# Assignment 2
# A Mathematical Essay on Logistic Regression

Pankaj Bhardwaj
*Cyber Physical Systems*
*Indian Institute of Technology Madras*
Chennai, India
bs20b024@smail.iitm.ac.in

***Abstract*—When the widely considered "unsinkable" Titanic sank, unfortunately only about 32were able to survive. In this analysis, we try to analyse if any subset of passengers had a greater chance of survival using logistic regression. We try to identify factors that played a major role in the survival of passengers. .**

***Index Terms*—Logistic regression, maximum likelihood estimation, confusion matrix, accuracy, precision, recall, F1 score, Titanic shipwreck**

## I. Introduction

Logistic regression is a statistical technique utilized for modeling binary classification problems, aiming to predict qualitative responses like Yes or No. It operates by estimating the probability of belonging to the first class. The fundamental mathematical function in logistic regression is the sigmoid function (x), which forms an S-shaped curve, mapping real values to a range between 0 and 1. To classify data points, a decision boundary is established on this sigmoid function by applying a threshold to its value.

In this article, we employ logistic regression to analyze the factors that impacted the survival of Titanic passengers. We posit that variables such as age, gender, and socio-economic status played a role in determining who had access to lifeboats and consequently survived the disaster. Our approach begins with an examination of the mathematical principles underpinning logistic regression. Subsequently, we apply this technique to the Titanic dataset. The process involves data exploration and analysis, model development, and performance comparison using a subset of the data. Ultimately, we employ the best-performing model to make predictions on the remaining data.

## II. Logistic Regression

In this section, we will examine the application of logistic regression for modeling a binary classification problem where the two classes are represented as 0 and 1. We aim to model the probability P(Y = 1 — X), which we will refer to as P(X) henceforth. This probability is estimated using a logistic function, which is defined as follows:

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{1}$$

From equation 1, it becomes apparent that when we have a small value for $0 + 1X$, the corresponding $P(X)$ is nearly 0, while a large value for $0 + 1X$ results in $P(X)$ approaching 1.

### A. Estimating the Regression Coefficients

The optimal regression coefficients can be estimated by maximum likelihood estimation. The likelihood function is given by

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y'_i=0} (1 - p(x'_i)) \tag{2}$$

The coefficients $\beta_0^*$ and $\beta_1^*$ are selected to maximize this likelihood function, and numerical techniques such as gradient descent can be employed for this purpose.

### B. Making Predictions

For every new data point X, a prediction can be made by estimating P(X) as

$$P(X) = \frac{e^{\beta_0^*+\beta_1^* X}}{1 + e^{\beta_0^*+\beta_1^* X}} \tag{3}$$

If this value surpasses a specific threshold (typically 0.5 when using a symmetric loss matrix), the class can be predicted as 1; otherwise, it is predicted as 0, and vice versa.

### C. Multiple Logistic Regression

Multiple logistic regression is employed when, rather than having a single input feature X, we have a vector of input features $X_1, X_2, \ldots, X_p$. In such cases, the model is formulated as follows:

$$P(Y = 1|X_1, X_2, \ldots, X_p) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}$$

Here, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with each of the input features $X_1, X_2, \ldots, X_p$. This model extends logistic regression to handle multiple input features simultaneously.

