A Mathematical Essay on Logistic Regression

Prajwal Dinesh Sahu
ME18B114
Dept. of Mechanical Engineering
Indian Institute of Technology, Madras
me18b114@smail.iitm.ac.in

***Abstract*—This paper gives a brief introduction of multivariate logistic regression and the basic mathematical theory behind it. We also look at some metrics to evaluate the performance of logistic regression models. Further, we apply logistic regression on a real world data-set containing the socio-economic details of the passengers who were on the RMS Titanic ship, which sank after colliding with an iceberg, resulting in the death of 1502 out of 2224 passengers. Through exploratory data analysis and predictive logistic regression model, we investigate trends in data to predict which passengers were more likely to survive based on their age, fares, passenger class, embarkation point etc. Then we evaluate our model using confusion matrix and F1 score. Our findings reveal some interesting trends.**

I. Introduction

There are multiple types of algorithm methods used in machine learning. One such popular and commonly used machine learning method is logistic regression. It is a supervised machine learning algorithm used to train a model to predict the behaviour of our data based on given parameters. It is used to estimate the relationship between a dependent (target) variable and one or more independent variables.

Logistic regression is a classification algorithm which models the probabilities with two possible outcomes. In this algorithm, the Logit function is used for calculating the log odds ratio, which in turn gives the likelihood of that particular outcome. A general equation for logistic regression can be written as -

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \qquad (1)$$

where,
p is the probability of event happening.
$\beta_0, \beta_1, ...\beta_n$ are the coefficients of the model.
To calculate these coefficients, we use Maximum Likelihood Estimation with log-likelihood as the cost function.

In this paper, we apply logistic regression to the data-set containing socio-economic data of all the passengers on the RMS titanic ship which sank on April 15, 1912. We are given various details of the passenger like their age, ticket fare, passenger class, embarkation point, sex etc. We are required to find out if it was just luck or is there is any relation between the above mentioned factors and the survival of the passenger.

The paper will first give a brief explanation of logistic regression. Then, we will try to solve the Titanic problem and model our data-set by
1. Cleaning the data.
2. Feature engineering for categorical variables.
3. Exploratory and visual analysis.
4. Checking for Correlation.
5. Scaling the Data. 5. Training the model.
6. Evaluating the model.
7. Interpreting the results.

II. Logistic Regression Theory

A. *Classification*

Classification is a process of categorizing a given set of data into classes. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, label or class.
Many real world problems require classification rather than regression. In the classification setting we have a set of training observations (x1, y1), . . . , (xn, yn) that we can use to build a classifier.

Linear regression is not useful in problems like if we are trying to predict the medical condition of a patient in the hospital and we have 3 possible diagnoses- stroke, drug overdose, and epileptic seizure based on the symptoms of the patient. We can label these diseases at 1,2 and 3. In this case, even if we create a metric to accurately convert the symptoms to numeric data, regression won't be able to classify the data as it cannot give the output as 1,2 or 3. Alternatively, we can create a threshold that if predicted value is between 0.5 to 1.5, we will assign it to label 1. This brings us to the basic idea behind classification algorithms which can give qualitative response.

B. *Logistic Regression*

*1) **The Logit Function**:* A Logit function, commonly known as the log-odds function, depicts probability values ranging from 0 to 1, as well as negative infinity to infinite. The function is the inverse of the sigmoid function, except

instead of the X-axis, it restricts values between 0 and 1. The Logit function is most often utilised in analysing probabilities since it occurs in the 0 to 1 range.

$$\text{logit}(p) = \sigma^{-1}(p) = \ln\left(\frac{p}{1-p}\right) \quad \text{for} \quad p \in (0,1) \tag{2}$$

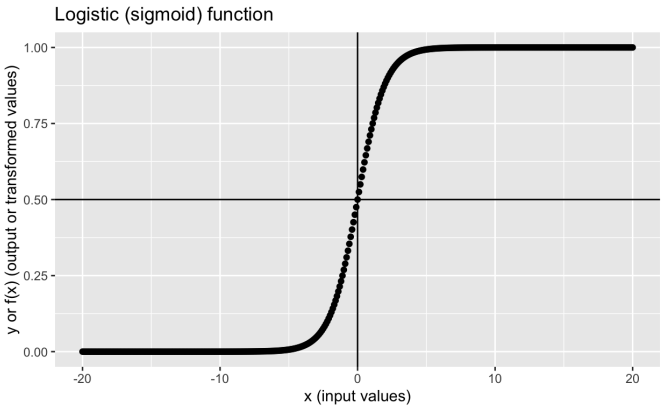

Fig. 1. Logit function in the domain of 0 to 1

*2) **Using Logistic Regression**:* Logistic regression is a technique used when the dependent variable is categorical. Logistic regression employs binomial probability theory in which there are only two values to predict: that probability (p) is 1 rather than 0, i.e. the event belongs to one group rather than the other.

