Titanic Survival Prediction Analysis

•••

By: Melvina Brummitt

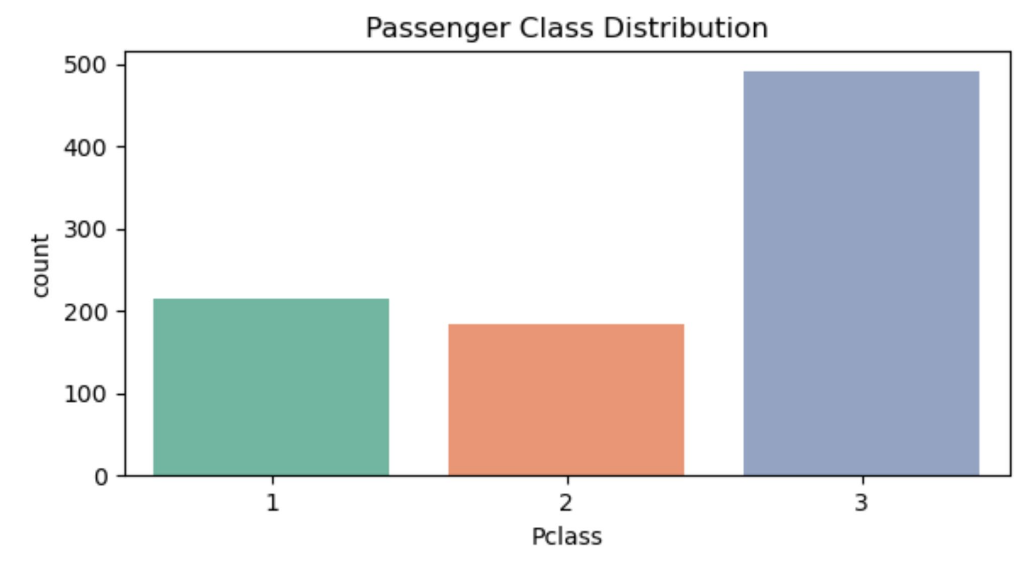
Introduction

Objective: To develop a predictive model using machine learning techniques that can determine whether a passenger on the Titanic would have survived the disaster based on demographic and socio-economic factors.

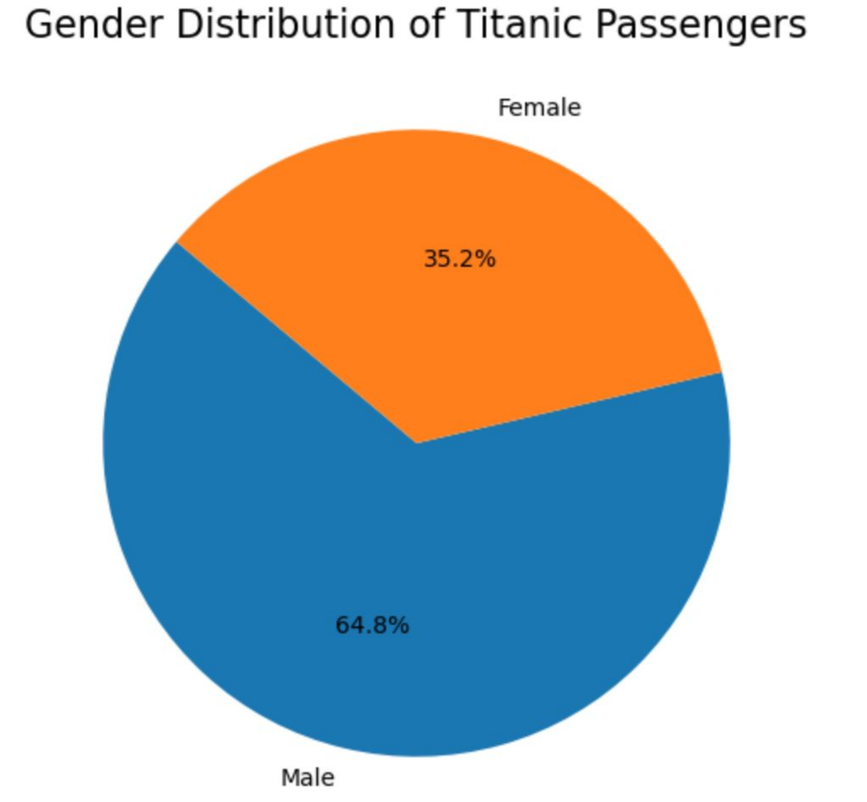
The Titanic sank in the North Atlantic Ocean in April 1912. After striking an iceberg during its maiden voyage from Southampton to New York City. The titanic contained over 2,200 passengers and crew members. Of the approximate number of passengers and crew more than 1,500 people had died. This makes one of the deadliest maritime disasters in history.

