Who could survive Titanic?

Logistic Regression - Team B11

Problem Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. Our train dataset contains the details of a subset of the passengers on board (count:891) and importantly, it reveals whether they survived or not. Test dataset (count:418) contains similar information but does not disclose the “ground truth” for each passenger.

Goal

Our goal is to determine whether there were any key characteristics shared by survivors? Were some passenger groups more likely to survive than others? Can we accurately predict survival? Using the patterns we found in the train data, we need to predict whether the other 418 passengers on board (found in test data) survived or not.

Exploratory Data Analysis

What is the distribution of numerical feature values across the data?

Total samples are 891 or 40% of the actual number of passengers on board the Titanic (2,224).
Survived is a categorical feature with 0 or 1 values.
Nearly 30% of the passengers had siblings and/or spouse aboard.
Fares varied significantly with few passengers (<1%) paying as high as $512.
Few elderly passengers (<1%) within the age range 65-80.

What is the distribution of categorical features across the data?

Names are unique across the dataset (count=unique=891)
Sex variable as two possible values with 65% male (top=male, freq=577/count=891).
Cabin values have several duplicates across samples. Alternatively, several passengers shared a cabin.
Embarked takes three possible values. S port used by most passengers (top=S)
Ticket features have a high ratio (22%) of duplicate values (unique=681)

Data Preprocessing

Dataset contains null values for a few columns. Therefore, we pre-processed the data in Python, where we imputed the missing values in the numerical columns with the mean values and the missing values in the categorical columns with the median values. Also, we removed columns like Name and Ticket number from the data as they were irrelevant from our problem perspective.

Methodology

To solve our problem, we executed logistic regression models on various platforms such as Python, Excel(StatTools) and SAS. We did this to confirm our findings, so that we can be confident about our predictions and insights. We ran multiple iterations to arrive at the most accurate and efficient logistic regression model.

Analyses, Outputs and Interpretations

After the 4 logistics regression iterations, we found the best regression model as shown in Figure 1.
Number of parents and children aboard(Parch), Embarked and Fare have $p>0.05$ and, hence they were disregarded from the calculation.On analyzing the Logistic Regression model, we can see that Passenger class, Sibling Spouse count, Age and Sex have proved to have a statistically significant $p<0.05$ effect upon Survival after multiple iterations of the regression model. We can also observe that passengers who have paid more fare for their cruise have a better chance of survival.

Who could survive Titanic?

Logistic Regression - Team B11

We also worked on different classifier models and compared their accuracy scores. We can observe that p-values and Wald values are significant for all variables.
As shown in Figure 2, we analyzed the Classification Matrix for accuracy(78.79%), Type- I (91) and Type- I I (98) errors. Model's prediction accuracy is 71.35% among the passengers that survived and 83.42% accuracy among the passengers who did not survive, with an overall accuracy of 78.79%.
We have also found the highest to lowest survival probability for each passenger in the train dataset.
Sex has highest positive coefficient, inversely as Pclass increases, probability of Survived=1 decreases the most. This way Age*PClass is a good artificial feature to model as it has the second highest negative correlation with Survived.

Figure: 1

Regression Coefficients	Coefficient	Standard Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coef)
Constant	2.45	0.42	5.78	0.00	1.62	3.28	11.62
Pclass	(1.17)	0.12	(9.79)	-	(1.41)	(0.94)	0.31
Sex	2.74	0.19	14.11	-	2.36	3.12	15.48
Age	(0.04)	0.01	(5.13)	0.00	(0.05)	(0.02)	0.96
SibSp	(0.36)	0.10	(3.44)	0.00	(0.56)	(0.15)	0.70

Figure: 2

Classification Matrix	1	0	Percent Correct
1	244	98	71.35%
0	91	458	83.42%

Summary Classification	Percent
Correct	78.79%
Base	61.62%
Improvement	44.74%

Model evaluation

All the independent variables selected were Passenger class, Sibling Spouse count, Age and Sex have proved to have a statistically significant effect upon Survival. From the excel regression model we discovered that, Higher class, young, independent and female passengers were more likely to survive.

References:

https://www.kaggle.com/c/titanic