Titanic Examination

Zachery Davis

Introduction:

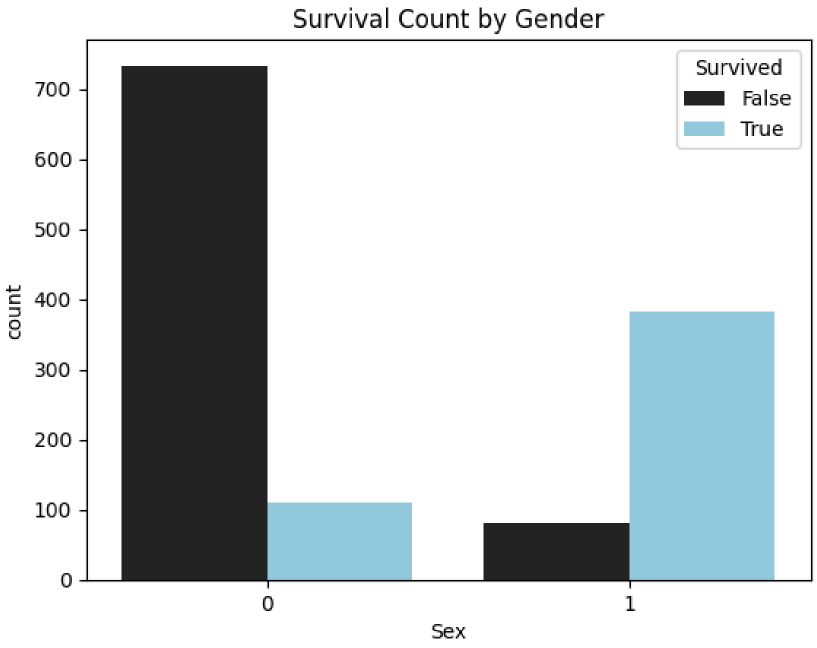
The sinking of the Titanic in 1912 is a very well-known tragedy, resulting in the loss of more than 1,500 lives. Many factors influence those who survive, whether it be social class, gender, and age. This project aims to build a classification model that predicts whether a passenger survived the disaster based on available features such as gender, age, ticket class, family size, and cabin information.

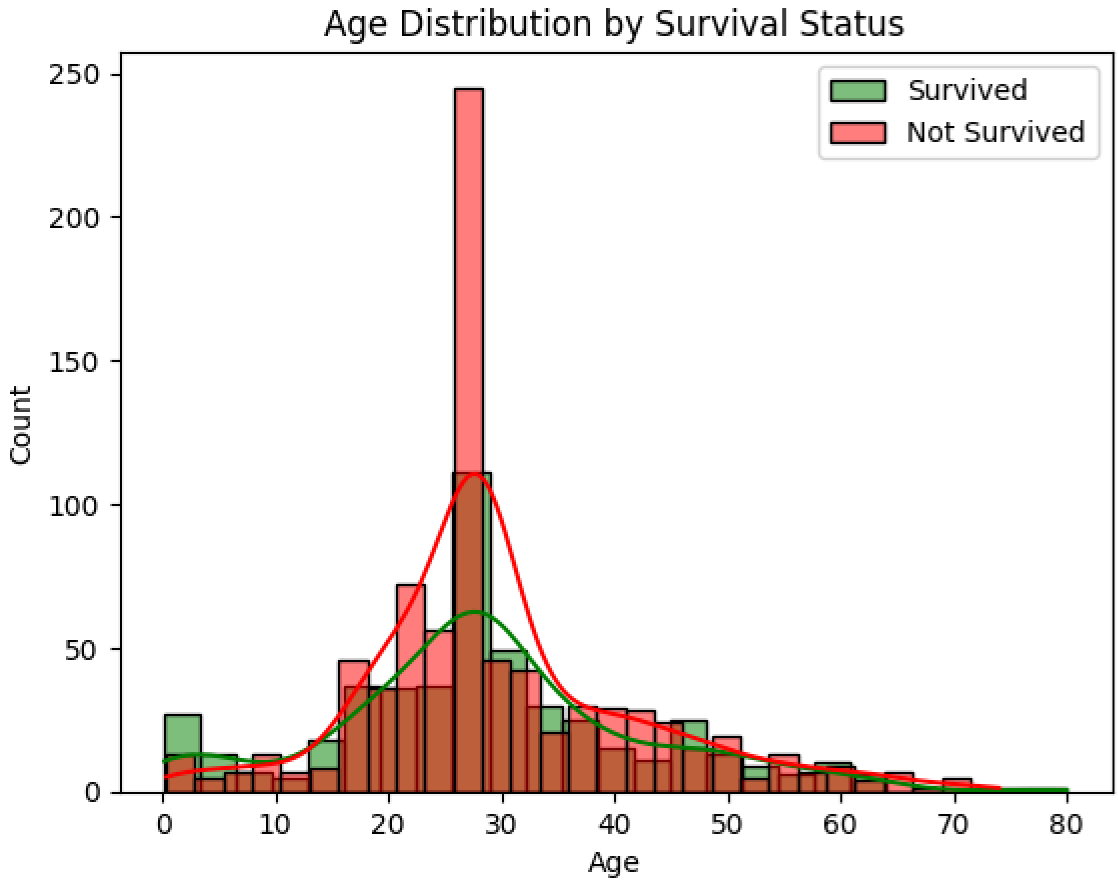
This problem of survival is of binary classification. In this project I seek to answer: Did women and children have a higher survival rate than others? Did socio-economic status, as indicated by ticket class, play a significant role in determining survival? Additionally, did the passenger's family size affect their chances of survival, and how did cabin location affect their likelihood of survival? Alongside these questions, this project also tackles the challenge of handling imbalanced data, as the dataset contains more passengers who did not survive than those who did.

