Predictive Modeling for Passenger Survival on the Titanic

Srikar Mukkamala
Student ID: 12141590

***Abstract*—This project involves the development of a predictive model to determine the likelihood of passenger survival on the Titanic. The dataset comprises two subsets: a training set with survival information for 891 passengers and a test set of 418 passengers without survival data. The goal is to predict the survival outcomes for the test set based on patterns observed in the training data. The implemented solution utilizes exploratory data analysis (EDA), feature engineering, and a Random Forest classifier. The predictive model is evaluated using stratified k-fold cross-validation, and its performance is assessed through ROC-AUC analysis.**

I. Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. In this project, I try to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (i.e. name, age, gender, socio-economic class, etc).

II. Code Structure

The project is implemented in a Python notebook using the Colab environment. It follows a systematic structure, including data loading, exploratory data analysis (EDA), handling missing values, feature engineering, and model development using a Random Forest classifier. The code is well-documented, utilizing comments and markdown cells for clear explanations.

III. Data Overview

A. Training Set

891 passengers with survival labels

B. Test Set

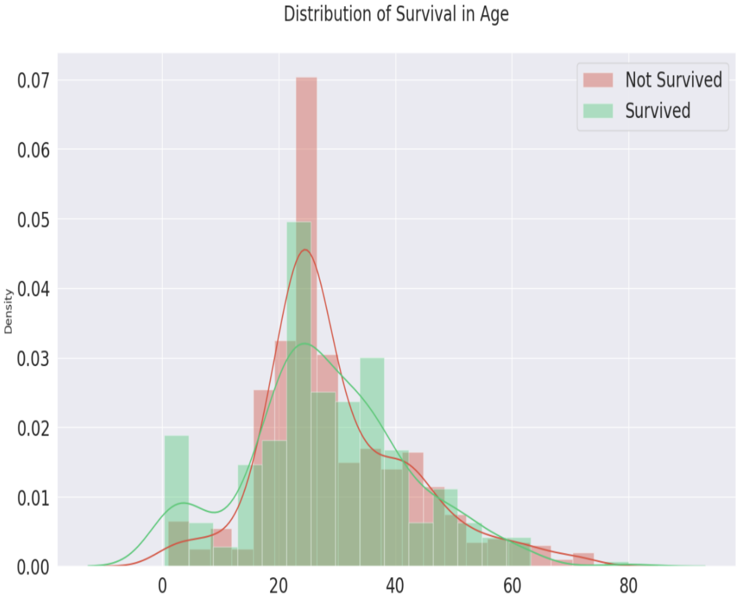
418 passengers without survival labels

C. Features

PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

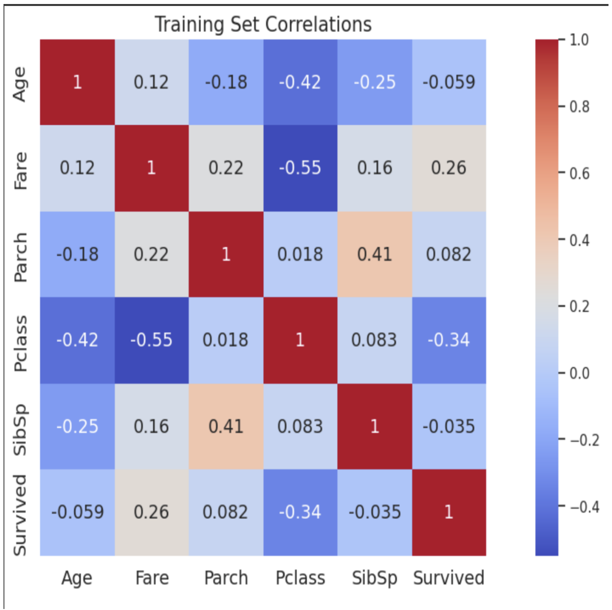
IV. Exploratory Data Analysis (EDA)

Comprehensive examination of data types, missing values, and basic statistics. Visualization of survival distribution, correlations, and target distribution in features. Analysis of continuous and categorical features to understand their impact on survival.



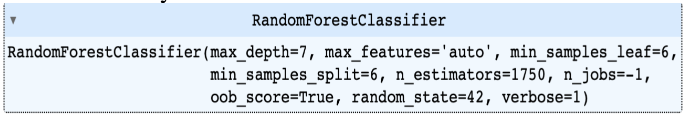
V. Feature Engineering

Binning of continuous features (Age, Fare) to capture patterns. Frequency encoding for Family Size and Ticket Frequency. Creation of new features: Family Survival Rate, Ticket Survival Rate, Survival Rate, etc. Label encoding and one-hot encoding for categorical features.



