DAL 2023-ASSIGNMENT 2

Kaja Vikas
Department of Engineering Design
Indian Institute of Technology, Madras
Chennai, India
ed20b026@smail.iitm.ac.in

***Abstract*—This paper presents a comprehensive analysis of the survival rates of passengers aboard the ill-fated Titanic. The dataset encompasses generic information about the passengers and survival outcomes. Through rigorous exploration and analysis, we endeavor to discern the pivotal factors that contributed to passenger survival. The dataset is subjected to feature engineering to enhance its suitability for further data analysis. One of our main aims in this paper is to study the aspects of logistic regression. We chose logistic regression over normal regression techniques to predict survival rates because logistic regression is specifically designed for binary outcomes, making it more suitable for modeling the probability of survival versus non-survival. Factors such as passenger class, age, gender, and embarkation point emerge as significant influencers in the likelihood of survival.**

I. Introduction

The tragic sinking of the RMS Titanic on April 15, 1912, stands as an indelible mark in maritime history, claiming the lives of over 1,500 individuals. This heart-wrenching event during the ship's maiden voyage serves as a stark reminder of the vulnerability of human endeavors in the face of natural forces. The dataset concerning the Titanic passengers offers a poignant glimpse into the lives of those on board, encompassing details such as age, gender, ticket class, cabin allocation, and the pivotal distinction of survival or loss. In this context, our primary aim is to discern the critical factors that influenced survival rates. By delving into this data, we seek to unravel the significant determinants of survival, shedding light on the intricate dynamics of this tragic event. The insights garnered hold potential not only for deepening our understanding of historical occurrences but also for informing future disaster preparedness and response strategies. To achieve this, we employ a logistic regression model. This statistical technique allows us to classify whether passengers in the test dataset survived or not, based on the patterns observed in the training data. Furthermore, it contributes to the broader narrative of disaster analysis and predictive modeling, reinforcing the invaluable lessons gleaned from this pivotal moment in history.

Logistic regression is a powerful statistical technique employed in predictive modeling, particularly when dealing with binary outcomes. It serves as a valuable tool to map a set of independent variables to a dependent binary variable, often representing a yes-or-no, success-or-failure scenario. This mapping is achieved through the utilization of the sigmoid function, which transforms a linear combination of input features into a probability score between 0 and 1. This inherent probabilistic nature makes logistic regression well-suited for scenarios like the Titanic dataset, where the goal is to predict passenger survival (binary: survived or not). Logistic regression can handle both categorical and numerical features, making it versatile for datasets with diverse attribute types. To evaluate the model's performance, we employ widely used metrics such as accuracy and the F1 score. Accuracy measures the proportion of correctly classified instances, while the F1 score balances precision and recall, crucial for imbalanced datasets like ours, where survival rates vary. Overall, logistic regression provides a transparent and reliable framework for predicting passenger survival, making it a prudent choice for our analysis.

We addressed missing values within the dataset through tailored strategies. Specifically, in the 'Age' column, we imputed absent values by computing the median age contingent on the passenger's title and the presence of parents or children. For the 'Embarked' column, we employed imputation by utilizing the most prevalent value, denoted as 'S', indicating embarkation from Southampton. To further refine the dataset for analysis, we conducted feature engineering—a systematic process involving the transformation and selection of pertinent attributes. Following model training, we rigorously evaluated its performance using established metrics, including accuracy and the F1 score. Accuracy provided insight into the proportion of correctly classified instances, while the F1 score, particularly valuable for imbalanced datasets like ours, offered a balanced assessment of precision and recall. Furthermore, we extended our analysis by applying the logistic regression model to the test dataset, leveraging the disclosed survival column to predict whether passengers in this dataset survived or not.

