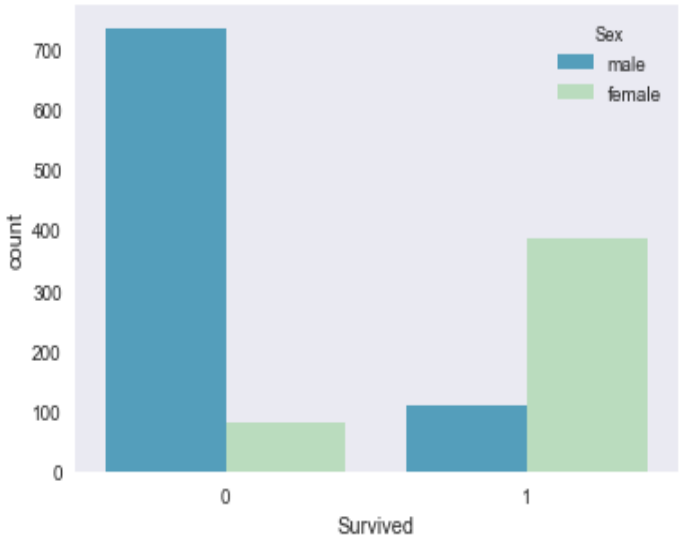
Dorcas Munoko
Big Data and Forecasting
2/12/2022

Introduction

The aim of the project was to build a tree classifier that predicts the probability of death in the Titanic. The train data contains 891 observations while the test data contains 418 observations, with main features including passenger id, survived, passenger class, age, cabin, sibling and spouse, parent and child, fare, sex, and embarked. The first step that I took in data pre-processing was to combine train and test datasets. I then checked for missing values in the combined dataset and replaced them with mean for numeric dataset and mode for categorical dataset. I also dropped features like cabin, name, ticket, and passenger Id, that were poor predictors of survival. I used log to transform the fare data column to uniform data distribution because the data was positively skewed. The last step in pre-processing was converting data type to integer to enable the model to run well.

Description of Features



Third class had the highest number of passengers (700), followed by first class (330), while second class had the lowest number of passengers (290). The ship had more male (843) than female (450), while more passengers embarked from city S (916). Approximately 800 did not survive, majority being male (730). Most of these deaths occurred in third class. I created a correlation matrix, and it indicates that survival had a negative correlation with age, passenger id, and passenger class. On the other hand, it had a positive correlation with fare, parent child, and sibling child.

Model, Results, and Conclusion

I built a decision tree model to estimate probability of survival in the Titanic. The variables I used include age, embarked, parent child, passenger class, sex, and sibling spouse. Sex (57%) and age (21%) were the strongest predictors of survival. I also carried out an accuracy analysis, and the model was 82% accurate in estimating the probability of survival in the Titanic. I carried out a 10-fold cross validation of the model, and it had a score of 82%. I compared the model to a simple probit classifier. Results indicate that the two classifier models had the same accuracy score in predicting the probability of survival (82%).