A Mathematical Essay on Logistic Regression

S ANIRUDDH
Mechanical Engineering
IIT MADRAS
Chennai,India
me18b185@smail.iitm.ac.in

***Abstract*—Logistic Regression is a supervised learning algorithm used in binary classification problems.In this paper, we study the mathematics behind Logistic Regression and apply this technique to a real world business problem.**

I. Introduction

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 linearly separable and the outcome is binary or dichotomous in nature.Logistic regression is usually used for Binary classification problems. Binary Classification refers to predicting the output variable that is discrete in two classes.

The given problem given has two data sets containing information about the passengers of the RMS Titanic Ship. One of the dataset namely train set has the survival information along with the above parameters.Our aim is to build a predictive model using the Logistic Regression technique to predict the survival of passengers in the test data set.

In Section 2, we will look through the theory behind the logistic regression, model and its various types.

In section 3, We briefly inspect the data set using exploratory data analysis techniques. We will first visualize the given data before applying any pre-processing techniques. Then we start applying our technique to the data set discussing about each step with reasons and after having obtained a model, we will finally look for insights within the data through our technique.

In section 4, we will conclude our study by discussing our observations and suggest some methods which will be helpful for the non profit organization in their mission.

II. Logic Behind Logistic Regression

A. Conceptual View

It's a classification algorithm, that is used where the response variable is categorical. The idea of Logistic Regression is to find a relationship between features and probability of particular outcome.

E.g. When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

This type of a problem is referred to as Binomial Logistic Regression, where the response variable has two values 0 and 1 or pass and fail or true and false. Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.

B. Why Logistic, why not Linear Regression?

In a classification problem, we will have a input feature X and the output feature Y will be 0 or 1 representing classes.

Linear regression model can generate the predicted probability as any number ranging from negative to positive infinity, whereas probability of an outcome can only lie between $0 < P(x) < 1$.

Also, Linear regression has a considerable effect on outliers. To avoid this problem, log-odds function or logit function is used.

C. Logistic Function

The Logistic Regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The Logistic Function is defined as:

$Logistic(X) = 1/(1 + exp(-X))$

The step from linear regression to logistic regression is kind of straightforward. In the linear regression model, we have modelled the relationship between outcome and features with a linear equation: $y = a + bx + cy + dz$

For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.

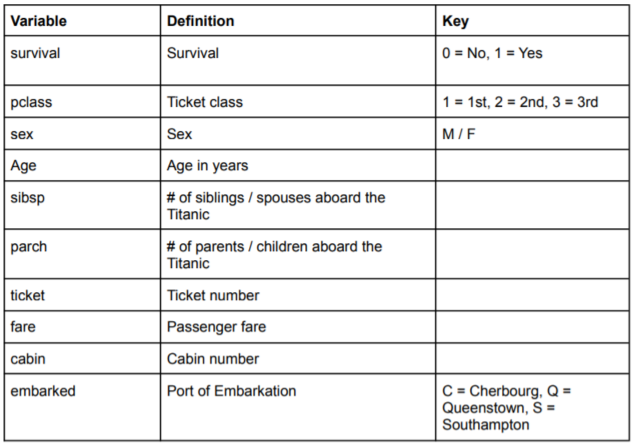
$p(y = 1) = 1/(1 + exp(-(a + bx + cy + dz))$

III. Working with Data

Before applying any pre processing technique let us inspect the given data set. The given dataset to work with is $'$train.csv$'$. It has 891 rows with the attributes- name of the passenger, number of siblings, number of parents or children, cabin, ticket number, fare of the ticket and the place where the person has embarked from. The raw data set has meta data and incomplete or missing values which will be filtered out in Pre processing which is described below.

Before the algorithm is built for this specific model, a few data exploration graphs have been made to analyze which features could be detrimental to the model and which could help us ameliorate our result.

We can observe that there are missing values in the selected features. We are not going to simply omit those rows containing the missing values because that would lead to loss of significant amount of data. Rather, we try to fill the missing values with NaN and then use the completely filled data set to get our model. Filling up NaN values is briefly discussed



Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	M / F
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Fig. 1. Attributes and their description

in the next sub section. The Categorical features has been encoded using label encoder.

Fig 1 gives us a brief outlook of the name of the features and what they depict.

