DAL ASSIGNMENT 2

Mohamed Afthab Bs
Department of Ocean Engineering
Indian Institute of Technology Madras
NA20B039@smail.iitm.ac.in

***Abstract*—In this analysis, we try to derive insights regarding the effects of the attributes like socio-economic status, age, class etc on the likelihood of survival of passengers onboard the titanic. We would approach this problem by training a logistic regression model on the training dataset available to us. We would then use our model on an unseen test dataset and try to predict the outputs. In this paper, we try to comprehend the mathematics behind logistic regression while applying it to a real world problem.**

I. Introduction

In this Paper, we conduct a logistic regression study Method to determine the relationship between survival rate of passengers on the Titanic based on factors such as age, gender, socioeconomic status, etc. This analysis will give gives us insight into the prejudices that may have existed at the time to determine who is most likely to survive. Logistic regression tries to solve classification problems by generating probabilities for a given set of input characteristics. In binary classification, a threshold is set. We classify items as a single class if the probability turns out to be greater than this value threshold and as another class if the probability is lower than the threshold. The training and testing datasets contain passenger data aboard the RMS Titanic. Data includes information such as name, age, gender, socioeconomic status, social class, etc. also contains data on number of siblings, ticket number, ticket price, and cabin number and port of embarkation. The training data contains one Additional information about surviving passengers disaster. We use this training data to create one the model predicts a passenger's chance of survival during the test data. We study the concepts behind logistic regression. We have next Use this technique to predict survival passengers of a given set of input features.

II. Logistic Regression

Logistic regression is a supervised machine learning algorithm used to solve classification problems for modeling probability of a certain event. Basically, it's better used only for binary classification problems. Single value from a set of predefined values that must be predicted using input feature.

$$y(x) = w^T x + \varepsilon$$

This is the easiest way to model the output. However, for classification, we seek to obtain probabilities that lie between 0 and 1. We use the sigmoid function to do this.

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

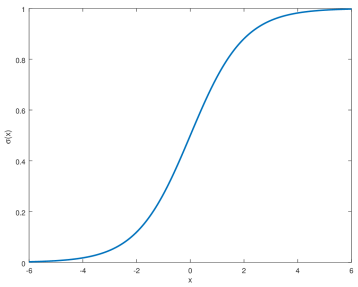

Fig. 1. Sigmoid function

Specifically,

$$p = \sigma(w^T x + \varepsilon)$$

Consider 2 classes, A and B for our binary classification.

$$\hat{y} = \begin{cases} A & \text{if } p < \text{Threshold} \\ B & \text{if } p \geq \text{Threshold} \end{cases}$$

A. Cost Function

Since the sigmoid equation is a nonlinear function, we get many local minima if we use linear regression cost function that minimizes the least square error. Instead, we derive another cost function from maximum likelihood The estimation method is called log loss.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i \cdot \log(\hat{Y}_i) + (1 - y_i) \cdot \log(1 - \hat{Y}_i)\Big)$$

where N represents number of datapoints.

B. Gradient Descent

Once we have determined our cost function, we use the gradient stepwise technique to iteratively adjust the parameters to obtain optimal solution. Slope is the direction in which one Maximum gain function. Go in reverse The direction of the gradient gives the maximum reduction. On each Again and again, we reduce costs by moving in this direction by a certain step size, influenced by the hyperparameter, learning rate α. This iteration continues until the stopping criterion is satisfied. If the learning rate value is too small, we would end up performing too many iterations. On the other hand, a large value would cause our output to oscillate and even diverge from the optimum point.

1) Parameter Initialisation: We initialise the initial parameters of our model to zero but it could be random values as well.

2) Stopping Criteria: For our model to converge to a local minima, we use certain stopping criteria. Some of the conditions we look for are:

- Change in cost function is less than a threshold
- Change in gradient is less than a threshold
- Change in parameters is less than a threshold
- Number of iterations has reached its threshold

III. The Problem

In this section we will perform data analysis using mathematical concepts described above. We will analyze these The impact of all the attributes of the titanic and attempted datasets predict the probability of the passenger surviving the test database.

A. Missing Values

1) Port of Embarkation: From the bar chart in **Fig. 2** we observe that there is a large number of passengers boarding from the port of Southampton. Therefore, we decide to impute all missing values with Southampton.

