Assignment-2: A Mathematical Essay on Logistic Regression

Shania Mitra
Roll Number: CH18B067
Department of Chemical Engineering
IIT Madras, Chennai
ch18b067@smail.iitm.ac.in

***Abstract*—In this study, we examine whether some groups of people on the RMS Titanic were more likely to survive than others, on the basis of Age, Socio-economic data and other factors. A logistic regression model is used to model the importance of these factors and predict the probability of individuals surviving**

***Index Terms*—Logistic Regression, Visualization, RMS Titanic**

I. 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. The lack of sufficient lifeboats, resulted in the death of 1502 out of 2224 passengers and crew. In this study, we aim to answer whether factors apart from luck were important for the survival of passengers.

Logistic Regression is a supervised machine learning algorithm that can be used to model the probability of a certain class or event. It is used when the data is thought to be linearly separable and the outcome is binary or dichotomous in nature.

In this study, we use logistic regression to model probability of survival as a function of age, socioeconomic factors, number of family members, port of embarkment etc.

We begin by gathering, cleaning and preparing the data, following which we perform exploratory analysis. Finally, we build statistical models and perform visualizations to provide quantitative and visual evidence of the relationships observed. In the next section, we highlight the key principles underlying Logistic Regression. In section 3, we discuss the insight and observations drawn from the data and the models. Finally, in section 4 we outline the salient features of the study and present further avenues of possible investigation.

II. Logistic Regression

The logistic function is a sigmoid function, which takes any real input t, and outputs a value between zero and 1. The standard logistic function is denoted as$\sigma(t)$. It must be noted that $\sigma(t) \in (0,1)$ for all t. The standard logistic function $\sigma : \mathbb{R} \to (0,1)$ is defined as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

Let us assume that t is a linear function of a single explanatory variable x. We can then express t as follows:

$$t = \beta_0 + \beta_1 x$$

And the general logistic function $p : \mathbb{R} \to (0,1)$ can now be written as:

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

In the logistic model, $p(x)$ is interpreted as the probability of the dependent variable Y equaling a success/case rather than a failure/non-case. It is clear that the response variables Y_i are not identically distributed: $P\left(Y_i = 1 \mid X\right)$ differs from one data point X_i to another, though they are independent given design matrix X and shared parameters β

If there are multiple explanatory variables, the above expression $\beta_0 + \beta_1 x$ can be revised to $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$. Then when this is used in the equation relating the log odds of a success to the values of the predictors, the linear regression will be a multiple regression with m explanators; the parameters β_j for all $j = 0, 1, 2, \ldots, m$ are all estimated. Again, the more traditional equations are:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

and

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

where usually $b = e$.

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 $Y \in \{0, 1\}$, we see that $\Pr(y \mid X; \theta)$ is given by $\Pr(y \mid X; \theta) = h_\theta(X)^y \left(1 - h_\theta(X)\right)^{(1-y)}$. We now calculate

PassengerId	0.000000
Survived	0.000000
Pclass	0.000000
Name	0.000000
Sex	0.000000
Age	19.865320
SibSp	0.000000
Parch	0.000000
Ticket	0.000000
Fare	0.000000
Cabin	77.104377
Embarked	0.224467
family_size	0.000000

dtype: float64

Fig. 1. Feature-wise percentage of missing values in the training set

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

$$\begin{aligned} L(\theta \mid y; x) &= \Pr(Y \mid X; \theta) \\ &= \prod_i \Pr\left(y_i \mid x_i; \theta\right) \\ &= \prod_i h_\theta\left(x_i\right)^{y_i} \left(1 - h_\theta\left(x_i\right)\right)^{(1-y_i)} \end{aligned}$$

Typically, the log likelihood is maximized,

$$N^{-1} \log L(\theta \mid y; x) = N^{-1} \sum_{i=1}^{N} \log \Pr\left(y_i \mid x_i; \theta\right)$$

which is maximized using optimization techniques such as gradient descent.

III. The Problem

In this study, we examine whether some groups of people on the RMS Titanic were more likely to survive than others, on the basis of age, socioeconomic data and other factors.

A. Data Preparation

The training dataset used in this study consists of 891 passengers and 12 features. Interpretation of the features is as follows:

- Survival: 0 if the passenger did not survive, 1 if the passenger survived
- Pclass: Class of the ticket - 1st, 2nd, 3rd
- Sex: Gender of the passenger - male, female
- Sibsp: Number of Siblings/Spouses
- Parch: Number of parents/children
- Ticket Number
- Fare: Fare paid for the ticket
- Cabin: Cabin Number
- Embarked: Port of Embarkment - C, Q, S

To impute the missing values, two approaches are considered: 1) Port of Embarkment wise medians are filled in in place of the empty values such that the distribution does not get distorted, as would be the case if we imputed using means 2) Missing Values are dropped

Columns PassengerID, Ticket and cabin were dropped since PassengerID is unique to every passenger, Cabin had 77% missing values and Ticket has 541 unique values.

