Analyzing survival in the Titanic shipwreck using Logistic Regression

Jeshlin Donna J
Department of Metallurgical and Materials Engineering
Indian Institute of Technology Madras
Chennai, India
mm20b029@smail.iitm.ac.in

***Abstract*—The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. During her voyage on April 15, 1912, the Titanic sank after colliding with an iceberg, killing approximately 1502 passengers and crew out of the 2224 onboard making it one of the most disastrous shipwrecks in history till date. An eye-opening observation that came forth from the sinking of Titanic is the fact that some individuals had a better chance at surviving than the others. Kids and women had been given foremost priority. Social classes were heavily stratified in the early twentieth century. Firstly, the aim is use and apply logistic regression and exploratory data analysis to uncover previously unknown or hidden facts in the data set available and predict the survival possibilities for the given test data. The task is to conclude the study of which types of individuals are more likely to live than the others.**

***Index Terms*—Feature engineering, Data Science, Logistic Regression, Model Evaluation, Exploratory Data Analysis**

I. INTRODUCTION

The disaster that occurred over a century ago ripped off many parts of the Titanic during the fateful night. Many classes of people of all ages and gender were present on that fateful night. Regrettably, there were not enough lifeboats present to rescue all the 2224 passengers onboard. The dead included a large number of men whose place in life-rafts were given to the women and children onboard.

Logistic regression is a machine learning technique used to describe data and to explain the relationship between binary dependent variables and one or more independent variables. It is a supervised classification algorithm and is the appropriate regression analysis to conduct when the dependent variable is of binary type. In a classification problem, the target variable(output) y, can take only discrete values for given set of features(inputs), X. Logistic regression is a regression model built to predict the probability that a given data entry belongs to the particular category and models the data using the sigmoid function.

The goal of this research paper is to correctly predict who are more likely to survive the Titanic given a set of information. A predictive model based on logistic regression was built using passenger data so people's genders, their ages, what class of ticket they belonged from and many other features contributed to whether they would be lucky enough to survive or tragically perish on the Titanic. Predictive analysis is a method of determining important and useful patterns in broad data sets by combining statistical approaches to determine significant and useful trends in large data. Survival is predicted using the logistic regression algorithm based on different feature combinations.

In this paper, we conduct exploratory data analysis to excavate different knowledge existing in available data set and to perceive the impact of every field with respect to the passengers' survival by the use of "Survival" field analysis in between each field of the data set. Data analysis was performed and the accuracy was tested. After analyzing the Titanic dataset, two predictions were generated. The first was information about what the survived passengers had in common that helped them survive the shipwreck, while the second was to predict if a particular person would have survived if they had been aboard the fateful ship by applying the tools of machine learning and statistical analysis.

II. LOGISTIC REGRESSION

Logistic regression is a statistical model that uses a logistic function to model binary dependent variables. The most common logistic regression models a binary variable, something that can take two values such as true/false, yes/no, and so on, where the two values are labeled "0" and "1". Whereas a multinomial logistic regression can model two or more dependant variables in a similar way. In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent predictor variables. The corresponding probability of the value labeled "1" can vary between 0 and 1 (both included). The logistic function then converts the log-odds to the probability. The logistic function is a sigmoid function(Fig.1.), which takes in the input variables, and outputs a value between zero and one. Logistic regression is particularly a useful analysis method for classification problems, when we are trying to determine which category a new sample data fits best into. [1,2]

A. Logistic Model

Let us try to understand logistic regression by considering a logistic model with two predictors, x_1 and x_2, and one binary (Bernoulli) response variable Y, whose probability of

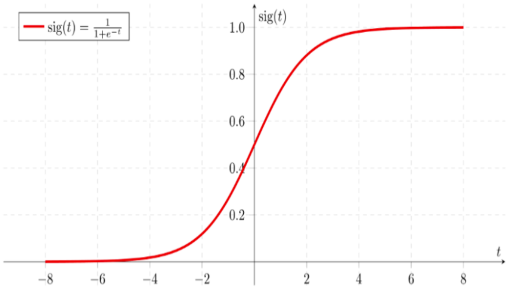


Fig. 1. Sigmoid Function

being true we denote as $p = P(Y = 1)$. We assume a linear relationship between the predictor variables and the log-odds (also called logit) of the event that $Y = 1$.