Section 2 introduces the Titanic dataset and the objective of the analysis and encompasses an in-depth exploration of the dataset, focusing on data visualization techniques to gain insights into passenger demographics and survival rates. Additionally, this section details the feature engineering process, which involves refining and selecting pertinent attributes to enhance the dataset's suitability for analysis. In Section 3, the paper delves into the application of logistic regression as the chosen modeling technique, highlighting its suitability for binary outcomes and explaining why it was preferred over other regression methods. The logistic regression model is trained and evaluated using the prepared dataset. Section 4 presents the results of the analysis, highlighting the accuracy and F1 score achieved by the logistic regression model. Feature importance analysis is also discussed, shedding light on the attributes that most significantly influence survival outcomes. Section 5 encapsulates the conclusion, summarizing the key findings and insights derived from the analysis. The section also outlines avenues for future research and potential improvements in predictive modeling techniques. Finally, Section 6 provides a comprehensive list of references, acknowledging the sources and studies that contributed to the paper's foundation and analysis.

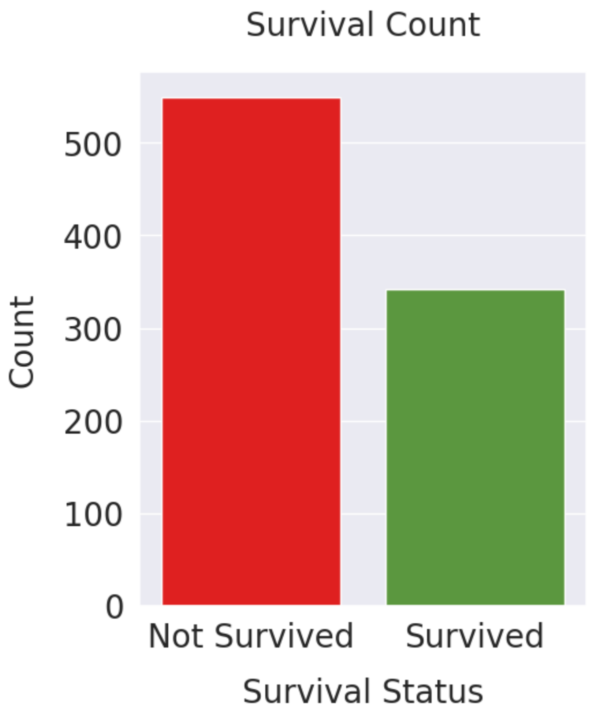
II. Datasets

The dataset comprises a comprehensive array of attributes for each passenger, offering a multifaceted view of those aboard the Titanic. These attributes include Passenger ID, which serves as a unique identifier for everyone, Survived, indicating whether the passenger survived (1) or did not (0), Pclass, denoting the class of their ticket, and Name, containing the names of the passengers. Sex designates the gender of the passenger, while Age reveals their respective ages. SibSp and Parch signify the number of siblings/spouses and parents/children aboard, respectively. Ticket holds ticket numbers, Fare denotes the fare paid, and Cabin divulges the cabin allocated to the passenger. Finally, Embarked records the port of embarkation.

Within the training dataset, the 'Cabin' attribute poses a significant concern, with 687 instances, 77% of the total, lacking values. Given this substantial absence of data, a prudent course of action would be to contemplate excluding the 'Cabin' feature from the analysis. This measure is essential to prevent potential distortion of the results. Turning attention to the 'Age' feature, it reveals a challenge with 177 missing values. This presents a notable hurdle in effectively addressing this variable. Conversely, the 'Embarked' feature demonstrates a small number of missing values, specifically only 2 instances. These gaps can be managed through appropriate imputation techniques.

