Project Report

The Titanic Survival Challenge

Student:
Arman Nik Khah
Arman.Nikkhah@UTDallas.edu

Advisor:
Prof. Rishabh Iyer

Department of Computer Science
February 1, 2024

Contents

1 Introduction

This report presents a comprehensive analysis of the Titanic Survival Challenge. Our project's goal is to construct a predictive model using passenger data from the Titanic disaster, aiming to identify determinants of survival. We delve into exploratory data analysis (EDA), data preprocessing, and machine learning to understand the patterns, relationships, and anomalies in the data, ultimately predicting survival outcomes.

2 Methodology

Our approach integrates several stages of data analysis and modeling. We begin with an extensive EDA to understand the data structure, followed by rigorous data preprocessing, including handling missing values and feature engineering. We then implement various machine learning algorithms to predict survival outcomes. Our methodology emphasizes the integration of statistical methods and machine learning techniques to derive meaningful insights and accurate predictions from the data.

3 Exploratory Data Analysis (EDA)

Our EDA aimed to uncover patterns, anomalies, and relationships within the Titanic dataset. This in-depth analysis was pivotal in guiding our subsequent data preprocessing and modeling strategies.

3.1 Data Insights and Interpretation

Passenger Demographics: A detailed inspection of passenger demographics revealed noteworthy patterns. The survival rate varied significantly across different age groups, genders, and socio-economic classes. Particularly, younger passengers and women exhibited higher survival rates. This observation aligns with historical accounts emphasizing the prioritization of women and children during lifeboat evacuations.

Socio-Economic Status: Our analysis uncovered a clear socio-economic divide in survival chances. Passengers from higher socio-economic classes (reflected in features like passenger class and fare) showed a higher likelihood of survival. This finding is indicative of the disparities prevalent in the early 20th century.

3.2 Visualization and Analysis

Distribution Plots: These plots were instrumental in visualizing the age and fare distributions, offering insights into the socio-economic status and age profile of passengers.

Correlation Analysis: Through heatmaps and correlation matrices, we explored the interdependencies between various features. This analysis was crucial in identifying features that directly or indirectly influenced survival chances.

Survival Rate Comparisons: Using bar graphs and pivot tables, we compared survival rates across different groups. This approach was particularly effective in highlighting disparities based on gender and passenger class.