This linear relationship can be written in the following mathematical form (where "ℓ" is the log-odds, b is the base of the logarithm, and β_i are parameters of the model):

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}. \tag{2}$$

By simple algebraic manipulation (and dividing numerator and denominator by $b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$), the probability that $Y = 1$ is given by

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} + 1} \tag{3}$$

$$= \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} \tag{4}$$

$$= S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2). \tag{5}$$

where S_b is the sigmoid function with base b.

The above formula shows that once β_i are fixed, we can easily compute either the log-odds that $Y = 1$ for a given observation, or the probability that $Y = 1$ for a given observation. The main use-case of a logistic model is to be given an observation (x_1, x_2), and estimate the probability p that $Y = 1$. In most applications, the base b of the logarithm is usually taken to be e. However, in some cases it is be easier to work with base 2 or base 10.

B. Decision Boundary

To predict which class a datapoint belongs to, a threshold can be set. Based on this threshold, the estimated probability is classified into classes. Such a decision boundary can be linear or non-linear.

C. Learning a logistic regression model

Learning a linear regression model involves estimating the values of the coefficients used in the representation using the training data.

Consider a generalized linear model function parameterized by θ, $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}} = \Pr(Y = 1 \mid X; \theta)$.

Therefore, $\Pr(Y = 0 \mid X; \theta) = 1 - h_\theta(X)$ and since $Y \in$\{0,1\},$\Pr(y \mid X; \theta)$ is given by $\Pr(y \mid X; \theta) = h_\theta(X)^y (1 - h_\theta(X))^{(1-y)}$.

We now calculate the likelihood function assuming that all the observations in the sample are independently Bernoulli distributed,

$$L(\theta \mid y; x) = \Pr(Y \mid X; \theta) \tag{6}$$

$$= \prod_i \Pr(y_i \mid x_i; \theta) \tag{7}$$

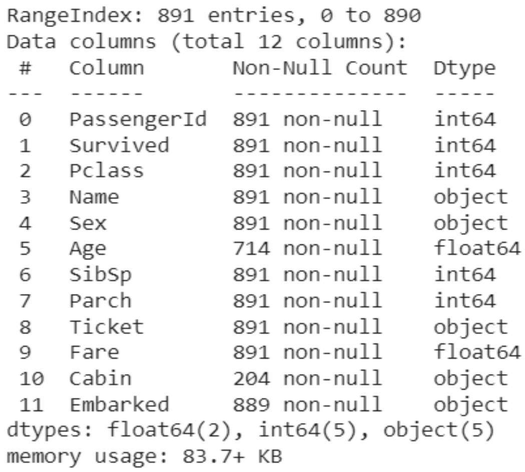
$$= \prod_i h_\theta(x_i)^{y_i} (1 - h_\theta(x_i))^{(1-y_i)} \tag{8}$$

Typically, the log likelihood is maximized i.e. $N^{-1} \log L(\theta \mid y; x) = N^{-1} \sum_{i=1}^{N} \log \Pr(y_i \mid x_i; \theta)$ is maximized using optimization techniques such as gradient descent.