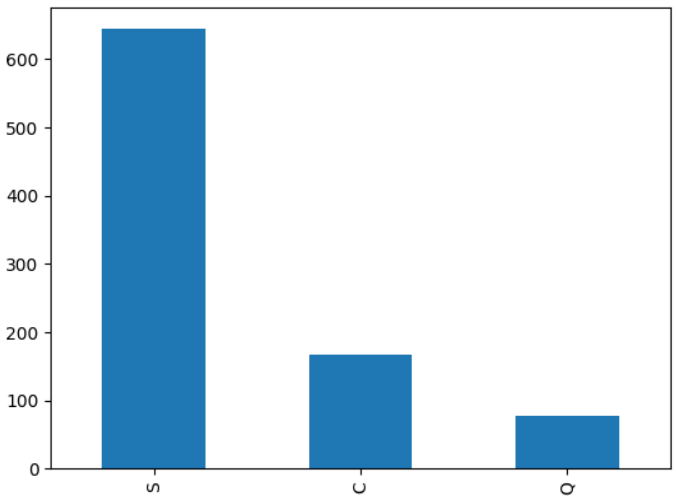

Fig. 2. Port of Embarkation Distribution

2) Age: From **Fig. 3**, we see that age does not follow a uniform distribution. Once we derive the titles of different passengers from the name column, we take mean values of age for each title and impute the missing ages of passengers

B. Exploratory Data Analysis

1) Passenger Class V/S Survival From the bar plot in **Fig. 4**, we can easily make out that first-class passengers were more likely to be saved during the rescue efforts. It seems the third class was the most risky travel choice.

2) Fare V/S Survival From **Fig. 5**, we come to the conclusion that a higher fare would significantly improve chances of survival. We can observe that the portion of people who opted for lower fares and died is extremely high.

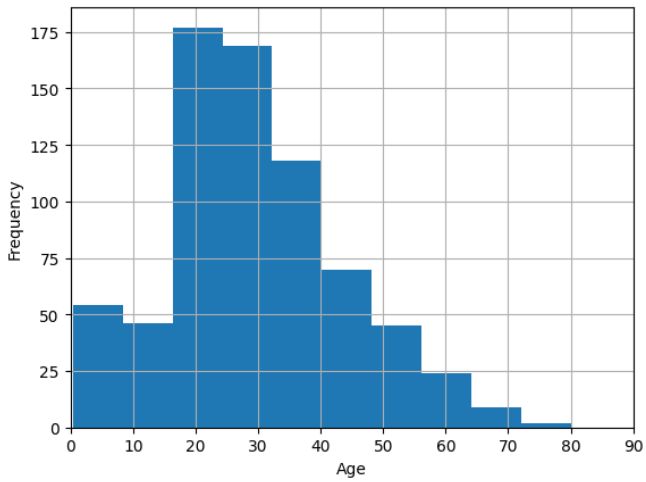

Fig. 3. Age Distribution

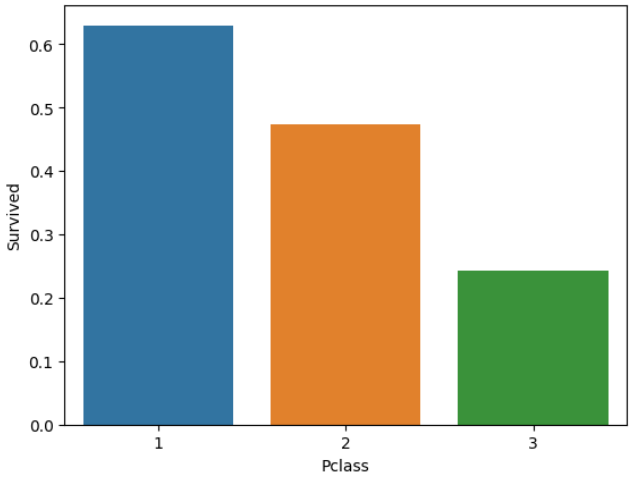

Fig. 4. Passenger Class V/S Survival

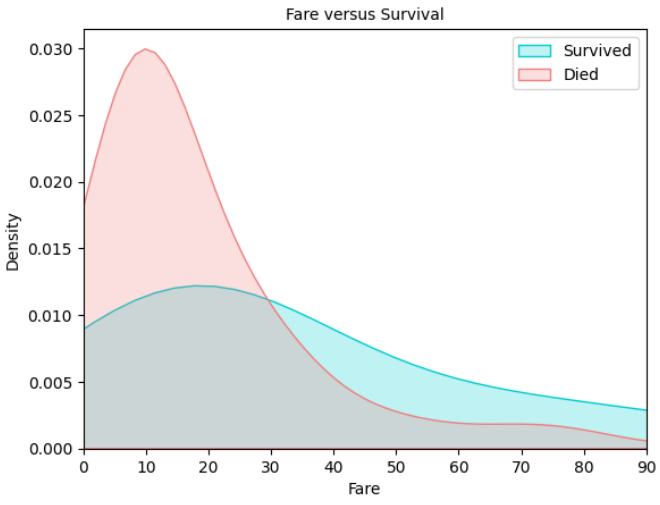

Fig. 5. Fare V/S Survival

3) Port of Embarkation V/S Survival From **Fig. 6**, we understand that people who boarded from Cherbourg, France, had the highest survival rate among the 3 ports of embarkation. This makes sense because the average fare by a Cherbourg passenger was 59.95 dollars (refer **table 1**, which was also the highest among all the three locations. It is safe to assume that the people who boarded from Cherbourg were richer and had higher survival rates.