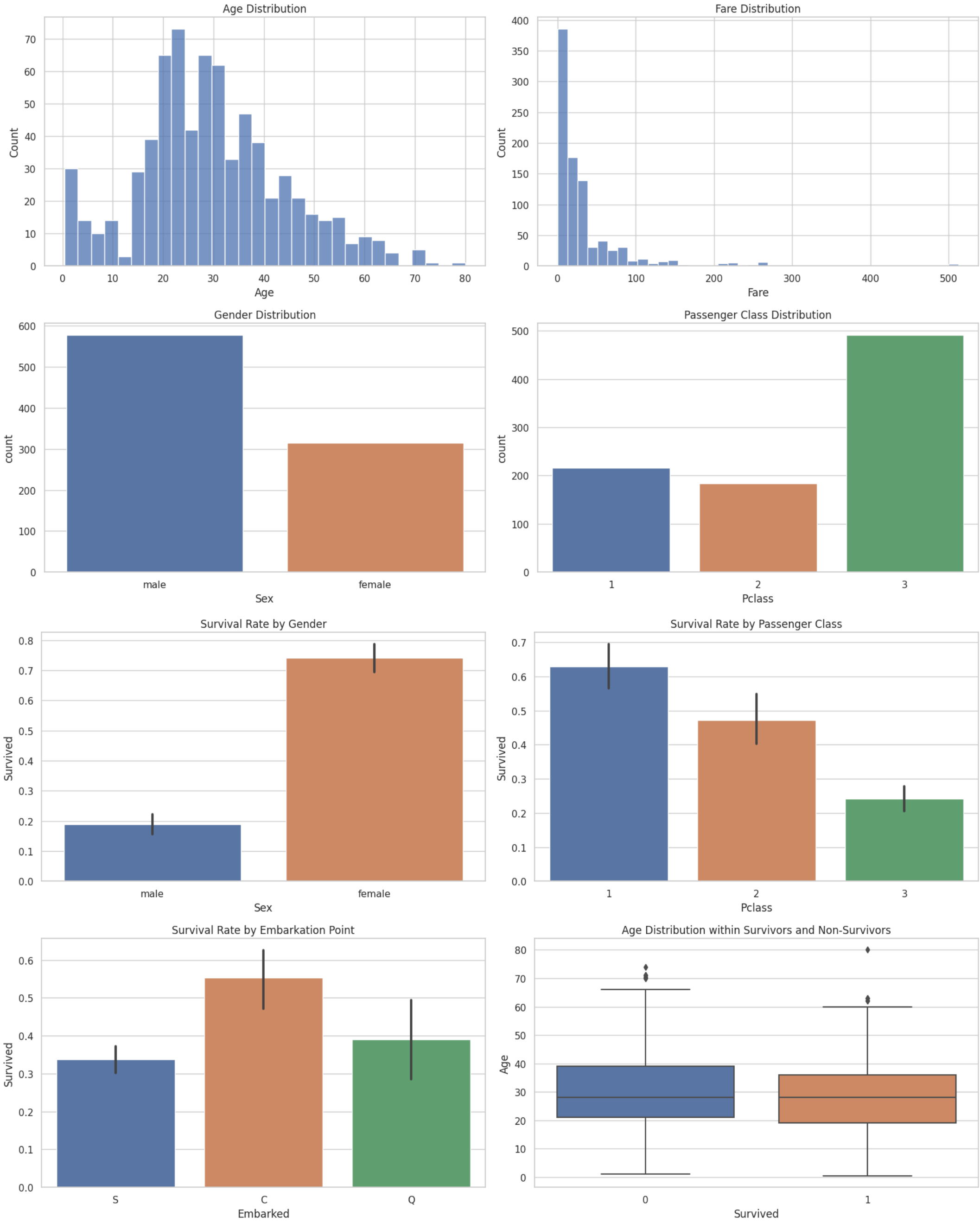
3.3 Anomalies and Patterns

Family Dynamics: An interesting anomaly was the influence of family size on survival rates. Larger families tended to have lower survival rates, possibly due to difficulties in organizing during the evacuation.

Embarkation Points: Differences in survival rates based on embarkation points suggested potential variations in the handling and allocation of lifeboats at different ports.

3.4 Statistical Summary

Descriptive Statistics: A comprehensive statistical summary, including measures of central tendency and dispersion, provided a quantitative backdrop for our analysis, ensuring that our interpretations were grounded in concrete data.


Age Distribution
Count
0
10
20
30
40
50
60
70
Age
0
10
20
30
40
50
60
70
80
Fare Distribution
Count
0
50
100
150
200
250
300
350
400
Fare
0
100
200
300
400
500
Gender Distribution
count
0
100
200
300
400
500
600
male
female
Sex
Passenger Class Distribution
count
0
100
200
300
400
500
1
2
3
Pclass
Survival Rate by Gender
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
male
female
Sex
Survival Rate by Passenger Class
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Pclass
Survival Rate by Embarkation Point
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
S
C
Q
Embarked
Age Distribution within Survivors and Non-Survivors
Age
0
10
20
30
40
50
60
70
80
0
1
Survived

Interpretation: The EDA provided valuable insights into the socio-economic, demographic, and familial factors influencing survival on the Titanic. It highlighted the crucial role of age, gender, and socio-economic status, which informed our subsequent modeling and feature engineering strategies.

4 Data Preprocessing

Data preprocessing is a critical step in our analytical process, determining the effectiveness of our machine learning models. This section outlines the strategies employed to prepare the Titanic dataset for analysis, along with interpretations of the impacts of these methods.

4.1 Handling Missing Values

The Titanic dataset presented challenges with missing values, particularly in the 'Age', 'Cabin', and 'Embarked' columns. We adopted distinct strategies for each:

- **Age**: We utilized median imputation, acknowledging the robustness of the median against outliers in age distribution.
- **Cabin**: The high proportion of missing values led to the extraction of deck information, translating a largely incomplete feature into a potentially insightful categorical variable.
- **Embarked**: Given its categorical nature and minimal missing values, we filled gaps with the mode, reflecting the most common embarkation point.

4.2 Feature Engineering

Two significant features were engineered:

- **FamilyNameEncoded**: This transformation captured family groupings, hypothesizing a potential impact of family size and structure on survival rates.
- **Deck**: Extracted from 'Cabin', this feature aimed to explore the influence of passenger location on survival, an aspect historically noted but not directly captured in the data.

4.3 Feature Transformation and Normalization

We employed One-Hot Encoding for categorical variables like 'Embarked', 'Deck', and 'Pclass'. This approach avoids the introduction of artificial ordinal relationships, preserving the nominal nature of these variables. Continuous variables, including 'Age', 'Fare', and 'Family Size', were normalized using MinMaxScaler, ensuring equal contribution to model training.

