**Predicting Titanic Survival**

Frannie K

# Question

Kaggle posted a competition to predict passenger survival on Titanic (<https://www.kaggle.com/c/titanic>). Dependent variable is the survival itself and a set of independent variables is a mix ore scale, ordinal and nominal variables, which are:

* Pclass – passenger class {1,2,3}
* Name – Name of passenger
* Sex – Sex of passenger
* Age – Age of passenger
* Sibsp - Number of siblings/spouses aboard
* Parch – number of parent/children aboard
* Ticket – Ticket number
* Fare – Passenger Fare
* Cabin – Cabin Number
* Embarked – Port of Embarkation {C,Q,S}

I have used exploratory analysis to reduce down the number of input variables and used discriminant analysis to determine membership of each observation.

# Data Treatment

Please refer to the Rmarkdown (Figure 1) for data cleansing steps undertaken. These include:

* Treatment of missing values
* Mending object class (e.g. factor, numeric, character.. etc)
* In SPSS, the data underwent another manual adjustment on type of data (scale, ordinal, nominal)



Figure 1

# Exploratory treatment

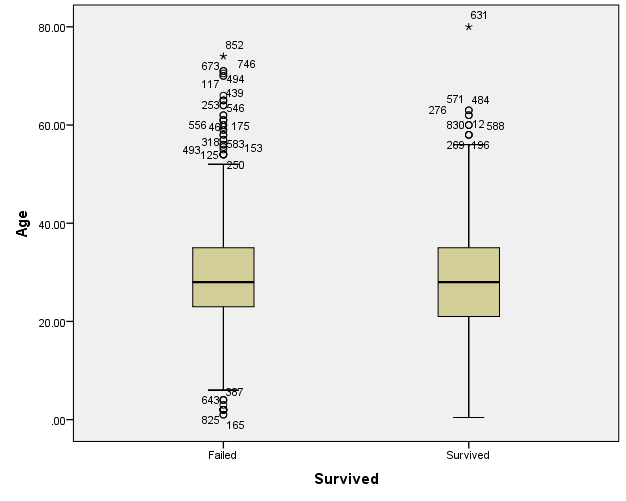
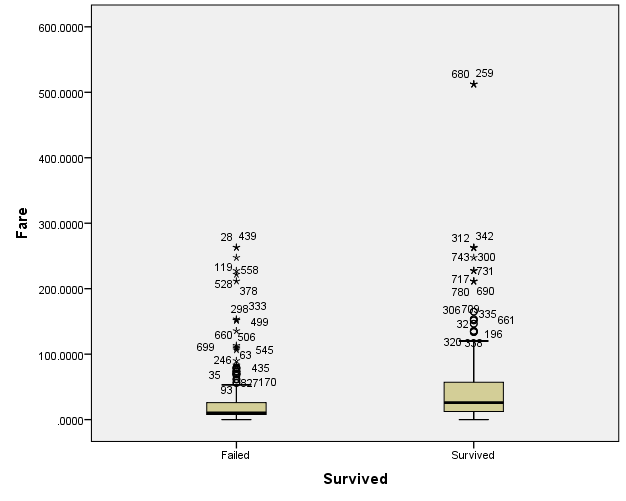
The independent variables are distinguished between1) parametric and 2) non-parametric (categorical) variables.