Consider a model with two predictors, 1x1 and x2, and one binary (Bernoulli) response variable, Y, which we call p=P(Y=1). We assume that the predictor variables and the log-odds (logit) of the occurrence that Y=1 have a linear relationship. The following mathematical form may be constructed for this linear connection

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3}$$

where l is the log-odds, b is the base of the logarithm, and β_{i} are parameters of the model

We can calculate odds and probability as

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1} \tag{4}$$

$$p = \frac{b^{\beta_0+\beta_1 x_1}}{b^{\beta_0+\beta_1 x_1}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1)}} \tag{5}$$

*3) **Maximum Likelihood Estimation**:* In logistic regression, we use Maximum Likelihood Estimation(MLE). The main aim of MLE is to find the value of our parameters for which the likelihood function is maximized. The likelihood function is nothing but a joint pdf of our sample observations and joint distribution is the multiplication of the conditional probability for observing each example given the distribution parameters.

Let a generalized linear model function parameterized by θ

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}} = \Pr(Y=1 \mid X;\theta) \tag{6}$$

The likelihood function for logistic regression is

$$L(\theta \mid y;x) = \Pr(Y \mid X;\theta) \tag{7}$$

$$= \prod_i \Pr(y_i \mid x_i;\theta) \tag{8}$$

$$= \prod_i h_\theta(x_i)^{y_i}(1-h_\theta(x_i))^{(1-y_i)} \tag{9}$$

*4) **Fitting the model**:* We use the log-likelihood as the cost function in case of logistic regression. It is defined as -

$$n^{-1}\log L(\theta \mid y;x) = n^{-1}\sum_{i=1}^{n}\log \Pr(y_i \mid x_i;\theta) \tag{10}$$

To find the coefficients of the our model, we use optimization methods like gradient descent or Iteratively reweighted least squares (IRLS). In IRLS method, we maximise log-likelihood of a Bernoulli distributed process. We have the following data -

Parameters -

$$\mathbf{w}^T = [\beta_0, \beta_1, \beta_2, \ldots] \tag{11}$$

Explanatory Variables -

$$\mathbf{x}(i) = [1, x_1(i), x_2(i), \ldots]^T \tag{12}$$

Expected value of the Bernoulli distribution -

$$\mu(i) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}(i)}} \tag{13}$$

We can iteratively use the following algorithm -

$$\mathbf{w}_{k+1} = \left(\mathbf{X}^T\mathbf{S}_k\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{S}_k\mathbf{X}\mathbf{w}_k + \mathbf{y} - \boldsymbol{\mu}_k\right) \tag{14}$$

where $\mathbf{S} = \text{diag}(\mu(i)(1-\mu(i)))$ and $\boldsymbol{\mu} = [\mu(1), \mu(2), \ldots]$ and

$$\mathbf{X} = \begin{bmatrix} 1 & x_1(1) & x_2(1) & \ldots \\ 1 & x_1(2) & x_2(2) & \ldots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

$\mathbf{y}(i) = [y(1), y(2), \ldots]^T$

Through this method, we can find the coefficients for our model and fit our data-set.

C. Evaluating the model

In model evaluation, we measure the extent to which the model fits the data.

*1) **Confusion Matrix**:* A confusion matrix is a table that shows the performance of a machine learning model. It displays how many of the model's predictions were right and how many were incorrect and which classes did the model succeed in and which did it fail in. It offers us an understanding of how the model works. If a model fails for a certain class, we may investigate it, figure out why, and try to improve the model.