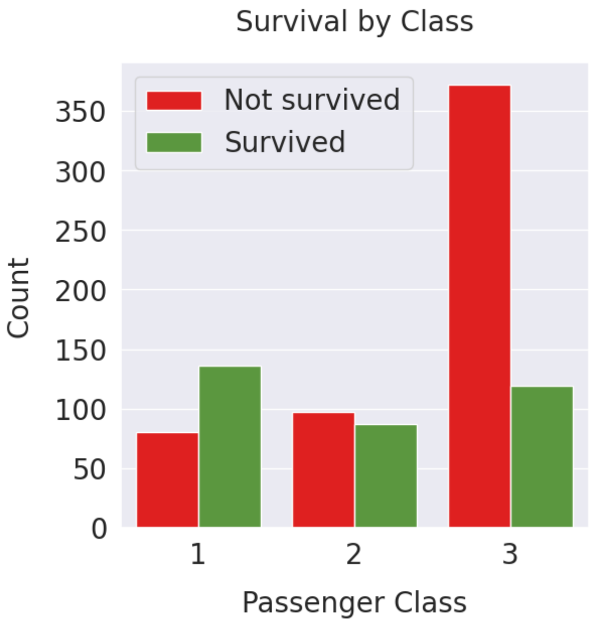
A. Data Visualization

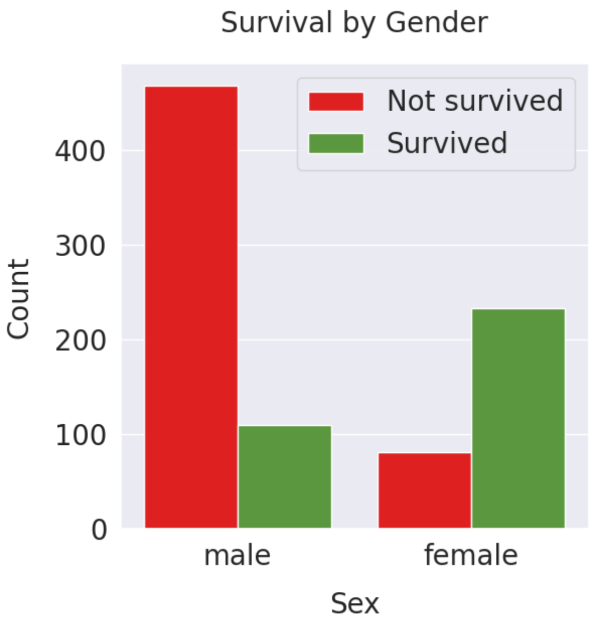
Upon examining the bar graph representing survival counts within the training dataset, a notable disparity emerges. It is evident that the number of passengers who survived is lower compared to those who did not. This observed distribution highlights a significant class imbalance within the dataset. The prevalence of this class imbalance prompts careful consideration in the selection of appropriate evaluation metrics for our predictive model. Considering this, metrics that demonstrate resilience to class imbalances become imperative. One such metric is the F1 score, which harmonizes both precision and recall. By synthesizing the precision's focus on true positives with recall's emphasis on capturing actual positives, the F1 score proves particularly adept in scenarios characterized by uneven class distributions. A discernible pattern emerges from our analysis of the bar graph representing passenger survival across different classes. It is evident that passengers in Class 3 experienced a significantly higher mortality rate compared to those in Classes 1 and 2. Class 3 exhibits a stark imbalance between the number of passengers who survived and those who did not, with a notably higher death count than survivors. Remarkably, Class 1 stands out as the sole category where the survival count surpasses the death count, indicating a higher likelihood of survival for passengers in this class. In contrast, Class 2 presents a balanced distribution, with survival and death counts appearing equivalent. It is worth noting that despite the preponderance of passengers in Class 3, Class 1 exhibits a higher survival count than Class 3. This observation underscores the complex interplay of factors influencing survival outcomes, with passenger class playing a pivotal role.



The bar graph illustrating survival by gender provides significant insights into the distribution of survivors based on this demographic attribute. Notably, the graph indicates a higher representation of males compared to females among the passengers. This imbalance is striking, with a clear majority being male. It is evident that the number of males who did not survive surpasses the count of those who did. This discrepancy highlights a substantial loss of male passengers. Conversely, among females, the survival count notably exceeds the death count, signifying a higher likelihood of survival for female passengers. A noteworthy finding is that despite the higher overall count of male passengers, the survival count of males is eclipsed by the survival count of females.