A. Dealing with missing values in the selected features

First of all, we need to check NaN values in a column. This can be done using $isna()$ function in Pandas library.Once we find there is a missing value in the selected feature column,there are many standard ways that can be used to fill the missing values in the column. Some of the functions that can be used are $fillna()$ and $interpolate()$. These functions fill the NaN entries with some values. But I found the mean of the column and filled the NaN values with mean of that column. Note, while calculating mean, python treats NaN values as zero. I preferred mean for the feature Age because mean was a good approximation compared to other statistical measures of the column. For the feature , we fill the missing value with mode. Because It is the best way for handling missing values in Categorical Variables. Now,we are done with filling the missing values in the selected features.

B. Handling Categorical Variables

Originally, the given data set had many categorical variables. The features such as $Name, TicketNumber, Fare andPassengerId$ have been dropped since these don't contribute to the model in an effective way. Also, the categorical variable $Cabin$ has been dropped since it has more than 600 missing values and it is not viable to fill this much large missing values as they could affect the efficiency of the model.

Now, we need to convert the remaining categorical values to a numerical form to include this feature in our model. This can be achieved by Label Encoder class of scikitlearn pre processing module. Fig 2 shows the correlation matrix of the features obtained after pre processing.

C. Visualising the data

- Fig 3 shows Age Vs Survived information.We can see that majority of persons survived lie in the range (0,60).

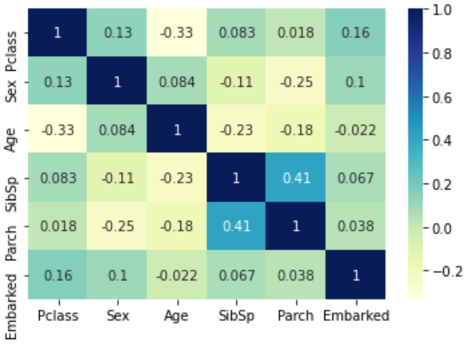


Fig. 2. Correlation matrix

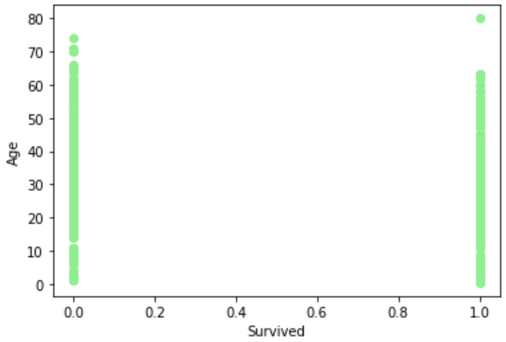


Fig. 3. Age of Survived Persons

We can infer that old age persons are more likely to not survive the sinking of ship.

- Fig 4 shows the class occupied by the passengers. We can infer that Class 3 is the most occupied Passenger Class.
- We can see from Fig 5 that Male constitute the majority of population.
- From Fig 6 and 7 we can see that with in the class of male majority of them didn't survive but within the class of females, majority of them survived.

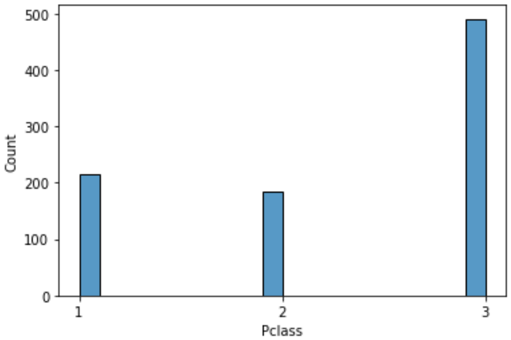


Fig. 4. Class occupied

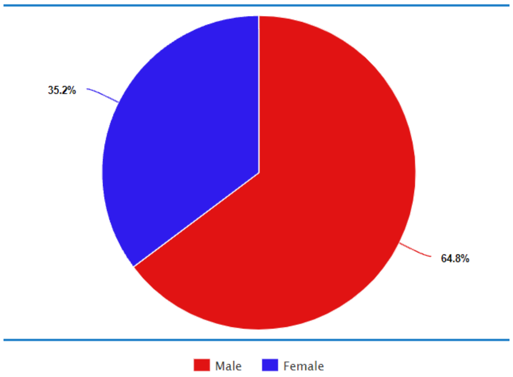


Fig. 5. Spread among Sex

D. Logistic Regression Model

Actually, we are given a prediction problem not a classification problem. Train data set is used to build a logistic regression model and the model is used to predict the survival of passengers in the test data.

