Titanic Kaggle Competition Machine Learning Project

1. Introduction

The Titanic Data Analysis project aims to predict passenger survival based on historical data using machine learning models. The dataset contains information about passengers, including their demographic details, travel class, and ticket fares. This project demonstrates the entire pipeline of a data science workflow, from data exploration and preprocessing to model training and evaluation.

Key Objectives:

- Analyze the dataset to gain meaningful insights.
- Preprocess the data and handle missing values.
- Use Feature Engineering new features to enhance model performance.
- Train and evaluate various machine learning models to predict survival.
- Use SHAP analysis to explain model predictions.
- Document the entire process in a comprehensive manner.

2. Concepts Overview

What is Data Exploration?

- Data exploration is the process of understanding the structure, content, and quality of a dataset before applying advanced analysis or modeling techniques.
- It helps identify patterns, relationships, missing values, and potential issues in the data.

What is Exploratory Data Analysis (EDA)?

- EDA involves using statistical and visual techniques to summarize the main characteristics of the data.
- **Why is EDA Important?**
    - Identifies trends, correlations, and anomalies.
    - Guides preprocessing decisions like handling missing values and outliers.
    - Ensures a better understanding of the dataset for informed model-building.

What is Feature Engineering?

- Feature engineering is the process of creating, modifying, or selecting the most relevant variables (features) from raw data to improve model performance.
- **Why is Feature Engineering Important?**
    - Transforms raw data into a format that machine learning models can understand.
    - Captures meaningful patterns and relationships in the data.
    - Enhances predictive power by adding domain knowledge.

Overview of Machine Learning Models Used

- **Logistic Regression**: A simple, interpretable algorithm for binary classification tasks.
- **Random Forest**: An ensemble method that combines multiple decision trees for better accuracy and robustness.
- **XGBoost**: A gradient boosting algorithm known for its efficiency and high performance on structured data.
- **Stacking Ensemble**: Combines multiple models to leverage their strengths, often yielding superior performance.

What is SHAP and Why is It Important?

- SHAP (SHapley Additive exPlanations) is a framework for explaining the predictions of machine learning models.
- **Why is SHAP Important?**
    - Helps understand how each feature contributes to a prediction.
    - Improves trust in machine learning models by making their decisions transparent.

- Identifies the most important features driving model prediction

2.1 Data Science Workflow

1. **Data Exploration**: Understanding the dataset, its structure, and distributions.
2. **Data Preprocessing**: Cleaning data, handling missing values, and transforming features.
3. **Feature Engineering**: Creating new meaningful features from existing ones.
4. **Model Training**: Training various machine learning models.
5. **Model Evaluation**: Comparing models based on accuracy, AUC-ROC, and other metrics.
6. **Explainability**: Using SHAP to interpret model decisions.

2.2 Dataset Description

- **Survived**: Target variable (0 = No, 1 = Yes).
- **Pclass**: Ticket class (1st, 2nd, 3rd).
- **Sex**: Gender of the passenger.
- **Age**: Age of the passenger.
- **SibSp**: Number of siblings/spouses aboard.
- **Parch**: Number of parents/children aboard.
- **Fare**: Ticket fare paid.
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

3. Implementation

Data Loading and Overview

- Describe loading the dataset and exploring its structure using functions like `head()`, `info()`, and `describe()`.
- **Visualizations**: Include images for the dataset overview (first five rows and missing value summary).

Exploratory Data Analysis (EDA)

- Summarize key insights from the dataset, such as distributions of numerical features and survival rates by various categories (e.g., gender, class, embarkation).
- **Visualizations**: Include histograms, bar charts, and pairplots.

Handling Missing Data

- Explain how missing values were handled, including imputation strategies for Age and `Embarked`, and the decision to drop or transform the `Cabin` column.
- **Visualizations**: Include the missing data heatmap.

Feature Engineering

- Detail the creation of new features like `Title`, `FamilySize`, `IsAlone`, and binning (`FareBin`, `AgeBin`).
- Explain why these features were relevant and how they improved the dataset for modeling.
- **Visualizations**: Include charts showing the distribution of engineered features.

Outlier Detection and Handling

- Describe techniques used to detect and handle outliers, such as boxplots and Z-scores.
- Highlight how capping or transforming features like `Fare` addressed extreme values.
- **Visualizations**: Include boxplots for numerical features.