5 Machine Learning Model Implementation

5.1 Overview

In this project, we employed a range of machine learning models, each with unique strengths and adaptabilities to our dataset characteristics. These models include Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Support Vector Machines (SVM), and Ensemble methods. We also explored a Fully Connected Neural Network (FCN) for its capacity to capture non-linear patterns.

5.2 Model Performance and Interpretation

5.2.1 Logistic Regression

Our Logistic Regression model achieved an accuracy of approximately 87.68%, with a balance of precision and recall, indicating its effectiveness in class separation. This model's strength lies in its simplicity and interpretability, offering insights through coefficients, crucial for understanding feature impact on survival likelihoods.

5.2.2 Random Forest

Random Forest, a more complex model, demonstrated an accuracy of 86.23%. It provided a depth of feature importance analysis, shedding light on socio-economic and demographic factors influencing survival rates. Its strength lies in handling categorical and continuous variables efficiently, offering a nuanced view of the data.

5.2.3 Support Vector Machines (SVM)

SVM's performance was notable, with an accuracy of 68.84%. However, its precision, recall, and F1-score metrics were less meaningful due to the absence of predicted positive samples in our test set, reflecting challenges in this model's application to our specific dataset.

5.2.4 Ensemble Models

Our ensemble model, combining Logistic Regression, Random Forest, and Gaussian Naive Bayes, achieved one of the highest accuracies of 86.96%. This model's robustness comes from leveraging the strengths of individual models and offering a comprehensive understanding of varied data patterns.

5.2.5 Fully Connected Neural Network (FCN)

The FCN, with a configured architecture suitable for our dataset, yielded an accuracy of 85%. While its performance was comparable to traditional models, the complexity and computational requirements suggest that simpler models might be more efficient for this specific task.

5.3 Insights and Implications

The diverse performances of these models underscore the complexity of the Titanic dataset. Logistic Regression's high performance suggests that linear relationships are significant in the data. In contrast, the success of Random Forest and Ensemble models indicates the presence of more complex, non-linear patterns. The variance in model effectiveness also highlights the importance of feature selection and engineering in predictive modeling, especially in datasets with socio-economic and demographic nuances like ours.

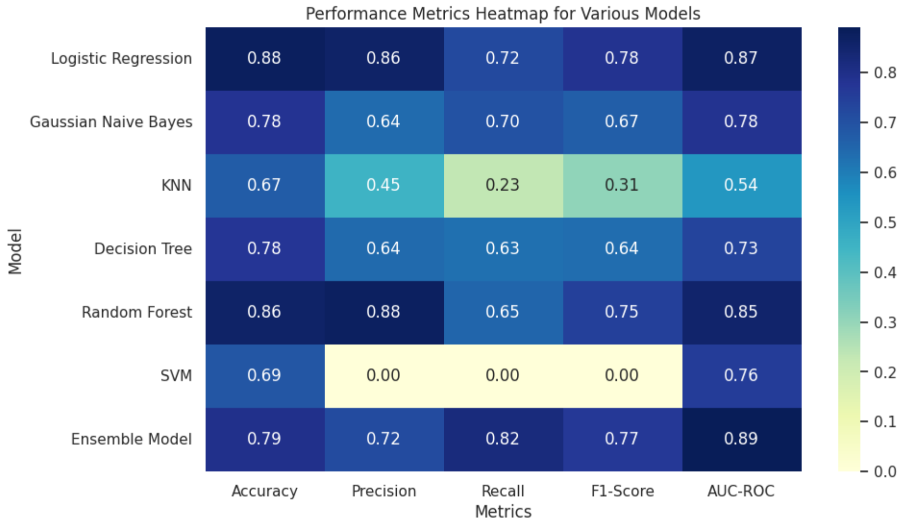
6 Evaluation

6.1 Model Performance Metrics

In-depth evaluations of multiple machine learning models were conducted, including Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machines (SVM), Ensemble Models, and a Fully Connected Neural Network (FCN). The key performance metrics considered were Accuracy, Precision, Recall, F1-Score, and AUC-ROC. These metrics provided a holistic view of each model's performance and their ability to predict survival on the Titanic.

