Data Analytics Lab: Assignment-2
A Mathematical Essay on Logistic Regression

Arjav Singh
Metallurgical and Materials Engineering
Indian Institute of Technology Madras
Chennai, India
mm20b007@smail.iitm.ac.in

***Abstract*—This study focuses on predicting what group of people were more likely to survive the most infamous shipwreck in history, the sinking of the RMS Titanic, than others based on Age, Socioeconomic data, and other factors. A logistic regression model is used to model the importance of these factors and predict the probability of individuals surviving.**

***Index Terms*—Introduction, Logistic Regression, Data, Problem, Conclusion**

I. Introduction

Classification is the process of mapping features to different data classes or categories based on the learning from the input data. Logistic Regression is on the Machine Learning algorithms under the Supervised Learning technique, used for predicting the categorical dependent variable using the set of independent variables. It is a significant Machine Learning algorithm because it can provide probabilities between 0 and 1 and classify new data using continuous or discrete datasets by choosing a cutoff value and classifying the outputs with probability greater than the cutoff as one class, below the cutoff as other.

In the given problem, Logistic Regression is utilized to classify people into groups based on age, socioeconomic data, and other factors. The objective is to determine which group is most likely to survive the RMS Titanic's tragic sinking.

II. Logistic Regression

In the field of statistics, the logistic or logit model serves as a statistical framework utilized for the characterization of event likelihood. It achieves this by expressing the logarithm of the odds for the event as a linear amalgamation of one or more independent variables. Within the scope of regression analysis, logistic regression, commonly called logit regression, finds application in estimating coefficients within the linear combination that defines a logistic model. In binary logistic regression, a singular binary dependent variable is present, depicted by an indicator variable, wherein the two possible values are denoted as ”0” and ”1.” Meanwhile, the independent variables can be binary (comprising two classes represented by indicator variables) or continuous (entailing real numerical values).

A. Formulation

A standard logistic function is a sigmoid function, which takes real input, t, and outputs a value between 0 and 1. This is interpreted as taking input log odds and having output probability for the logit. The standard logistic function $\sigma : \mathbb{R} \to (0, 1)$ is defined as follows:

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

Let us assume that t is a linear function of a single explanatory variable x. We can then express t as follows:

$$t = \beta_0 + \beta_1 x$$

And the general logistic function $p : \mathbb{R} \to (0, 1)$ can now be written as:

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

In the logistic model, the probability denoted as $p(x)$ is interpreted as the likelihood of the dependent variable Y being a success or case rather than a failure or non-case. It becomes evident that the response variables Y_i exhibit variability in their distribution; that is, $P(Y_i = 1 \mid X)$ differs from one data point X_i to another while maintaining independence given the design matrix X and shared parameters β.

One can now introduce the logit function, represented as g, which serves as the inverse of the standard logistic function, σ. The logit function's properties are evident through the following relationships:

$$g(p(x)) = \sigma^{-1}(p(x)) = \operatorname{logit}(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right)$$
$$= \beta_0 + \beta_1 x$$

Equivalently, after applying the exponential function to both sides, we obtain the odds:

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

In the case where t is a linear combination of multiple explanatory variables, we can express t as:

$$t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$$

When this is used in the equation relating the log odds of the success to the values of the predictors, the linear regression converts to multiple regression with m explanatory variables, the parameters β_j for all $j = 0, 1, 2, \ldots, m$ are all estimated. Again the more versatile equation is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

and

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m)}}$$

where usually $b = e$.

B. Parameter Estimation

Consider a generalized linear model function parameterized by θ,

$$h_\theta(X) = \frac{1}{1 + e^{-\theta^T X}} = \mathrm{P}(Y = 1|X;\theta)$$

Therefore,

$$\mathrm{P}(Y = 0|X;\theta) = 1 - h_\theta(X)$$

and since $Y \in \{0, 1\}$, we see that $\mathrm{P}(y|X;\theta)$ is given by

$$\mathrm{P}(y|X;\theta) = h_\theta(X)^y (1 - h_\theta(X))^{1-y}$$

We now calculate the likelihood function assuming that all the observations in the sample are independently Bernoulli distributed,