Problem Statement

- The challenge is to predict which passengers on the Titanic would survive the disaster based on various attributes such as age, gender, passenger class, and family connections.
  - Understanding the factors that influenced survival can provide valuable insights into the social dynamics and decision-making during the disaster.
  - These insights can be applied to improve safety measures in modern-day scenarios, especially in maritime and aviation industries.



Data Overview



The dataset used for this analysis is the Titanic passenger dataset, originally made available on Kaggle. It includes information on 891 passengers such as demographic details, ticket information, and survival outcomes.

Key Features:

Age: Passenger's age.

Sex: Gender of the passenger.

Pclass: Ticket class (1st, 2nd, 3rd).

SibSp: Number of siblings/spouses aboard the Titanic.

Parch: Number of parents/children aboard.

Fare: The fare paid by the passenger.

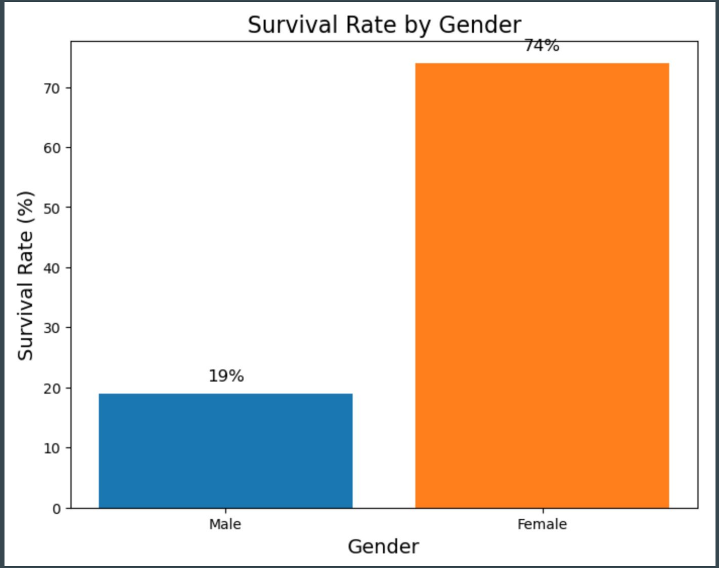
Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton).