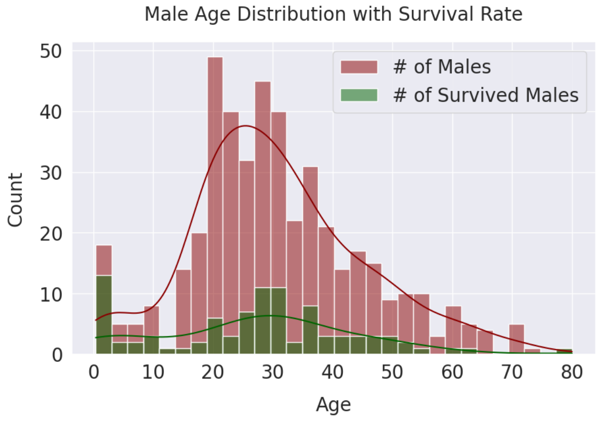
The bar graph depicting the distribution of males by age and their corresponding survival rates imparts significant insights. A conspicuous concentration of males falls within the age bracket of 20 to 40 years. This demographic cohort constitutes a substantial portion of the male passengers. Upon closer

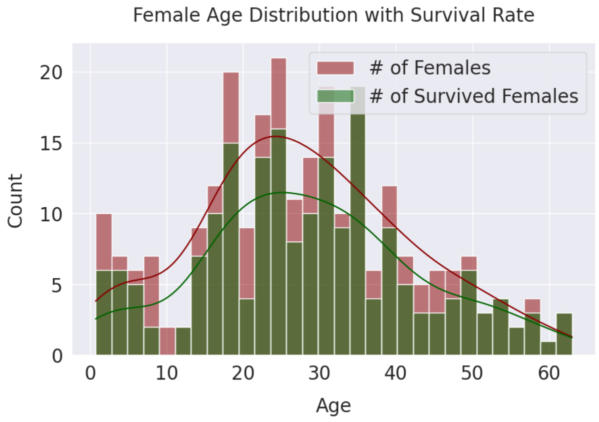




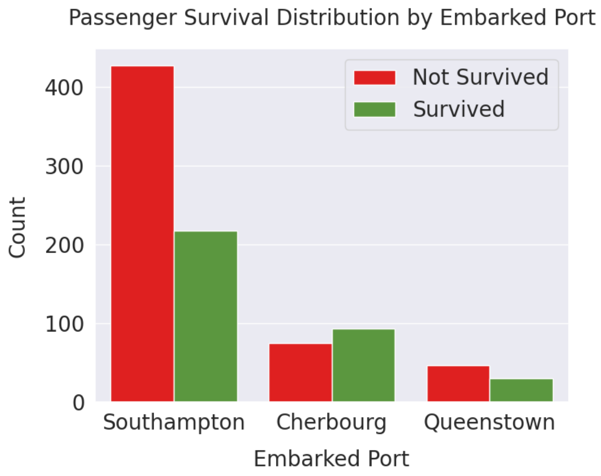
examination, a noteworthy pattern emerges. Within this age range of 20 to 40 years, the proportion of males who did not survive is notably higher compared to other age groups. This observation underscores a significant loss of male passengers in their prime years. An intriguing finding arises when considering males below the age of 4 years. In this age group, there is a notably higher percentage of survivors relative to the total number of males. This suggests a greater likelihood of survival for young boys and infants. This observation highlights a protective aspect, indicating that efforts were made to ensure the safety of the youngest passengers.

The bar graph illustrating the distribution of females by age and their corresponding survival rates imparts valuable insights. Notably, there is a discernible concentration of females within the age range of 20 to 40 years. This demographic group constitutes a huge portion of the female passengers. Upon closer examination, a salient pattern emerges. Within this age bracket of 20 to 40 years, the survival rates among females are remarkably consistent. Regardless of age within this range, the percentage of females who survived is quite balanced. This observation suggests a uniform likelihood of survival for females in their prime years. Furthermore, it is noteworthy to observe that, across all age groups, the percentage of females who did not survive is lower compared to the percentage who did survive. This indicates a higher overall survival rate among females, regardless of age. The analysis of female age distribution and survival rates reveals that while there is a concentration of females in the 20-40 age range, survival rates among females are notably consistent across various ages within this bracket. Additionally, females, in general, demonstrate a higher likelihood of survival. This observation underlines the priority given to ensuring the safety of female passengers, regardless of their age.