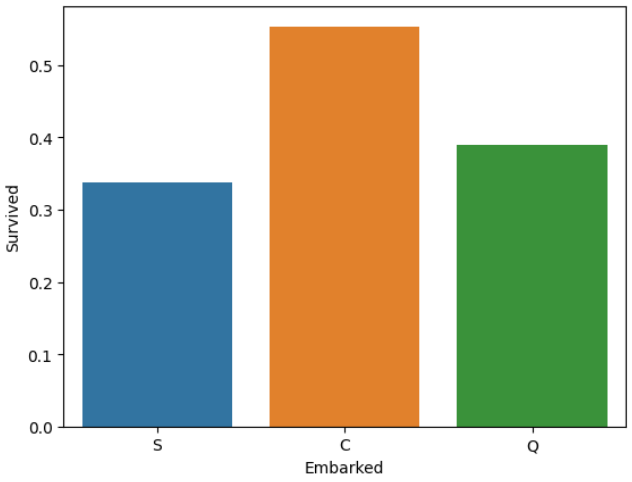

Fig. 6. Port of Embarkation V/S Survival

TABLE I
MEAN FARE AND SURVIVAL RATE OF PASSENGERS

Embarked	Fare	Survived
C	59.954144	0.553571
Q	13.276030	0.389610
S	27.079812	0.336957

4) Gender V/S Survival: From **Fig. 7**, we understand that females had a much greater chance of survival than males.

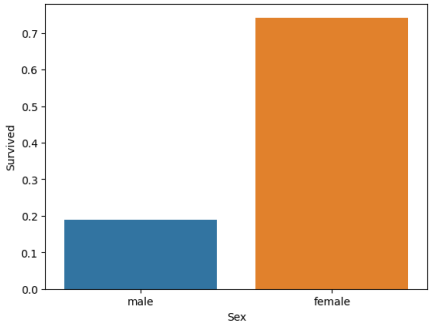

Fig. 7. Gender V/S Surviva

5) Age V/S Survival: From the plot in **Fig. 8**, we can observe that children were given more preference during the rescue operations.

C. *Feature Engineering*

1) Titles: During the time of the tragedy, a person's titles were used to determine the person's rank and status in society.

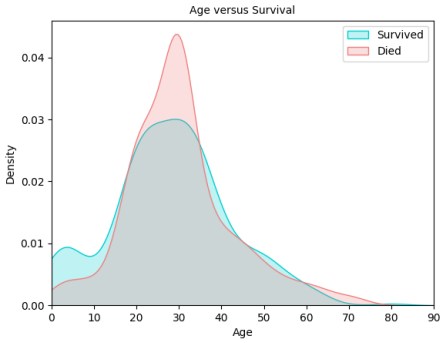

Fig. 8. Age V/S Survival

These factors did come into play in the rescue operations. Therefore, we try to include this aspect in our model as well.

2) Family Size: Use siblings, parents and Counting children, we estimate family size for each passenger. Using this information, we create a separate "Alone" column for know whether passengers are traveling alone or in groups. Our hypothesis is that a group passenger had The chances of survival are higher than those traveling alone. We try integrate this into our model. We remove columns 'SibSp' and 'Parch' because we derive useful information from them. 'Cabin' and 'Tickets' Columns are also removed because they don't add much value our analysis.

D. *Model Fitting*

After data cleaning, pre-ownership and feature engineering, We trained a logistic regression model on the cleaned training dataset using sklearn library. Training accuracy dataset is **81.6%** .To understand our performance better model, we created the confusion matrix. Matrix are presented below in **Table 2**.

TABLE II
TABLE II: CONFUSION MATRIX FOR THE LOGISTIC REGRESSION MODEL

	Prediction	
Label	**Positive**	**Negative**
Positive	476	73
Negative	91	251

IV. Conclusions

From our analysis, we can conclude that status in society, income, ticket class, gender, family role all in one one way or another affected the survival rate of the passengers. The question of which part of the rich population benefits more The benefits and incentives still exist. In this analysis, we applied logistic regression techniques to Real-world data and results demonstrate that logistic regression can also be used on real-world complex datasets.

REFERENCES

[1] Aurelian Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, pp. 112–180.

[2] Christopher M. Bishop, Pattern Recognition and Machine Learning

[3] Joanne Peng, An Introduction to Logistic Regression Analysis and Reporting