**Lab report for the logistic regression project**

The aim of this report is to analyze what kind of factors influenced the survival of passengers that were on board the Titanic. More specifically it will look into whether 4-year-old Sue's and 20-year-old Kate's chances of surviving Titanic would have increased if Kate's spouse and Sue's father Leonardo had accompanied the pair. Logistic regression is used to determine the effect of the different predictors.

**Methods and Results**

*Data preparation and cleaning*

Based on the initial exploration and analysis, the factors initially selected to be part of the model are the individual's gender (Sex), ticket class (Pclass), Age, whether they were accompanied by siblings or spouses (SibSp), or parents or children (Parch), fare of the ticket (Fare), and the port of embarkation (Embarked). All of these were determined to have some effect on the chances of survival.

Data cleaning included replacing missing values of the port of embarkation with the mode of Southhampton, and the age with the median of 28 years. In order to increase the prediction accuracy and simplify the analysis, the values of 'Age' and 'Fare' were grouped together (see Table 1).

**Table 1.** Selected predictors

| ***Predictor*** | ***Definition*** | ***Notes*** |
|---|---|---|
| Survived | Survival | 0 = No, 1 = Yes |
| Pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| Age_group | Age group | 0-13, 14-24, 25-38, 39-64, 65+ |
| Sibsp | Number of siblings or spouses accompanying the passenger onboard | Amount (numerical) |
| Parch | Number of parents or children accompanying the passenger onboard | Amount (numerical) |
| Fare_type | Price group of the ticket | low = 0-10£, medium = 11-79£, high = ≥80£ |
| Embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

*Results*

Based on the initial data exploration and analysis a logistic regression model was built where the dependent variable is survival and independent variables are the individual's gender, ticket class, age group, how many siblings or spouses, or parents or children they had with them, their place of embarkation, and the fare type.

The logistic regression model (see Table 2) had a significantly better model fit compared to the null model (AIC of the model = 796.53, -2LL of the model = 768.53, AIC of null model = 1188.66, -2LL of the null model = 1186.66). The LogLikelihood of the regression model is also significantly higher (-384.26) than in the null model (-593.33). The model explained 35% of the variance (McFadden R^2 =0.35). Out of all passengers, 38% survived (342 out of 549 passengers).

The final model correctly predicts survival in 73% of the cases and non-survival in 87% of the cases. The overall correct prediction rate was 81%.

**Table 2.** Logistic regression model

| *Predictors* | *Estimate* | *OR* | *CI* | *Z-value* | *P-value* |
|---|---|---|---|---|---|
| (Intercept) | 4.87004 | 130.33 | 36.90 – 489.47 | 7.394 | **<0.001** |
| Sex (male) | -2.80408 | 0.06 | 0.04 – 0.09 | -13.338 | **<0.001** |
| Pclass (2) | -0.90747 | 0.40 | 0.23 – 0.71 | -3.147 | **0.002** |
| Pclass (3) | -1.95002 | 0.14 | 0.07 – 0.27 | -5.805 | **<0.001** |
| Age_group (14-24) | -2.11658 | 0.12 | 0.05 – 0.28 | -4.851 | **<0.001** |
| Age_group (25-38) | -1.99649 | 0.14 | 0.06 – 0.30 | -4.875 | **<0.001** |
| Age_group (39-64) | -2.56281 | 0.08 | 0.03 – 0.18 | -5.727 | **<0.001** |
| Age_group (65+) | -3.81237 | 0.02 | 0.00 – 0.15 | -3.294 | **0.001** |
| SibSp | -0.47078 | 0.62 | 0.48 – 0.80 | -3.585 | **<0.001** |
| Parch | -0.21162 | 0.81 | 0.62 – 1.04 | -1.620 | 0.105 |
| Fare_type (medium) | 0.23499 | 1.26 | 0.69 – 2.30 | 0.769 | 0.442 |
| Fare_type (high) | 0.57500 | 1.78 | 0.63 – 5.09 | 1.080 | 0.280 |
| Embarked (Q) | -0.04892 | 0.95 | 0.44 – 2.03 | -0.126 | 0.900 |
| Embarked (S) | -0.43764 | 0.65 | 0.40 – 1.04 | -1.795 | 0.073 |
| Observations | 891 | | | | |
| AIC | 796.53 | | | | |

