DAL 2023 - Assignment 2

Pranoy K P
Engineering Design
Indian Institute of Technology Madras
ed20b045@smail.iitm.ac.in

***Abstract*—In this study, we tackle the challenge of predicting survival outcomes for passengers aboard the RMS Titanic during its ill-fated maiden voyage in 1912. Leveraging the provided datasets, we aim to discern factors influencing passenger survival, such as age, gender, socioeconomic class, etc. With the training dataset providing the ground truth for 891 passengers, we employ machine learning techniques to build a predictive model. This research aims to shed light on the dynamics of survival during maritime disasters and contribute to our understanding of the Titanic tragedy's enduring impact.**

I. INTRODUCTION

In this paper, we employ logistic regression to explore the correlation between the survival rates of Titanic passengers and various factors such as age, gender, socioeconomic status, and more. This analysis aims to provide insights into potential biases that may have influenced survival outcomes during that era.

Logistic regression addresses classification problems by calculating probabilities based on input characteristics. In binary classification, a threshold is established, and items are categorised into one class if the probability exceeds this threshold and into another class if it falls below it.

Our dataset comprises passenger information from the RMS Titanic, encompassing details like name, age, gender, socioeconomic status, social class, and additional data such as the number of siblings, ticket information, cabin number, and port of embarkation. The training dataset includes valuable information about the passengers who survived the disaster, which we use to construct a predictive model for assessing a passenger's likelihood of survival in the test dataset. Throughout the paper, we delve into the fundamental principles of logistic regression and apply this technique to predict the survival probabilities of passengers based on a set of input features.

The paper aims to study the concepts behind logistic regression and leverage it to delve into the Titanic passenger dataset. We then use this technique to predict the passengers' survival probability from the given set of input features. By modelling the survival probabilities, we can better understand the tragic events surrounding the Titanic disaster.

II. LOGISTIC REGRESSION

A. Definition

Logistic regression is a supervised machine learning algorithm used to tackle classification problems. Logistic regression models a relationship between predictor variables and a categorical response variable. The different types of logistic regression, depending on the nature of the categorical response, are:

1) **Binary Logistic Regression:** Used when the response is binary (two possible outcomes).
2) **Nominal/Multiclass Logistic Regression:** Used when three or more classes exist.

The logistic regression model converts the continuous output from the linear regression function into a categorical output by employing a sigmoid function. This sigmoid function, or the logistic function, takes any set of real-valued independent variables as input and maps them to a value from 0 to 1.

$$z(x) = w^T x + b$$

Here, x represents the input matrix, w is the weights or coefficient matrix, and b is the bias term, also known as intercept.

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

Now, we use the sigmoid function where the input will be z and find the probability between 0 and 1.

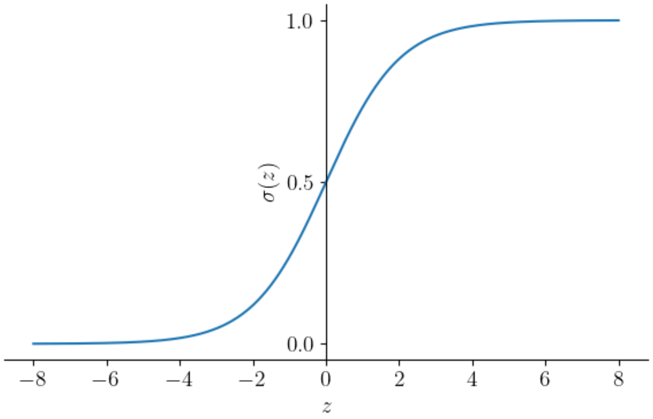

Fig. 1: Sigmoid Function

Consider two classes, A and B, for our binary classification problem. The classification rule is as follows:

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

B. Cost Function

The cost function measures the performance of a machine learning model for a data set. The cost function quantifies the error between predicted and expected values and presents that error as a single real number. The cost function employed is Log Loss or Cross Entropy. It is defined as:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(h_\theta(x_i)) + (1 - y_i)\log(1 - h_\theta(x_i))]$$

