Predicting Titanic Survival Using Machine Learning

Mohibkhan Pathan

November 17, 2024

Abstract

The Titanic disaster is one of the most tragic and well-known events in modern history. Using machine learning, we analyze the Titanic dataset to predict passenger survival based on demographic and socioeconomic factors. This paper details the application of logistic regression for binary classification, evaluates the model's performance, and interprets its results. The study demonstrates that gender and socioeconomic class are the most significant predictors of survival, achieving consistent accuracy of 80%. Insights from this project highlight the effectiveness of machine learning in understanding historical datasets and suggest directions for further improvement.

1 Introduction

The sinking of the RMS Titanic on April 15, 1912, resulted in over 1,500 fatalities. With limited lifeboats and chaotic evacuation procedures, passenger survival was influenced by factors such as gender, age, socioeconomic class, and family relationships. The Titanic dataset, widely used in data science and machine learning, provides detailed information about passengers, enabling predictive modeling to assess survival probabilities. This paper describes a machine learning approach using logistic regression to predict survival outcomes.

2 Dataset Overview

The Titanic dataset, sourced from Kaggle, consists of two files: a training set and a test set. The training set includes survival labels (0 for non-survivors and 1 for survivors), while the test set omits these labels. Each dataset contains variables such as gender, age, passenger class (Pclass), number of siblings/spouses aboard (SibSp), number of parents/children aboard (Parch), ticket fare (Fare), and port of embarkation (Embarked). The dataset also contains missing values and imbalanced classes, which require preprocessing.

3 Methodology

3.1 Data Preprocessing

Preprocessing involved handling missing values, encoding categorical variables, and dropping irrelevant columns. Missing values in the `Age` and `Fare` features were imputed using

their median values, while the most common value (`'S'`) was used to fill missing values in the `Embarked` column. The `Sex` and `Embarked` variables were label-encoded, with `Sex` converted to binary values (0 for Female, 1 for Male) and `Embarked` converted to numerical labels (0 for Cherbourg, 1 for Queenstown, 2 for Southampton).

3.2 Model Training

Logistic regression, a widely used algorithm for binary classification, was employed to predict survival outcomes. The model was trained on the preprocessed training set, with `Survived` as the target variable. Cross-validation was performed to evaluate the model's generalizability.

4 Results

The logistic regression model achieved an accuracy of 80% on the training data and a mean cross-validation accuracy of 79%. Key evaluation metrics are summarized as follows:

- **Precision:** The precision for non-survivors (class 0) was higher than for survivors (class 1), reflecting the model's ability to handle the majority class effectively.
- **Recall:** Recall was also higher for non-survivors, indicating that the model correctly identified most non-survivors but struggled slightly with survivors.
- **F1-Score:** The F1-score balanced precision and recall, with values of 84% for class 0 and 73% for class 1.

5 Feature Importance

Analysis of feature importance revealed that gender (`Sex`) and socioeconomic class (`Pclass`) were the most significant predictors of survival. Female passengers and first-class passengers had higher survival probabilities. Other features, such as `SibSp` and `Embarked`, contributed moderately, while features like `Fare` and `Age` had minimal impact.

6 Discussion

The results highlight the importance of demographic and socioeconomic factors in predicting survival. The model performed well on the training and cross-validation datasets, demonstrating its reliability. However, the imbalanced nature of the dataset resulted in lower recall for survivors. Addressing this issue through techniques such as oversampling, undersampling, or using class weights could improve performance.

7 Conclusion

This study demonstrates the application of logistic regression for survival prediction on the Titanic dataset. The model achieved consistent accuracy and provided interpretable insights into the factors affecting survival. Future work could explore advanced models,

such as Random Forest or Gradient Boosting, and address class imbalance to enhance prediction performance further. This project underscores the potential of machine learning in extracting meaningful patterns from historical datasets.