Titanic Machine Learning Project

Overview

This project involves building machine learning models to predict passenger survival on the Titanic. The dataset used is the well-known Titanic dataset from Kaggle, and the project includes data preprocessing, feature engineering, model training, evaluation, and comparison.

Objectives

1. Explore and preprocess the Titanic dataset.
2. Engineer new features to improve model performance.
3. Train and evaluate multiple machine learning models.
4. Compare model performance based on various metrics.

Dataset

The dataset contains information about Titanic passengers, including demographic details, ticket information, and survival status. Key columns include:

- Survived: Target variable (1 for survived, 0 for did not survive).
- Pclass: Passenger class.
- Name: Passenger name.
- Sex: Gender of the passenger.
- Age: Age of the passenger.
- SibSp: Number of siblings/spouses aboard.
- Parch: Number of parents/children aboard.
- Ticket: Ticket number.
- Fare: Fare paid for the ticket.
- Cabin: Cabin number.
- Embarked: Port of embarkation.

Feature Engineering

New features were created to enhance model performance:

1.

isChild:

- **Description**: Indicates whether a passenger is a child (age < 18).
- **Justification**: Children might have a higher survival probability due to priority in rescue efforts.

Title:

- **Description**: Extracted titles (e.g., Mr., Miss, Rev., Master) from the Name column.
- **Justification**: Titles provide additional information about social status and gender.

Deck:

- **Description**: Extracted deck information from the Cabin column.
- **Justification**: Deck location could correlate with survival likelihood.

FamilySize:

- **Description**: Calculated as the sum of SibSp and Parch.
- **Justification**: Family size could influence survival chances as families might prioritize staying together.

Models

Three machine learning models were trained and evaluated:

- **Logistic Regression**
- **Random Forest Classifier**
- **Support Vector Machine (SVM)**

Evaluation Metrics

Models were evaluated using:

- Precision
- Recall
- F1-score
- Accuracy
- ROC-AUC score

Results

The following table summarizes the performance of the models:

Metric	Logistic Regression	Random Forest	SVM
Precision	0.82 (class 0), 0.75 (class 1)	0.84 (class 0), 0.78 (class 1)	0.69 (class 0), 0.64 (class 1)
Recall	0.87 (class 0), 0.68 (class 1)	0.88 (class 0), 0.71 (class 1)	0.89 (class 0), 0.34 (class 1)
F1-score	0.84 (class 0), 0.71 (class 1)	0.86 (class 0), 0.75 (class 1)	0.78 (class 0), 0.44 (class 1)
Accuracy	79%	82%	68%
ROC-AUC	0.7703	0.7965	0.6128

Combined Model Performance Report

Below is a consolidated performance report for the three models:

Logistic Regression

- **Precision**: High precision for class 0 indicates reliable prediction of non-survivors.
- **Recall**: Strong recall for class 0 but lower for class 1 suggests better identification of non-survivors.
- **F1-Score**: Balanced performance for both classes.
- **ROC-AUC Score**: Moderate discriminative ability (0.7703).

Random Forest Classifier

- **Precision**: Excellent precision for both classes, especially class 0.
- **Recall**: Strong recall for class 0, decent for class 1.
- **F1-Score**: Best overall F1-scores across models.
- **ROC-AUC Score**: Highest among all models (0.7965).

Support Vector Machine (SVM)

- **Precision**: Fair precision but lags behind other models.
- **Recall**: Highest for class 0, very low for class 1.
- **F1-Score**: Lowest performance, especially for class 1.
- **ROC-AUC Score**: Lowest among all models (0.6128).

Conclusion

The Random Forest model performed the best with the highest accuracy (82%) and ROC-AUC score (0.7965). Feature engineering significantly contributed to improving model performance by adding relevant predictive features.

Acknowledgments

- Kaggle for providing the Titanic dataset.
- Scikit-learn and Pandas for machine learning and data manipulation.

License