Assignment 2: Theory of Logistic Regression

Shashwat Patel
Metallurgical and Materials Engineering
Indian Institute of Technology,Madras
mm19b053@smail.iitm.ac.in

***Abstract*—Logistic regression is a classical classification method in statistical machine learning that uses a logistic function to model binary dependent variable. The main idea behind logistic regression is to use sigmoid function to nonlinearize multivariate linear regression. In this paper, we explain the theory behind logistic regression and try to understand the effect or parameters key to the survival of a person had they been on the famous Titanic ship. The survival prediction has been done by using Logistic regression.**

***Index Terms*—Titanic, Logistic regression, Sigmoid function, Confusion matrix, Cost function, odds ratio**

I. Introduction

RMS Titanic was a British cruise ship said to be the largest cruise ever made in the history of world. It sank in the North Atlantic Ocean on 15 April 1912 . It collided with an iceberg during its maiden journey from Southampton to New York City. With more than 2200 passengers on board, nearly half of them died after the unprecedented mishap making the sinking at the time one of the deadliest of a single ship. The disaster was met with worldwide shock and outrage at the huge loss of life, as well as the regulatory and operational failures that led to it. Public inquiries in Britain and the United States led to major improvements in maritime safety. With much public attention in the aftermath, the disaster has since been the material of many artistic works and a founding material of the disaster film genre. [1]

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether an email is spam or not spam or whether a high school student will pass or fail the exam.

Sigmoid Function:

$$S(x) = \frac{1}{1 + e^{-x}}$$

It is the go-to method for binary classification problems (problems with two class values). The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular outcomes. It uses sigmoid function (Eq 1.) to nonlinearize the linear regression. It is used when the dependent variable is categorical.

The RMS Titanic sinking disaster was met with worldwide shock and outrage at huge loss of life. The data of this infamous incident is used to identify what parameters were key to the survival of the person had they been on the ship. Whether a person survives or not prediction is done by using the logistic regression.

This paper majorly deals with the theory of logistic regression and the mathematics behind it. We try to understand the parameters that were important for survival in the Titanic disaster. Two datasets have been given: train.csv and test.csv, using the train.csv dataset we train the logistic regression model and using this model we predict the survivability of a person in test.csv dataset.

II. Logistic Regression

Logistic regression is a classification based algorithm which works on discrete data instead of continuous data. After learning from the data, it classifies what the result is in two groups: True(1) and False(0), if the problem is a binary classification problem. Generally logistic regression is used for binary classification problems but it can be used for multiclass classification as well by using the complex extensions to it.

Say, we have to filter emails as spam(1) or not spam(0), to predict which class a data belongs, a threshold is set. Based upon this threshold, the obtained estimated probability is classified into classes. If predicted value ≥ 0.5, then email is classified as spam else as not spam.

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:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

The logistic curve looks like this:

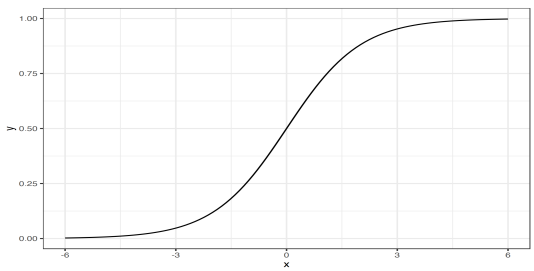

Fig. 1. Logistic function

In the linear regression model, the relationship between outcome and features is given with a linear equation:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_p x_p^{(i)} \quad (2)$$

For classification problems probabilities are used, so we put the RHS of Eq. 2 in the logistic function. [2] This forces the output to assume only values between 0 and 1.

$$P(y^{(i)} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_p x_p^{(i)})}} \quad (3)$$

Since the outcome in logistic regression is a probability between 0 and 1, the coefficients do not influence the probability linearly [2]. On reformulating the Eq. 3, we get:

$$log(\frac{P(y=1)}{P(y=0}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (4)$$

The Eq. 4 is known as "log odds" and the term inside the log() function is known as odds(probability of event divided by probability of no event). On applying exp() function to both sides of Eq. 4, we get:

$$\frac{P(y=1)}{P(y=0} = odds = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} \quad (5)$$

Then, we compare what happens when we increase one of the feature values by 1 and keeping all other feature values the same.

$$\frac{odds_{x_j+1}}{odds} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j (x_j+1) + \cdots + \beta_p x_p}}{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_p x_p}} \quad (6)$$

$$\frac{odds_{x_j+1}}{odds} = e^{\beta_j} \quad (7)$$

Eq. 6 and 7 shows that a change in a feature by one unit changes the odds ratio by a factor of e^{β_j}.

