DAL 2023 Assignment 2: A Mathematical Essay on Logistic Regression

Arun Palaniappan
Department of Mechanical Engineering
Indian Institute of Technology Madras
Chennai, India
me20b036@smail.iit.ac.in

***Abstract*—This essay provides an overview of the mathematical formulations and applications of Logistic Regression. First, the fundamentals of Logistic Regression are explored, as well as the specifics necessary to comprehend the research. The problem statement and data-set are introduced in a detailed way. The data-set under consideration is one of the sinking of the RMS Titanic, to explore which groups of passengers were more likely to survive than others. To accomplish this efficiently, the data has been thoroughly evaluated, and both visual and quantitative insights have been provided in this essay. Finally, the results of the Logistic Regression model application are reported to better understand which features are essential when associating socioeconomic status, age and gender with the survival during the wreck. The "Data" section has been improved by adding One Hot Encoded vectors instead of label encoding.**

***Index Terms*—logistic regression, precision, recall, titanic, gender, socioeconomic groups, visualization**

I. Introduction

Logistic regression is a powerful statistical technique employed to model and predict the association between various factors or variables. It falls under the category of supervised learning algorithms, primarily used for binary classification tasks, such as spam detection. The objective of logistic regression is to determine the optimal line, referred to as the logistic curve, that best characterizes this relationship.

This is accomplished by an intelligent optimisation algorithm. We employ a *likelihood* function, which is a method of quantifying how close are the predicted probabilities of a class and the ground truth. As a result, we strive to maximize this likelihood function in order to find the best-fit model. To test how our model will perform on unforeseen data-sets in the future, we build a train-validation split, then train the model (minimize the Loss function) on the train split, and then check the prediction accuracy on the validation split.

The RMS Titanic, a British passenger liner operated by the White Star Line, met a tragic fate in the North Atlantic Ocean on April 15, 1912. This disaster occurred during the ship's maiden voyage from Southampton, England, to New York City, USA, when it struck an iceberg. Regrettably, it marked one of the deadliest maritime disasters in history, claiming the lives of over 1,500 out of approximately 2,224 passengers and crew members. Despite its reputation as an "unsinkable" vessel, the insufficient number of lifeboats onboard contributed to the loss of many lives. A passenger's chances of survival hinged on their proximity to available lifeboats, with factors such as age and socioeconomic status likely influencing their fate in this regard.

In this paper, a comprehensive understanding of logistic regression fundamentals is provided. The paper also undertakes a detailed exploration of the targeted problem and elucidates how the resulting insights are underpinned by a combination of visual representations and quantitative data.

II. Logistic Regression

In this section, we will describe the mathematical details of logistic regression. logistic regression is a supervised learning algorithm used for modeling the relationship between two types of variables, the target(dependent) and features(independent). Unlike linear regression, target of logistic regression is a categorical variable. The following is an overview of the jargons and mathematics utilised in the logistic regression approach.

- *Features/Independent variables:* The features or independent variables are used as predictor features to predict the target variable. They can be classified into three two categories i.e., continuous and categorical features.
- *Target/Dependent variables:* Dependent variables, known as target variables, exhibit either continuous or discrete traits. When implementing logistic regression for a specific problem, the target variable takes on a discrete distribution.

A. Assumptions in logistic regression

Logistic regression is a powerful statistical technique, but its validity and reliability rests upon certain assumptions that must be met for accurate and meaningful results. Here, we will delve into the key assumptions underlying logistic regression:

- *Binary target variable:* Binary Logistic regression assumes that the dependant/target variable is categorical unlike linear regression where it is continuous.
- *Independence:* The data points used in logistic regression should be independent of each other.
- *No or Little Multicollinearity:* Multicollinearity refers to a high correlation between independent variables in the regression model. This can make it challenging to isolate the individual effects of each variable, leading to unstable coefficient estimates.