The bar graph depicting passenger survival distribution by embarkation port reveals distinct patterns. Southampton saw the highest number of embarkations, markedly surpassing Cherbourg and Queenstown. However, survival rates vary significantly among these ports. For instance, Southampton experienced a notably lower survival rate, with only about one-third of passengers surviving. In contrast, Cherbourg exhibited a higher survival count compared to the number of casualties. Conversely, Queenstown displayed a lower survival count relative to the number of fatalities, indicating a comparatively lower survival rate. Notably, both the death and survival count for Queenstown were lower compared to passengers from other ports. These observations highlight the influence of embarkation port on survival outcomes and warrant further investigation into the underlying factors at play.



B. Feature Engineering

In the process of feature engineering, several key decisions were made to enhance the dataset's suitability for analysis. Notably, the PassengerId, Ticket, and Name attributes were systematically removed from both the training and test datasets. PassengerId and Ticket were identified as unique identifiers without inherent predictive value for survival outcomes. Additionally, the Cabin column was systematically removed from both datasets. This decision was driven by a significant observation: the Cabin feature exhibited an extensive 77% of missing values. Given this substantial absence of data, the column was deemed impractical for meaningful analysis and subsequently excluded. Therefore, these columns were deemed non-contributory to the analysis. This feature selection process ensures that the dataset is refined to include attributes with higher predictive potential, streamlining it for subsequent machine learning modeling.

As an extension of the feature engineering process, titles were extracted from the Name attribute to capture social or marital status. This information could be influential in predicting survival outcomes. Less common titles were grouped under 'Other', and synonymous titles were unified. The titles were then converted to numerical values. Missing titles were conservatively filled with 0. This enhancement enriches the dataset with refined attributes for subsequent machine learning modeling.

Missing values in the Embarked column were addressed by filling them with the mode value 'S', signifying embarkation from Southampton for passengers with unrecorded embarkation information. Next, categorical values in the Embarked column ('S', 'C', 'Q') were mapped to numerical equivalents (0, 1, 2) using a predefined dictionary. Furthermore, gender information was encoded numerically by mapping 'male' to 1 and 'female' to 0 in the Sex column. To ensure data integrity, missing fare values were imputed with the mean fare value from the training dataset. Finally, Fare values were converted to integers, streamlining their representation while retaining relevant information.

To refine the dataset further, a new binary column named HasParch was introduced. This column classifies passengers based on whether they had parents or children aboard, denoted by a value of 1 if true, and 0 otherwise.

Subsequently, a calculated median age was assigned to different combinations of name_title and HasParch. This comprehensive grouping ensured a more nuanced approach to estimating missing age values.

To execute this strategy, a custom function fill_age was defined. In cases where a passenger's age was absent, the function leveraged the available information of name_title and HasParch to ascertain the most appropriate median age. In instances where specific median ages were unavailable, the function defaulted to the overall median age derived

article amsmath

III. Logistic Regression: An In-Depth Overview

A. Algorithm

Logistic regression is a statistical model used for binary classification tasks. It's particularly well-suited for problems where the dependent variable is categorical and has two possible outcomes, such as 'yes' or 'no', '1' or '0', or in our case, 'survived' or 'not survived'.

The logistic regression algorithm operates by modeling the probability that a given instance belongs to a particular category. It accomplishes this through a transformation using the logistic function, also known as the sigmoid function:

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

Here, $P(Y = 1)$ represents the probability of the event of interest (in our case, survival), while $\beta_0, \beta_1, ..., \beta_n$ are the coefficients associated with the intercept and input features $X_1, X_2, ..., X_n$ respectively.

B. Assumptions

1) **Linearity in Log-Odds**: The relationship between the independent variables and the log-odds of the dependent variable should be linear.
2) **Independence of Errors**: The observations should be independent of each other.
3) **No Multicollinearity**: The independent variables should not be highly correlated with each other.
4) **Large Sample Size**: Logistic regression performs best with a large sample size.