To estimate the values of the coefficients β, cost function J has to be minimized. [3]

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} -y_i log(h(x_i)) - (1 - y_i) log(1 - h(x_i)) \quad (8)$$

Here, $h(x_i)$ is same as Eq. 3.

To minimize J, gradient descent algorithm is used on every coefficient. [3]

repeat until convergence {

$$\beta_j := \beta_j - \alpha \frac{\partial J(\beta)}{\partial \beta_j} \quad (9)$$

}

Here α is the learning rate that needs to be set explicitly. Every β_j has to be updated simultaneously.

$$\frac{\partial J(\beta)}{\partial \beta_j} = \frac{1}{n} \sum_{i=1}^{n} (h(x^{(i)}) - y^{(i)}) x_j^{(i)} \quad (10)$$

On putting the value of $\frac{\partial J(\beta)}{\partial \beta_j}$ in Eq. 9 we get

$$\beta_j := \beta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} (h(x^{(i)}) - y^{(i)}) x_j^{(i)} \quad (11)$$

After convergence, we get the estimated values of the coefficients. These coefficients can also be estimated using the theory of maximum likelihood.

To measure the accuracy of the coefficient estimates, their standard error is computed. Z-statistic associated with a coefficient β_j is defined as $\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$ and so a large values of Z-statistic indicates that the feature associated with β_j is significant and this rejects the null hypothesis indicating that there is indeed an association between the feature variable and the dependent variable. [4]

There are different evaluation metrics to assess the accuracy of the model, Akaike Information Criteria(AIC) is an important indicator of model fit. The smaller the value of AIC, better is the model fit. It is more helpful in model selection. [4]

Confusion matrix is the most crucial metric commonly used to evaluate classification models. The confusion matrix avoids "confusion" by measuring the actual and predicted values in a tabular format [5]. In Figure 2, Positive class = 1 and Negative class = 0. Different metrics can be derived from

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

Fig. 2. Confusion Matrix

confusion matrix like accuracy, precision, F-score, sensitivity and specificity.

Receiver Operator Characteristic (ROC) determines the accuracy of a classification model at a user defined threshold value. It determines the model's accuracy using Area Under Curve (AUC). The area under the curve (AUC) represents the performance of the ROC curve. Higher the area, better the model. [4]

III. The Problem

Sinking of Titanic is considered as one of the most devastating disaster. There weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew [1]. While there was some element of luck involved in surviving, some groups of people were more likely to survive than others. Using the data given in "train.csv", the task is to prepare a logistic regression model and predict whether a person survived or not given in "test.csv" dataset. The dataset consists of name, age, gender, passenger class, fare, ticket no. , cabin no. , port of embarkment, no. of siblings

boarding the ship, no. of parents/children boarding the ship and the person survived or not.

*Data Cleaning:*77% of rows have missing data in the cabin column, so ”cabin” column was removed. 1 row of missing data in ”Fare” column was imputed with median fare value of passengers who embarked from Southampton and were in 3rd passenger class. Mean was not used due to lot of outliers and would likely overestimate the fare. 2 rows of missing data in ”Embarked” column, both person were travelling on the same ticket and so they would have embarked from same place and most of the people had travelled from Southampton only, so Southampton was imputed as embarkment place for those persons. 263 rows had Age data missing, for imputation median age has been used.

A new column ”family_size” has been created by adding the ”Parch” and ”SibSp” column. Another column, ”Alone” is created which describes whether the person was travelling alone or was in some group. ”Age_Group” column has been created where ages of people have been grouped and has been made for visualization purposes. On analysis it is found that a group of people travelled on a same ticket. Some of these groups were families and some were not, so a column ”No_per_ticket” is made which consists of number of people who travelled on same ticket.

In Figure 3, a linear relationship between fare and number of people on same ticket can be seen, previously it was considered that fare data consisted of fare of single person only but Figure 3 leads us to believe that the fare data consists of total fare of a group. With that a new column called ”Fare_per_person” is created by dividing Fare with No_per_ticket column. Figure 4 makes much more sense because fare in same passenger class should almost be same, these fare differences are due to different embarkment places or large groups might have got some discount on fare .

Fig. 3. Total Fare(Passenger class 3 only)

Figure 5 shows the box-plot of fare per person based on passenger class. There are some outlier but data makes more sense here as fares of passenger class 3 are lesser than class 2 and fares of class 2 are lesser than class 1.Surprisingly, 17 people travelled at zero fare and all of them embarked from Southampton. 4 people were travelling on ticket ”PC 17755”

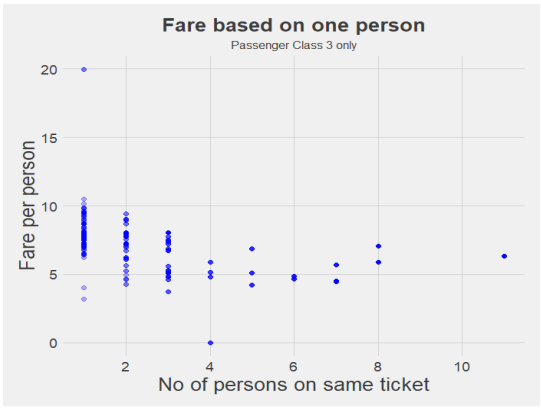

Fig. 4. Fare per person(Passenger class 3 only)

and they gave the maximum fare of 128. All these 4 people embarked from Cherbourg, which was not even the starting point of the Titanic.

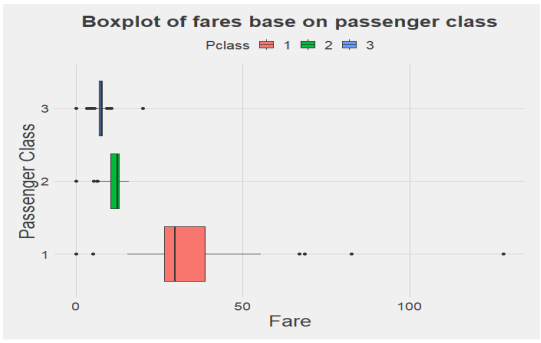

Fig. 5. Fare per person box-plot)

Visualization: some certain insights helps in finding factors which were helpful in survivability of a person. Figure 6 shows

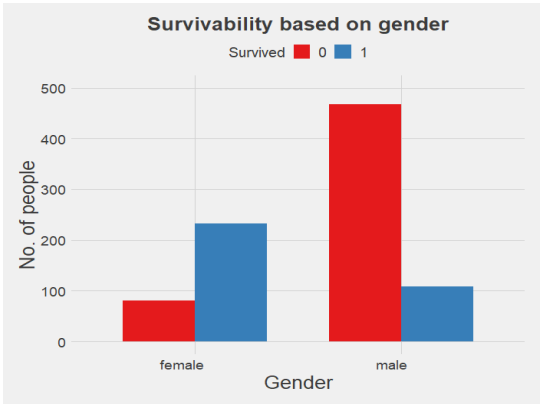

Fig. 6. Survivability dependency on gender

that female population in general tended to survive more than the male population. Around 74% of female population on the ship survived while around 19% of male population survived. It can be inferred that females were prioritised more during the saving process on the ship. The Figure 7 shows that number

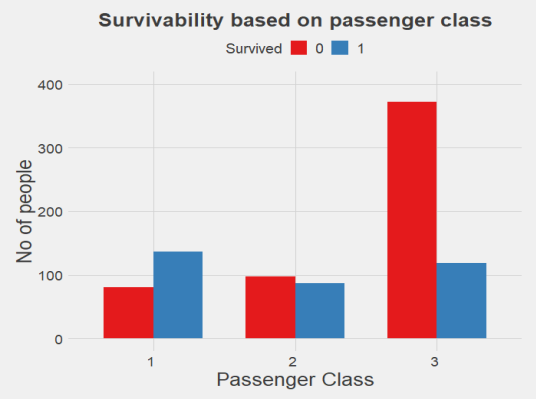

Fig. 7. Survivability dependency on Passenger class

of passengers in class 1 survived the most, and class 3 had least amount of survivors. From figure 7 it can be inferred that richer people in general were prioritised more for saving, as class 1 had highest fare while class 3 had least amount of fare. Figure 8 shows the survivability of passengers in different