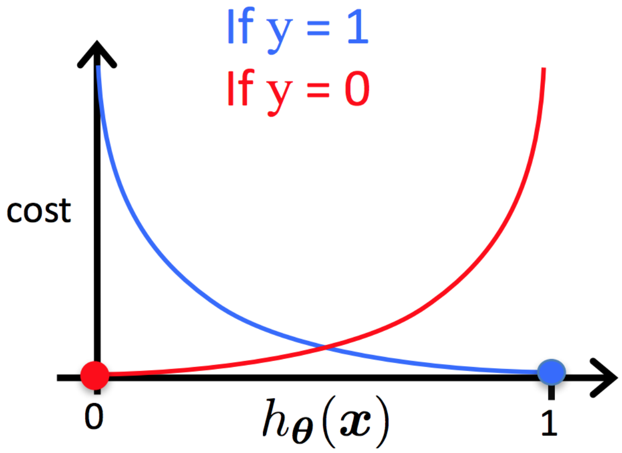

Fig. 2: Cost Function

C. Gradient Descent

Gradient Descent is one of the most commonly used optimisation algorithms to train machine learning models by minimising errors between actual and expected results. The gradient points in the direction of the steepest increase in the function, and moving in the opposite direction leads to the most significant reduction in the cost. During each iteration, we reduce the cost moving along this direction, and the size of each step is determined by the hyperparameter known as the learning rate α. This iterative process continues until a stopping criterion is met.

The choice of the learning rate is crucial. If it's too small, it would require excessive iterations to converge. Similarly, the output can oscillate and diverge from the optimal solution if it's too large. Therefore, finding an appropriate learning rate is crucial to gradient descent optimisation.

III. THE PROBLEM

This section will apply the earlier mathematical principles to conduct our data analysis. Our objective is to assess the influence of all the variables within the Titanic dataset and make an attempt to estimate the likelihood of passengers' survival in the test dataset.

A. Imputation for Missing Values

1) Port of Embarkation: As depicted in the bar chart in Fig. 3, it's evident that the majority of passengers boarded from Southampton. So, we have chosen to fill in all the missing values with Southampton as the port of embarkation.

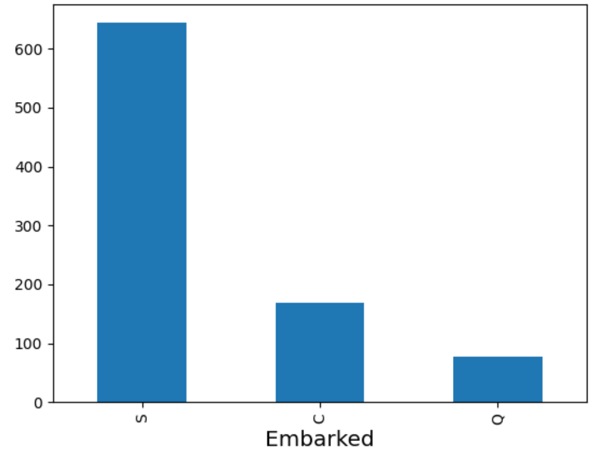

Fig. 3: Port of Embarkation

2) Age: Fig. 4 reveals that the distribution of ages is not uniform. We employ a strategy based on passenger titles extracted from the name column to address missing age values. We compute the mean age for each title and use this information to impute the missing ages for passengers accordingly.

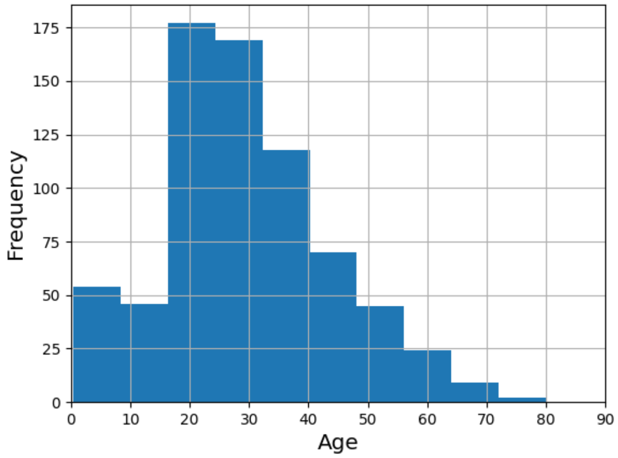

Fig. 4: Age Distribution

B. Exploratory Data Analysis

This is a crucial step in understanding the data before building a logistic regression or any other machine learning model. EDA helps you gain insights into your dataset, identify patterns, detect anomalies, and make informed decisions about data preprocessing and feature selection.