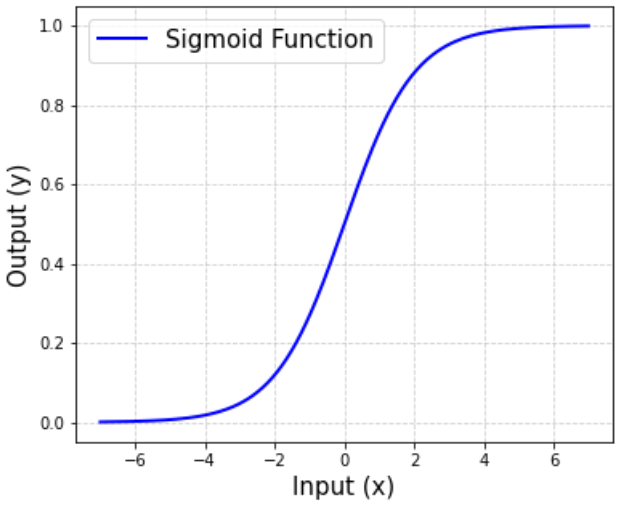


Fig. 1. Sigmoid Function for Logistic Regression

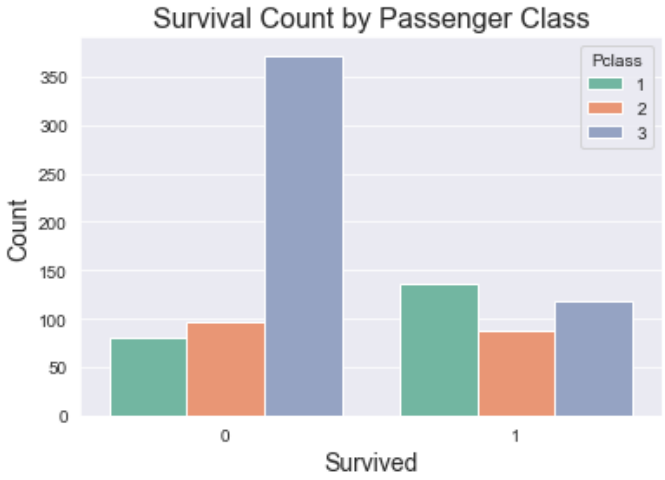


Fig. 2. Passenger Class Vs Survived

## III. Data

We've been given a dataset containing information about Titanic passengers, including details like their age, ticket information, class, family size, cabin assignment, and embarkation location. The data is mostly clean, with a total of 891 entries, representing 40% of the actual number of Titanic passengers. The survival rate in our sample is 38%, slightly higher than the historical incident rate of 32%. The dataset is predominantly composed of male passengers, making up 65% of the sample. Most passengers, over 75%, are traveling without family members. There are very few elderly individuals aged above 65 in the dataset.

The dataset contains both numeric and categorical data. Age and Fare are continuous numeric variables, while SibSp and Parch are discrete numeric variables. On the other hand, sex, Pclass, embarkation location, and survival status are categorical variables. Our target variable of interest is the survival status, categorized as Yes or No.

| Variable Name | Description & Key |
|---|---|
| survival | Survival & 0 = No, 1 = Yes |
| pclass | Ticket Class & 1 = $1^{st}$ Class, 2 = $2^{nd}$ Class, 3 = $3^{rd}$ Class |
| sex | Sex & M/F |
| sibsp | No. of siblings / spouses aboard the Titanic & M/F |
| cabin | Cabin Number |
| sex | Sex & M/F |
| Age | Age in years |
| fare | Passenger fare |
| embarked | Port of Embarkation & C = Cherbourg, Q = Queenstown, S = Southampton |

Due to inconsistencies, errors, and typos, the passenger names column is not considered for analysis. Similarly, the Ticket and Cabin columns, which contain alphanumeric values, are also excluded. It's worth noting that multiple individuals may share the same cabins. Additionally, there are missing values in the Cabin, Age, and Embarked columns.

## IV. The Problem

The objective of this article is to understand if there exists any relationship between passenger details like age, sex, and socio-economic status and the survival rates using the train dataset. The subsequent step would be to try and predict survival of the passengers in the test data set. The steps followed to explore the same are:

- Data cleaning and imputation
- Data visualization and exploratory analysis
- Building a model to fit the data
- Quantifying the performance of the model
- Using the model to predict survival

### A. Visualization

The data we received is mostly clean, but some data preprocessing is required to ensure a stable model training process and improve generalization performance. We have identified missing values that need to be filled with appropriate data. Specifically, in the Age column, approximately 177 entries of the data is missing. Upon examining the Age distribution, we observed that it exhibits a right-skewed pattern.A Heat Map is plotted to check correlation between numerical features

There are a total of 177 null values in the Age column. To handle these missing age values, we utilized an imputation method by examining the dependency of age on the Pclass variable. As we've previously observed that the survival rate in the 3rd class is considerably lower, we started by visualizing this distribution by creating a boxplot that displays the relationship between age and Pclass.