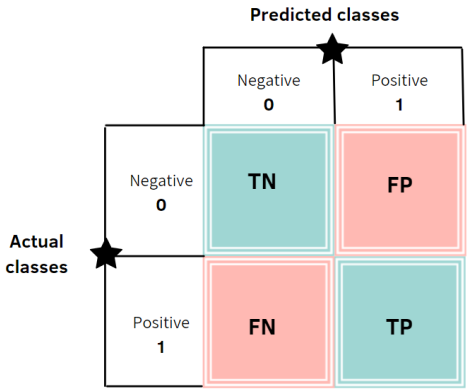

Fig. 2. Sample Confusion Matrix

There are various metrics based on confusion matrix. Some of them are listed below

*2) **Precision and Recall**:* Precision is the ratio between the True Positives and all the Positives. For our model, it will be the measure of persons who we correctly identified as survived out of all the people who survived.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{15}$$

Recall is the measure of our model correctly identifying True Positives. For all the patients who actually survived, recall tells us how many we correctly predicted.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{16}$$

*3) **F1 Score**:* F1 score is the measure which combines both precision and recall. It is calculated as the harmonic mean of precision and recall -

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

D. Multivariate Logistic Regression

In practice, we often have more than one predictor. We extend the simple logistic regression model so that it can directly accommodate multiple predictors. The ordinary expression can be modified as -

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{18}$$

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{19}$$

To estimate the coefficients of multivariate logistic regression, we use the same Maximum Likelihood Estimation methods.

III. The Problem

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In this challenge, we have to build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (ie name, age, gender, socio-economic class, etc).

A. The Variables

Survival - Survival in the crash, 0 = No, 1 = Yes
Pclass - Ticket class of the Passenger, 1 = 1st, 2 = 2nd, 3 = 3rd class.
Sex - Male or Female
Age - Age in years
Sibsp - Number of siblings / spouses aboard the Titanic.
Parch - Number of parents / children aboard the Titanic.
Ticket - Ticket number.
Fare - Passenger fare.
Cabin - Cabin number.
Embarked - Port of Embarkation out of Cherbourg, Queenstown, and Southampton.

B. Cleaning the Data

We check for null values in all columns.(Fig. 3)
We find that the a significant percentage of cabin data missing. So we drop that column. The missing data on age can be imputed.

C. Cleaning and Visualizing the numeric variables

*1) **Age**:* The age distribution has clear trends with Pclass, as visible from the boxplot.

The median values class-wise are-
Class 1 - 37
Class 2 - 29
Class 3 - 24

We impute these values for the missing data.
To visualize the trend, we make a lineplot between Age and the proportion of passengers who survived of that particular age. (Fig. 5 and 6.)

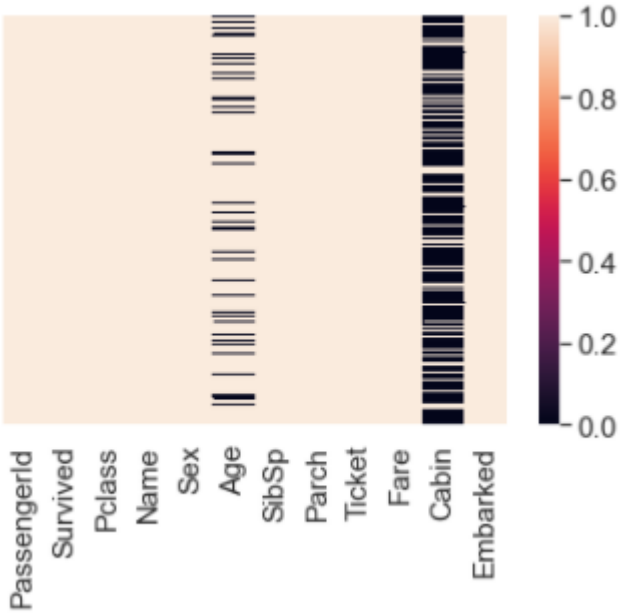

Fig. 3. Null values heatmap

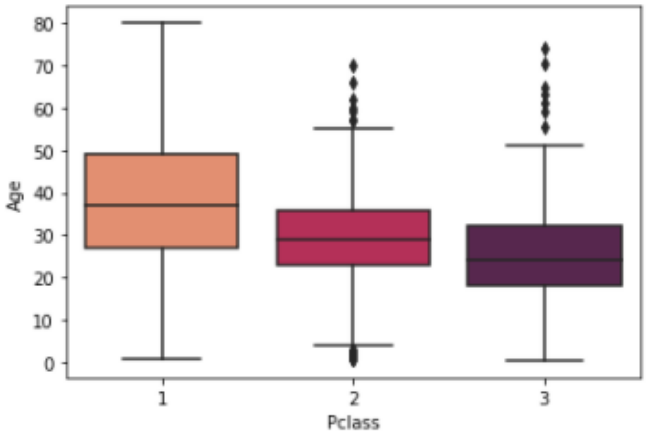

Fig. 4. Age grouped by Pclass

*2) **Fare**:* The fare is well-distributed, with the mean fare being 32.204.