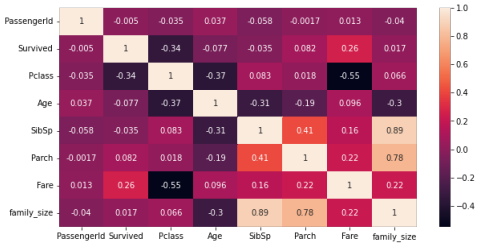

Fig. 2. Pearson Correlation Coefficient among all numeric features

Survived	Pclass	Age	SibSp	Parch	Fare	family_size
0	2.531876	30.626179	0.553734	0.329690	22.117887	1.883424
1	1.950292	28.343690	0.473684	0.464912	48.395408	1.938596

Fig. 3. Average Value of all features for passengers by survival status

Additional features introduced include: Family size = Parch + SibSp + 1

B. Exploratory Analysis

In this section we look at the trends between various features:

Among the 891 passengers in the training set provided, 549 were found dead while 342 managed to survive.

In Figure 2 we see the correlation matrix among the features. In particular, we see that there is negative correlation among PClass and Survived indicating that passengers in higher PClass had lower survival rates. Further, from the negative correlation between fare and PClass we understand that higher Pclasses have lower fares. Thus, passengers who paid higher fares are more likely to survive. We also observe that as number of siblings/spouses increase, number of parents and children also increases.

In Figure 3, we see difference in mean values for each Survival State. We can clearly see that people who survived on an average were in lower PClasses and paid higher fares. This may have happened due to more expensive cabins being located at more strategic locations in the ship, allowing the people in them to escape more easily. This can further be seen in figure 5, in which we observe that in the 1st Class, the survival rates are much higher than death rates. For 2nd Class, they are almost the same whereas for the 3rd Class it is much lower. This may also be attributed to the fact that the number of people in 3rd Class was much higher than those in 1st or 2nd Class, making it difficult for people to escape from the stampede. People who escaped also had a lower average age, which aligns with the fact that younger people due to higher agility find it easier to escape. This can be seen from Figures 1, 3 and 9

In Figure 4, we can see that even though there are fewer females, the survival rate among females is much higher.

From Figure 6 it can be seen that maximum passengers on the ship were individuals. However, the survival rate among individuals was very low. In contrast, the survival among

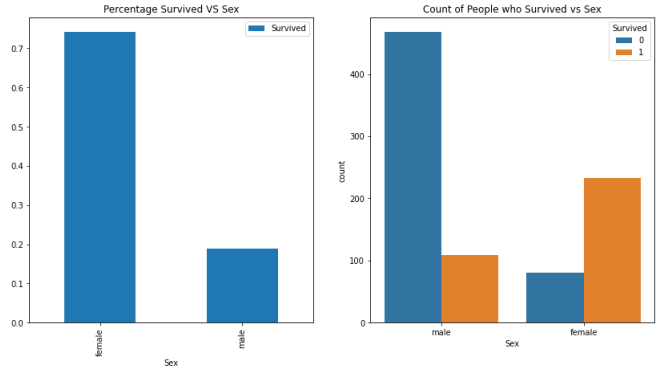

Fig. 4. (a) Percentage of Males and Females that survived (b) Counts of Males and Females that survived

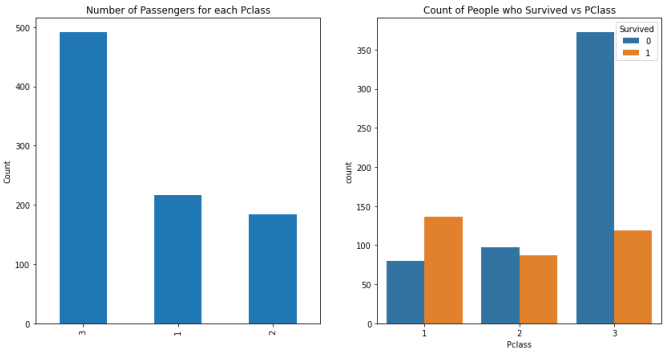

Fig. 5. (a) Number of Passengers by PClass (b) Count of people who survived by PClass

families with 2, 3 and 4 people was higher than death rates, possibly due to members among the family helping eachother to escape. In large families, however, death rates were very high possibly because of family members not wanting to escape without eachother and the lack of lifeboats restraining their escape.

In Figure 7 it can be seen that most of the people on the ship were from S, however, survival rates were highest among passengers from Port C.