III. Using Logistic Regression to Solve the Problem

A. Description of the Dataset

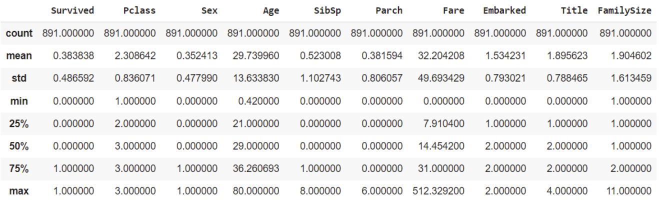
The training dataset comprises of 891 observations and 10 characteristics/variables. Out of which 1 ('Survived') is the dependent variable and the rest 9 are independent variables.



```
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Fig. 2. Information about the dataset

It was noted that there is notably a large difference between 75th %tile and maximum values (Fig.3.) in most of the variables. This suggests that there are extreme values(outliers) in the data.



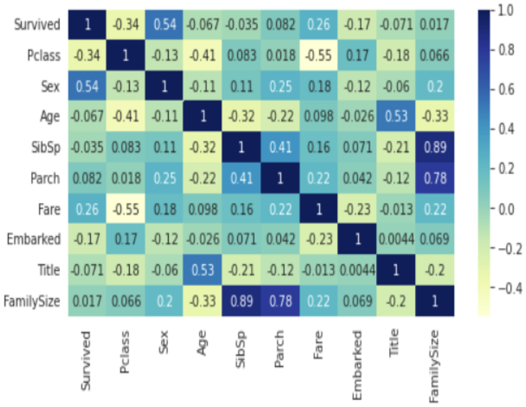
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title	FamilySize
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	0.352413	29.739960	0.523008	0.381594	32.204208	1.534231	1.895623	1.904602
std	0.486592	0.836071	0.477990	13.633830	1.102743	0.806057	49.693429	0.793021	0.788465	1.613459
min	0.000000	1.000000	0.000000	0.420000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	0.000000	2.000000	0.000000	21.000000	0.000000	0.000000	7.910400	1.000000	1.000000	1.000000
50%	0.000000	3.000000	0.000000	29.000000	0.000000	0.000000	14.454200	2.000000	2.000000	1.000000
75%	1.000000	3.000000	1.000000	36.260693	1.000000	0.000000	31.000000	2.000000	2.000000	2.000000
max	1.000000	3.000000	1.000000	80.000000	8.000000	6.000000	512.329200	2.000000	4.000000	11.000000

Fig. 3. Statistics of the dataset

B. Preparing and cleaning the data

While the dataset contains a wealth of information, we shall limit our analysis to only the required variables in the data which will be helpful for our analysis. The columns 'Ticket', 'Name' and 'Passenger ID' have been logically found to be irrelevant to our area of analysis and hence have been ignored. Whereas new variables called 'Title' and 'FamilySize' have been created to capture certain correlations of the data better.

The column/variable 'Cabin' was discarded as it has a significantly high number of missing/NaN values. Few number of missing/NaN values were also found in the columns/variables 'Age' and 'Embarked', they have been handled using imputation via the SimpleImputer from the Datawig library. In the preprocessing stage, the categorical variables 'Sex', 'Embarked' and 'Title' were mapped unto numerical values.



	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Title	FamilySize
Survived	1	-0.34	0.54	-0.067	-0.035	0.082	0.26	-0.17	-0.071	0.017
Pclass	-0.34	1	-0.13	-0.41	0.083	0.018	-0.55	0.17	-0.18	0.066
Sex	0.54	-0.13	1	-0.11	0.11	0.25	0.18	-0.12	-0.06	0.2
Age	-0.067	-0.41	-0.11	1	-0.32	-0.22	0.098	-0.026	0.53	-0.33
SibSp	-0.035	0.083	0.11	-0.32	1	0.41	0.16	0.071	-0.21	0.89
Parch	0.082	0.018	0.25	-0.22	0.41	1	0.22	0.042	-0.12	0.78
Fare	0.26	-0.55	0.18	0.098	0.16	0.22	1	-0.23	-0.013	0.22
Embarked	-0.17	0.17	-0.12	-0.026	0.071	0.042	-0.23	1	0.0044	0.069
Title	-0.071	-0.18	-0.06	0.53	-0.21	-0.12	-0.013	0.0044	1	-0.2
FamilySize	0.017	0.066	0.2	-0.33	0.89	0.78	0.22	0.069	-0.2	1

Fig. 4. Correlation heat-map between the various features of data

It was also found that the 'Pach' and 'SibSp' variables had a significant correlation with the variable 'FamilySize'(Fig.4.), hence these columns/variables were removed from the data before training.

The dataset was also split into the training and validation parts with the validation data being used to calculate the accuracy and performance of the model before making predictions on the test data.