1) Fare vs Survival: Drawing from the information presented in Figure 5, it is evident that a higher fare substantially increased the likelihood of survival. The dataset highlights a significant portion of individuals with lower fares who, unfortunately, did not survive. This might shed light on the hierarchy of social classes. As indicated by the fare class, socioeconomic status played a notable role in determining survival outcomes during the Titanic's unfortunate event.

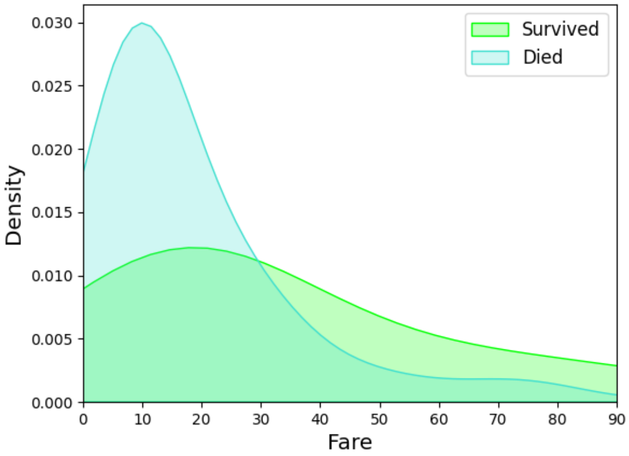

Fig. 5: Fare vs Survival

2) Age vs Survival: The plot in Fig. 6 highlights that children received preferential treatment during the rescue operations. This makes sense because the lives of innocent children would have been put above the lives of adults.

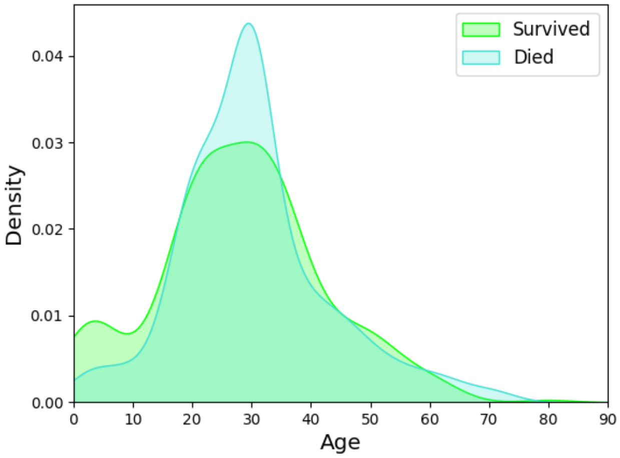

Fig. 6: Age vs Survival

3) Passenger Class vs Survival: By examining the bar plot in Figure 7, it becomes apparent that passengers in the first-class category had a higher probability of surviving the rescue operation. This suggests that choosing the third class was associated with a higher level of risk during the journey.

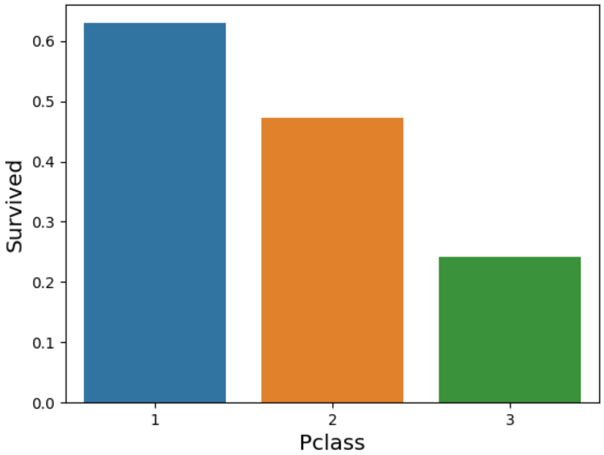

Fig. 7: Passenger Class vs Survival

4) Port of Embarkation vs Survival: As indicated by Figure 8, passengers who embarked from Cherbourg displayed the highest survival rate compared to passengers from the other two ports. This relationship corresponds with the observation that Cherbourg passengers had the highest average fare, amounting to 59.95 (as shown in Table 1). This suggests that the more affluent demographic who boarded from Cherbourg were more likely to survive the tragedy.