VI. Model Development

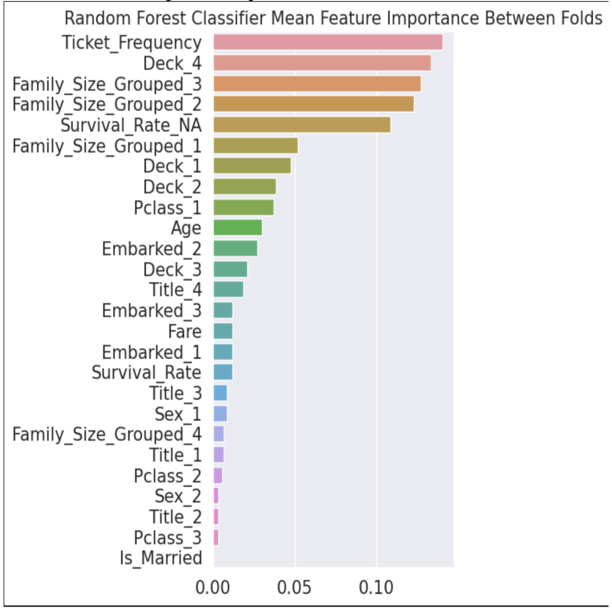
Utilization of a Random Forest classifier for prediction. Hyperparameter tuning for optimal model performance. Implementation of stratified k-fold cross-validation for model evaluation. Calculation of ROC-AUC scores for assessing model accuracy.

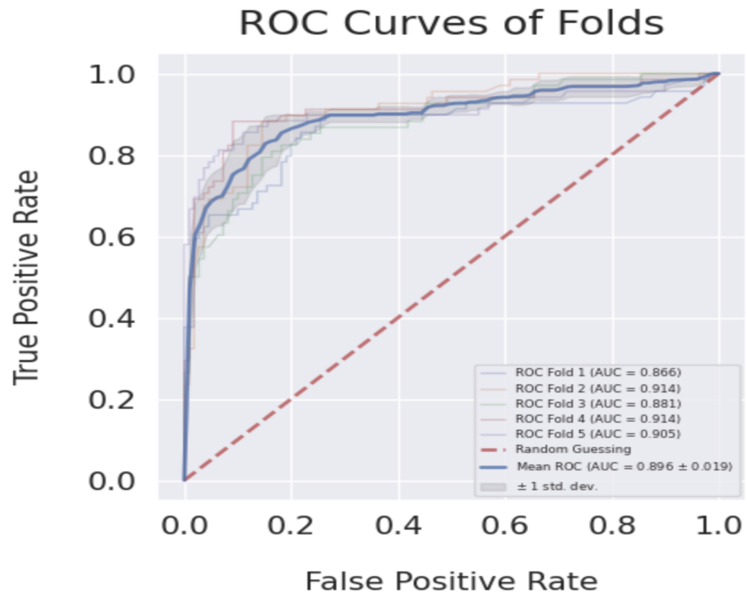


```
▾                         RandomForestClassifier
RandomForestClassifier(max_depth=7, max_features='auto', min_samples_leaf=6,
                       min_samples_split=6, n_estimators=1750, n_jobs=-1,
                       oob_score=True, random_state=42, verbose=1)
```

VII. Results

Achieved a strong ROC-AUC score through Random Forest classification. Feature importance analysis for better interpretability. Robust handling of missing data through imputation. Effective transformation of categorical features for model compatibility.





VIII. Conclusion

The developed model demonstrates promising predictive capabilities for passenger survival on the Titanic. The combination of feature engineering techniques and a Random Forest classifier yields accurate and reliable predictions. The project provides valuable insights into the factors influencing survival, contributing to the broader field of predictive modeling in data science.

The GitHub Link for the whole project is available here.

IX. Recommendations and Future Work

Explore additional feature engineering techniques for enhanced model performance. Investigate alternative machine learning algorithms for comparison. Conduct further analysis on misclassified instances to identify potential improvements. Consider ensemble methods to combine multiple models for increased robustness.

X. Acknowledgments

The author acknowledges the dataset source and relevant libraries used in the analysis and modeling process.

References

[1] A. Singh, S. Saraswat and N. Faujdar, ”Analyzing Titanic disaster using machine learning algorithms,” 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017, pp. 406-411, doi: 10.1109/CCAA.2017.8229835.

[2] Anasuya Dasgupta, Ved Prakash Mishra, Sanjiv Jha, Bhopendra Singh, Vinod Kumar Shukla, ”Predicting the Likelihood of Survival of Titanic's Passengers by Machine Learning”, 2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), pp.52-57, 2021.

[3] Burcu DURMUŞ, Öznur İŞÇİ GÜNERİ, ”Analysis and detection of Titanic survivors using generalized linear models and decision tree algorithm”, International Journal of Applied Mathematics Electronics and Computers, vol.8, no.4, pp.109, 2020.

[4] Shaurya Khanna, Shweta Bhardwaj, Anirudh Khurana, ”Titanic Data Analysis by R Data Language for Insights and Correlation”, Emerging Trends in Expert Applications and Security, vol.841, pp.73, 2019.

[5] https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8
[6] https://www.math.lsu.edu/system/files/Titanic