A Mathematical Essay on Logistic Regression

Aryan Pandey
Department of Ocean engineering
Indian Institute of Technology, Madras
Chennai, India
na19b030@smail.iitm.ac.in

***Abstract*—Problems where we need to classify data into a given set of labels is one that is of extreme importance. To handle such problems simple regression techniques are not enough and this has given rise to a new set of ways to handle such problems. The main focus of this study is one such method, Logistic Regression, where we will also look at its application on one of the most famous classification problems - Titanic Survival Prediction. The main aim of this problem is to predict whether or not a given person would have survived the Titanic disaster.**

***Index Terms*—Logistic Regression, Classification**

I. Introduction

The Titanic was a British Passenger ship which sank in the North Atlantic Ocean on 15th April 1912 after striking an iceberg during her maiden voyage from Southampton, UK to New York City, United States. Out of the 2,224 passengers on board, more than 1500 died. In our problem, we aim to better understand the demography of the people on board along with the relation of certain individual traits to their survival chance. Further we move on and try to look into what exactly were the important traits of a person which decided whether or not the person survived.

Trying to model the chances of survival as a binary classification problem is something that would be of interest, since we would be able to figure out what kind of people are the most likely to survive in the event of such a disaster. The main reason why over 1500 people died in this disaster was simply because there weren't enough lifeboats on the ship. It is interesting to try and figure out through our analysis, which sections of society were able to survive this disaster and which ones succumbed to it. The data-set given to us contains details of 891 of the people who were on board the ship along with whether or not they survived the shipwreck. The details given are the ticket class of the passenger, their gender, age and how many siblings or spouses they had onboard the ship, their ticket number, the fare of the journey, the cabin number if the were allotted one and where they embarked on the journey form.

In order to do this whole modelling, we fit a Logistic Regression model to the data-set given using the sklearn package in python. Since this is a classification problem, we try to model this using one of the most basic and simple to understand or interpret classification models. We split the data-set into a training data-set and a validation data-set in a 85:15 ratio in a stratified manner. This ensures that we can validate our results and make sure that the model has not over-fit to it. All the training is done on the training set and any results reported are on the validation set.

For our evaluation we use the F1-score metric which has been explained in later sections of the paper. We achieve an F1-score of 0.7736 using the Logistic Regression model. The confusion matrix for the same has also been shown in the later sections of the paper. Through this paper we better understand the demography of the people travelling onboard the Titanic and also build out a list of features that turned out to be the contributing factor to the survival rates onboard the ship.

Section II gives a detailed view of the data-set and some visualisations of the features in the data-set. In Section III we explain the mathematical aspects of Logistic Regression. Section IV talks about how we applied Logistic Regression to the problem at hand and makes a few statements about the features which were an important contributing factor to the survival of the passengers.

II. Datasets

The dataset provided to us is the classic Titanic Dataset. In this dataset we are given details about whether or not the person survived, the Ticket Class of the person which serves as a proxy for the socio-economic status of the person, the gender of the person, their age, number of siblings and spouses, number of family relations (in terms of mother, father and children), the ticket number, ticket fare, the cabin number if a cabin was allotted and the port form which the passenger embarked.

The age of the person was fractional is the age was less than 1. If the age was an estimated age, it takes the form xx.5 . While counting siblings and spouses, any mistresses and fiances were ignored. While counting the number of family relations, children who were travelling only with a nanny were considered as having no family relations on board. The dataset has a high number of missing values in the Age and Cabin features. In order to solve this, we first find out the median age of the passengers across a passenger class and impute any missing values with the median age of that person's class. Fig. 1 shows a boxplot representing the Age plotted against the passenger class. In order to fill the missing values, any missing values in Passenger Class 1 are imputed with a value of 24, Passenger Class 2 with 39 and Passenger Class 3 with 29. In order to fill the Cabin, the missing values are treated as if there were no cabin allotted to those individuals. Moreover, to ease the preprocessing, we take the first letter to

demonstrate the importance of the cabin assigned. Therefore, the values in the Cabin column are A, B, C and N, representing Cabins in the A-series, B-series, C-series and No Cabin. As another preprocessing step we also extract the designation of the person from their name. These follow as ”Mr.”, ”Mrs.” and so on.