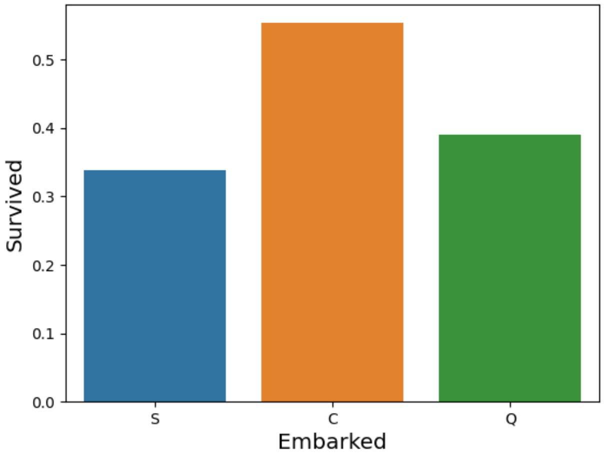

Fig. 8: Port of Embarkation vs Survival

TABLE I: Mean Fare and Survival Rate of Passengers

Embarked	Fare	Survival Rate
C	59.954	0.554
Q	13.276	0.390
S	27.080	0.337

5) Sex vs Survival: As indicated in Fig. 9, females had a considerably higher chance of survival than males. This can be attributed to the historical practice of prioritising women and children during emergencies. This courteous

approach led to a notably higher survival rate among female passengers than their male counterparts.

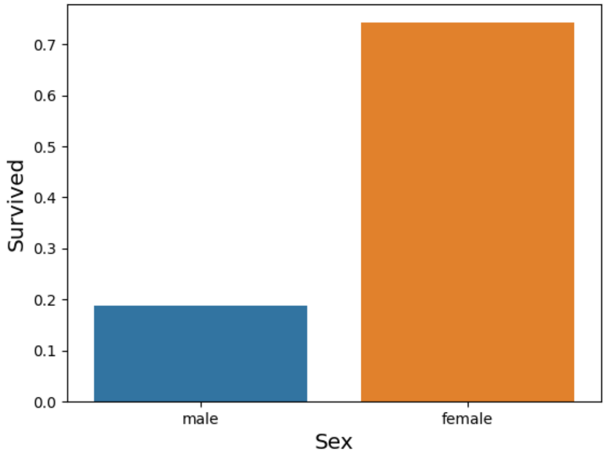

Fig. 9: Sex vs Survival

6) Family size vs Survival: The data suggests that passengers travelling in groups had a notably higher likelihood of survival than those travelling alone.

TABLE II: Family Size vs Survival

Size	Survived
Groups	0.505650
Alone	0.303538

C. Feature Engineering

1) Titles: During the tragedy, a person's titles indicated their rank and societal status, factors that played a role in the rescue efforts. Consequently, we incorporate this aspect into our model.
2) Family Size: Utilising data on the number of siblings, parents, and children, we estimate the family size for each passenger. To account for this, we introduce a separate column labelled 'Alone' to track whether a passenger was travelling alone or with companions. We realise that group travellers are more likely to survive than solo passengers. We incorporate this information into our model.
3) Column Removal: We eliminate the 'SibSp' and 'Parch' columns since we have extracted valuable information from them. Additionally, we drop the 'Cabin' and 'Ticket' columns as they contribute very little to our analysis.

D. Model Fitting

Following data cleaning, preprocessing, and feature engineering, we employed the sklearn library to train a logistic regression model on the refined training dataset. The training accuracy of the model stands at 81.59%. To gain a deeper insight into our model's performance, we constructed a confusion matrix presented in Table 3 above.

TABLE III: Confusion Matrix for Logistic Regression Model

	Positive (Actual)	**Negative (Actual)**
Positive (Prediction)	476	73
Negative (Prediction)	91	251

IV. CONCLUSIONS

In this study, we applied logistic regression techniques to real-world data, and the outcomes illustrate the applicability of logistic regression to complex, real-world datasets. Based on our analysis, it can be inferred that factors such as societal status, income, ticket class, gender, and family roles all influenced passengers' survival rates during the Titanic disaster. This leads us to question how different segments of the affluent population were impacted and whether certain benefits and incentives persisted.

REFERENCES

[1] Aurelian Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, O'Reilly Media, Inc., pp. 33–143

[2] Bishop Christopher M.. 2006. Pattern Recognition and Machine Learning. Springer. pp.179-224.