Exploring Data Insights through Logistic Regression Analysis: A Comprehensive Study on Titanic Dataset

Boneshwar V K
dept. of Biotechnology - IDDD Cyber Physical Sytems
Indian Instituteof Technology, Madras
Chennai, India
bs20b012@smail.iitm.ac.in

***Abstract*—The Titanic disaster, which occurred a century ago on April 15, 1912, resulted in the tragic loss of approximately 1500 passengers and crew members. This event continues to intrigue researchers and analysts, prompting them to delve into the factors that influenced the survival of some individuals while leading to the demise of others. The dataset includes features such as age, sex, passenger class, and fare paid, as well as whether or not the passenger survived the sinking. Its objective is to ascertain the relationship between variables such as age, gender, passenger class, and fare, and their impact on the likelihood of passenger survival. It is uncertain whether these variables directly influenced survival rates. This paper explores the application of logistic regression in predicting passenger survival. Specifically, the research compares the correlation of the data columns and describes accuracy percentages on the dataset.**

***Index Terms*—Logistic regression, Exploratory Data Analysis, Titanic**

I. Introduction

In the field of machine learning, analysts have harnessed historical data and events to extract valuable insights. Among these events, the Titanic disaster is a widely recognized maritime tragedy. The Titanic, a British ocean liner, met its tragic end in the North Atlantic Ocean following a collision with an iceberg. While established facts exist regarding the disaster's cause, various conjectures surround the survival rates of its passengers. Over time, data pertaining to both survivors and casualties has been systematically collected and made publicly accessible via the Kaggle.com platform [1].

This dataset has undergone rigorous examination and analysis, employing diverse machine learning algorithms such as Random Forest and Support Vector Machines (SVM). Multiple programming languages and tools, including Weka, Python, R, and Java, have been applied for algorithm implementation. The research paper primarily centers on using Python for executing Logistic Regression algorithm.

The primary objective of this research is to conduct a comprehensive analysis of the Titanic disaster, aiming to establish correlations between passenger survival and their respective characteristics through the application of logistic regression. Specifically, this study evaluates the performance of these algorithms based on their accuracy percentages on a designated test dataset.

II. Logistic Regression

Logistic Regression is a fundamental statistical method utilized in predictive modeling and classification tasks. It is particularly well-suited for scenarios where the target variable is binary, representing outcomes such as survival or non-survival in the case of the Titanic dataset. Unlike linear regression, Logistic Regression models the probability that a given instance belongs to a particular category. The core principle involves fitting a logistic function to the data, effectively mapping the input features to a range between 0 and 1. For the single variable logistic regression, we use the following equation,

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

In the context of the Titanic dataset, Logistic Regression enables us to discern the influence of various factors like fare, passenger class, boarding place, and gender on the likelihood of survival. By estimating the probability of survival for each passenger, we gain valuable insights into the interplay between these features and the ultimate outcome. This method empowers us to make informed inferences and predictions regarding passenger survival. Usually we perform multinomial logistic regression where there are several important classes to be considered while performing regression. Here, n is the number of classes or number of features.

$$\log\left(\frac{P(Y=k)}{P(Y=K)}\right) = \beta_{0k} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Furthermore, Logistic Regression provides interpretable coefficients, offering a clear understanding of the impact each feature has on the odds of survival. This attribute proves invaluable for uncovering the underlying dynamics within the dataset.

In summary, Logistic Regression forms a pivotal component of our analytical framework for the Titanic dataset. It allows us to model the probability of survival, providing a comprehensive understanding of the influential factors and their contribution to the passengers' outcomes. This methodology lays the foundation for robust predictive modeling and insightful analysis in this study.

III. Data

This section provides an overview of the dataset, focusing on one of the most notorious maritime disasters in history—the sinking of the RMS Titanic. On April 15, 1912, during its maiden voyage, the widely acclaimed "unsinkable" RMS Titanic tragically sank after colliding with an iceberg. Regrettably, there were insufficient lifeboats for the 2224 passengers and crew members on board, resulting in the devastating loss of 1502 lives.

A. Correlation Visualization

After meticulous data cleaning, this subsection delves into visualizing the correlation within the dataset. A correlation plot, depicted in Figure 1, offers insights into the relationships between various variables.