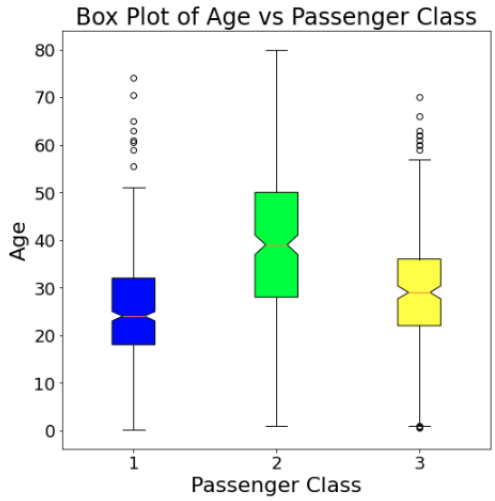

Fig. 1. Age Variation across different Passenger Classes

In order to better understand the data given to us, we first try and understand how the survival rates vary across the different passenger classes. Fig. 2 shows a plot which demonstrates how many people survived and how many didn't across different passenger classes. Clearly from this we can see that the fraction of people who survived in the first class is the highest, whereas it reduces slightly in the second class and the survival rates for the third class are abysmally low.

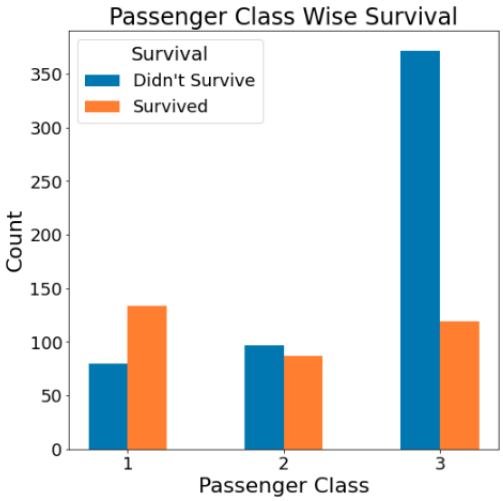

Fig. 2. Passenger Class Wise Analysis of Survival vs Non Survival

We also try to better understand the demography of the people there by seeing the gender distribution across the different passenger classes. Fig 3 shows this distribution. One can see that in general the number of males travelling onboard the ship is higher than the number of females travelling onboard the ship. In First Class, we can see that the ratio of Male to Female passengers is almost one whereas in second class the ratio is almost 1.5 and in third class the ratio is well above 2.

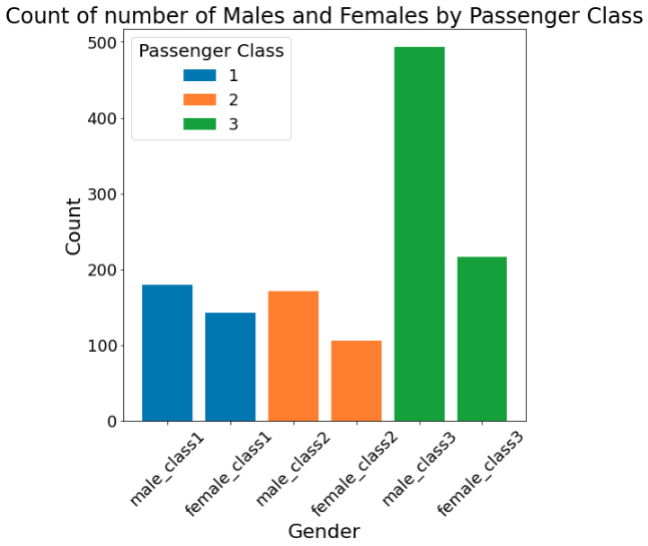

Fig. 3. Number of Males and Females in each Passenger Class

In Fig.4 we see a Pearson Correlation Matrix demonstrated as a heatmap. This shows us the Correlation between any two features. For us the main interest here is to see which features are heavily correlated to the Survival of a person. On inspection we can see that the Gender of a person as well as the Passenger Class and Fare play an important role in deciding the survival chances since they are all correlated to the survival chances. Since no feature here has a severely low correlation to Survival, we won't be dropping any features and will let the model decide via it's training as to which features turn out to be more important. An interesting thing to note here is some other correlations. The Cabin allotted to a person is highly correlated with the Passenger Class. This means that the first class passengers are allotted the best cabins while the third class passengers go cabinless.