C. Building the Linear Regression Model

A logistic regression model has been built using the inbuilt LogisticRegression method from the Skikit-Learn library. The features- 'Pclass', 'Sex', 'Age', 'Fare', 'Embarked', 'Title' and 'FamilySize' were taken as the independent input variables for the model. These features were used to predict the binary dependant variable- 'Survived'.

D. Visualizations and insights from the analysis

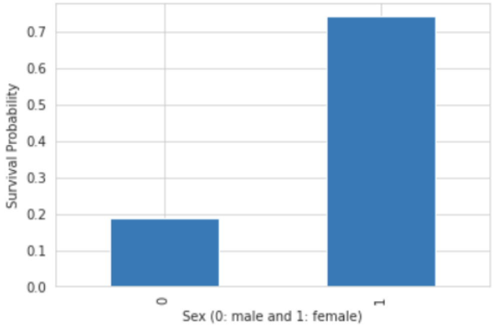


Fig. 5. Plot of survival probability for both sexes

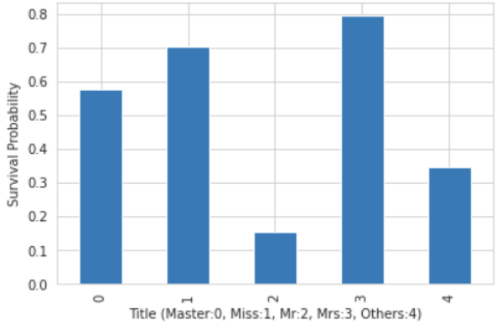


Fig. 6. Plot of survival probability for different name titles

From (Fig.5. and Fig.6.) we infer that onboarders with the title 'Mrs' and 'Miss' i.e. women have higher probability of survival when compared to men. It was also found that women had a survival rate of 74%, while men had a survival rate of only about 19%. The survival rate for the female passengers is very high compared to the men, this is predominantly because of the strict maritime tradition of evacuating women and children first in situations of danger.

From (Fig.7.) we infer that the first class has the highest probability of survival followed by the second and third classes. This fact may have to do with the positioning of the seats of each passenger class and hence the ease of evacuation or escape during the shipwreck.

From (Fig.8.) we infer that the survival probability varies with the port of embarkation as:

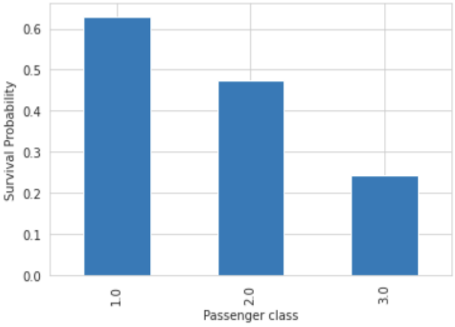


Fig. 7. Plot of survival probability for the three passenger classes

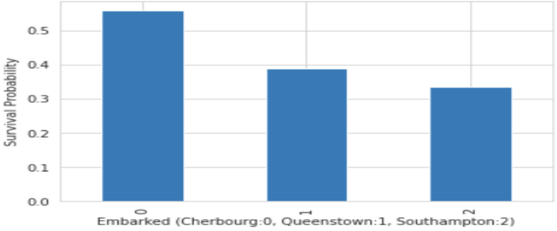


Fig. 8. Plot of survival probability for the three ports of embarkation

Cherbourg>Queenstown>Southampton. Passengers who boarded in Cherbourg, France, appear to have the highest survival rate. Passengers who boarded in Southhampton were marginally less likely to survive than those who boarded in Queenstown. This is probably related to passenger class, or maybe even the order of room assignments (e.g. maybe the passengers who boarded earlier were more likely to have rooms closer to the deck).

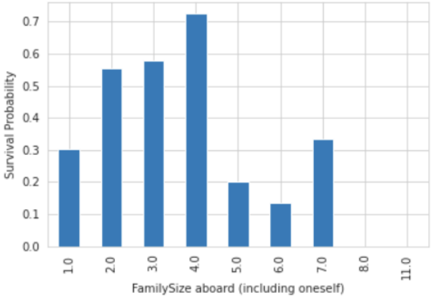


Fig. 9. Plot of survival probability for various family-sizes of the on-boarders

From (Fig.9.), it is a general inference that the individuals traveling alone without their family(i.e. individuals with the value of the variable 'FamilySize' being 1) were more likely to die in the disaster than those with their family aboard(with the exception of 'FamilySize' being 5, 6 or 7). Given the era, it is also likely that individuals traveling alone were male.

From (Fig.10.) we infer that the age distribution for survivors and deceased is actually very similar. One notable difference is that, of the survivors, a larger proportion were children. The passengers evidently made an attempt to save children by giving them a place on the life rafts.

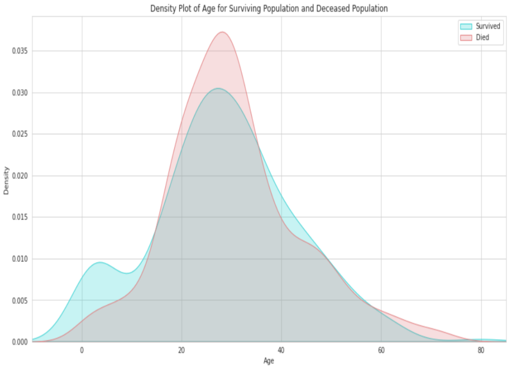


Fig. 10. Density Plot of Age for Surviving Population and Deceased Population

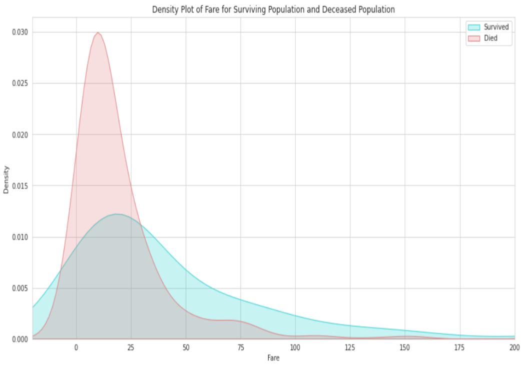


Fig. 11. Density Plot of Fare for Surviving Population and Deceased Population

From (Fig.11.), we can infer that those who's tickets costed less had a lesser survival rate. This evidently has to do with the fact that the seats which costed more were well-positioned and made it easier to escape the shipwreck from.