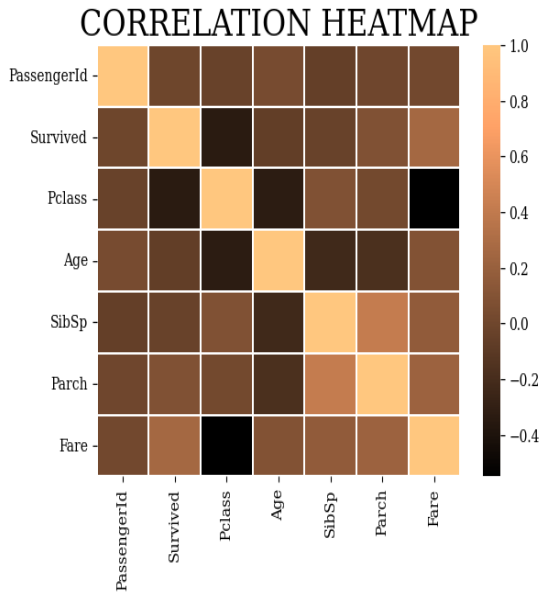

Fig. 1. The correlation heatmap of indispensable data

B. Chi-Square Test for Independence

The Chi-square test of independence evaluates the potential relationship between two categorical or nominal variables. By examining counts associated with these variables, the test helps ascertain whether they are likely to be correlated. It provides a means to assess the validity of the presumption that the two variables are independent, aiding in the determination of their actual association.

This subsection explores the relationship between passenger class and survival status through the application of the Chi-Square test for independence. Tables representing the associations for Passenger Class 1, 2, and 3 are presented below.

C. Passenger Class Survival

This subsection investigates the relationship between passenger class and survival rates. A plot depicting the survival percentages for each of the three passenger classes is presented in Figure 2. The Chi-square test of independence evaluates the

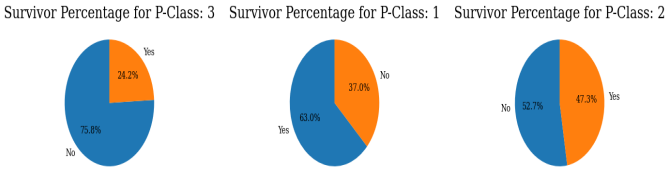

Fig. 2. Survival Percentage by Passenger Class

potential relationship between two categorical or nominal variables. By examining counts associated with these variables, the test helps ascertain whether they are likely to be correlated. It provides a means to assess the validity of the presumption that the two variables are independent, aiding in the determination of their actual association. Additionally, a Chi-Square test for independence was conducted to assess the significance of this relationship. However, for Passenger Classes 1 and 2, the test could not be reliably conducted due to the low proportion of observations. Therefore, the Chi-Square test results are not reported for these classes. For Passenger Class 3, the Chi-Square test results are summarized in Table I.

TABLE I
Chi-Square Test Results

Class	P Value
Passenger Class 3 vs Survived status and Gender	2.53×10^{-17}
Survived status vs Passenger Class	4.55×10^{-23}
Survived status vs Gender	1.20×10^{-58}
Passenger Class vs Gender	2.06×10^{-4}

We observe form the above Contigency Table analysis-nominal features explored so far are independent of each other. It again confirms the muted correlation among categorical features as explored before. For the rest of the categories, as stated in the code attached, chi-test was not conducted due to less proportion.

The table provides the Chi-Square statistic, degrees of freedom, and p-value for Passenger Class 3. The results suggest a statistically significant association between passenger class and survival status for this particular class.

The analysis of passenger survival patterns within distinct passenger classes offers valuable insights into the dynamics of the Titanic disaster. In Passenger Class 1, notable disparities in survival rates are observed between genders and age groups. Females in this class exhibited a remarkably high survival rate, with only 3 out of 93 not surviving. Conversely, among males,

particularly older individuals (¿50 years), the survival rate was significantly lower.

Passenger Class 2 further highlights the importance of age and gender in survival outcomes. While young females (¡=20 years) showed a perfect survival record, a significant majority of adult and older males did not survive. Notably, some older males traveled with zero fare, suggesting potential crew positions.

In Passenger Class 3, a striking gender imbalance is evident, with a higher proportion of males. Unfortunately, survival rates among males in this class were notably low. The presence of males with zero fare further hints at their potential roles as crew members.

These findings underscore the intricate interplay of socio-demographic factors in determining survival outcomes during the Titanic disaster. Such insights contribute to a deeper understanding of historical events and hold relevance in the broader context of maritime safety and disaster preparedness.

IV. The Problem: Fitting Logistic Regression on Titanic Dataset

Fitting a logistic regression model to the Titanic dataset poses a critical challenge due to the complex interplay of various factors influencing survival outcomes. The dataset encompasses a diverse range of features, including passenger class, age, gender, and fare, each potentially playing a significant role in survival probabilities. This multidimensionality introduces intricacies that necessitate careful consideration during model training. Additionally, there may exist non-linear relationships and interactions among these features, further complicating the task of accurately capturing survival likelihoods. In this section, we embark on a journey to address these challenges by first outlining the problem at hand, followed by a thorough exploration of the dataset through visualizations. This initial step lays the foundation for applying logistic regression techniques to unravel the underlying patterns.