In the previous figures, we could easily infer that the features $Age, Sex, PClass, Embarked, Sibsp, Parch$ affected the survival of passengers.One way of establishing the relation of the feature with the survival prediction is to inspect the correlation matrix and observe which of the features are correlated. But this is not sufficient, because we are only able to infer that how the features are correlated with Survival. we can only predict whether they are directly proportional or inversely related. This does not give any insight about how a small change in any feature can bring about a change in the output feature. Also, building up a model paves way for giving a quantitative and visual evidence to support the relation. So, we use these features to build a logistic regression model.

How to validate the accuracy of the model as the ground truth of the test class is not known?
We can solve this problem by using 80 percent of the train data set for the model building and use the remaining 20 of the train data set for checking the accuracy of the model. Logistic Regression model is built with the training set using the sklearn library.

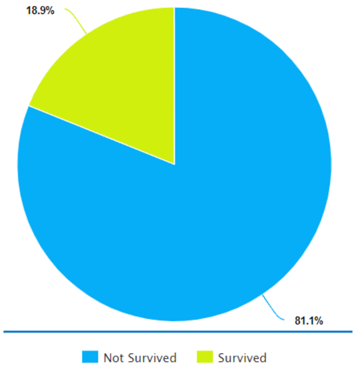


Fig. 6. Survival among Males

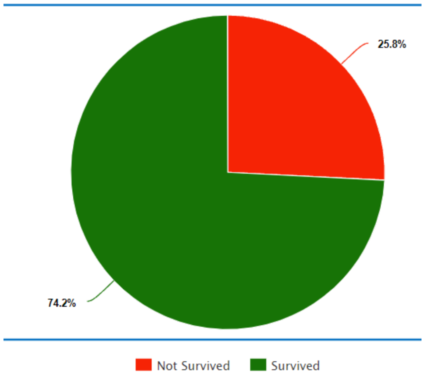


Fig. 7. Survival among Females

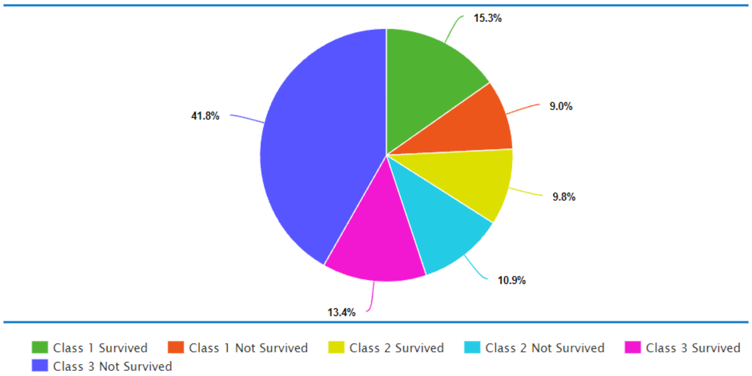


Fig. 8. Survival of passengers spread across passenger class

We will now check the coefficients of our model. Logistic regression worked on categorical values of Embarked, Sex and Pclass but also with numeric values of greater range like that of Age, SibSP. Probability of all attribute values crucial for prediction of survival have been calculated

Intercept: 5.40697804
Coefficients : [-1.17791199 -2.6193015 -0.03989492 -0.33808123 0.02956076 -0.21356421]

The Negative coefficients of Sex, Embarked and P class imply that keeping other variables constant, passenger with value of sex tending towards 0 (i.e for female), who have boarded from SouthHampton(2) and belonging to third class were more likely to surive. Also, we need to note that survivors belonging to first and third class were comparable in size.

The accuracy of this model can be evaluated from the confusion matrix generated for the 20 percent of the train data set.

Model accuracy is 80.3 percent, which implies the model is very good in terms of the accurateness. We can now use this model to predict the outcomes of test data.

In the next section, we will observe the insights obtained from the test data.