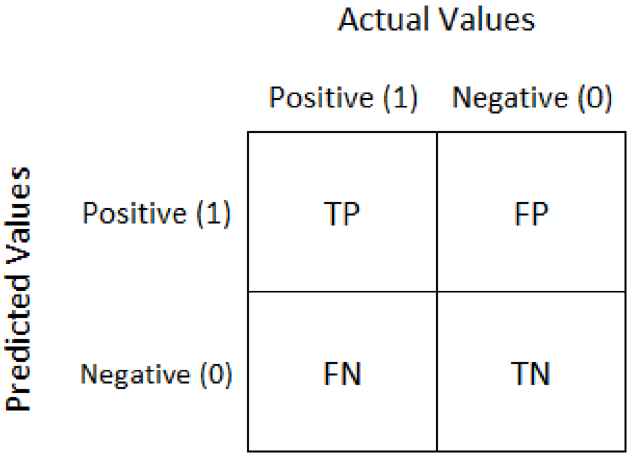
$$L(\theta|y;x) = \mathrm{P}(Y|X;\theta) = \prod_{i=1}^{N} \mathrm{P}(y_i|x_i;\theta)$$

$$L(\theta|y;x) = \prod_{i=1}^{N} h_\theta(x_i)^{y_i} (1 - h_\theta(x_i))^{1-y_i}$$

Typically, the log-likelihood is maximized,

$$\log L(\theta|y;x) = \sum_{i=1}^{N} \log \mathrm{P}(y_i|x_i;\theta)$$

Which is maximized using optimization techniques such as gradient descent to get the parameters' values.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 1. Confusion Matrix.

C. Metrics for model evaluation

1) **Confusion Matrix**: It is used to summarize the performance of a classification algorithm on a set of test data for which the true values are previously known. Sometimes it is also called an error matrix. Terminologies of the Confusion matrix (Figure 1) are:
   - **True Positive**: TP means the model predicted yes, and the actual answer is also yes.
   - **True negative**: TN means the model predicted no, and the actual answer is also no.
   - **False positive**: FP means the model predicted yes, but the actual answer is no.
   - **False negative**: FN means the model predicted no, but the actual answer is yes.

   The rates calculated using the Confusion Matrix are:

   a) **Accuracy**: (TP+TN/Total) tells about overall how classifier Is correct.
   b) **True positive rate**: TP/(actual yes) it says about how much time yes is predicted correctly. It is also called "sensitivity" or "recall."
   c) **False positive rate**: FP/(actual number) says how much time yes is predicted when the actual answer is no.
   d) **True negative rate**: TN/(actual number) says how much time no is predicted correctly, and the actual answer is also no. It is also known as "specificity."
   e) **Misclassification rate**: (FP+FN)/(Total) It is also known as the error rate and tells about how often our model is wrong.
   f) **Precision**: (TP/ (predicted yes)) If it predicts yes, then how often is it correct.
   g) **Prevalence**: (actual yes /total) how often yes condition actually occurs.

2) **ROC curve (Receiver Operating Characteristic)**: The Receiver Operating Characteristic (ROC) curve is a

useful tool for assessing a model's performance by examining the trade-offs between its True Positive (TP) rate, also known as sensitivity, and its False Negative (FN) rate, which is the complement of specificity. This curve visually represents these two parameters.

The Area Under the Curve (AUC) metric to summarize the ROC curve concisely. The AUC quantifies the area under the ROC curve. In simpler terms, it measures how well the model can distinguish between positive and negative cases. A higher AUC indicates better classifier performance.

In essence, AUC categorizes model performance as follows:

- If AUC = 1, the classifier correctly distinguishes between all the Positive and Negative class points.
- If 0.5¡ AUC ¡ 1, the classifier will distinguish the positive class value from the negative one because it finds more TP and TN than FP and FN.
- If AUC = 0.5, the classifier cannot distinguish between positive and negative values.
- If AUC =0, the classifier predicts all positive as negative and negative as positive.

III. Problem

The problem at hand is centered around examining a hypothesis positing differential survival rates among distinct groups of individuals during the tragic sinking of the RMS Titanic. Logistic regression will be employed to investigate this hypothesis, incorporating various features such as age, socioeconomic indicators, and other relevant factors for analysis.

A. *Exploratory Data Analysis and Feature Generation*

The training dataset used in this study consists of 891 passengers and 12 features. Interpretation of the features is as follows:

- Survival: 0 if the passenger did not survive, 1 if the passenger survived.
- Pclass: Class of the ticket - 1st, 2nd, 3rd
- Sex: Gender of the passenger - male, female.
- Sibsp: Number of Siblings/Spouses.
- Parch: Number of parents/children.
- Ticket Number.
- Fare: Fare paid for the ticket.
- Cabin: Cabin Number.
- Embarked: Port of Embarkment - C, Q, S.