As highlighted above, in Figure 8 we observe that passengers who paid higher fares tend to to have better survival rates. This can be attributed to two possible reasons: (a) They were alloted cabins at more strategic locations (b) They boarded the ship from Port C and were hence given seats in strategic locations, giving them fast access to lifeboats

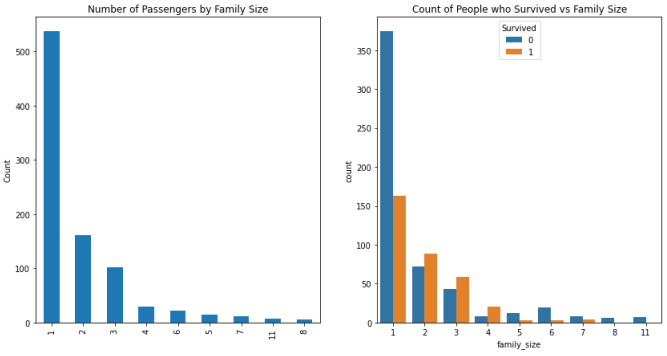

Fig. 6. (a) Number of Passengers by Family Size (b) Count of people who survived by Family Size

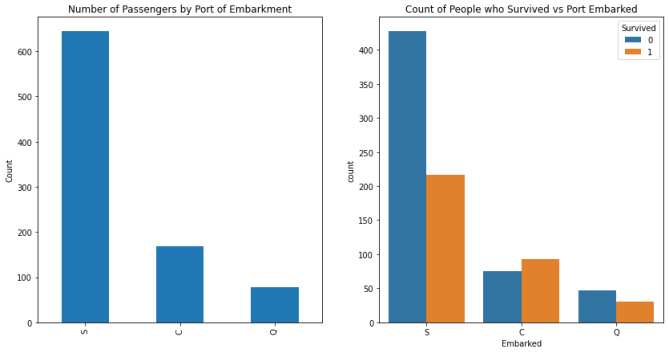

Fig. 7. (a) Number of Passengers (b) Count of people who survived by Port of Embarkment

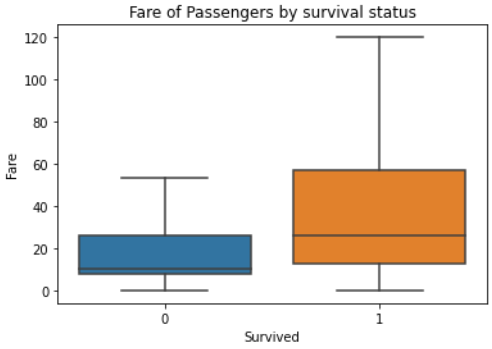

Fig. 8. Fare of Passengers vs Survival Status

C. Model

To train the model and gauge its performance, the dataset was split into train and validation sets with an 80-20 split. Following are the metrics achieved by the Logistic Regression Model:

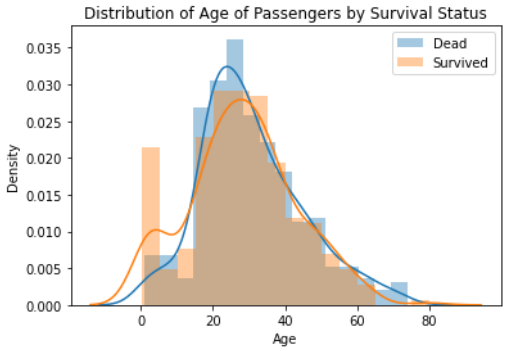

Fig. 9. Distribution of Ages of Passengers by Survival Status

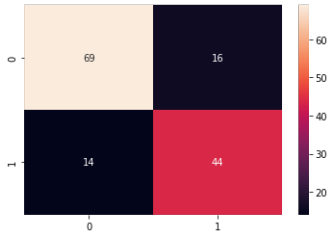

Fig. 10. Confusion Matrix for Logistic Regression Model; On the Y axis we have True labels and on the X axis we have predicted labels

	precision	recall	f1-score	support
0	0.84	0.84	0.84	90
1	0.73	0.72	0.72	53
accuracy			0.80	143
macro avg	0.78	0.78	0.78	143
weighted avg	0.80	0.80	0.80	143

Thus, for the people who survived, we achieve an f1-score of 0.72 while for the people who were unable to survive, we achieve an f1-score of 0.84. Support refers to the number of samples in each class.

For the 143 validation samples, from the confusion matrix in Figure 10 we can see that majority of the samples that survived were predicted correctly. However, many that lived were thought to be dead by the model, possibly due to the large number of people who died.

Predictions were made on the test dataset provided. However, the performance could not evaluated due to the absence of corresponding labels.

IV. Conclusions

In this study, we observed that among the passengers in the RMS Titanic, females were more likely to survive as compared to males. Further, passengers who were in PClass 1, which is more expensive, were more likely to survive than passengers in other classes. Finally, passengers who were younger and had family sizes between 2 and 4 were more likely to survive due to help from family and higher agility. In the future, non-linear models can be used to model the survival rates better

References

[1] https://www.kaggle.com/c/titanic/overview
[2] https://en.wikipedia.org/wiki/Logistic_regression
[3] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
[4] https://machinelearningmastery.com/logistic-regression-for-machine-learning/