III. Models

In this section we will develop a thorough understanding of the Logistic Regression framework, the math behind it, any assumptions and the evaluation metrics being used here.

A. Variables

The independent variables used to train any classical machine learning framework are of two types: Categorical and Continuous. Continuous variables are those variables which are continuous in nature over their whole domain. Categorical Variables are the discrete variables that are present. Some Categorical variables present themselves in textual form (eg. Gender) which we need to convert to a numeric form so that

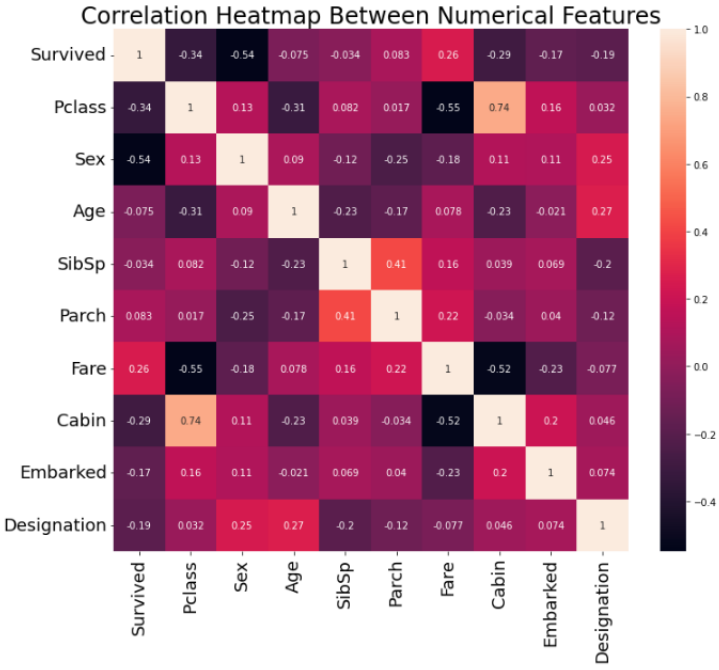

Fig. 4. Correlation Plot for the features in the Dataset

our model can understand it. We do this using a Label Encoder which assigns a single number to represent each unique value of the variable. In case of a Logistic Regression framework, the dependent variable is usually a Binary Variable (meaning that it takes two discrete values).

B. The Algorithm

The logistic function for the Logistic Regression Algorithm to calculate the logits is as shown below:

$$logit(Y) = ln(odds) = ln(\frac{\pi}{1-\pi}) = \alpha + \beta X \tag{1}$$

The above equation assumes that there is a single feature in our dataset. In case we have n features, the formulation changes to as shown below:

$$logit(Y) = ln(odds) = ln(\frac{\pi}{1-\pi}) = \alpha + \beta_1 X_1 + ...\beta_n X_n \tag{2}$$

In the Logistic Regression framework, we try to predict the chances of an event happening as a regression problem and then apply a function to convert these chances to a probability value. In order to do this we will need a function that takes in all real number as its domain and gives a value from [0,1] as it's range. This function is given by the Sigmoid function. The Sigmoid function definition is as given below:

$$S(x) = \frac{1}{1 + exp(-x)} \tag{3}$$

When applied to our Logistic Regression Framework for n features, it evaluates to the expression given below:

$$P(Y = 1|X) = \frac{1}{1 + exp(-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i))} \tag{4}$$

Since this is a binary classification problem:

$$P(Y = 0|X) = 1 - P(Y = 1|X) \tag{5}$$

$$P(Y = 0|X) = \frac{exp(-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i))}{1 + exp(-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i))} \tag{6}$$

All the Coefficients represented by β_i are calculated via the least squares minimisation based on the Maximum Likelihood Estimates.

C. Assumptions of Logistic Regression

Logistic regression does not make many of the key assumptions of linear regression and general linear models that are based on ordinary least squares algorithms – particularly regarding linearity, normality, homoscedasticity, and measurement level. First, logistic regression does not require a linear relationship between the dependent and independent variables. Second, the error terms (residuals) do not need to be normally distributed. Third, homoscedasticity is not required. Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale. However, some other assumptions still apply.

- Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
- Logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.
- Logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.
- Logistic regression assumes linearity of independent variables and log odds. although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.
- Logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of (10*5 / .10) which is 500.

D. Evaluation Metric

While there are a bunch of metrics which one can use to evaluate a classifciation model, we use the F1-Score to evaluate the Logistic Regression model that we train. The F1-Score is defined as:

$$F_1 = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

Where Precision is calculated by dividing the true positives by anything that was predicted as a positive. Recall (or True Positive Rate) is calculated by dividing the true positives by anything that should have been predicted as positive.

IV. Modelling

In order to apply the Logistic Regression Framework to the given dataset we use the sklearn package. We first split the dataset into a Training and Validation Split in an 85:15 ratio. After training the model on the Training Split for a maximum of 1000 iterations, we observe that the F1-Score achieved by the model is 0.7736. In order to better understand what the model has learnt, we plot the feature importance of the model which is shown in Fig. 5

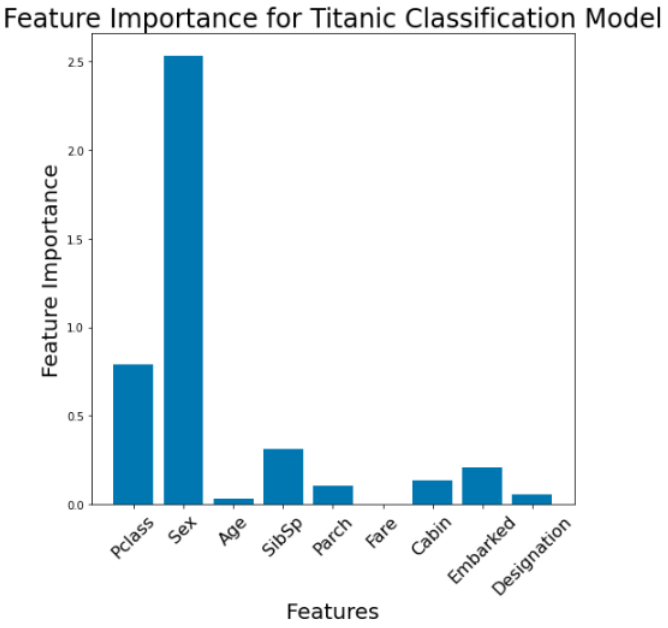

Fig. 5. Feature Importance from the Model

As one can see from the Figure, the model places a heavy emphasis on the Gender of a person, followed by the Passenger Class of the person. In order to understand this better, we also take cases where we change the gender and designation of a person (from male to female and vice versa) keeping all other parameters the same and observe the model's predictions. Shown in Fig. 6 are the results of this experiment

	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked	Designation
0	3.0	Male	22.0	1.0	0.0	7.2500	7.0	2.0	Mr.
1	1.0	Male	38.0	1.0	0.0	71.2833	2.0	0.0	Mr.
2	3.0	Female	22.0	1.0	0.0	7.2500	7.0	2.0	Mrs.
3	1.0	Female	38.0	1.0	0.0	71.2833	2.0	0.0	Mrs.

```
Male Survival:  [0. 0.]
Female Survival:  [1. 1.]
```

Fig. 6. Gender Bias in the Model

As you can see there is some sort of inherent bias present in this model, since we can see that even if we keep all other parameters same and change the gender of the person, the survival of the person changes as predicted by the model.

V. Conclusions

A. Contributions

- A person's gender was one of the key contributing factors in their survival. Due to the protocol of saving Women and Children first, priority was given to them for the limited lifeboats present onboard the ship.
- It is also observed that people travelling in first class had higher survival rates.
- A model trained on this dataset assumes an inherent bias due to these factors when trying to predict the survival of a person onboard the ship.

B. Avenues for Further Research

From this we can see that one possible source of further research is to build a bias free model here while still retaining the essence of the problem. Moreover, we could also make an attempt at fitting non-linear decision boundaries by making the features themselves non-linear

References

[1] Article on the Titanic Disaster: https://en.wikipedia.org/wiki/Titanic
[2] Assumptions of Logistic Regression: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/
[3] F1 Score: https://en.wikipedia.org/wiki/F-score