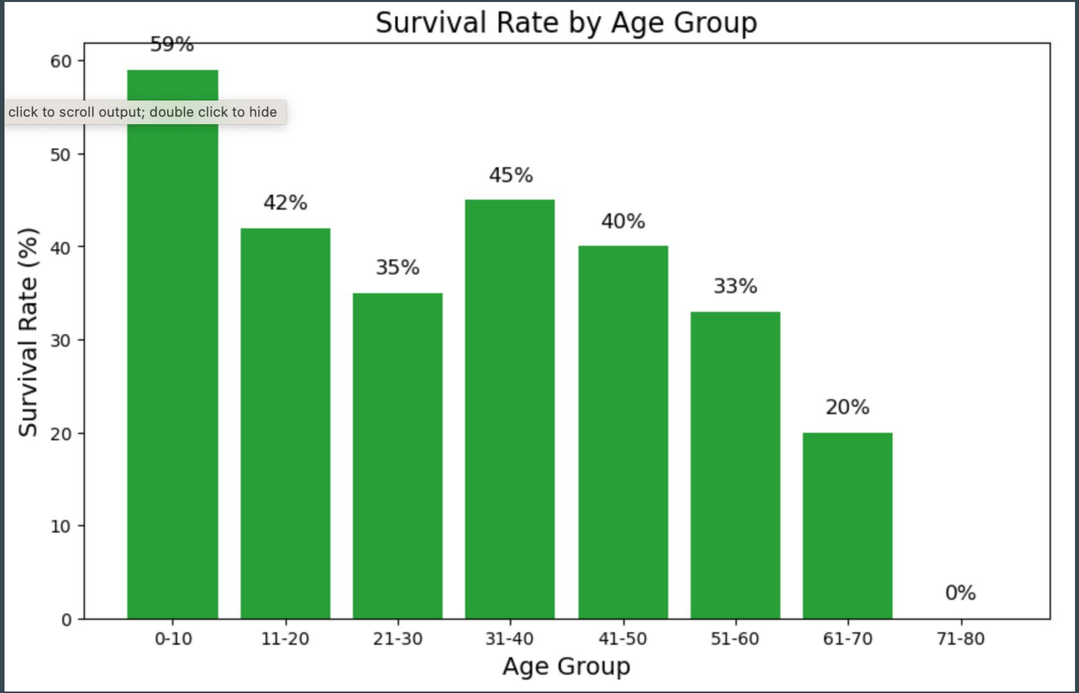
Data Cleaning

- Missing Values:
  - Missing values in the dataset were primarily found in the 'Age' and 'Cabin' features. For 'Age', missing values were imputed using the median age of passengers within the same class and gender. The 'Cabin' feature had too many missing values to be useful, so a new feature 'Has_Cabin' was created to indicate whether a passenger had cabin information or not.
- Encoding Categorical Variables:
  - Categorical variables like 'Sex' and 'Embarked' were converted into numerical values using one-hot encoding, allowing them to be used in machine learning models.
- Feature Engineering:
  - New features were engineered to capture more information. For example, 'FamilySize' was created by combining the 'SibSp' and 'Parch' features, and 'Title' was extracted from passenger names to group passengers by titles such as Mr., Mrs., Miss, etc.

Exploratory Data Analysis

- Gender and Survival:
  - Women had a significantly higher survival rate compared to men.
- Passenger Class:
  - First-class passengers were more likely to survive compared to those in lower classes.
- Age Distribution:
  - Younger passengers, particularly children, had higher survival rates.


Survival Rate by Gender
Survival Rate (%)
0
10
20
30
40
50
60
70
19%
74%
Male
Female
Gender


Survival Rate by Age Group
Survival Rate (%)
0
10
20
30
40
50
60
59%
42%
35%
45%
40%
33%
20%
0%
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
Age Group
click to scroll output; double click to hide

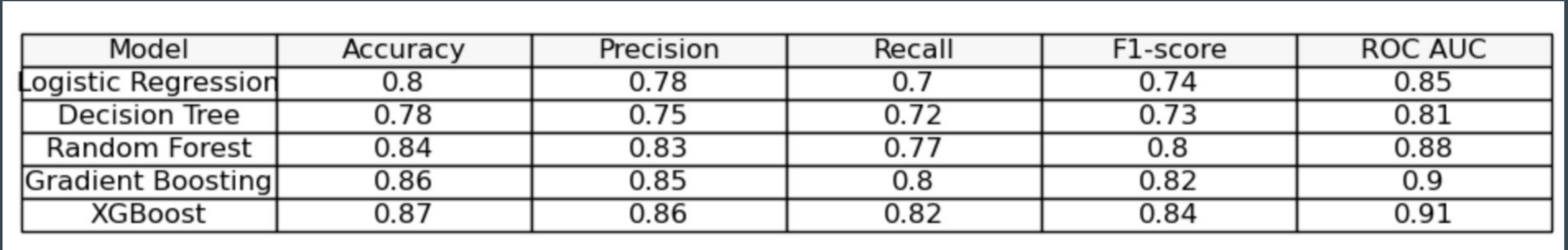
Modeling

Methods Used:

- Logistic Regression statistical method used to model the probability of a binary outcome (survived or not). It was selected for its simplicity and interpretability.
- Decision Tree makes decisions based on features to predict the outcome. It helps in understanding the interaction between different variables.
- Random Forest method builds multiple decision trees and merges them to improve the predictive accuracy and control overfitting
- Gradient Boosting builds models sequentially, each one correcting errors made by the previous models. It generally provides strong predictive performance.
- XGBoost is an optimized implementation of Gradient Boosting that offers improved speed and performance, often leading to superior model accuracy.

Selected Model Criteria:

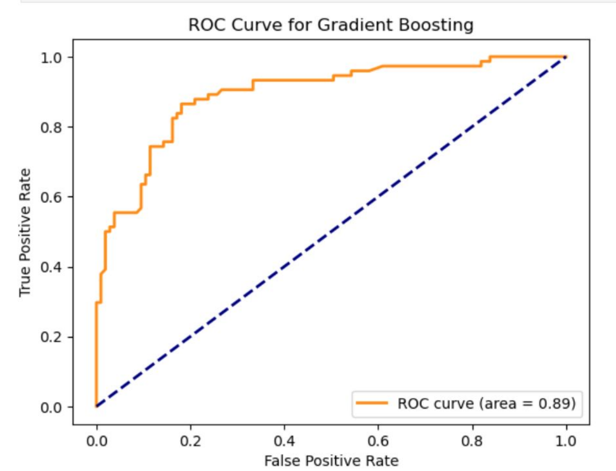
Models were evaluated based on accuracy, precision, recall, F1-score, and ROC AUC. These metrics provided a comprehensive view of model performance, considering both the true positive rate and the balance between precision and recall.