E. Model performance on the validation dataset

The accuracy of our Logistic Regression model has been found to be 82.12%. And the confusion matrix of our model(Fig.12.) on the validation dataset showed 28 misclassified datapoints.

F. Using the model to make predictions on the test data

The test data provided was first cleaned and organized which was followed by data pre-processing using similar methods employed on the test data. Then, the model was used to make predictions(Fig.13.) on the test data to find if a passenger would survive the shipwreck given personal information about the same.

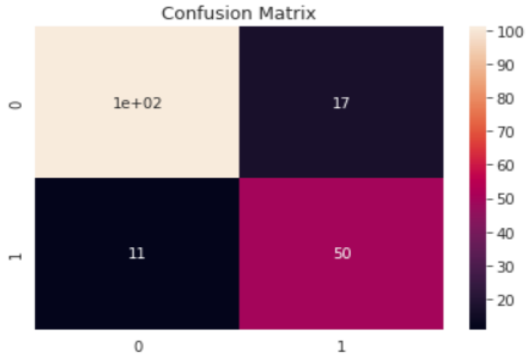


Fig. 12. Confusion matrix of our model on the validation dataset

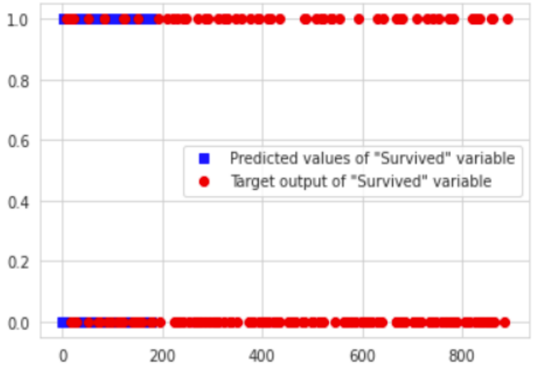


Fig. 13. Predicted and Target outputs on the validation dataset

IV. CONCLUSIONS

We analyzed the role of passenger data in affecting the chances of survival in the shipwreck. The survival probability of women and children were found to be significantly higher than that of the men. Women had a survival rate of 74%, while men had a survival rate of only about 19%. The survival rate for the female passengers and children is higher compared to the men because of the strict maritime tradition of evacuating women and children first in situations of danger. It is a general inference that the individuals traveling alone without their family were more likely to die in the disaster than those with their family aboard. It was also found that the first class has the highest probability of survival followed by the second and third classes, this is due to the positioning of seats of each class in the ship and hence the ease of evacuation during the shipwreck. Passengers who boarded in Cherbourg appear to have the highest survival rate. This is probably related to the order of room assignments in accordance to the ports of embarkation. It is to be noted that those who's tickets costed less had a lesser survival rate, this evidently has to do with the fact that the seats which costed more were well-positioned in such a way that it made it easier to escape from the shipwreck.

Further improvements in the model are highly possible. While imputing for the missing values in the 'Age' and 'Embarkation' variables, SimpleImputer was directly used, instead of this, different methods of imputation can also be tried out based on further research on the data. In the present model, outliers found in the data were not accounted for, further work can done on addressing the outliers present in the data. In order to further improve the overall result, an extensive hyper-parameter tuning should also be done to improve the accuracy of the model for it to better fit the data. It is possible to further improve it by doing ensemble learning or finding better machine learning algorithms that fit the data better.

REFERENCES

[1] Peng, Joanne Lee, Kuk Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[2] Hilbe, J.M. (2009). Logistic Regression Models (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420075779

[3] Kleinbaum, David G., K. Dietz, M. Gail, Mitchel Klein, and Mitchell Klein. Logistic regression. New York: Springer-Verlag, 2002.

[4] S. Cicoria, J. Sherlock, M. Muniswamaiah, L. Clarke, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster", Proceedings of Student-Faculty Research Day CSIS, pp. 1-6, May 2014.