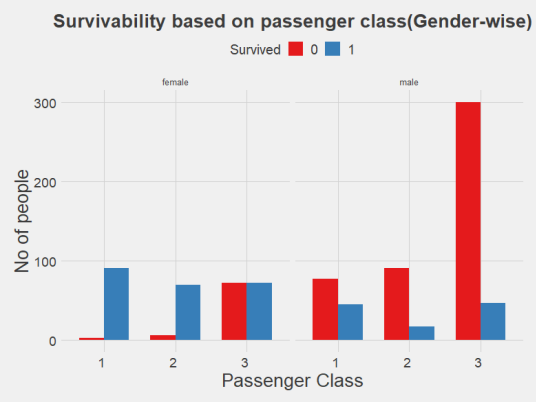

Fig. 8. Survivability dependency on Passenger class(gender-wise)

classes gender-wise. It shows that females in every passenger class tended to survive more than the male population.

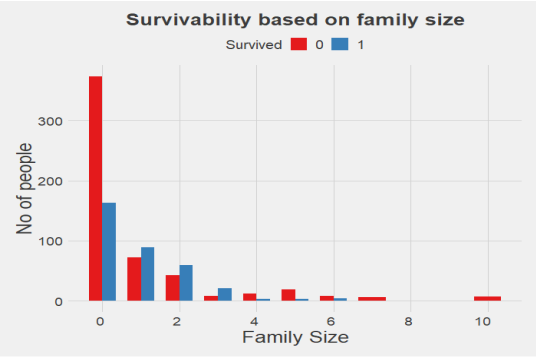

Fig. 9. Survivability dependency on family size

Figure 9, shows that generally people with larger family size survived less. It can be clearly seen that people with family size more than 6 did not survive at all while more number of people travelling alone without any family member survived.

Many of the people who survived were from Southampton, the starting point of the Titanic. It makes sense that most of the people embarked from Southampton only, so much more people survived that embarked from Southampton.

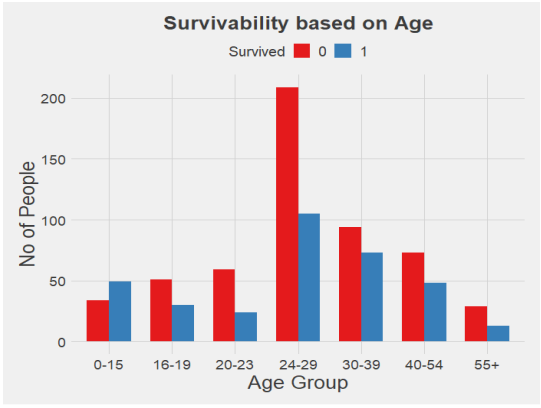

Fig. 10. Survivability dependency on age

Figure 10 shows the dependence of age on survivability. Younger people had high amount of people surviving from the disaster. Younger people were prioritised for saving. Very few 55+ (elderly) people survived the mishap.

Logistic Regression Model: Using the data, 2 models are built. The 1st model uses variables that were originally present like pclass, Sibsp, parch, age, fare etc. The 2nd model uses some data which was originally present and some variables which were made like family size, alone, fare_per_person etc.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6199  -0.6089  -0.4176   0.6187   2.4514

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.064159   0.472813   8.596  < 2e-16 ***
Pclass2     -0.919468   0.297326  -3.092  0.00199 **
Pclass3     -2.150048   0.297720  -7.222 5.13e-13 ***
Sexmale     -2.719444   0.200977 -13.531  < 2e-16 ***
Age         -0.038517   0.007855  -4.903 9.43e-07 ***
SibSp       -0.321794   0.109193  -2.947  0.00321 **
Parch       -0.093329   0.118856  -0.785  0.43232
Fare         0.002339   0.002469   0.947  0.34346
EmbarkedQ   -0.056267   0.381471  -0.148  0.88274
EmbarkedS   -0.434226   0.239530  -1.813  0.06986 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  785.04  on 881  degrees of freedom
AIC: 805.04

