Who Survived on the Titanic - Logistic Regression Analysis with Titanic Data

➢ Basic Description

Age			Sex		
Survived	Adult	Child	Survived	female	male
no	405	38	no	154	709
yes	255	58	yes	308	142

Table 1. Number of survivals in term of Age and Sex

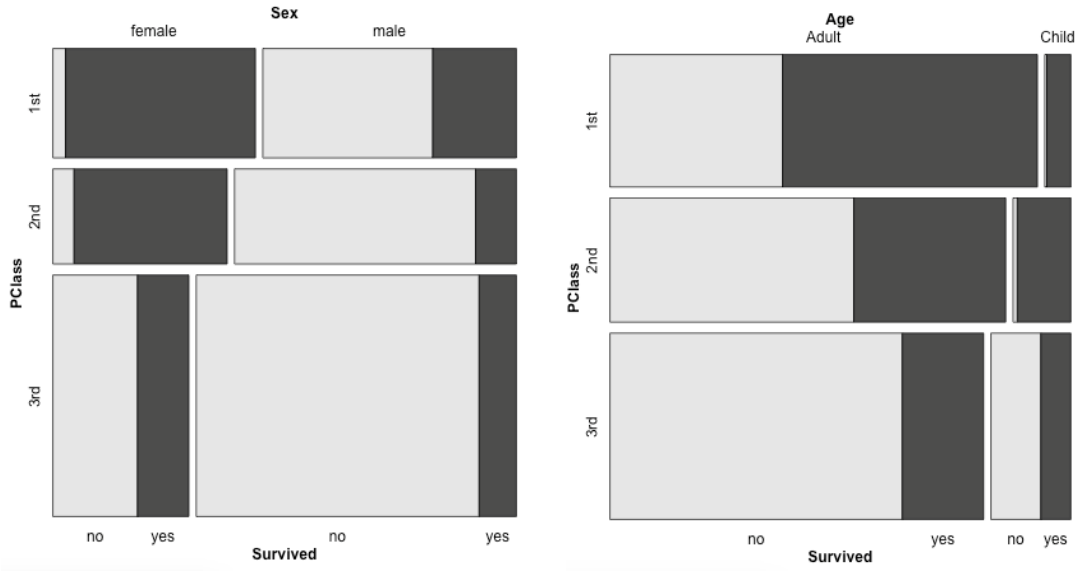

Figure 1. Mosaic graph for Age and Sex

Basically, we see that women were more likely to survive than men. Childs were more likely to survive than adults. Odds of a female surviving is 2.16 and of a male surviving is 0.20, so the odds ratio in favor of survival if passengers were women versus a man on the Titanic was 10.82. The relative likelihood of a female versus a male surviving was 4.10, which is just the proportion of female who survived divided by the proportion of males who survived. So female was almost 4.1 times as likely to survive. In terms of age variable, the odds of a child surviving is 1.52 and of an adult surviving is 0.62, so the odds ratio in favor of survival if the passenger were a child versus an adult on the Titanic was 2.42. The relative likelihood of a child versus an adult surviving was 1.56. In this sense, it seems female and the child may have a higher survived rate.

➢ Logistic Regression

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.759662   0.397567   9.457  < 2e-16 ***
PClass2nd   -1.291962   0.260076  -4.968 6.78e-07 ***
PClass3rd   -2.521419   0.276657  -9.114  < 2e-16 ***
Sexmale     -2.631357   0.201505 -13.058  < 2e-16 ***
Age         -0.039177   0.007616  -5.144 2.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1025.57  on 755  degrees of freedom
Residual deviance:  695.14  on 751  degrees of freedom
  (557 observations deleted due to missingness)
AIC: 705.14

Number of Fisher Scoring iterations: 5
```

```
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-3.0869  -0.6453  -0.4643   0.4599   2.3346

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.845505   0.598061   8.102 5.41e-16 ***
PClass2nd          -1.486038   0.587018  -2.532 0.011357 *
PClass3rd          -4.038030   0.544289  -7.419 1.18e-13 ***
Sexmale            -3.702774   0.507177  -7.301 2.86e-13 ***
Age                -0.044854   0.008179  -5.484 4.16e-08 ***
PClass2nd:Sexmale -0.089869   0.656052  -0.137 0.891043
PClass3rd:Sexmale  2.256406   0.581805   3.878 0.000105 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1025.57  on 755  degrees of freedom
Residual deviance:  664.84  on 749  degrees of freedom
  (557 observations deleted due to missingness)
AIC: 678.84