The initial step in the analysis involved the assessment of the survival rate following the shipwreck. As indicated in Figure 2, the data illustrates that a mere 38.4% of individuals managed to survive. Subsequently, an examination was conducted to ascertain the relationship between all features and the 'Survived' variable to categorize individuals into groups demonstrating a higher likelihood of survival.

The analysis commenced with an exploration of the influence of gender on survival rates. As depicted in Figure 3, a conspicuous pattern emerges, highlighting that females

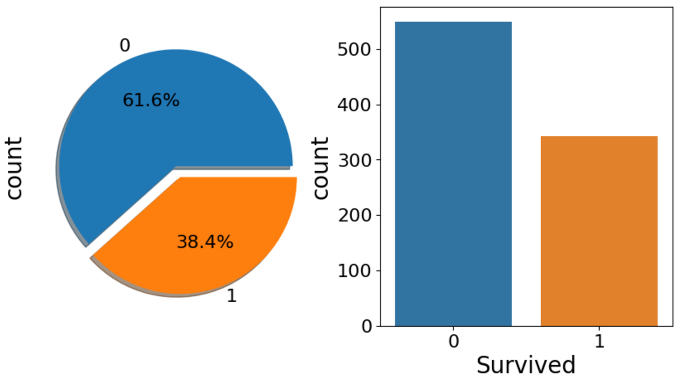


Fig. 2. Percentage of People Survived.

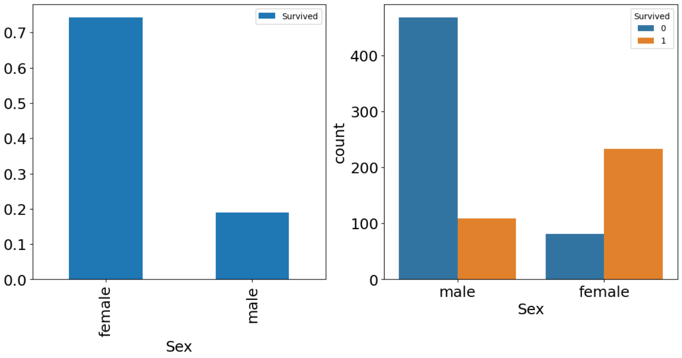


Fig. 3. Relation of Survival with Sex of the passengers.

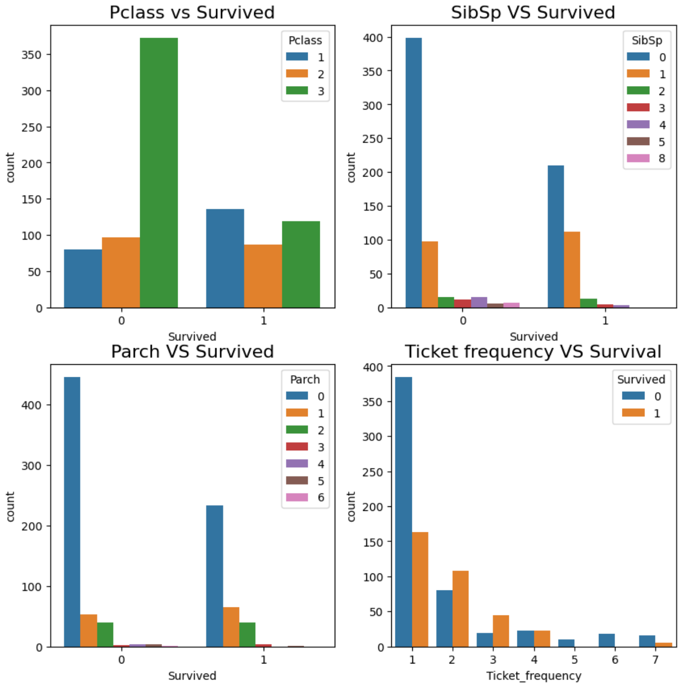


Fig. 4. Relation of Survival with different features.