Number of Fisher Scoring iterations: 5
```

Fig. 11. 1st Logistic Regression Model

In Figure 11, the 1st logistic regression model shows that passenger class, gender, Age are factors that affected the survivability the most. Surprisingly, fare does not seem to be that important factor for survivability but in Figure 7, it was seen that passenger in class 1 had higher survival rate as compared to other classes and class 1 fare was the highest. Place of embarkment does not play that much role in survival. The 2nd model shows that number of family members as well

as whether the person was alone or in group also affected the survivability.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4599  -0.6173  -0.4205   0.6083   2.4729

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.951760   0.593278   6.661 2.72e-11 ***
Pclass2        -0.721217   0.360850  -1.999 0.045645 *
Pclass3        -1.842937   0.387349  -4.758 1.96e-06 ***
Sexmale        -2.683162   0.201285 -13.330  < 2e-16 ***
Age            -0.037582   0.007907  -4.753 2.01e-06 ***
EmbarkedQ       0.021825   0.382377   0.057 0.954484
EmbarkedS      -0.375331   0.243365  -1.542 0.123011
family_size    -0.286340   0.080211  -3.570 0.000357 ***
Alone1         -0.474903   0.228976  -2.074 0.038077 *
Fare_Per_Person  0.013203   0.011966   1.103 0.269876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  781.89  on 881  degrees of freedom
AIC: 801.89

Number of Fisher Scoring iterations: 5
```

Fig. 12. 2nd Logistic Regression Model

The AIC of model 2 is lesser than the model 1, so using model 2 is better. Figure 13, shows the confusion matrix. The

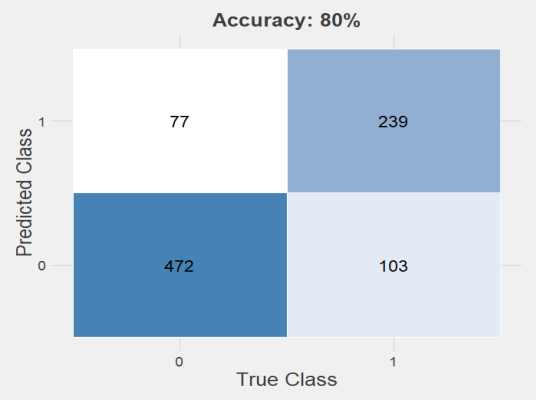

Fig. 13. Confusion Matrix

accuracy of the model is 80%. 77 passengers who did not survive were predicted to survive while 103 passengers who survived were predicted dead. On applying the model to the test dataset "test.csv", the model predicted that out of 418 people 158 people survived and 260 people were predicted dead.

The model suggests that certain factors like passenger class, gender, age, family size, person was alone in the ship or was in group are some key parameters that affected the survival of person. Fare per person, Place of embarkment did not effect the survival very much.

IV. Conclusion

From the model, it is learnt that certain parameters like age, family size, passenger class, gender, whether the person was alone on the ship or was in group were the key for survival. More females survived as compared to males. Passenger in class 1 survived more in comparison to passengers in class 2 or 3. Younger population survived more in comparison to elderly. People who were travelling alone on the ship had higher chance of survival. Our model had an accuracy of 80% . When the model was applied on test dataset it was found that 158 people out of 418 people survived.

Further improvements in the model are possible. The "Name" column in the dataset was not used but feature extraction can be used there. Title of passenger name can be extracted and that can be used as a variable. "Cabin" column can be used to for further insights. Introducing new feature can help as well i.e. using log(age) instead of age can help. Instead of using only logistic regression other models like random forest, support vector machine and further sophisticated models like neural networks can also be used.

References

[1] https://en.wikipedia.org/wiki/Titanic

[2] Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. https://christophm.github.io/interpretable-ml-book/.

[3] https://www.internalpointers.com/post/cost-function-logistic-regression

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

[5] https://cran.r-project.org/web/packages/caret/caret.pdf