As expected, the fare varies with passenger class, with means being -
Class 1 - 84.15
Class 2 - 20.66
Class 3 - 13.67

*3) **PClass**:* We will check whether there is any relation between the proportion of people survived and their Passenger class by drawing a countplot grouped by class.

The 1st class survived more, as evident from mean of survived values grouped by Pclass -

The median values class-wise are-
Class 1 - 0.63

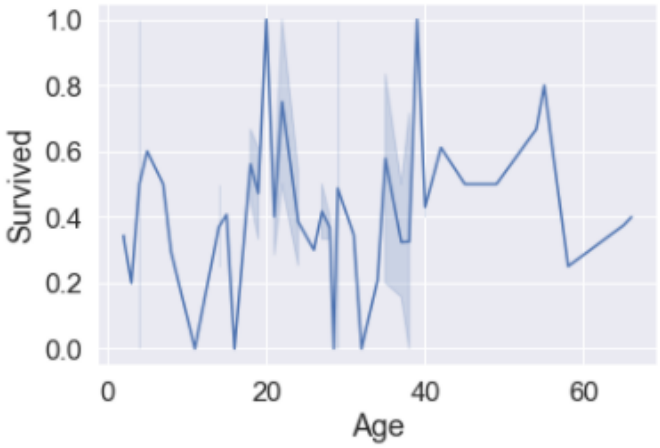

Fig. 5. Survival vs Age

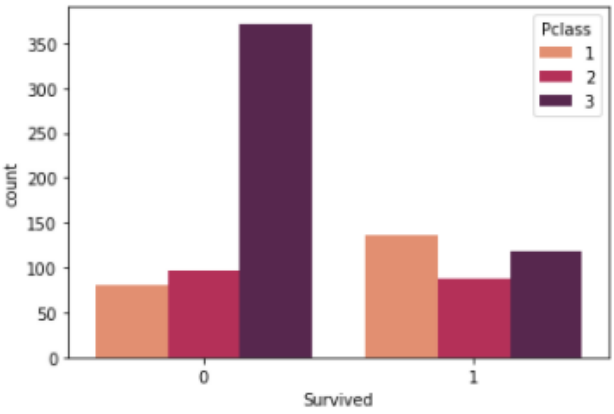

Fig. 6. Survival grouped by Passenger Class

Class 2 - 0.47
Class 3 - 0.24

*4) **Siblings and Parents**:* We can combine the siblings and parents columns to figure out if a passenger is travelling alone or with relatives. Then we plot the data to compare survival based on co-passengers.

It seems that passengers who travelled alone had less survival chance.(Fig. 7)

D. Feature Engineering in Categorical Variables

In this section, we will visualize and create dummies for the categorical variables.

*1) **Sex** :* We check whether survival depends upon gender or not, by making a countplot.

The mean of survived variable is -
Female - 0.742
Male - 0. 188

Clearly, females had a better chance of survival. We convert the categorical features into numeric by creating male and