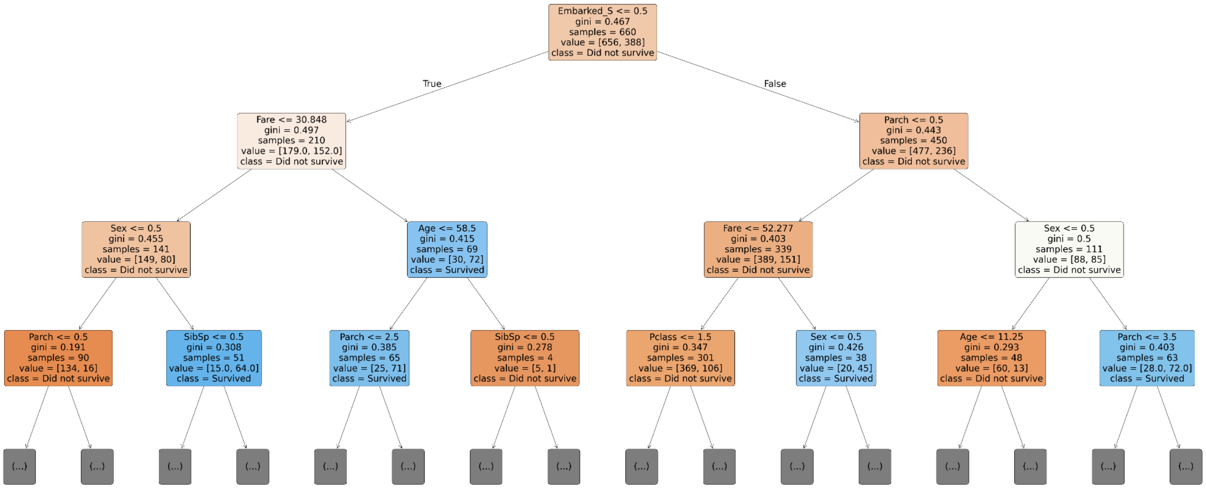
Pre-processing;

To prepare the Titanic dataset for analysis, the two separate CSV files of training and test were combined along with the survival data for the testing set. Then, the missing values for three passengers for Fare and Embarked were dropped. Missing values in the Age column were filled using the median age. Sex was converted into an integer and Survived to boolean values. To examine the cabin data a subset of the data has to be taken since there are so many missing values for this feature.

Visualization:







Visualization revealed that gender, class, and family size strongly correlated with survival on the Titanic. Women and higher-class passengers had significantly higher survival rates, due to "women and children first" procedures and physical proximity to lifeboats. Family size also influenced survival odds, with those in small groups having an advantage over those alone or with large families, highlighting the social ties and logistical challenges. These patterns informed model selection, leading to Random Forests as the primary choice for classification due to their ability to capture complex, non-linear relationships. While Logistic Regression struggled with linearity assumptions, and Decision Trees risked overfitting, Random Forests balance interpretability with robust accuracy by averaging predictions across multiple trees. Model evaluation using accuracy and F1 scores indicated an overall accuracy above 80%. The confusion matrix showed the model handled non-survivors well but required tuning to classify survivors better. By carefully selecting features, visualizing patterns, and refining model parameters, the Random Forest model effectively captured the historical factors influencing survival on the Titanic.

Storytelling:

This project has revealed that survival on the Titanic was heavily influenced by social structures and individual attributes such as gender, class, and age. Women and first-class passengers, even in a disaster scenario, had distinct advantages, reflecting historical social biases. The significance of family size in survival odds also adds depth to our understanding, as social ties may have played both protective and restrictive roles. Through data analysis and modeling, we were able to identify these patterns clearly, answering our initial question about which factors most impacted survival. The Random Forest model, in particular, underscored these findings by assigning high-importance scores to gender, age, and class features.

Impact Section:

While this analysis primarily serves as an academic exercise, it also provides a view of social implications. By examining survival factors, we gain insights into how historical societal structures shaped life-or-death outcomes in crises. This knowledge underscores the need for ethical considerations when examining demographic impacts, as certain factors should not dictate survival probabilities. Additionally, though this project has no direct negative impact, its interpretation can inform discussions around social equity in current emergency procedures. Finally, predictive modeling based on sensitive attributes must always be handled responsibly, ensuring respect for historical and ethical boundaries.

References:

Code: https://github.com/Zachery-Davis/Data-Mining-Project-2
Dataset: Titanic - Machine Learning from Disaster | Kaggle