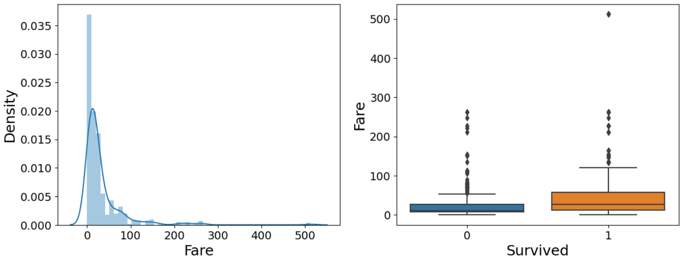


Fig. 5. Relation of Survival with Fares.

exhibited a notably higher likelihood of survival than males. Approximately 70% of females survived, whereas the survival rate among males was markedly lower, at only 20%.

Subsequently, the remaining features were explored, as depicted in Figure 4. The analysis revealed that individuals in passenger class 3 exhibited the lowest likelihood of survival, whereas those in passenger classes 1 and 2 had better survival rates. This observation finds support in the fare distribution, with passengers who paid higher fares being more inclined to survive, as illustrated in Figure 5.

Likewise, passengers who traveled with a solitary companion, forming groups of 2, were more likely to survive than other group sizes. A similar pattern was observed in ticket frequency and family size (created by merging the 'SibSp' and 'Parch' features), as depicted in Figure 6.

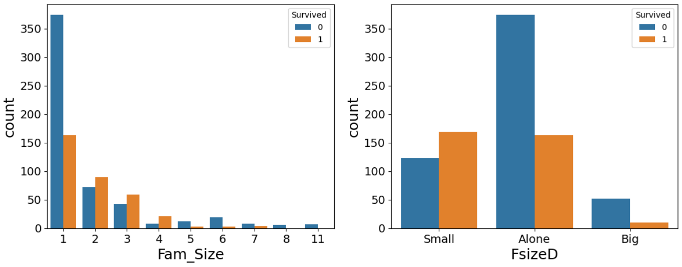


Fig. 6. Relation of Survival with Family Size and Family Type.

A noteworthy observation pertains to missing data within certain variables, as illustrated in Figure 7. Among the 12 features under consideration, it is observed that 'Age' exhibits a notable absence of data, accounting for approximately 19.8%. Similarly, the 'Cabin' feature presents a substantial proportion of missing data, encompassing approximately 77.1%. Moreover, the' Embarked' variable contains only two missing data points.

The Pearson Correlation coefficients between various features were examined to address missing data points. It was observed that the 'Age' feature exhibited the highest absolute correlation with 'Pclass' and 'Sex.' Consequently, a novel feature, denoted as 'title,' was created. Subsequently, the missing values within the 'Age' feature were imputed using the mean values derived from grouping the 'Pclass,' 'title,' and 'Sex' features. The transformation of the 'Age' feature through this engineering process is illustrated in Figure 8 and

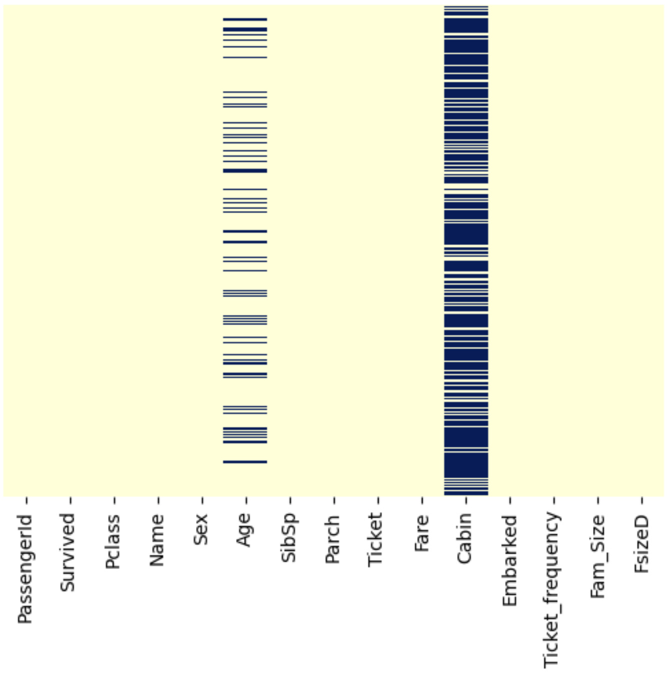


Fig. 7. Heatmap of Missing data.