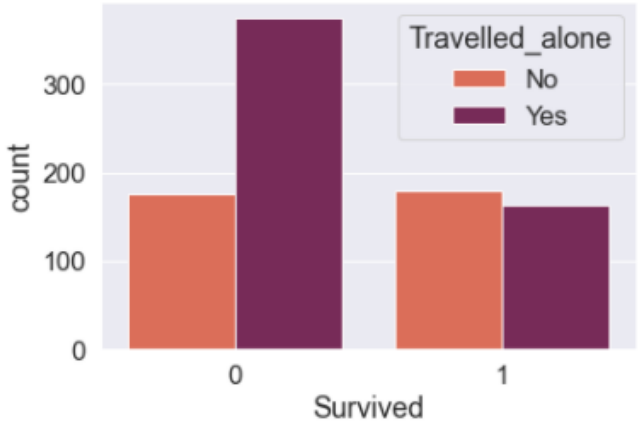

Fig. 7. Survival grouped by number of relatives onboard

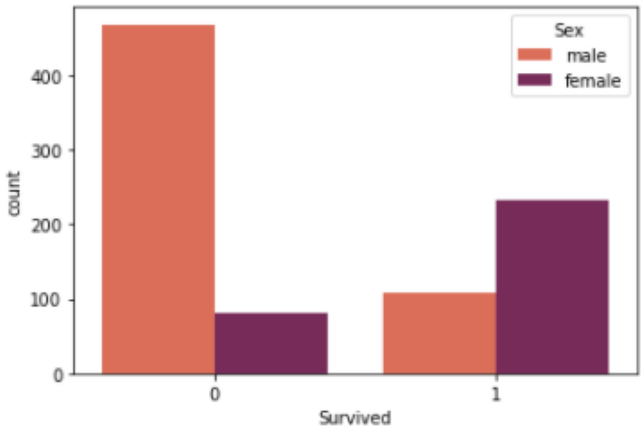

Fig. 8. Survival grouped by Sex

female columns.
(Fig. 8)

*2) **Embarked**:* There are 3 embarkation points- Cherbourg, Queenstown, and Southampton. To check whether the survival depends on embarkation point, we create the following count-plot.
(Fig. 9)

The embarkation pointwise mean of survived variable is -
Southampton - 0.553
Cherbourg - 0.389
Queenston - 0.336

Then we create dummy columns to account for these trends in our model.

*3) **Name and Ticket**:* These variables can be dropped.

E. Checking Correlation

In our final set of variables, we will check if there is any instance of high correlation among them, by making a heatmap.

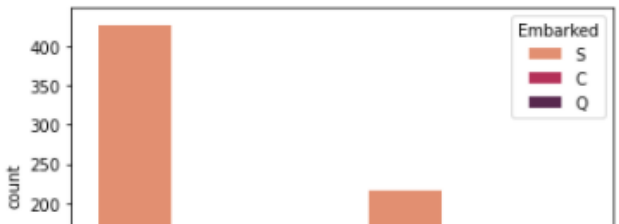

Fig. 9. Embarked Countplot

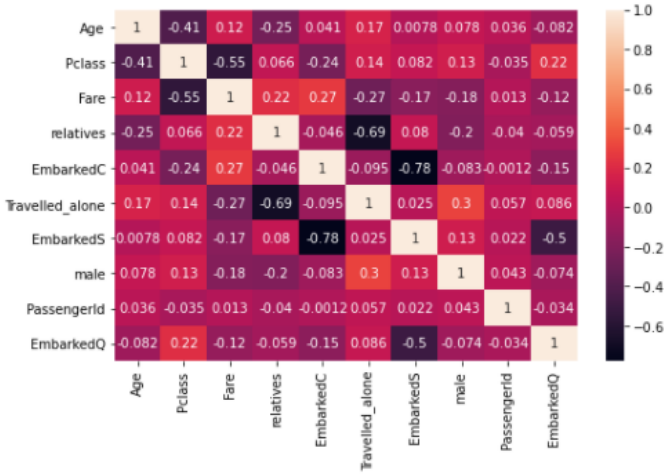

Fig. 10. Correlation Heatmap

No high correlation is witnessed among our variables.(Fig. 10)

F. Scaling the Data

Scaling refers to the process of putting values in the same range or scale such that no one variable dominates the others. Many Machine Learning algorithms use distance between two data points and if the features have varying magnitudes.

We have used scikit-learn's StandardScaler method to scale our data. StandardScaler removes the mean and scales each feature to unit variance. This operation is performed feature-wise in an independent way.

