Titanic - Machine Learning from Disaster

Lab Report
Name: Marzuk Entsie Sanni
Submission Date: 1st January 3, 2025

---

1. Introduction

Objective

The goal of this project was to develop a predictive model to determine passenger survival based on the Titanic dataset. This exercise strengthened understanding of the end-to-end machine learning pipeline, including data exploration, preprocessing, feature engineering, model selection, and evaluation.

---

2. Data Exploration and Visualization

2.1 Dataset Overview

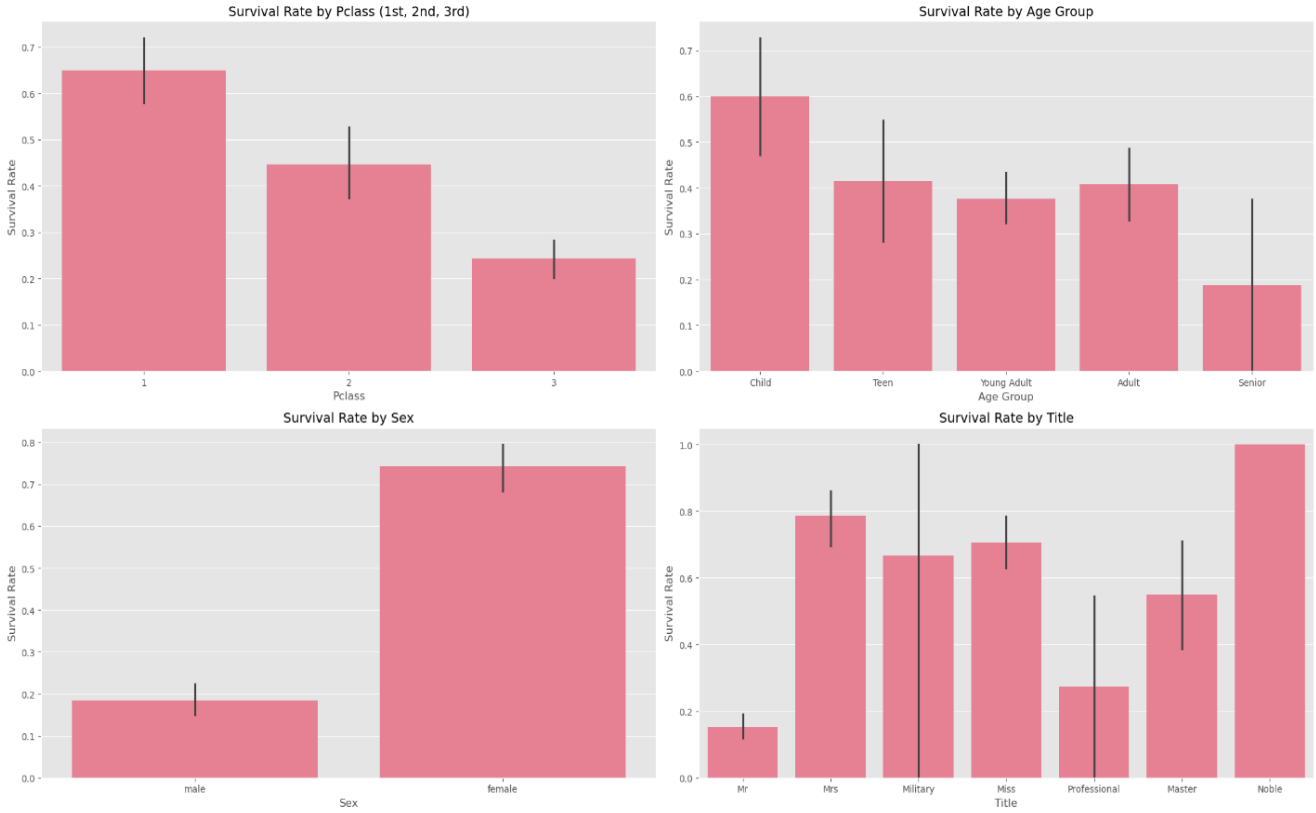
- **Source**: Titanic dataset (train.csv, test.csv).
- **Key Features**: Survival, Pclass, Age, Sex, SibSp, Parch, Ticket, Fare, Cabin, Embarked.

2.2 Key Statistics

- Summary statistics for numerical features (mean, median, standard deviation).

2.3 Visualizations

- Survival rates by:
    - Gender
    - Title
    - Age Groups
    - Sex



---

3. Data Cleaning and Preprocessing

3.1 Handling Missing Values

- Approach for handling missing values in Age, Embarked, and other columns.

3.2 Encoding Categorical Variables

- Conversion of categorical variables like Sex and Embarked to numerical values.

3.3 Normalization/Scaling

- Scaling techniques used for numerical features.

3.4 Train-Validation Split

- Details of the split (e.g., 80-20 split).

---

4. Feature Engineering

4.1 New Features and Justification

- Description of new features created (e.g., FamilySize, IsAlone, Title, AgeGroup).

- FamilySize : This feature captures the total number of family members aboard the Titanic (calculated as SibSp + Parch + 1). My justification for this is that family presence could significantly influence survival rates, as families might prioritize saving each other during the disaster
- IsAlone: Traveling alone might have impacted survival chances, as isolated individuals might have been less likely to receive assistance.
- AgeGroup: Categorized passengers into age groups (Child, Teen, Young Adult, Adult, Senior). Survival likelihood could vary across age demographics, with children and seniors possibly prioritized during evacuation.
- Title: Titles reflect social status and gender roles, both of which likely influenced survival rates.

4.2 Feature Selection

- Techniques used (e.g., correlation matrix, feature importance scores).
- Final selected features for model training.

---

5. Model Selection and Training

5.1 Models Trained

- Logistic Regression
- Random Forest
- Support Vector Machines

5.2 Cross-Validation and Metrics

- Table comparing model performance on validation data (Accuracy, Precision, Recall, F1-score, ROC-AUC).

	Accuracy	Precision	Recall	F1	ROC-AUC	CV Score
Logistic Regression	0.842105	0.756757	0.756757	0.756757	0.810284	0.791209
Random Forest	0.807018	0.692308	0.729730	0.710526	0.835732	0.789011
SVM	0.789474	0.658537	0.729730	0.692308	0.833801	0.802198

---

6. Model Optimization

6.1 Hyperparameter Tuning

- RandomizedSearchCV results for each model.

6.2 Final Optimized Model Performance

- Comparison of optimized model metrics.

	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.815789	0.735294	0.675676	0.704225	0.779396
Random Forest	0.842105	0.806452	0.675676	0.735294	0.798877
SVM	0.789474	0.658537	0.729730	0.692308	0.773956

---

7. Testing and Submission

7.1 Test Set Predictions

- Description of the best model used for test data predictions.

7.2 Submission File

- Marzuk_submission.csv.

---

8. Conclusion

Key Insights

- Factors that most influenced survival (e.g., gender, Title, Age, Fare).
- I faced a challenge of improving my models performance and I was able to improve it by generating new features and also adjusting threshold in a attempt to improve the model.

Future Work

- Possible improvements or additional analyses (e.g., ensemble models, deeper feature engineering).