From the boxplot, it was evident that older individuals were predominantly in Pclass 1 and so forth. To impute the missing age values, the following code was employed. Additionally,

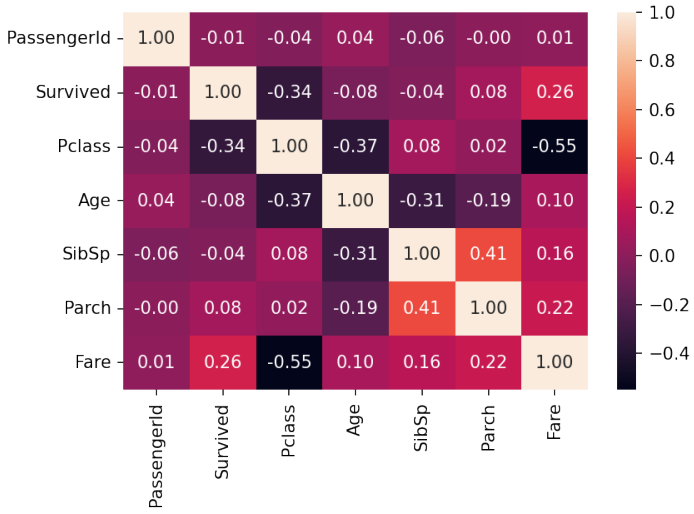


Fig. 3. Sigmoid Function for Logistic Regression

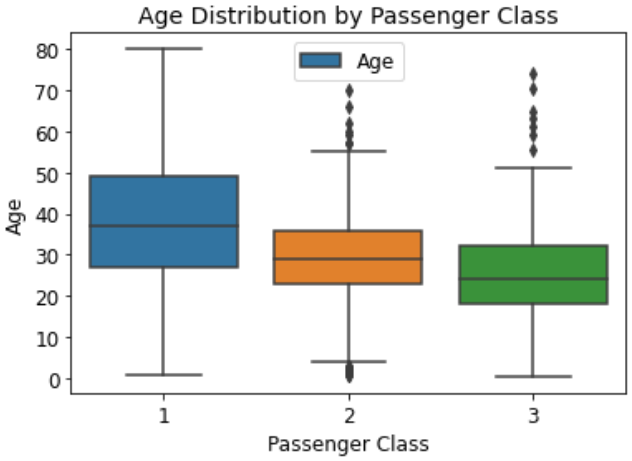


Fig. 4. BoxPlot Showing Age Distribution Over Passenger Class

the medians for each Pclass were determined to be Pclass 1 = 37, Pclass 2 = 29, and PClass 3 = 24.

An if-else structure was created in the past tense to define a function.This function returned the age, with its value being set to the median of the corresponding Pclass.

### B. Applying Logistic Regression

We constructed a predictive logistic regression model using the provided data. After calculating the correlation coefficients for each feature, we observed that Pclass = 1 had the highest positive correlation, while sex = male had the most significant negative correlation.

The purpose of employing logistic regression was to validate our assumptions and decisions regarding feature engineering and completion goals. This validation was accomplished by computing the coefficients of the features within the output function. Positive coefficients were associated with an increased likelihood of survival, while negative coefficients were linked to a decreased likelihood of survival.