IV. The Model

A. *Train Test Split*

We have the following X and y for our final model, which we divide our data into training and testing data..
X= PClass, Age, SibSp, Parch, Fare, Male, Female, Embarked, Relatives
Y = Survived

B. *Training the Model*

Using scikit-learn's LogisticRegression() function on our X and y variables, we develop a multivariate logistic regression model. The coefficients -

	Coefficient
Pclass	-0.896801
Age	-0.555191
Fare	0.177938
relatives	-0.625110
Travelled_alone	-0.276369
male	-0.622823
female	0.622823
EmbarkedS	-0.244168
EmbarkedC	-0.192774
EmbarkedQ	0.010602

Fig. 11. Coefficients

C. *Predictions and Metrics*

*1) **Training Data**:* The confusion matrix for our training set in shown below (Fig.12)
Based on the matrix, we can calculate the following -

*2) **Testing Data**:* The confusion matrix for our testing set in shown below (Fig.14)
Based on the matrix, we can calculate the following -

V. Conclusions

From our observations, we can conclude the following -

- **Passenger class** - A higher proportion of people who survived were in class 1. The class 3 passengers survived the least.
- **Sex** - More number of female passengers survived in the crash than the male counterparts.

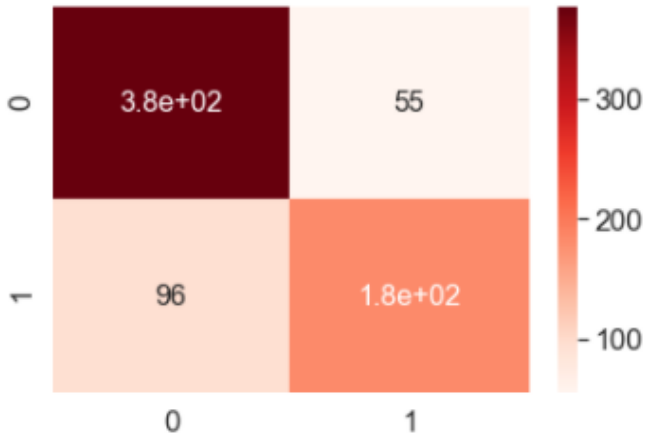

Fig. 12. Confusion Matrix for training data

	precision	recall	f1-score	support
0	0.81	0.87	0.84	432
1	0.77	0.69	0.73	280
accuracy			0.80	712
macro avg	0.79	0.78	0.78	712
weighted avg	0.79	0.80	0.79	712

Fig. 13. Scores for training data

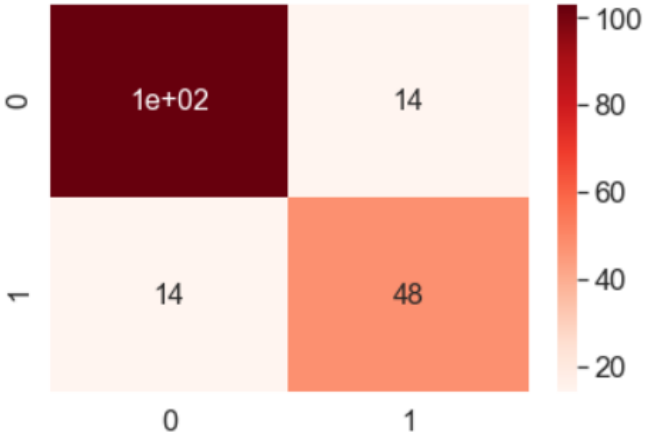

Fig. 14. Confusion Matrix for testing data

	precision	recall	f1-score	support
0	0.88	0.88	0.88	117
1	0.77	0.77	0.77	62
accuracy			0.84	179
macro avg	0.83	0.83	0.83	179
weighted avg	0.84	0.84	0.84	179

Fig. 15. Scores for testing data

- **Embarkation Point** - Out of the 3 embarkation points, Cherbourg, Queenstown, and Southampton, we find that the passengers from Southampton survived in higher proportion, followed by Cherbourg and then Queenstown.
- **Travelling alone** - According to the observations, passengers who travelled alone had lower chances of survival.
- **Age** - From the data, we observed that elder passengers bought class 1 tickets which increased their survival.
- **Fare** - Similar to age, we can report that higher the fare, higher is the passenger class and higher is survival chance.

Thus, survival was not entirely based on luck, there were various factors that contributed to it, as shown by both the qualitative and the quantitative analysis.

REFERENCES

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning: With applications in r. New York, NY: Springer, 2021.

[2] "Logistic regression," Wikipedia, 21-Sep-2021. [Online]. Available: https://en.wikipedia.org/wiki/Logisticregression. [Accessed: 27-Sep-2021].

[3] J. Brownlee, "What is a confusion matrix in machine learning," Machine Learning Mastery, 14-Aug-2020. [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine-learning/. [Accessed: 27-Sep-2021].