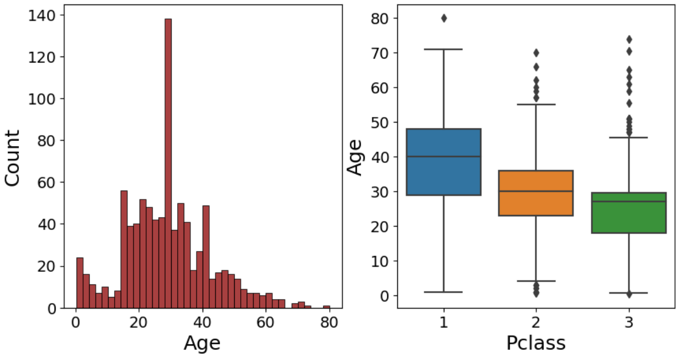


Fig. 8. Age data distribution and Age vs Pclass before data engineering.

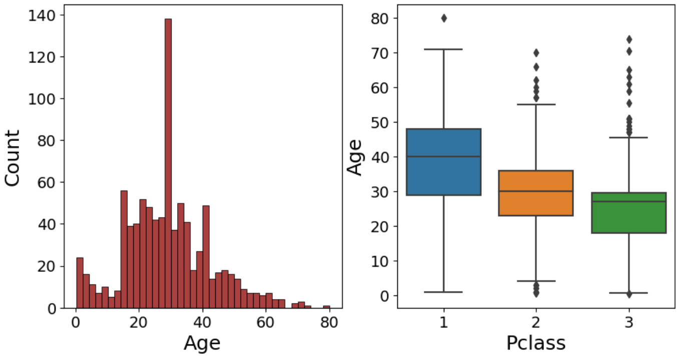


Fig. 9. Age data distribution and Age vs Pclass after data engineering.

Figure 9, depicting its distribution before and after the feature engineering steps.

In the context of the ’Cabin’ feature, a new category labeled as ’Z’ was introduced to address missing data points, while the outlier category ’T’ was replaced with class ’A.’ The resultant distribution of passenger class and survival likelihood concerning cabin data, following the feature engineering procedures, is illustrated in Figure 10.

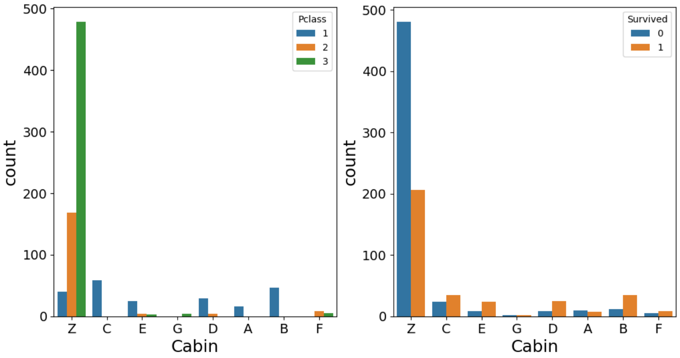


Fig. 10. Passenger class vs Cabin data and Survival chance vs Cabin data after data engineering.

Further analysis of the cabin data to the ’Pclass’ and ’survived’ features revealed notable insights. It was observed that a significant proportion of individuals assigned to the ’Z’ cabin class predominantly belonged to Pclass 3 and exhibited lower chances of survival. Conversely, individuals from other cabin classes were primarily associated with Pclass 1 and demonstrated a higher likelihood of survival. Among the specific cabin classes ’A,’ ’B,’ ’C,’ ’D,’ ’E,’ and ’F,’ it was discerned that class ’A’ exhibited the lowest survival rates, while the other cabin classes displayed comparatively better survival probabilities. This finding supports the previous intuition that ’Pclass’ may contribute to increased chances of survival, as inferred from the analysis of the cabin data.

The ’Embarked’ feature comprises three distinct categories, namely ’S,’ ’C,’ and ’Q,’ as depicted in Figure 11. Notably, most passengers boarded from the ’S’ class, with a significant portion belonging to ’Pclass 3.’ In contrast, passengers from ’C’ seem to have a higher survival rate. This observation may be attributed to the successful rescuing of all passengers from ’Pclass 1’ and ’Pclass 2’ who boarded from this port. The ’Embark S’ location is the primary departure point for most affluent passengers. Nevertheless, the survival prospects for passengers embarking from ’S’ are relatively low, primarily due to the unfortunate fate of a substantial portion of ’Pclass 3’ passengers, with approximately 81% of them not surviving. Port ’Q’ had nearly 95% of its passengers originating from ’Pclass 3.’