E. Insights from the test data outcome

First of all,let us check gender wise survival.
We observe that out 418 passengers, only 155 survived and

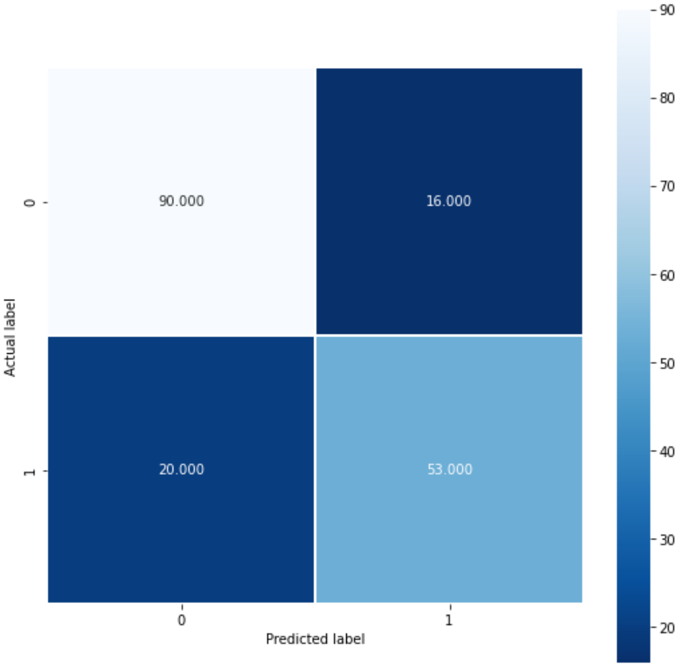


Fig. 9. Confusion matrix

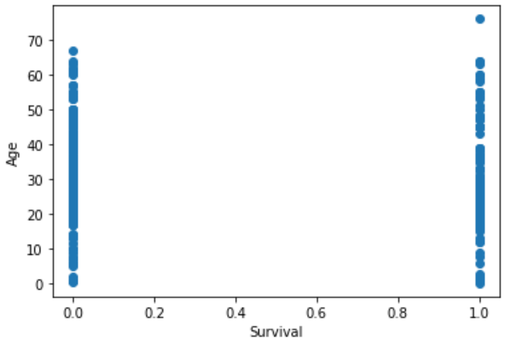


Fig. 10. Survival of passengers in the test data set spread across their Age

among these 155, 144 of them are females.
This confirms our model coefficient obtained above that females are more likely to survive.

Now, let us check how Embarked feature is affecting the outcome.Similarly, we can see that majority of the survivors have boarded from Southhampton.It is actually 8o out of 155 survivors have boarded from Southhampton. This confirms our model coefficient obtained above that People who have Embarked from Southhampton are more likely to survive.

Let us consider survival by age.We can see that people with age in range 20 to 60 years are more likely to survive. Fig 3 plotted from the train data also reflects the same.

Survival with respect to Pclass feature establishes that first and third class had comparable survivors and second class passengers were less likely to survive. The same is confirmed from test data.

F. Conclusion

Logistic Regression is a widely used technique in various branches of science and technology. In this paper, Logistic Regression has been used to predict outcomes of a data set pertaining to a real world business problem. After applying the Logistic Regression technique to the given data set, we were able to establish a correlation between Sex, Age, Pclass and Embarked features with the survival of passengers.
We can clearly see that Sex and Pclass which is an indicator of socio-economic status has an impact on the survival of the passengers. Conclusion drawn from any model depends on the accuracy of the model which in turn depends on the pre processing involved in the data set.The categorical variable cabin has been dropped due to lack of information and if cabin wise information has provided with, we could more specifically predict which section of the ship would more likely to survive and also we found that most of the features were not useful in classification.For example, the number of sibling/spouses and the number of parents/children did not help with classification in any of the three models. Knowing the number of relatives aboard did not help with classification, but perhaps, if we were given the links between passengers then we'd be able to infer more about the survival rate. Since family units tend to all die or all survive, knowing the family links would have been useful. However, since our model has around 80 percent accuracy, we can say with good amount of confidence outcomes of the test data are reliable to a good extent.

References

[1] Farag, Nadine, and Ghada Hassan. Predicting the Survivors of the Titanic Kaggle, Machine Learning From Disaster. ICSIE '18 Proceedings of the 7th International Conference on Software and Information Engineering , May 2018, dl.acm.org/citation.cfm?id=3220282.

[2] https://en.wikipedia.org/wiki/Logistic regression

[3] K. Singh, R. Nagpal and R. Sehgal, "Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset," 2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2020, pp. 320-326, doi: 10.1109/Confluence47617.2020.9057955.

[4] A. Singh, S. Saraswat and N. Faujdar, "Analyzing Titanic disaster using machine learning algorithms," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 406-411, doi: 10.1109/CCAA.2017.8229835. 89.

[5] http://cs229.stanford.edu/proj2012/LamTang-TitanicMachineLearningFromDisaster.pdf