- *Log odds condition:* Logistic regression assumes linearity between independent variables and log odds. Although this analysis does not require a linear relationship between the dependent and independent variables, it does require that the independent variables be linearly related to the log odds.
- *Sufficient sample size:* This assumption ensures stable coefficient estimates, greater statistical significance, and reliable model performance. A larger sample size also enhances the generalizability of results.

B. Types of logistic regression

- *Binary logistic regression:* When the target variable has only two possible classes, binary logistic regression is employed.
- *Multi-class logistic regression:* In multiclass logistic regression, the ”one-vs-all” (also known as ”one-vs-rest”) approach is a common strategy for extending binary logistic regression to handle problems with more than two classes. In this approach, we break down the multiclass classification problem into multiple binary classification subproblems, where each subproblem involves distinguishing one class from the rest of the classes. For instance, if we have three classes (P, Q, and R), we would create three binary classifiers: P vs. (Q & R), Q vs. (P & R), and R vs. (P & Q). When making predictions for a new sample, we apply all three classifiers to the sample. Each classifier provides a probability score or prediction. The class associated with the classifier that gives the highest score becomes the predicted class for the sample.

C. Algorithm

Our aim is to maximize the probability of a data point being classified as the ground truth. We could model the probability $p(X)$ using a linear regression model. But the problem in this case is that there will be some values of X for which $p(X) < 0$ and $p(X) > 0$ which is not acceptable.

Hence we use a non-linearity called sigmoid, on top of the linear regression which can map all values generated by linear equation to a range of $[0, 1]$. This is also called the *logistic function*.

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{1}$$

After performing simplification, we obtain equation for the odds of an event.

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \tag{2}$$

By taking the logarithm of both sides of equation we arrive at,

$$log(odds) = \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3}$$

The left hand side is referred to as log odds or logit. Thus, the logistic regression model has a logit that is linear in X.

To estimate the coefficients of the model, we maximize the likelihood function. This is also called the *Maximum Likelihood Estimate(MLE)*. We aim to derive estimations for coefficients in a way that the predicted probability $\hat{p}(x_i)$ closely aligns with their actual class. In simpler terms, our objective is that, when we substitute these estimations into the model for $p(X)$ as defined before, it produces a value close to one for the correct class and a value close to zero for the incorrect class. This intuition can be formed using a mathematical equation called a likelihood function that needs to be maximised:

$$\arg\max_{\beta} \ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{4}$$

Hence the MLE coefficients are obtained in this process.

D. Evaluation Metrics

We can create a table of all possible cases to evaluate the model,

TABLE I
ALL POSSIBLE OUTCOMES

All possible cases		True class	
		(1)	(0)
Predicted class	(1)	True Pos. (TP)	False Pos. (FP)
	(0)	False Neg. (FN)	True Neg. (TN)

We now define three quantities which can be used to evaluate the model.

- *Precision:* It is defined as the ratio of True positives and the total positive predictions,

$$P = \text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- *Recall:* It is defined as the ratio of True positives and the total true positives,

$$R = \text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- *F_1-score:* It is defined as the harmonic mean of precision and recall,

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

- *ROC-AUC score:* The ROC-AUC (Receiver Operating Characteristic Area Under the Curve) is a graphical representation of a binary classification model's performance. It plots the True Positive Rate against the False Positive Rate across various probability thresholds for classifying positive and negative instances.

III. The Data

A data-set with information on survival across various socioeconomic groups, age and gender are provided to us. The ultimate aim of the study is to identify key relationships between a person's survival and several factors describing the person. The significance of various features as determined by the logistic regression model is also explored in this study.

A. Data description

In this section, we describe the structure of the data-set. The train data-set has 891 entries and the test data-set had 418 entries. Some of the important columns from the data are:

- *Survived:* This column indicates whether or not the passenger survived. If it is 1, it indicates the person survived and 0 indicates did not survive.
- *Pclass:* The class on the Titanic in which the passenger was travelling. Represents first, second and third class.
- *Sex:* Indicates the gender of the passenger.
- *Age:* Indicates the age of the passenger.
- *SibSp:* Number of siblings / spouses aboard the Titanic.
- *Parch:* Number of parents / children aboard the Titanic.
- *Fare:* The cost of the ticket purchased by the passenger.
- *Cabin:* The passenger's cabin number.
- *Embarked:* Port of Embarkation, C = Cherbourg, Q = Queenstown & S = Southampton.

