HM 3 Writeup

Xiang Li

Introduction

The sinking of the Titanic in 1912 remains one of the most tragic maritime disasters in history. Understanding the factors that influenced survival on the Titanic has long been of interest to researchers and historians. In this study, we aim to predict survival probabilities using predictive modeling techniques and compare the performance of different models. We explore the importance of various features such as gender, age, and others in determining survival outcomes.

Data Description

The dataset used in this analysis contains information about Titanic passengers, including features such as passenger class, gender, age, number of siblings/spouses aboard, number of parents/children aboard, passenger fare, and port of embarkation. The target variable is binary, indicating whether a passenger survived or not. The dataset was cleaned by removing missing values and transforming categorical variables into numerical representations.

Variable Selection and Data Transformations

The selection of variables was based on prior knowledge and historical context. We included features such as passenger class and gender, which are known to have influenced survival probabilities due to the "women and children first" policy. Age was also considered important, as younger and older passengers may have had different chances of survival. Additionally, variables like family size (SibSp and Parch), fare, and port of embarkation were included for their potential impact on survival.

Modeling Approach

We employed several modeling techniques to predict survival probabilities:

1. **Decision Tree (CART) Model:** A baseline decision tree model was built to classify passengers into survived or not survived categories. This model provided insights into the importance of different features in predicting survival.
2. **Pruned Tree:** To mitigate overfitting, we pruned the decision tree using cost complexity pruning and evaluated its performance on the test set.
3. **Probit Model:** A Probit regression model was used to estimate survival probabilities based on the features. This model allowed us to analyze the marginal effects of each feature on survival probability.
4. **Random Forest Model:** Finally, we employed a random forest classifier to predict survival probabilities. Random forests aggregate multiple decision trees to improve predictive accuracy and robustness.

Results and Discussion

The baseline decision tree model achieved an accuracy of 81.3% on the test set, with gender and age emerging as the most important predictors of survival. The Probit model provided insights into the marginal effects of features, indicating that being male was associated with a lower probability of survival. After pruning, the decision tree model's accuracy improved to 89.5%, demonstrating the effectiveness of pruning in reducing overfitting. The random forest model achieved an accuracy of 85% and identified gender and age as significant predictors of survival, consistent with previous findings.

In conclusion, our study demonstrates the effectiveness of predictive modeling techniques in predicting survival probabilities on the Titanic. Decision trees, Probit regression, and random forests offer valuable insights into the factors influencing survival outcomes. While decision trees provide interpretability, Probit regression offers insights into marginal effects, and random forests offer improved accuracy.