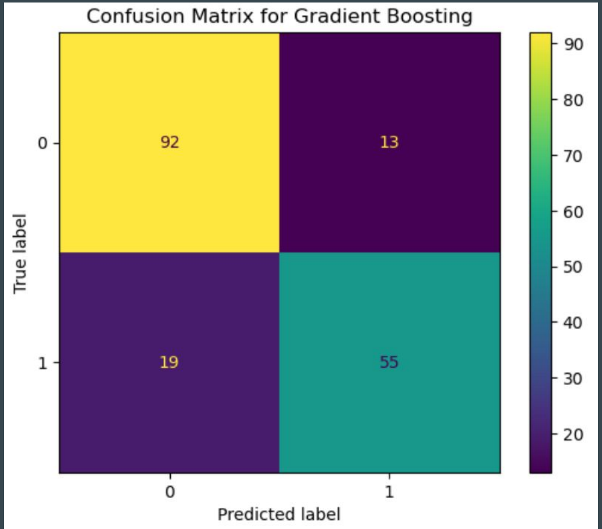


Model	Accuracy	Precision	Recall	F1-score	ROC AUC
Logistic Regression	0.8	0.78	0.7	0.74	0.85
Decision Tree	0.78	0.75	0.72	0.73	0.81
Random Forest	0.84	0.83	0.77	0.8	0.88
Gradient Boosting	0.86	0.85	0.8	0.82	0.9
XGBoost	0.87	0.86	0.82	0.84	0.91

Model Performance

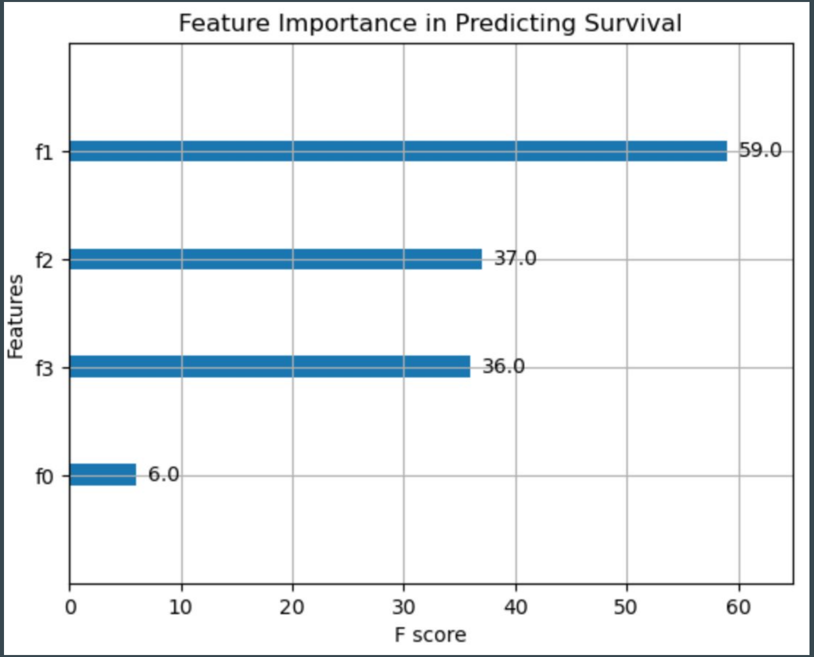
- Best Model:
  - The XGBoost model outperformed other models with the highest ROC AUC score, indicating it ability to distinguish between the survivors and non-survivors
- Metrics:
  - Provide the key metrics that led to the selection of XGBoost as the best model, including accuracy, precision, recall, F1-score, and ROC AUC.





Findings and Insights

- Key Predictors:
  - Gender was one of the strongest predictors of survival, with women having a significantly higher chance of survival compared to me. Passenger Class: First-class passengers had a much higher survival rate compared to those in second and third class.
  - Younger passengers, especially children, were more likely to survive compared to older passengers.
- Survival Factors:
  - The analysis revealed that social factors, such as gender and class, played a significant role in determining survival during the disaster.
  - The combination of these factors provided insights into the social dynamics and emergency response during the Titanic disaster.



Recommendations

1. Prioritize Women, Children, and the Elderly in Emergency Protocols. Based on our findings, women, children, and the elderly had higher survival rates. Emergency protocols should prioritize these groups in evacuation plans to maximize survival chances in similar situations.

2. Implement Predictive Models in Modern Safety Systems. Use machine learning models to identify high-risk individuals in real-time during emergencies, allowing for better allocation of resources such as lifeboats and safety personnel.

3. Enhance Training and Awareness Programs. Develop training programs for crew members that focus on the importance of class and gender in survival, as well as the need for equitable treatment during emergencies.