6.2 Model Comparisons and Interpretations

- **Logistic Regression:** Exhibited high accuracy and a balanced precision-recall trade-off. The model was particularly effective in distinguishing between survival classes, as indicated by its AUC-ROC.
- **Gaussian Naive Bayes:** While it had a lower accuracy and precision, its recall was comparable to Logistic Regression, suggesting effectiveness in identifying true positives.
- **K-Nearest Neighbors (KNN):** Showed lower performance across all metrics, possibly due to its sensitivity to irrelevant features and the nature of the dataset.
- **Decision Tree:** Demonstrated moderate performance, but its lower AUC-ROC hinted at less effectiveness in differentiating between classes.
- **Random Forest:** Achieved high accuracy and precision, along with a good recall and F1-score, indicating its robustness in classification tasks.
- **Support Vector Machines (SVM):** Had reasonable overall accuracy and AUC-ROC. However, the lack of predicted positive samples in the test set impacted its precision, recall, and F1-score, limiting its interpretability.
- **Ensemble Model:** This model, combining Logistic Regression, Random Forest, and Gaussian Naive Bayes, showcased the highest AUC-ROC and recall values, surpassing other models in balancing precision and accuracy.
- **Fully Connected Neural Network (FCN):** Demonstrated comparable accuracy to other models, but did not outperform logistic regression or ensemble methods, indicating that traditional machine learning models were sufficient for the dataset's complexity.



6.3 Heatmap Visualization

A heatmap was created to visually compare the performance metrics of these models. This visualization facilitated an easy comparison across different metrics, allowing for informed decision-making regarding model selection.

6.4 Interpretation of Results

The comparative analysis revealed key insights:

- Models like Logistic Regression and Random Forest were effective in capturing both simple linear and complex non-linear relationships within the data.
- The underperformance of KNN and SVM highlighted the challenges these models faced with the dataset's feature space and distribution.
- The Ensemble Model's balanced performance suggested that a combination of different algorithms was beneficial in capturing diverse patterns in the data.

Overall, the evaluation underscored the complexity of the Titanic dataset, with different models capturing various aspects effectively. The choice of model depends on specific application requirements, such as the need for precision over recall or a balance between both.

7 Feature Importance Analysis

We analyzed feature importance for both Logistic Regression and Random Forest models. Important features included socio-economic status (Pclass), gender (PersonType), and age, indicating their significant impact on survival rates. Deck and FamilyNameEncoded also played a notable role, reflecting the influence of cabin location and family dynamics.

8 Comparative Model Analysis

We compared various models including AdaBoost, GradientBoosting, Random Forest, and XGBoost. XGBoost and GradientBoosting showed the highest accuracy and F1-Scores, indicating a balanced precision-recall trade-off. The performance metrics across these models varied, with XGBoost slightly outperforming others in precision and AUC-ROC.