In the concluding Exploratory Data Analysis phase, the ’Age’ variable is discretized into bins of width 5. Likewise, the ’Fare’ variable is discretized into four bins. This binning process transforms these continuous variables into categorical representations, which the machine learning model can more effectively process. Additionally, categorical variables are encoded into dummy variables, imparting meaningful numerical values to these categorical attributes.

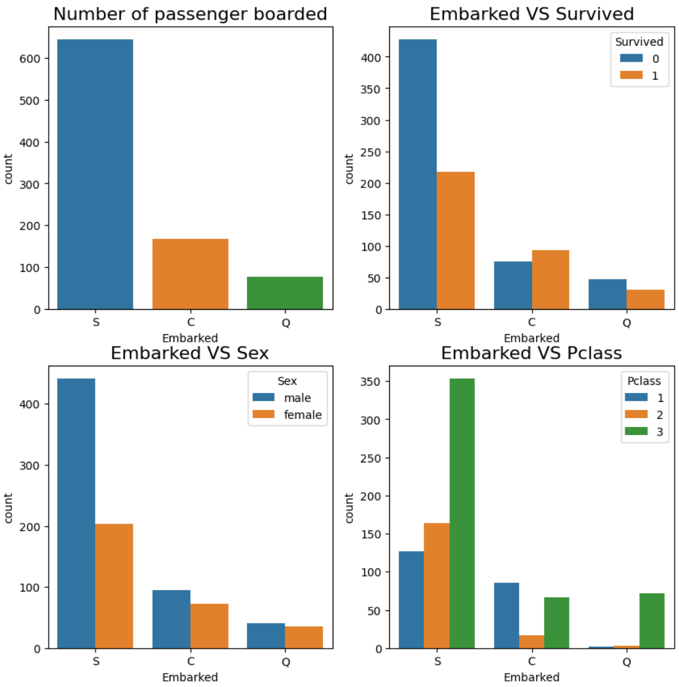


Fig. 11. Embarked feature analysis.

B. Statistical Logistic Regression Modelling

The Sklearn library in Python is employed to implement Logistic Regression after preparing all independent variables. Subsequently, the data is split into train and validation sets with a ratio of 10:3.

After fitting the model, the analysis revealed that the model’s accuracy is 80.22%. The rest of the values are available in the table below.

Class	Precision	Recall	F1-score	Support
0 (Not Survived)	0.79	0.90	0.84	154
1 (Survived)	0.83	0.68	0.74	114
Accuracy	0.80 (Total: 268)			

IV. Implementing the Model to the Test Data

The provided test data have all the features of the train dataset. The methodology is followed for Exploratory Data Analysis, from data pre-processing to missing data handling for features including ’Age’ and ’Cabin.’ The transformation of the ’Age’ feature through this engineering process is illustrated in Figure 12 and Figure 13, depicting its distribution before and after the feature engineering steps.

The model has effectively made predictions regarding the survival likelihood of all passengers, utilizing the parameters on which it was initially trained. This predictive performance is illustrated in Figure 14, which presents the percentage of

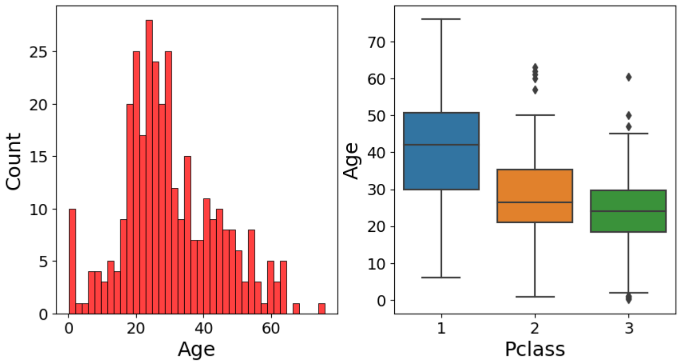


Fig. 12. Test Age data distribution and Age vs Pclass before data engineering.