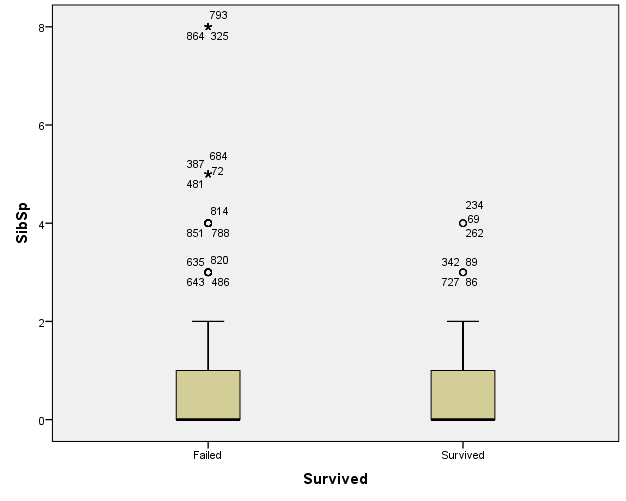
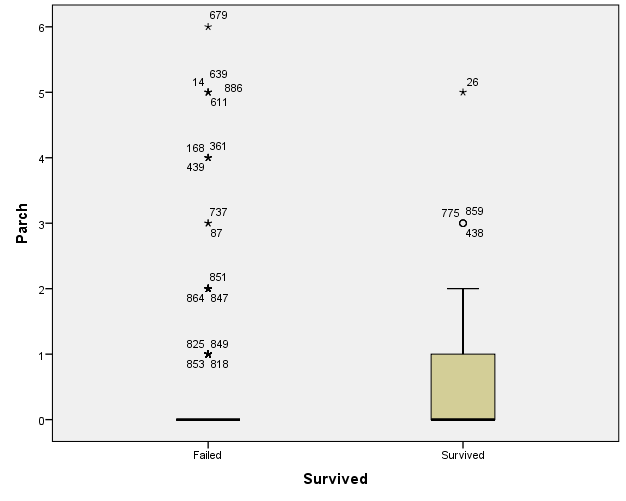
1. parametric variables consist of Age and Fare but as later pointed out, Fare is highly correlated with Pclass and therefore is substituted by categorical Pclass variable. For these, we analyse descriptive stats and ANOVA to test the significance of correlation with Survival rate. Table 1 displays the descriptive stats. It is noted the mean age of all passengers in training group is 29.4, the mean fare they paid across all passenger class is $32 and on average each passenger had less than 1 sibling/spouse and less than 1 parent/children associated.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | | | | | | |
|  | N | Range | Minimum | Maximum | Mean | Std. Deviation | Variance | Skewness | | Kurtosis | |
| Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Age | 891 | 79.58 | .42 | 80.00 | 29.3616 | 13.01970 | 169.512 | .510 | .082 | .994 | .164 |
| Fare | 891 | 512.3292 | .0000 | 512.3292 | 32.204208 | 49.6934286 | 2469.437 | 4.787 | .082 | 33.398 | .164 |
| SibSp | 891 | 8 | 0 | 8 | .52 | 1.103 | 1.216 | 3.695 | .082 | 17.880 | .164 |
| Parch | 891 | 6 | 0 | 6 | .38 | .806 | .650 | 2.749 | .082 | 9.778 | .164 |
| Valid N (listwise) | 891 |  |  |  |  |  |  |  |  |  |  |

Table 1

Before going into statistical significance test, we also look at whether the five-number summary and deviation metrics differ by Survival. This is shown through the boxplots below. From initial review, Age is more dispersed in Failed but the mean value and interquartile range (IQR) seem largely similar. Fare seems slightly higher in Survived category on average and by the data points in the above upper 1.5IQR. SibSp seems largely the same across the Survival status and Parch mean was the same across the category at zero but those with 1-2 parents or children survived more and anything more 2 tend to die more. Based on this, I decide to remove SibSp from the input variable set as it seems the variable has little impact on the Survival rate.

1. Nonparametric variables such as Pclass, Sex and Embarked are analysed through median, mode and chi-squared test to measure the significance of correlation with Survival rate.

# Prediction

However, we know from various studies, over-fitting a model does not always result in better forecasting. Moreover, in real-world application, data can be scarce and input variables may be limited. Therefore, I also look at significance of correlations between independent variables and also with dependent variable to filter only the key variables for modelling.

To determine the test group’s survival, I’ve used discriminant analysis. The model can intake all types of input variables and the dependence variable is categorical as is the case here.