Titanic Dataset Analysis: Predicting Passenger Survival Using Logistic Regression

Divya Dhole
University of Arizona
Divyadhole@arizona.edu

Abstract

This report presents an analysis of the Titanic dataset using machine learning techniques. The primary goal is to predict passenger survival based on various features such as age, gender, and ticket class. We employ logistic regression for this binary classification task and evaluate the model's performance using standard metrics.

1 Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. This project aims to analyze the factors that influenced passenger survival and build a predictive model. By examining various passenger attributes, we seek to understand the patterns that contributed to survival rates and develop a model that can accurately predict survival outcomes.

2 Dataset

The dataset used for this analysis is the Titanic dataset, which contains information about 891 passengers aboard the Titanic. Each entry includes various features such as:

- Passenger class (Pclass)
- Sex
- Age
- Number of siblings/spouses aboard (SibSp)
- Number of parents/children aboard (Parch)
- Ticket fare
- Port of embarkation (Embarked)

The target variable is 'Survived', indicating whether a passenger survived (1) or not (0).

3 Methodology

3.1 Data Preparation

- Created a data directory to store the dataset
- Downloaded the Titanic dataset from a GitHub repository
- Loaded the dataset using pandas

3.2 Data Preprocessing

- Removed less relevant features: 'Name', 'Ticket', 'Cabin'
- Handled missing values:
  - Filled missing 'Age' values with the median age
  - Dropped rows with missing 'Embarked' values
- Encoded categorical variables:
  - Mapped 'Sex' to numerical values (male: 0, female: 1)
  - Mapped 'Embarked' to numerical values (C: 0, Q: 1, S: 2)

3.3 Model Training

- Split the dataset into training (80%) and testing (20%) sets
- Trained a Logistic Regression model using scikit-learn

3.4 Evaluation Metrics

- Evaluated model performance using:
  - Accuracy
  - Classification report (precision, recall, F1-score)

4 Results and Discussion

The Logistic Regression model achieved the following performance on the test set:

Model Evaluation Report:

- Class 0 (Non-survivors):

  Precision: 0.81 Recall: 0.87 F1-score: 0.84 Support: 106 samples

- Class 1 (Survivors):

  Precision: 0.78 Recall: 0.70 F1-score: 0.74 Support: 73 samples

- Overall Performance:

  Accuracy: 0.80 (80Total samples: 179

- Macro Average:

  Precision: 0.80 Recall: 0.78 F1-score: 0.79

- Weighted Average:

  Precision: 0.80 Recall: 0.80 F1-score: 0.80

- Overall Accuracy: 0.80

Key observations:

- The model achieved an overall accuracy of 80% on the test set.
- The model performed slightly better in predicting non-survivors (class 0) with higher precision and recall.
- There is a balanced performance across both classes, indicating that the model is not heavily biased towards one class.

The visualizations (not shown in this report due to limitations) provided additional insights:

- A bar chart showing the count of survivors vs. non-survivors
- A count plot of survival by gender
- A line plot of average age by survival status

These visualizations help in understanding the relationships between various features and survival rates.

5 Conclusion

This analysis successfully classified Titanic passengers into survivors and non-survivors using a Logistic Regression model. The model's performance, with an 80% accuracy, demonstrates its effectiveness in predicting survival based on the given features.

Key findings include:

- The model's ability to predict both survivors and non-survivors with reasonable accuracy
- The importance of features such as gender, passenger class, and age in determining survival probability

.