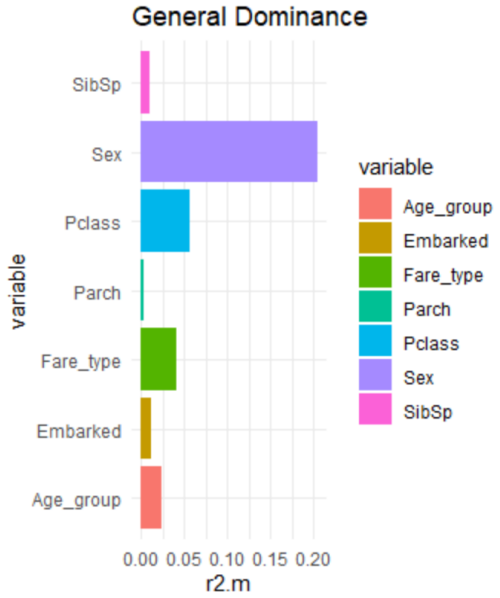
**Table 3.** The null model

| *Predictors* | *Estimate* | *OR* | *CI* | *Z-value* | *P-value* |
|---|---|---|---|---|---|
| (Intercept) | -0.47329 | 0.62 | 0.54 – 0.71 | -6.87 | **<0.001** |
| Observations | 891 | | | | |
| AIC | 1188.7 | | | | |

From the logistic regression model it can be observed that being a female, part of a younger age group, traveling in the first class, traveling alone, embarking from Cherbourg as well as paying for a higher ticket fare are improving individuals' chances of survival, while other factors have a negative effect on survival. The number of parents or children on board, as well as the port of embarkation and fare type, are not statistically significant ($P > 0.05$).

Dominantly the most influential predictor is gender (see Figure 1). The fare type, as well as ticket class, seemed to have a medium effect on survival, with the age group's effect being a little bit more dominant than the port of embarkation and presence of family members which have minimal effect on survival.

**Figure 1.** Predictors' dominance analysis



The regression equation can be calculated from the model. This helps answer the research question of whether Sue's and Kate's chances of survival would have been enhanced if Leonardo had been onboard. What is known is that both Kate and Sue are females, they had third-class tickets, Sue is 8 years old while Kate is 20, they embarked from Southampton (S), and the fare was 8 pounds for each of them.

So using the logistic regression equation we can calculate the following:

**Table 4.** Survival equations for Sue and Kate

| *Case* | *Regression Equation* | *Chances of surviving* |
|---|---|---|
| Sue without Leonardo | 4.87004 - 2.80408 (sex)*0 - 1.95002 (Pclass) * 1 - 0.47078 (SibSp) * 0 - 0.21162 (Parch) * 1 - 0.43764 (Embark) * 1<br><br>=2.27076 | 91% |
| Sue with Leonardo | 4.87004 - 2.80408 (sex)*0 - 1.95002 (Pclass) * 1 - 0.47078 (SibSp) * 0 - 0.21162 (Parch) * 2 - 0.43764 (Embark) * 1<br><br>= 2.05914 | 89% |
| Kate without Leonardo | 4.87004 - 2.80408 (sex)*0 - 1.95002 (Pclass) * 1 - 2.11658 (Age_group) * 1 - 0.47078 (SibSp) * 0 - 0.21162 (Parch) * 1 - 0.43764 (Embark) * 1<br><br>= 0.15418 | 54% |
| Kate with Leonardo | 4.87004 - 2.80408 (sex)*0 - 1.95002 (Pclass) * 1 - 2.11658 (Age_group) - 0.47078 (SibSp) * 1 - 0.21162 (Parch) * 1 - 0.43764 (Embark) * 1<br><br>= -0.3166 | 42% |

**Discussion**

As is shown in Table 4, Sue and Kate's chances of survival wouldn't have increased with Leonardo on board. With his presence on the Titanic, especially Kate's chances to survive would have instead quite clearly decreased. Leonardo on board affects more on Kate's chances more since accompanying spouses have a bigger (negative) effect than parents as the dominance analysis shows (see Figure 1).