Number of Fisher Scoring iterations: 5
```

Table 2. Logstic regression model I & II

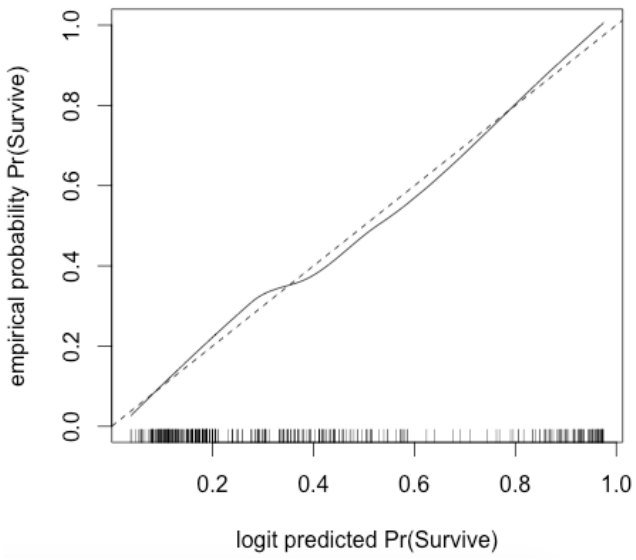

Figure 2. Plot for GLM model

We can see from the left model, that all predictors, passenger class, age and gender, are statistically significant (with $p<0.001$). The passenger class is recorded as 1st, 2nd and 3rd, so negative coefficient -1.29/-2.52 means that as the passengers' class increases, the probability of surviving decreases. In the second model, we use interaction term of 'class and sex', which is statistically significant only for the passengers in the 3rd class. Based on the data, we could also use the predict function to see GLM fits. According to the result, the model predicts that there are 97.1% chances survive for the first passenger, who is a first class female. There are 99.1% chances survive for the second passenger,

who is a first class child. In figure 2, we propose assessing a model's predictive accuracy by constructing a plot with the predicted probability of survival on the x-axis, and the empirical proportion of survivors with that predicted probability on the y-axis. The empirical proportion is computed by running a lowess regression of the model's predicted probability against the binary (1/0) survival variable. The logit model actually does a good job of predicting the probability of passenger survival in the test data.

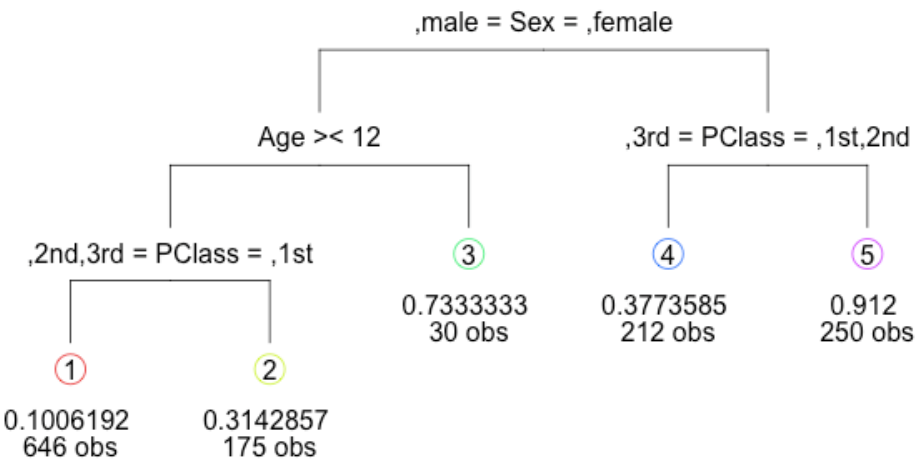

Figure 3. Maptree

In the cooperative map tree graph, it gives a clear description of surviving rate in terms of age, sex and classes. The surviving of men passengers were mostly determined by age and the surviving of women passengers were more likely determined by class. The first and second class female passengers achieved a very high survived rate, so does the child male passengers who were under 12 years old.