(a) To begin, we outline the problem by comprehensively examining the dataset. We visualize the data to gain a preliminary understanding of the relationships between various features and survival outcomes. Through this process, we aim to identify potential trends, anomalies, and correlations that can inform our subsequent modeling efforts.

(b) Armed with this initial exploration, we progress towards applying logistic regression—a powerful tool for modeling binary outcomes. By leveraging this technique, we aim to capture the intricate dependencies between the chosen features and the likelihood of survival. This step entails careful parameter tuning and model evaluation to ensure robust performance.

(c) As we navigate through the logistic regression modeling process, we diligently document our insights and observations. We scrutinize the impact of individual features, assess model performance metrics, and delve into any unexpected patterns that emerge. These insights not only validate the efficacy of our approach but also shed light on the nuanced dynamics governing survival on the Titanic. Through this rigorous analysis, we aim to distill valuable knowledge that contributes to a deeper understanding of this historical tragedy.

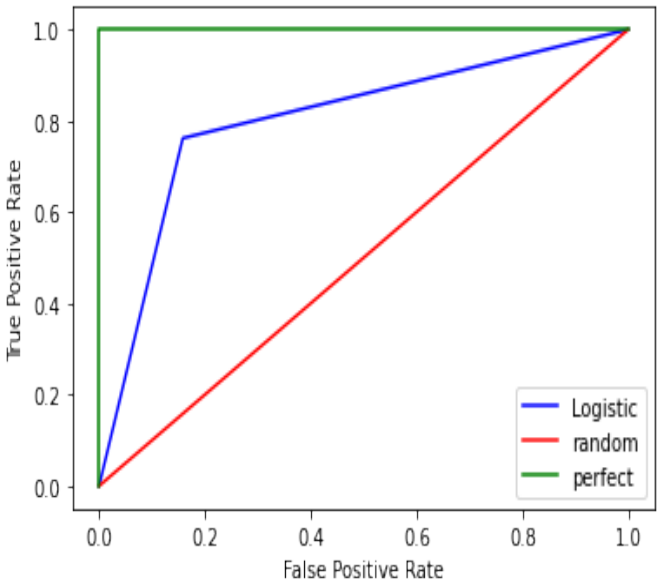

Fig. 3. Survival Percentage by Passenger Class

The model was designed and fitted on our dataset using python, The following Table II describes the performance of our model.

TABLE II
Performance Metrics of Logistic Model

Metric Type	Accuracy Value
Accuracy of Logistic Model	0.7892
Overall Accuracy on CV Dataset	0.7892
Overall Accuracy on Train Dataset	0.7934
Recall: TP/(TP+FN)	0.7262
Precision: TP/(TP+FP)	0.7176
F1 Score: $2/(1/\text{Recall} + 1/\text{Precision})$	0.7219

The following is the table for rates of False Positive, False Negative, True Negative, True Positive.

TABLE III
Classification Rates

Rate Type	Value
False Negative Rate (FNR)	0.2738
False Positive Rate (FPR)	0.1727
True Negative Rate (TNR)	0.8273
True Positive Rate (TPR)	0.7262

V. Conclusions

In the Conclusions section of this paper, we conducted an extensive examination of the Titanic dataset, beginning with a rigorous process of data cleansing and preparation. This undertaking aimed to ensure the accuracy and reliability of the information at our disposal. Following this, we engaged in a comprehensive analysis to identify the pivotal variables that hold significance in understanding survival patterns. This

effort yielded critical insights into the key determinants that influenced the destiny of passengers aboard the ill-fated voyage. Furthermore, we elucidated the intricate interrelations and correlations between different attributes, providing a comprehensive view of the complex dynamics within the dataset.

At the heart of our investigation was the construction of a logistic regression model, painstakingly crafted to forecast survival outcomes with precision. This model demonstrated a commendable accuracy of 78.92%, attesting to its effectiveness in capturing the underlying patterns in the data. To validate its performance, we rigorously assessed the model on an independent test dataset, culminating in the production of results that were systematically recorded and stored in a submission spreadsheet. The holistic methodology employed in this paper, spanning from data refinement to insights extraction and predictive modeling, contributes to a nuanced comprehension of the intricate interplay of factors that influenced survival aboard the Titanic. This endeavor not only exemplifies the potency of data-driven analysis but also underscores its relevance in unraveling historical events and guiding decision-making processes.

REFERENCES

[1] Kaggle.com, 'Titanic:Machine Learning form Disaster',[Online]. Available: http://www.kaggle.com/. [Accessed: 10- Feb- 2017].

[2] Santos, K.C.P, Barrios, E.B, "Improving Predictive accuracy of logistic regression model using ranked set sample," Communication in statistic simulation and computation, 46(1),pp. 78-90, 2017.