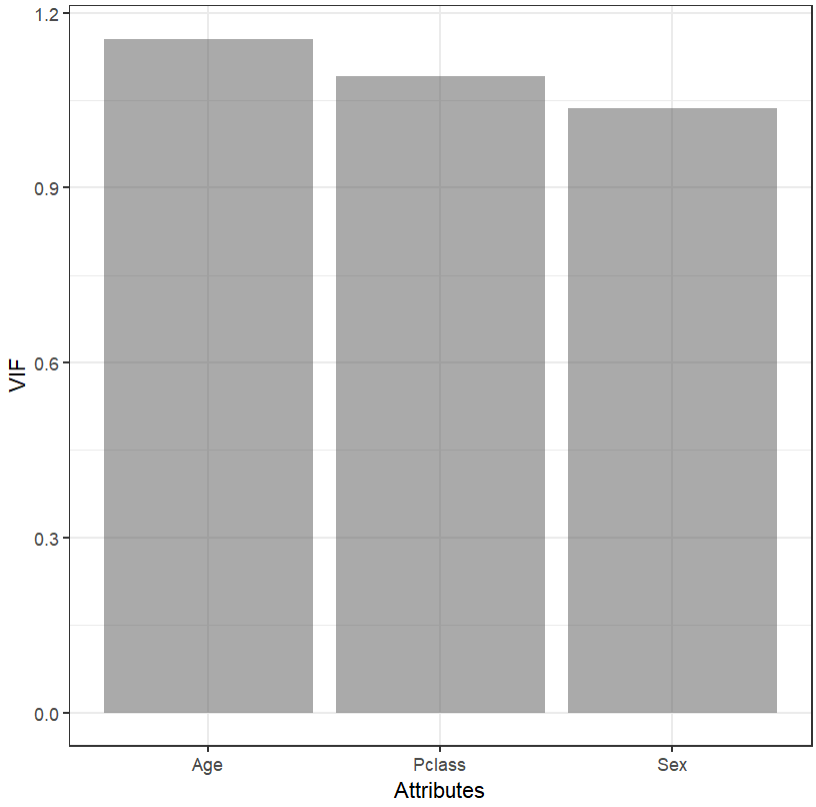

Figure 3: VIF Values for attributes used in the fitted logistic regression model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
Age         -0.036985   0.007656  -4.831 1.36e-06 ***
---
```

Figure 4: Table showing the coefficients and p-values for the predictors. sexFemale and Pclass1 are referencial categories

References

Geller, Judith B. (October 1998). Titanic: Women and Children First. W. W. Norton & Company. p. 197