C. Evaluation Metrics

- **Accuracy**: The proportion of correctly classified instances. It's calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

- **Precision**: The ratio of true positives to the sum of true positives and false positives. It measures the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall (Sensitivity)**: The ratio of true positives to the sum of true positives and false negatives. It measures the ability to identify all positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

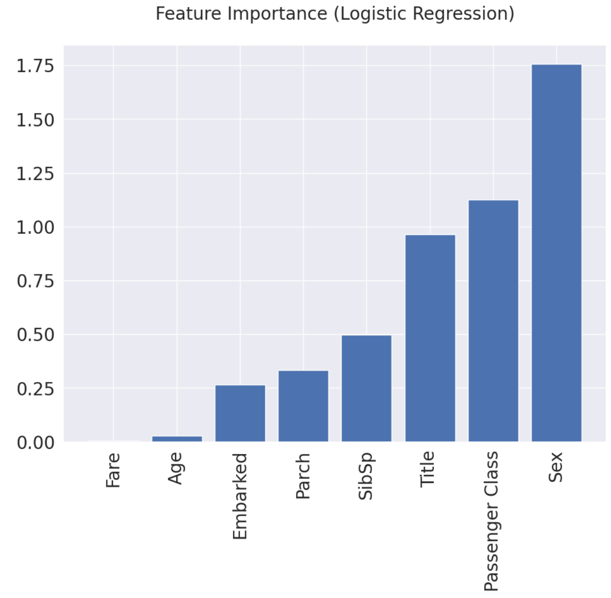
- **F1 Score**: The harmonic mean of precision and recall, providing a balanced assessment.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

D. Maintaining the Integrity of the Specifications

IV. Results

The logistic regression model achieved an accuracy of 83.05% on the training data, indicating that it correctly predicted survival outcomes for a significant proportion of passengers. The F1 score, which balances precision and recall, is a measure of a model's accuracy. In this case, the logistic regression model attained an F1 score of 0.77, reflecting a robust performance in classifying survival outcomes.



A feature importance analysis was conducted to understand the influence of different attributes on the model's predictions. From the feature importance plot, it is evident that the most influential factor affecting survival rate is gender, with passenger class coming in as the second most important. Martial status, represented by title, holds the third position, followed by the number of siblings/spouses (SibSp), and the number of parents/children (Parch). On the other hand, Embarked location also shows some influence, while age and fare do not appear to have a significant impact on passenger survival.

V. Conclusion

The analysis of the Titanic dataset through data visualization and feature importance has provided valuable insights into the factors influencing passenger survival rates.

1) **Key Influential Factors**:
   - Gender emerged as the most critical factor, with female passengers having a significantly higher chance of survival.
   - Passenger class also played a pivotal role, indicating that individuals in higher classes had better odds of survival.
   - Martial status, represented by title, showed a notable influence, underscoring the importance of social standing.
2) **Limited Influence of Age and Fare**: Surprisingly, age and fare did not exhibit a substantial impact on survival rates. This suggests that other factors took precedence in determining who survived.

VI. Future Work

1) **Feature Engineering**: Further exploration of the dataset for potential feature engineering could yield additional attributes that may enhance predictive models.
2) **Advanced Modeling Techniques**: Employing more complex machine learning algorithms such as Random Forests or Gradient Boosting could potentially improve predictive accuracy.
3) **Incorporating Additional Data**: Introducing external datasets, such as historical weather data or additional passenger information, could provide a more comprehensive understanding of the factors affecting survival.
4) **Temporal Analysis**: Considering the evolving conditions during the Titanic disaster, a temporal analysis could reveal how the situation changed over time and its impact on survival rates.
5) **Ethnicity and Nationality**: Exploring the influence of ethnicity and nationality on survival outcomes could offer a deeper sociocultural perspective.

By undertaking these future steps, we aim to further refine our understanding of the tragic events aboard the Titanic and potentially develop more accurate predictive models for similar scenarios in the future.

References

[1] https://www.ibm.com/topics/logistic-regression: :text=Resources-,What

[2] https://www.geeksforgeeks.org/understanding-logistic-regression/.

[3] https://en.wikipedia.org/wiki/Logistic$_{r}egression.$

[4] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc.