DAL Assignment 2: A Mathematical Essay On Logistic Regression

C Vamshi Krishna Reddy
CH20B112
EE5708: Data Analytics Laboratory
IIT Madras, TamilNadu, India
ch20b112@smail.iitm.ac.in

***Abstract*—This document serves as a model for showcasing the application of Logistic Regression in answering questions like "What kind of people were more likely to survive in the most famous Titanic ship?", "Is there any effect of gender or age or class on the chance of survival?" etc., Using the given labeled titanic dataset, we have to predict the survival on a new dataset.**

***Index Terms*—Classification, logistic regression, statistical modeling**

I. Introduction

Classification constitutes a wide range of problems tackled with the help of Machine Learning apart from Regression. These models are basically supervised models where the algorithms use labeled datasets and create patterns between input and output data. These algorithms are used to predict the outcomes of new input data. In this paper, we are using a Supervised Learning model which is Logistic Regression.

Logistic regression, a widely-used mathematical model, offers simplicity and interpretability in predicting outcomes and finding applications in diverse fields from environmental and biological sciences to social sciences and business. Logistic Regression is used commonly for classification problems. The discussion of logistic regression is expanded upon in the section that follows, and then it is used to analyze and resolve a real-world problem. The most popular method for estimating a logistic regression's parameters is "maximum likelihood estimation (MLE)". Unlike linear least squares, this does not have a closed-form expression. For binary or categorical responses, logistic regression by MLE serves a similar fundamental function that linear regression by ordinary least squares (OLS) does for scalar responses.

This paper aims to use Logistic regression in explaining the effect of age, gender, passenger class, and other socio-economic classes of a person on his/her chance of survival.

The paper provides a detailed overview of logistic regression by solving a real-world problem and gives insights into the relation between variables like Age, Sex, Ticket class, no of siblings/spouses aboard the Titanic, no of parents/children aboard the Titanic, Passenger fare, Port of Embarkation and Survival of the passenger.

II. Logistic Regression

Regression is a widely employed application in machine learning, especially in supervised settings. It essentially helps us understand how certain independent factors connect to a particular outcome or result. In simpler terms, it's like a tool used in machine learning to make predictions, particularly for continuous results. The core idea is to teach algorithms how different independent variables are linked to an outcome or dependent variable. Once the model grasps this relationship, it becomes capable of predicting what might happen when faced with new, unexpected data or completing missing parts in existing data.

Logistic regression is one of the most commonly used constituent of the supervised machine learning model. It is also regarded as a discriminative model because it makes a conscious attempt to distinguish between different classes (or) categories. Contrary to generative algorithms, it is not capable of producing information about the class that it is attempting to predict.

Based on the type and number of classes of output variables, Logistic regression is classified into three subgroups: Binary logistic regression, Multinomial logistic regression, and Ordinal logistic regression.

• **Binary logistic regression:** In this logistic regression, the dependent variable has only two possible outcomes (say 0 or 1). This is the most commonly used approach in logistic regression and is the most common classifier for binary classification. One popular example is the classification of emails into spam and not spam.

• **Multinomial logistic regression:** In this logistic regression, the dependent variable has more than two possible class labels, but these values do not have any specific order. This can be considered as a simple extension of binary logistic regression. One example could be the political party to which a person votes.

• **Ordinal logistic regression:** This logistic regression model also has three or more possible outcomes for the dependent variable, but in this case, these values have a defined order. The most common example is a review system where the rating ranges from 1 to 5.

A. Mathematical Equation of Linear Regression

Logistic regression can be expressed through a mathematical equation given below:

$$f(x) = \frac{1}{1 + e^{-(W^T x + b)}} \tag{1}$$

Here X refers to a set of input features, and W refers to the weights corresponding to individual features and Y refers to the Output variable (or) result.

B. Cost Function

The cost function acts as a measure of how much the model's predictions deviate from actual outcomes. Log loss is used as a loss function for Logistic Regression, the cost function is:

$$J = \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log(h(x^{(i)})) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right] \tag{2}$$

C. Evaluation Model

In classification models like Logistic regression, confusion matrix and f1 score are used to evaluate the performance of the model. Here since our problem is a Binary Logistic regression we have a confusion matrix of 2 X 2 dimension, And F1-Score is defined as the harmonic mean of recall and precision values of the classification problem. The formula for F1-Score is as follows:

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{3}$$

where Precision measures how many of the "positive" predictions made by the model were correct. And Recall measures how many of the positive class samples present in the dataset were correctly identified by the model.

III. The Problem

The objective is to develop a model to predict if a passenger survives or not based on factors like age, gender, passenger class, and port of Embarkation. The process initiates with data cleaning and then exploratory data analysis where we find out factors that affect the survival of passengers. Subsequently, the correlation between these determinants and the response variable, Survival is established.

A. Data Cleaning

The given dataset has a large number of missing values mainly in cabin and age columns. They should be eliminated or modified using common methods such as mean/median/mode estimation-based filling, distribution-based filling, etc. Certain features like 'PassengerId', and 'Ticket' which do not add any additional value to the problem statement were removed, and columns like 'Pclass', 'Sex', and 'Embarked' were turned into categorical features using Label Encoder.

From Figure 1 it is evident that the count of null-values are almost half percent which was incredibly high in the cabin column. As a result, it was decided not to include the cabin column for training the model. The age column which has around 20 percent missing values was decided to impute missing values of Age using the salutation of the passenger's name Like "Mr.", "Miss.", "Mrs.", "Master" and others and this has shown good impact on model performance.

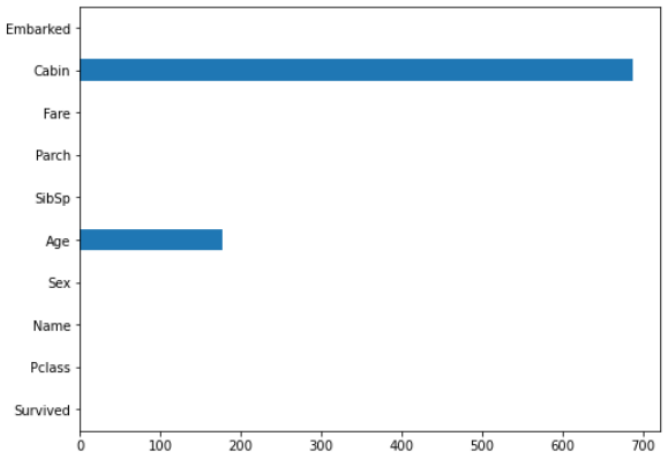

Fig. 1. Bar Graph distribution of Null Values.

B. Exploratory Data Analysis and Visualisations

The provided dataset includes passenger information like name, age, gender, socio-economic class, etc. These details of the passengers on board are important and will be used to train a model that can predict whether the passenger survived or not, based on the "ground truth" i.e. known data. The test file data set contains similar information but does not have information on whether the passenger had survived or not.

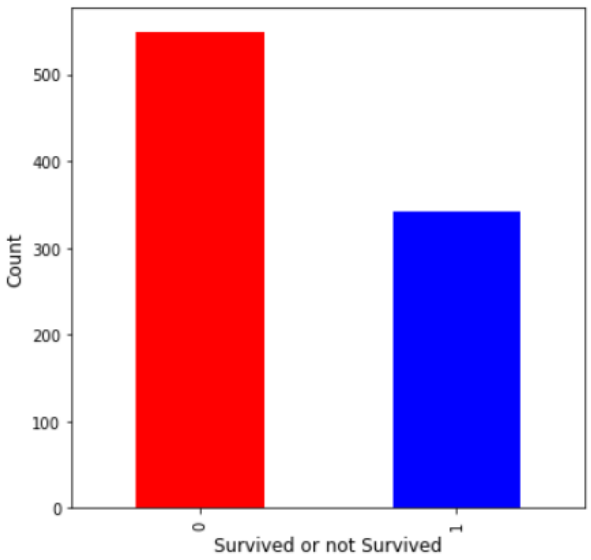

Fig. 2. Count of passengers survived or not survived.

So our motive is to develop a model using a train data set and then predict the Survival column of the test data set. In the dataset, we are provided with columns sibsp, parch which refer to number of siblings/spouses aboard the titanic and the number of parents/children aboard the Titanic. We added a new feature Family which contains the total no of sibsp and parch.

Figure 2 shows a bar graph plot that tells the count of survived and non-survived passengers. From that figure, it is evident that the given dataset is Balanced dataset. And also we infer that almost 60 percent of passengers had died.