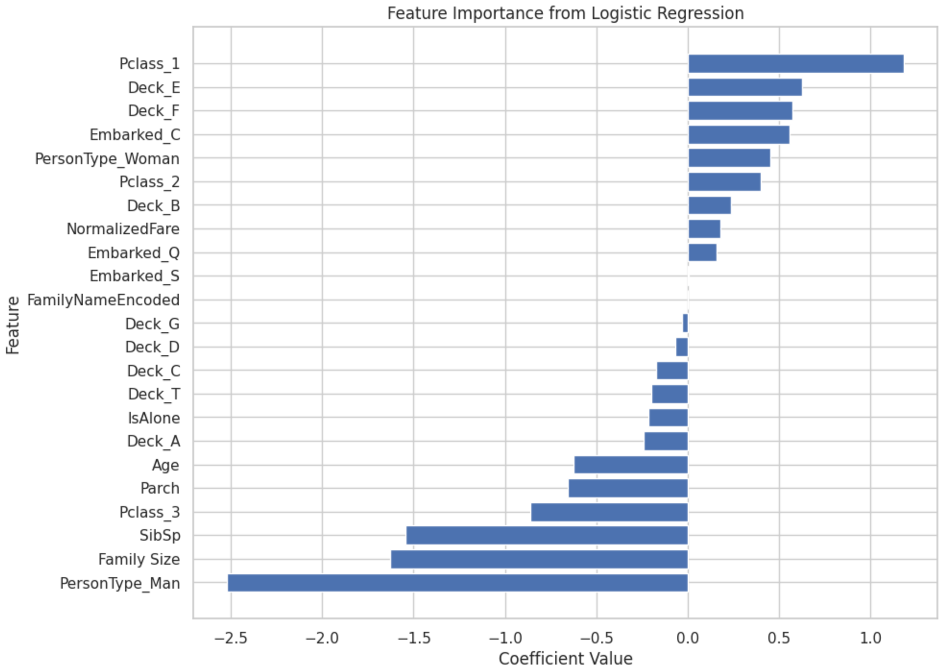
9 Conclusions

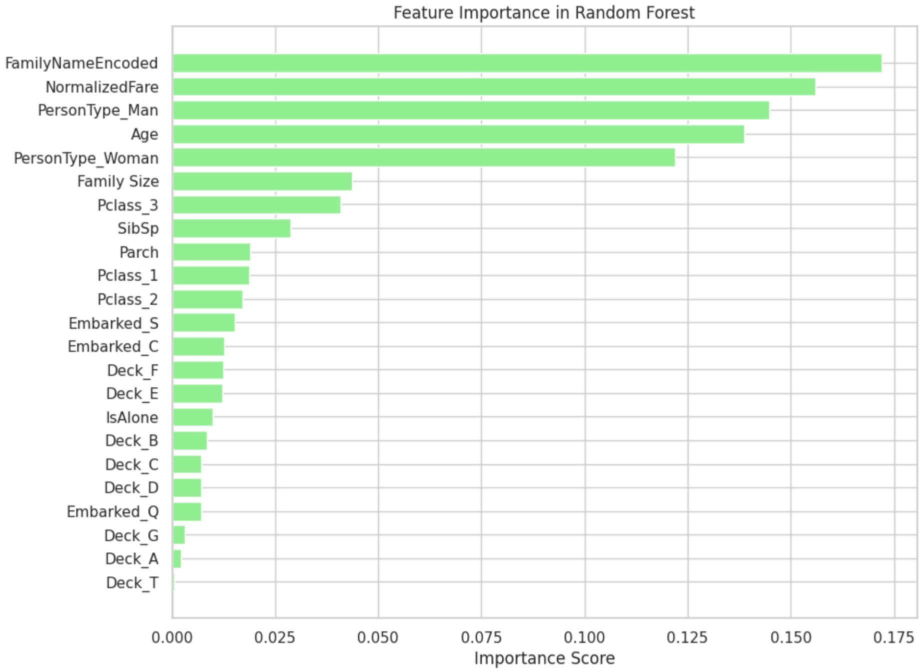
9.1 Interpretation of Results

Our study's comprehensive analysis, leveraging diverse machine learning models, offers profound insights into the Titanic disaster's survival dynamics. The final models, including Logistic Regression, Random Forest, SVM, and advanced ensemble methods like XGBoost, provided varied perspectives on the feature importance and model performance.

Gender, Age, and Socio-Economic Status: Consistently across models, gender, age, and socio-economic status emerged as critical determinants of survival. The Logistic Regression model highlighted the negative impact of being a man (PersonType_Man) on survival chances, aligning with historical accounts of prioritizing women and children for lifeboat access. The importance of passenger class (Pclass) and normalized fare, proxies for socio-economic status, underscored the disparity in survival based on class divisions.

Family Dynamics: The Random Forest model's emphasis on 'FamilyNameEncoded' suggests family affiliations played a crucial role, potentially indicating better survival coordination among family groups or reflecting socio-economic correlations within families.


Feature Importance from Logistic Regression
Pclass_1
Deck_E
Deck_F
Embarked_C
PersonType_Woman
Pclass_2
Deck_B
NormalizedFare
Embarked_Q
Embarked_S
FamilyNameEncoded
Deck_G
Deck_D
Deck_C
Deck_T
IsAlone
Deck_A
Age
Parch
Pclass_3
SibSp
Family Size
PersonType_Man
Feature
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
Coefficient Value


Feature Importance in Random Forest
FamilyNameEncoded
NormalizedFare
PersonType_Man
Age
PersonType_Woman
Family Size
Pclass_3
SibSp
Parch
Pclass_1
Pclass_2
Embarked_S
Embarked_C
Deck_F
Deck_E
IsAlone
Deck_B
Deck_C
Deck_D
Embarked_Q
Deck_G
Deck_A
Deck_T
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance Score

Model Performance Comparison: The Logistic Regression and Ensemble Models outperformed others in accuracy, while XGBoost excelled in precision and AUC-ROC. This indicates XGBoost's superior ability to correctly identify true positives and avoid false positives. However, the Ensemble Model's balanced performance across metrics demonstrates the effectiveness of combining different model strengths.

9.2 Insights into Feature Interactions

Cabin and Deck Analysis: While certain deck locations like Deck_D and Deck_E were associated with higher survival rates, possibly due to proximity to lifeboats, the overall impact of specific deck locations was less critical than initially assumed.

Embarkation Points: The influence of embarkation points on survival rates might indicate logistical or administrative differences in how passengers were managed based on boarding locations, though this requires further investigation.

9.3 Future Research Directions

Our findings pave the way for future research to explore deeper into socioeconomic factors, potentially using even more advanced machine learning techniques. The unexpected importance of certain features, like family names, opens new avenues for understanding social dynamics aboard the Titanic. Additionally, further investigation into the impact of embarkation points and cabin locations on survival could yield novel insights.