B. Data cleaning and imputation

While cleaning the data, we had to handle missing values and categorical classes. They were dealt in the following way:

- The data-set contained missing values only in the predictor variables. Precisely about 20% of the *Age* data was missing, which was imputed using mean. The *Cabin* column had around 75% of data missing, hence that column was removed from the feature set.
- Categorical classes were encoded using One Hot Encoding. Each distinct category is converted into a binary column, also known as a dummy variable, and a binary value (0 or 1) is assigned to indicate the presence or absence of that category. These dummy columns combine to form a new set of features that effectively represent the categorical variable in a format suitable for use in machine learning models. As we can see, the "Sex" column is transformed from TABLE II to TABLE III.

TABLE II
Sex column of data

Sex
male
female
male
female
male

TABLE III
Sex columns of one hot encoded data

Sex_m	Sex_f
1	0
0	1
1	0
0	1
1	0

One-hot encoding allows algorithms to interpret and process categorical data by providing a numerical representation that eliminates misleading ordinal relationships and ensures accurate predictions across unseen data-sets.

C. Data Visualization

1) Correlation plot: Fig. 1 shows the heat map of the correlation values between variables in the data-set(variables like name, passenger ID and ticket were dropped since they did not provide any useful information):

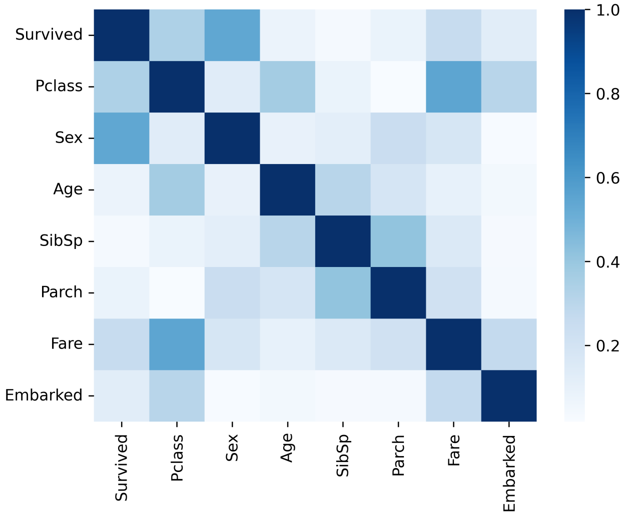

Fig. 1. Correlation heat-map within the variables

From the correlation plot we can conclude that:

- One obvious correlation is the *Passenger class* and the *ticket fare*. Higher the passenger class, higher is the ticket fare.
- The *survival* is correlated with the *Passenger class* and the *Sex* of the passenger.

2) Gender, Pclass and Survival: From Fig. 2, it is very evident that, passengers having class-1 tickets, class-2 tickets, class-3 tickets had a decreasing order of survival rate, and in each class the survival rate of females are very high compared to males.

3) Fare and Survival: Passengers who bought costlier tickets had a sightly higher survival rate than people with cheaper tickets as observed from Fig. 3.

4) Gender, Embarkment and Survival: The Fig. 4 shows the gender distribution of passengers embarking from a specific port. The majority of passengers clearly departed from

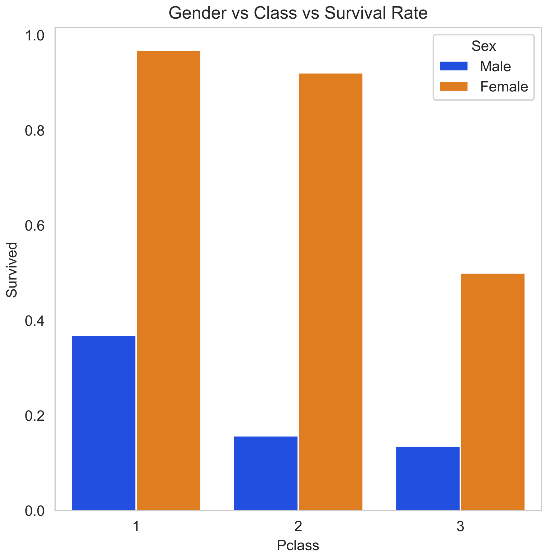

Fig. 2. Gender vs Pclass vs Survival

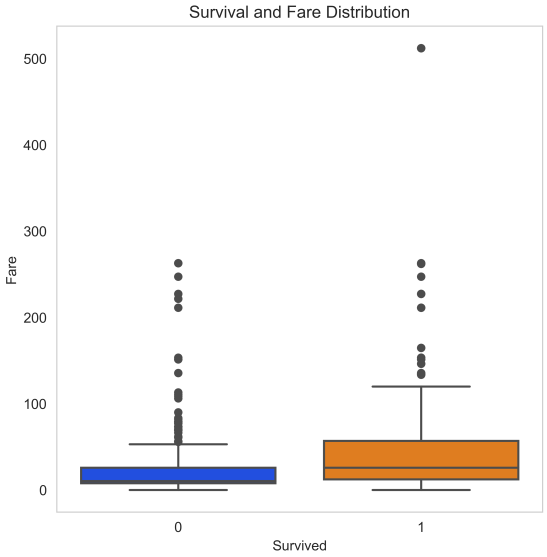

Fig. 3. Fare vs Survival

Southampton, with the fewest departing from Queenstown. Male passengers outnumber females, however females appear to have a higher survival rate.

IV. The Problem

A. Outline

We have done the analysing, cleaning and visualizing the data, as it has a lot of missing values and unclean data points. Then we decide which variables must be used in our logistic regression model according to its importance. Then we impute the missing data using imputation techniques. A logistic regression model is then fit on the data and inferences are made from the model coefficients.

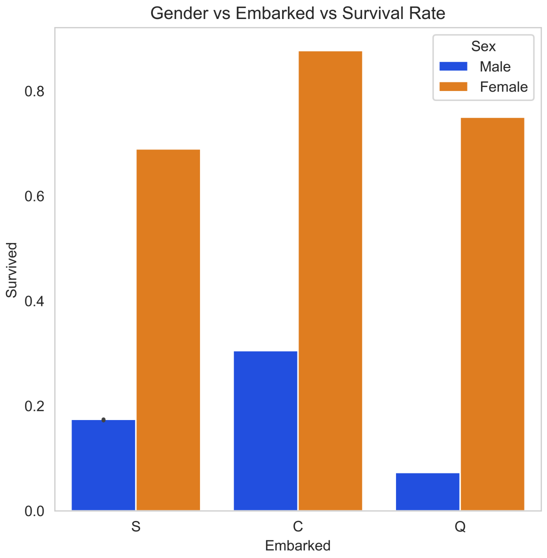

Fig. 4. Gender vs Embarkment vs Survival

B. Logistic regression modelling

We accomplish this by using *scikit-learn* package in python. The predictor variables that we use are *'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare' & 'Embarked'* We now instantiate a logisitc model, for prediciting the *'Survival'*. We use F1-score and ROC-AUC score as our evaluation criteria on the validation data-set. After testing on the Validation Split, the logistic model for *'Survival'* obtains an F1-score of 0.84 and ROC-AUC score of 0.92. While improving the paper, we implemented One Hot Encoding for the categorical variables and hence the F1-score reduced to 0.78, but our model has become more robust now without any label encoder bias.

C. Confusion matrix

This matrix generates and gives the count to all the possible outcomes in the validation set(10% of the data).

TABLE IV
Confusion matrix

All possible cases		True class	
		(1)	(0)
Predicted class	(1)	18	13
	(0)	13	41

D. ROC-AUC score

We can notice that the area under the curve (Fig. 5) is close to 1, and the curve tends towards the LHS top corner, which implies it has a high true positive rate and low false positive rate, hence more clearer differentiation between classes.

E. Feature importances

The Fig. 6 clearly shows that passenger sex has the greatest influence on whether or not a person survives, following that is the passenger class (implies a person who could afford a

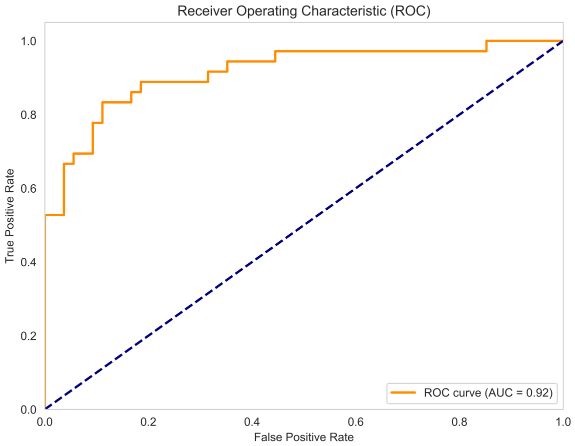

Fig. 5. ROC-AUC

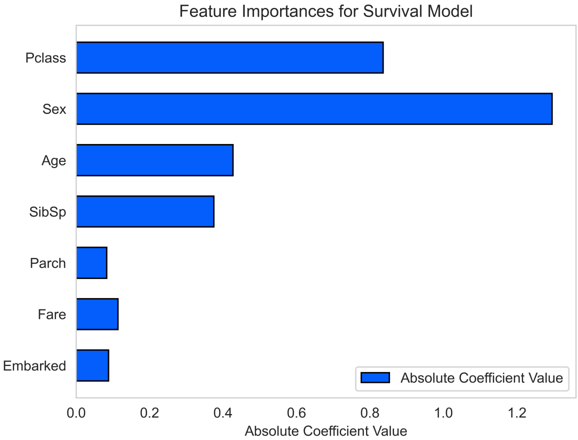

Fig. 6. Feature importances

high class ticket is more likely to be saved). This confirms the relation between socioeconomic status and survival.

We can also observe that SibSp and Parch also have played a role in model. This could be because when a person has a family aboard he/she is more likely to be given a life boat together and hence saved.

V. Conclusion

The conclusions that can be drawn from all the above analysis are:

- The most important characteristics according to Feature Importances are the sex of the passenger and the passenger class.
- This is evident from history that, usually women and children are brought to safety first, hence they have a high survival rate.
- It is also evident that a person who is in a superior socioeconomic group is more likely to be saved.
- The plots shown, support all the above quantitative evidence in a visual way.
- The model has been run on the test-set and the prediction are provided within the code file.

Future research could use *One-Hot Encoding* for sex of passenger and port of embarkment, that is creating dummy columns for denoting the presence of a class to avoid any unnecessary bias. We have observed that, there is class imbalance in the survival column. This can create bias inherent bias in the model, hence we can adopt class-balancing techniques to overcome this bias. We could also explore on using regularised techniques(L1 and L2). We can also apply Principle Component Analysis(PCA) to find important combination of features to explain the variance in the data.

References

[1] ”Sinking of the Titanic.” Wikipedia, Wikimedia Foundation, 8 Sept. 2023, *https://en.wikipedia.org/wiki/Sinking_of_the_Titanic* (accessed 10 Sept. 2023.)

[2] ”Logistic Regression.” Wikipedia, Wikimedia Foundation, 31 Aug. 2023, *https://en.wikipedia.org/wiki/Logistic_regression* (accessed 10 Sept. 2023)

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistical Learning : with Applications in R*. New York :Springer, 2013.