Titanic Survival Prediction Project

Introduction and Background

In 1912, the RMS Titanic tragically sank after hitting an iceberg during its maiden voyage, resulting in a massive loss of life. This project seeks to analyze passenger data from the Titanic to predict survival outcomes based on various passenger characteristics, such as socio-economic status, age, and gender. By leveraging data science techniques, the objective is to build a predictive model that highlights the most influential factors in determining a passenger's likelihood of survival.

Problem Statement

The primary aim of this project is to develop a machine learning model that accurately predicts the survival status of Titanic passengers. By identifying key factors that contributed to survival, this analysis could provide valuable insights into demographic and socio-economic elements that influence outcomes in life-threatening scenarios.

Objectives

The project objectives are as follows:

1. Data Exploration and Preprocessing: Understand and clean the dataset by handling missing values, transforming categorical features, and identifying essential predictors like socio-economic class and family relations.

2. Feature Engineering: Enhance model accuracy by creating and selecting features that carry the most predictive power.

3. Model Selection and Training: Test various classification algorithms (e.g., Decision Tree, Random Forest, Logistic Regression) to identify the best model for survival prediction.

4. Hyperparameter Tuning: Improve model performance through tuning, such as using RandomizedSearchCV to optimize parameters.

5. Model Evaluation: Assess the model’s accuracy and reliability, comparing predictions with baseline accuracy.

Data Preprocessing

The dataset included several critical features, such as:

- **Pclass**(Passenger socio-economic class)

- **Age** (Passenger age)

- **Sex** (Gender)

- **SibSp** (Number of siblings/spouses on board)

- **Parch** (Number of parents/children on board)

- **Fare** (Fare paid)

- **Embarked** (Port of embarkation)

The following steps were taken in data preprocessing:

- **Missing Value Handling:** Features with missing values, like Age and Embarked, were imputed using the mean and mode, respectively, to maintain dataset completeness.

- **Feature Encoding:** Categorical variables such as Sex and Embarked were encoded to numerical values using OneHotEncoding, making them suitable for model input.

- **Scaling:** Continuous variables, such as Age and Fare, were scaled to ensure that the model treats each feature on a comparable scale.

Exploratory Data Analysis

The analysis revealed several insights into the survival distribution:

- **Gender:** Female passengers had a significantly higher survival rate than males.

- **Passenger Class:** Higher class passengers (Pclass = 1) had better survival rates, likely due to their proximity to lifeboats and faster access to resources.

- **Family Size:** Passengers with a moderate family size had a higher chance of survival compared to those traveling alone or with large families.

Feature Engineering

Additional features were created to enhance model accuracy:

- **Family Size:** The sum of SibSp and Parch provided insights into survival likelihood, with mid-sized families showing a survival advantage.

- **Title Extraction:** Titles like Mr., Mrs., and Miss were extracted from names, adding predictive value based on passenger demographics.

Model Selection and Training

Several machine learning algorithms were evaluated, including:

- **Decision Tree**
- **Random Forest**
- **Support Vector Machines (SVM)**

After comparing performance, Random Forest was selected as the final model due to its high accuracy and stability on this dataset.

Hyperparameter Tuning

Using **RandomizedSearchCV**, the Random Forest model's hyperparameters, such as the number of estimators and max depth, were fine-tuned, resulting in optimized predictive performance.

Model Evaluation

The final model achieved an accuracy of **80.47%,** while the test set prediction accuracy was **80.13%.** These results indicate a strong and reliable model that successfully captures the key determinants of survival.

Conclusion

The Titanic Survival Prediction project successfully utilized data science methodologies to develop a predictive model with significant accuracy. The findings underline the importance of demographic and socio-economic factors in survival outcomes and reflect the user's ability to implement complex data preprocessing, feature engineering, and model optimization techniques to achieve high predictive accuracy.