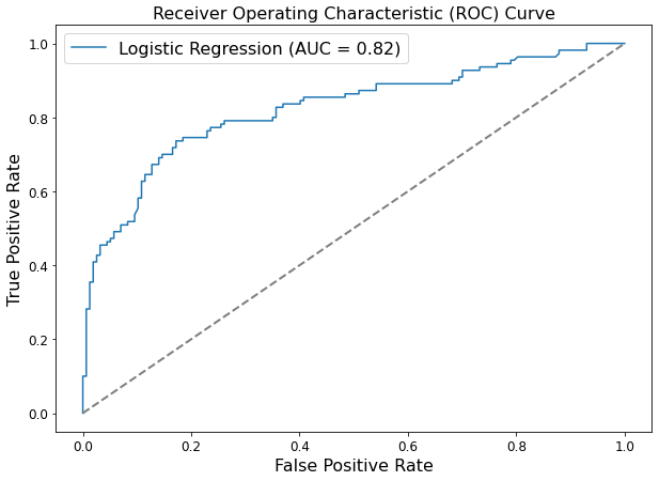


Fig. 5. ROC Curve for Logistic Regression Model

### C. Observations and Insights

In my logistic regression model, I obtained a confusion matrix that provided valuable insights into its performance. The matrix revealed that there were 77 true positive (TP) instances, indicating that the model correctly predicted positive outcomes when they were indeed positive. Additionally, there were 134 true negative (TN) cases, showcasing the model's ability to accurately identify negative outcomes when they were genuinely negative. On the flip side, the model also produced 23 false positive (FP) results, signifying instances where it incorrectly predicted positive outcomes when the data indicated otherwise. Furthermore, there were 33 false negative (FN) instances, representing situations where the model failed to recognize positive outcomes that were present in the data. To visually represent this performance evaluation,A visualization plot of the confusion matrix using sns was made

When we tested our predictive model with a train dataset, we achieved an accuracy rate of 79.02%, indicating that our model correctly predicted 79.02% of the inputs. Additionally, the area under the ROC curve was measured at 82%. Applying the logistic regression model on the given test dataset, it was observed that out of the total dataset, 268 individuals were predicted to have survived, representing approximately 64.1% of the total. whereas 150 individuals were predicted to have not survived, accounting for roughly 35.9% of the total. When we tested our predictive model with a train dataset, we achieved an accuracy rate of 79.02%, indicating that our model correctly predicted 79.02% of the inputs. Additionally, the area under the ROC curve was measured at 82%.

## Conclusion

Using a logistic regression model, it is possible to predict the survivors of the Titanic shipwreck with an accuracy rate of approximately 79%.Specifically, when we control for other

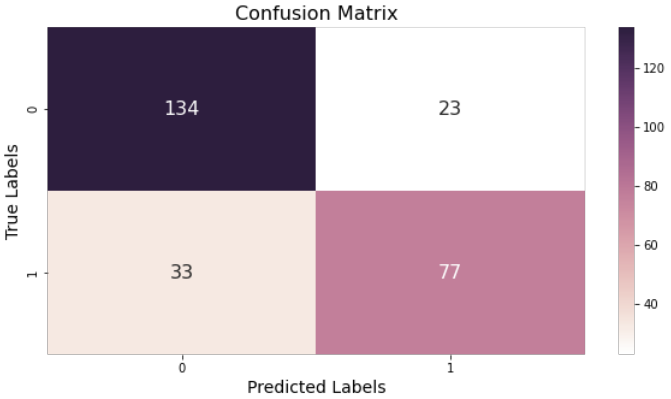


Fig. 6. Confusion Matrix

predictor variables, females exhibit a higher likelihood of survival compared to males. Furthermore, older individuals are less likely to survive, even when we consider other factors. Additionally, individuals from lower social classes also have a decreased likelihood of survival when all other predictor variables are held constant. Conversely, our analysis indicates that females, infants, passengers with high-class tickets, and those traveling with their families had a significantly greater chance of surviving the disaster. These findings support the widely recognized ”women and children first” protocol that was followed during the rescue procedure, which prioritized the safety of women and young passengers.

## ACKNOWLEDGMENT

I would like to thank professor Dr. Gaurav Raina and IIT Madras for providing me this opportunity to work on this project. I would also like to express my gratitude towards my classmates and friends who helped me with this project.

## REFERENCES

[1] A. Thampi, S. Armour, Zhong Fan and D. Kaleshi, ”A logistic regression approach to location classification in OFDMA-based FFR systems,” in 2013 IEEE 14th International Symposium on ”A World of Wireless,Mobile and Multimedia Networks” (WoWMoM), Madrid, Spain, 2013 pp. 1-9

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani,“Classification” in An Introduction to Statistical Learning, New York,Springer, 2013

[3] Komarek, P. (2004). Logistic regression for data mining and highdimensional classification. Carnegie Mellon University.

[4] ”Titanic Data Science Solutions — Kaggle,”