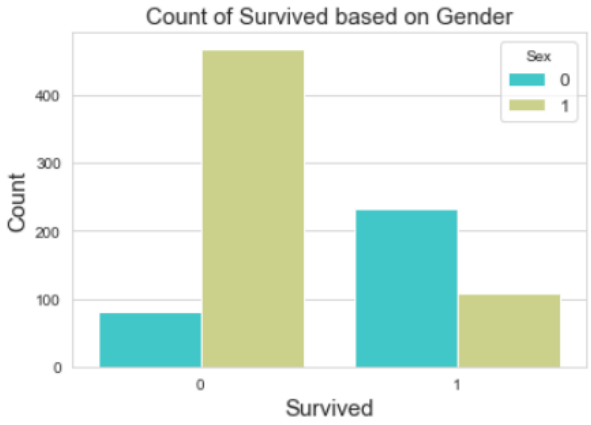

Fig. 3. Survival Count based on gender.

Figure 3 shows that approximately 65 percent of the passengers were male while the remaining 35 percent were female. Nonetheless, the percentage of female survivors was higher than the number of male survivors. More than 80 percent of male passengers died, as compared to around 70 percent of female passengers.

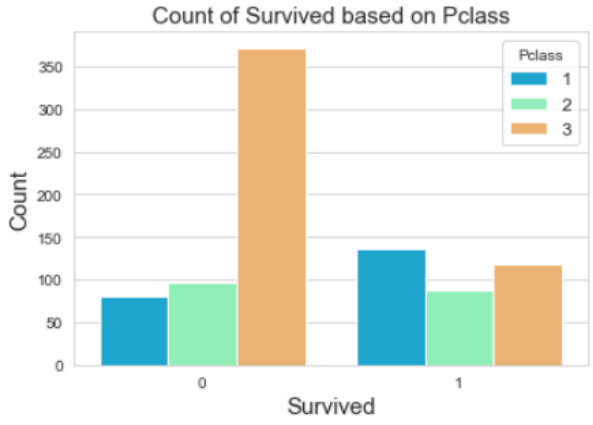

Fig. 4. Survival Count based on Passenger class.

Figure 4 shows that third class had the highest number of passengers who had died, followed by class 2 and class 1. The number of tourists in the third class was more than the number of passengers in the first and second classes combined. The survival chances of class-1 passengers were higher than those of class-2 and class-3 passengers.

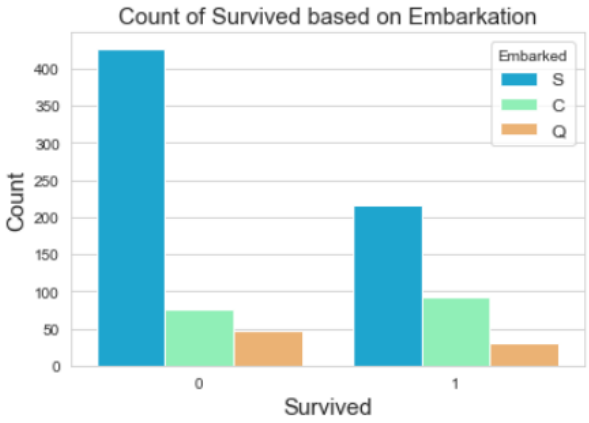

Fig. 5. Survival Count based on Port of Embarkation.

Figure 5 shows the Survival Count based on Port of Embarkation. There are three possible values for Embark — Southampton, Cherbourg, and Queenstown. More than 70 percent of the passengers boarded from Southampton. Just under 20 percent boarded from Cherbourg and the rest boarded from Queenstown. People who boarded from Cherbourg had a higher chance of survival than people who boarded from Southampton or Queenstown.

C. Correlation Heat Map

The heat map is plotted using a clean data set to identify the correlation between different features with the Survival column. From this heat map of correlation, we can see that the Survival column is primarily correlated with passenger class, and price of the ticket. Along with these columns, Age, gender, and Family(which is the sum of sibsp and Parch columns) columns also affect the survival column.

From this heat map of correlation, we can see that Survival highly correlated with Pclass, Gender, Fare, port of Embarkation. We can see that Survival is negatively correlated with Pclass which tells that the passengers of high class had high chance of survival.

D. Model

The logistic Regression model is developed on this updated data set. This updated data set comprises columns ”Survived”, ”Age”, ”Family”, ”Fare”, ”Sex”, ”Embarked”, and ”Pclass”. If we want, we can use the entire train data set to train the model and then use the model on the test data set. But this does not give us any idea about model performance. So we have decided to split the training data set using a train test split from the sklearn library with a test size of 0.2. This trained model is tested by fitting the predicted values and comparing them with the known values of the Survival column. Using the classification report from sklearn metrics we can see that our model has a precision of 0.83 and an f1 score of 0.82.

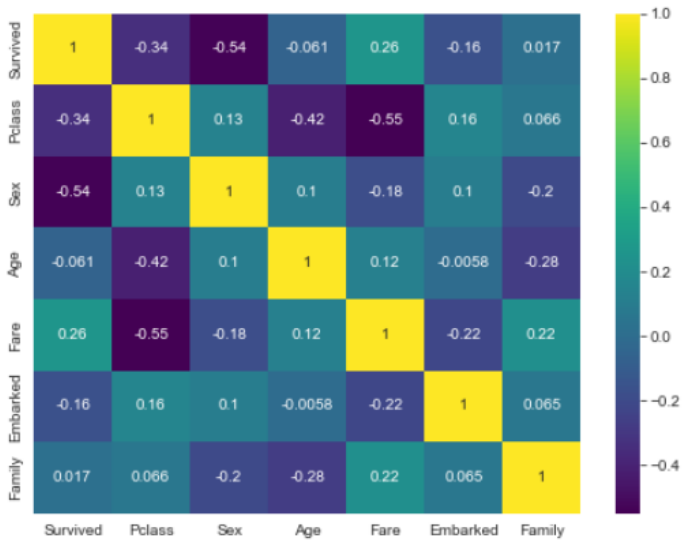

Fig. 6. Heap map of correlation.

	precision	recall	f1-score	support
0	0.83	0.87	0.85	105
1	0.80	0.74	0.77	74
accuracy			0.82	179
macro avg	0.81	0.80	0.81	179
weighted avg	0.81	0.82	0.81	179

Fig. 7. Heap map of correlation.

As our model performed reasonably well on the validation data, we can now use this model on the testing data set. Here we need to note that our model takes only a certain number of columns mentioned earlier as input so we need to perform the same column operations as training data set i.e. dropping ”Name”, ”PassengerId”, and” Ticket” columns, and label encoding of ”Sex”, ”Pclass” and ”Embarked” columns. Now we can successfully run the developed Logistic regression model to predict the survival column of the test data set.

E. Insights and Observations

From the correlation map we can see that there is a positive correlation between Fare and Survived i.e. passengers who bought expensive tickets most likely survived compared to normal tickets. We can also observe a negative correlation between Pclass and Surived i.e. Higher class is more likely to survive compared to the 3rd class. This can also be explained by the negative correlation between Fare and Pclass as Class 1 tickets are generally more expensive compared to class 2 and class 3 tickets. One more interesting observation is age has a negative correlation with Pclass, this can be informally explained as it takes time for a person to become wealthy enough to travel in 1st class considering the person belongs to the middle class. And also From Figure 3, we can see female survival rate is higher than male survival rate.

IV. Conclusion

We know that the collision of the Titanic with an iceberg has Unfortunately resulted in the deaths of 1502 out of 2224 passengers and crew. This is due to the insufficient availability of lifeboats for everyone onboard. While there might be some element of luck involved in surviving, On deeply analyzing the data it is evident that some groups of people were more likely to survive than others. The correlation matrix makes it evident that people in 1st class are more likely to survive compared to 2nd or 3rd class. Moreover, people who buy high-fare tickets are more likely to survive. Age column had a negative correlation with Survival i.e. younger people are more likely to survive. People who boarded from Cherbourg had a higher chance of survival than people who boarded from Southampton or Queenstown. Females also have a significantly higher chance of survival compared to men. The female survival rate is about 75 percent whereas for males it is around 20 percent. These can be explained by the humane practice of prioritizing women and children during the evacuation process which is reflected in the data. Our final model developed on clean data (after removing unnecessary columns and transforming the categorical columns into features) has a precision of 0.83 and an f1 score of 0.82. These evaluation parameters of the model assure us that the generated model is good and can be used on real datasets to predict if the passenger had survived or not based on the input variables.

References

[1] Bishop, Christopher M., Pattern Recognition and Machine Learning, 2006.
[2] Park, Hyeoun-Ae, “An Introduction to Logistic Regression: From Basic Concepts to Interpretation”.
[3] https://medium.com/analytics-vidhya/your-guide-for-logistic-regression-with-titanic-dataset-784943523994.
[4] https://www.geeksforgeeks.org/understanding-logistic-regression/.
[5] https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.