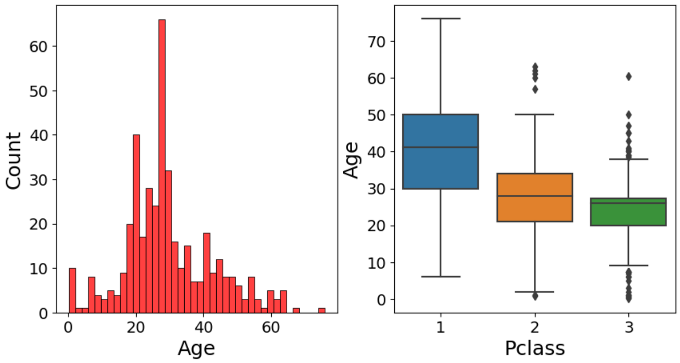


Fig. 13. Test Age data distribution and Age vs Pclass after data engineering.

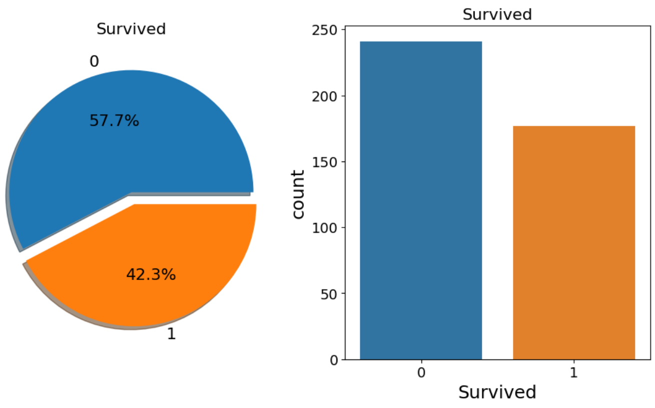


Fig. 14. Predicted percentage of survivors from the test data.

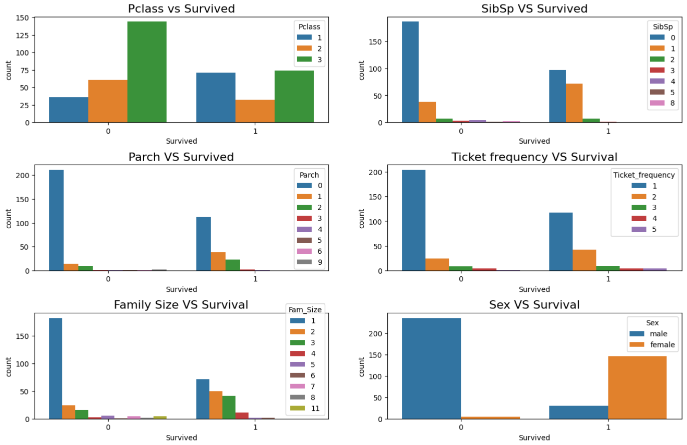


Fig. 15. Features vs Predicted survival chances.

survivors within the test data. Additionally, Figure 15 visually portrays the distribution of features as a barplot, highlighting their correlation with the predicted survival chances.

V. Conclusion

In this study, it is observed that among the passengers on the RMS Titanic, females exhibited a higher likelihood of survival when compared to males. Additionally, passengers occupying Pclass 1, which corresponds to a higher ticket price, demonstrated a greater survival probability than those in other passenger classes. Moreover, younger individuals with family sizes ranging from 2 to 4 members were found to have an increased likelihood of survival, likely attributed to assistance from family members and greater physical agility.

When applied to the test data set, the current model provided similar insights as mentioned above, with an accuracy of 80%, and we observed a similar behavior between features and the survival chance. The study suggests that in the future, the application of non-linear models may offer improved insights into modeling survival rates.

References

[1] JavaTpoint, “Logistic Regression in Machine Learning,” JavaTpoint, [Online]. Available: https://www.javatpoint.com/logistic-regression-in-machine-learning.

[2] ”Logistic Regression - Model Fitting,” Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression#Model_fitting.

[3] ”Performance Measurement in Logistic Regression,” Medium, [Online]. Available: https://meettank29067.medium.com/performance-measurement-in-logistic-regression-8c9109b25278.

[4] “Kaggle Titanic Competition: Missing Values.” *Python in Plain English*, Available: https://python.plainenglish.io/kaggle-titanic-competition-missing-values-f3280267b361.

[5] “Kaggle Titanic Challenge: Create Them Features.” *Python in Plain English*, Available: https://python.plainenglish.io/kaggle-titanic-challenge-create-them-features-a324ba577812. [Accessed: September 14, 2023].