Survival Prediction on the Titanic Dataset

Prepared by: Vishnu N
Date: 21/11/2024
Affiliation: Data Science Enthusiast at Government College of Technology
LinkedIn: https://www.linkedin.com/in/vishnu-n-121108293?utm_source=share&utm_campaign=share_via&utm_content=profile&utm_medium=android_app

Abstract

The Titanic disaster of 1912 remains one of the most infamous maritime tragedies. This project aims to predict passenger survival using machine learning models trained on the Titanic dataset. The dataset was pre-processed to handle missing values, normalize numerical features, and encode categorical variables. Various models, including Logistic Regression, Random Forest, and AdaBoost, were employed and optimized using GridSearchCV. The Random Forest model achieved the highest accuracy of 83.4%, followed by AdaBoost at 85.6% with hyperparameter tuning. This project demonstrates the power of data-driven insights in understanding historical events and predicting outcomes.

Introduction

The sinking of the Titanic is a historical tragedy that has fascinated researchers and the public for decades. Among the 2,224 passengers and crew, only 710 survived, raising questions about the factors influencing survival. Using the Titanic dataset, this project aims to predict survival probabilities based on passenger demographics, ticket class, and other features. By applying machine learning techniques, we explore the relationship between these factors and survival, identify key predictors, and develop a model to provide reliable predictions.

The dataset, sourced from Kaggle, consists of 891 records with features such as passenger age, gender, class, and fare. The primary objective is to preprocess the data, apply machine learning models, and optimize their performance through hyperparameter tuning.

Exploratory Data Analysis (EDA)

The dataset was analyzed to identify key patterns and relationships between variables:

1. **Gender and Survival**:
   - Females had a significantly higher survival rate than males.
   - **Visualization**: A count plot showed females were more likely to survive.
2. **Passenger Class and Survival**:
   - Passengers in first class had a higher survival rate compared to those in lower classes.

   - **Visualization**: A bar plot depicted survival rates by class.

3. **Age Distribution**:
   - The average passenger age was 29.7 years.
   - Children and younger passengers had higher survival probabilities.
4. **Missing Values**:
   - 177 missing values in the Age column were filled with the mean.
   - 2 missing values in Embarked were filled with the mode.

Key graphs were plotted using Seaborn and Matplotlib to visualize survival distributions across features like age, gender, and class.

Methodology

The project followed a structured approach:

1. **Data Preprocessing**:
   - Removed irrelevant columns (PassengerId, Name, Ticket, Cabin).
   - Normalized Age and Fare using MinMaxScaler and StandardScaler.
   - Encoded categorical variables (Sex, Embarked) using one-hot encoding.
2. **Feature Engineering**:
   - Ensured all features were numerical for compatibility with ML models.
   - Created dummy variables for categorical features.
3. **Model Selection and Training**:
   - **Models Used**: Logistic Regression, Random Forest, AdaBoost.
   - **Training**: Split the dataset into 75% training and 25% testing sets.
4. **Hyperparameter Tuning**:
   - Used GridSearchCV to find the best parameters for each model.
   - Scoring was based on accuracy and other relevant metrics.
5. **Prediction Function**:
   - Implemented a function to predict survival based on user-input features, making the model interactive.

Results

The performance of different models was evaluated based on accuracy:

Tabular Representation of best Parameters:

Model	Best Parameters (via GridSearchCV)	Accuracy on Test Set
Logistic Regression	{'C': 1.0, 'solver': 'lbfgs', 'penalty': 'l2'}	82.0%
Random Forest	{'criterion': 'entropy', 'max_depth': 10, 'n_estimators': 50}	83.4%
AdaBoost	{'algorithm': 'SAMME.R', 'learning_rate': 0.01, 'n_estimators': 200}	85.6%

Insights:

- Ensemble methods like Random Forest and AdaBoost outperformed Logistic Regression.
- Hyperparameter tuning significantly improved model performance, especially for AdaBoost.

Discussion

Key Observations:

1. **Important Features**:
    - Gender, passenger class, and fare were the most influential factors for survival.
    - Females and first-class passengers had higher survival rates.
2. **Model Performance**:
    - AdaBoost achieved the highest accuracy due to its iterative boosting mechanism.
    - Random Forest performed well but slightly underperformed AdaBoost.
3. **Challenges**:
    - Limited dataset size restricted model training.
    - Missing values required careful imputation to avoid data loss or bias.

Limitations:

- The dataset lacks additional features like health conditions or travel companions, which might further influence survival.

Future Work:

- Experiment with more advanced models (e.g., XGBoost, Gradient Boosting).
- Use feature selection techniques to identify the most impactful features.

Conclusion

This project successfully predicted Titanic passenger survival using machine learning techniques. By preprocessing the data and applying models like Logistic Regression, Random Forest, and AdaBoost, we achieved an accuracy of up to 85.6%. The results highlight the significance of gender, class, and fare in survival predictions.

Future iterations could incorporate larger datasets, additional features, or advanced models for further improvement. This project demonstrates the value of machine learning in understanding historical events and deriving actionable insights.