Data Preprocessing

- Discuss encoding categorical variables using one-hot encoding and scaling numerical features.
- Explain the importance of splitting the data into training and testing sets for unbiased evaluation.

Model Training and Testing

- Detail the training process for each model (Logistic Regression, Random Forest, XGBoost, Stacking Ensemble).
- Explain the use of hyperparameter tuning (RandomizedSearchCV).

Model Evaluation and Comparison

- Present the evaluation metrics for each model, including accuracy, AUC-ROC, confusion matrix, and classification report.
- **Visualizations**: Include confusion matrices, AUC-ROC curves, and comparison bar charts.

Explainability Analysis (SHAP)

- Discuss SHAP analysis results for Logistic Regression, Random Forest, and XGBoost.
- Summarize how SHAP insights aligned with feature importance analysis.
- **Visualizations**: Include SHAP summary plots and beeswarm plots.

4. Results and Findings

- Summarize the key findings from the analysis:
    - The best-performing model (Stacking Ensemble) achieved an accuracy of 82.12% and an AUC-ROC of 0.905.
    - Feature engineering and SHAP analysis identified key predictors like `Fare`, `Pclass`, and `Sex`.

3.1 Data Loading and Overview

- Load the dataset and display its structure and initial rows.
- Key functions used: `pd.read_csv()`, `data.info()`, `data.describe()`, and `data.head()`.

```
First five rows of the dataset:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

3.2 Exploratory Data Analysis (EDA)

- Analyze distributions of numerical features using histograms.
- Examine relationships between categorical features and survival using bar plots.
- Visualize missing data using a heatmap.

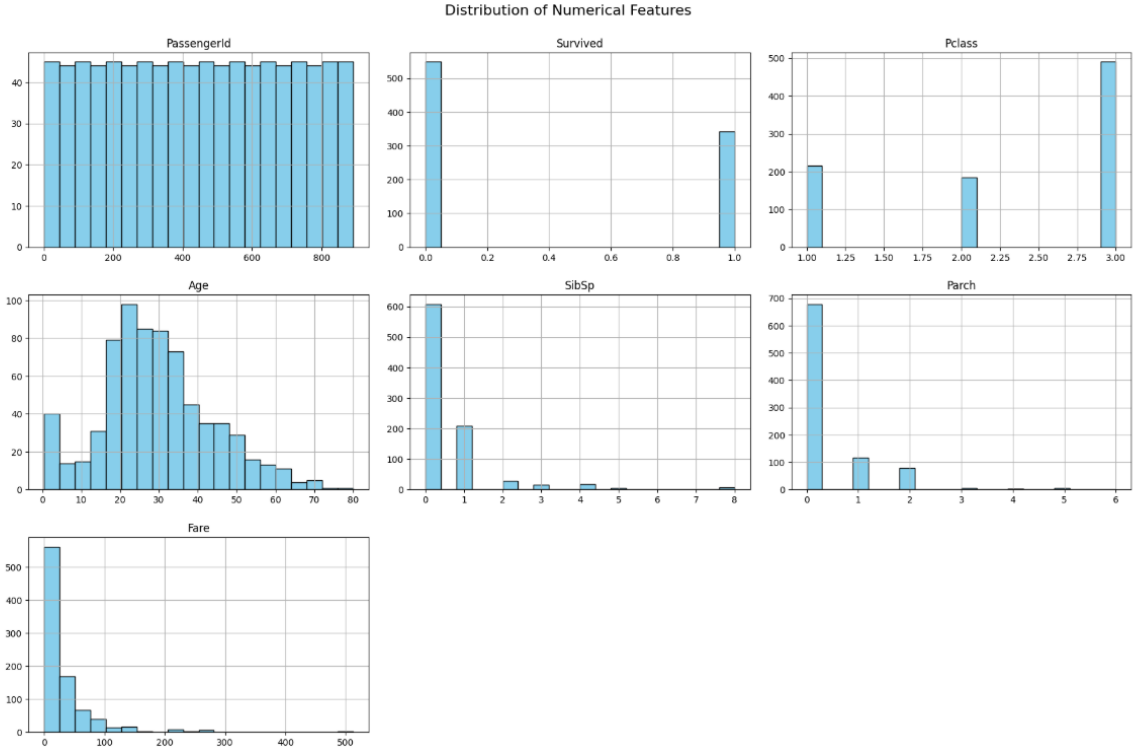
Visualizations:

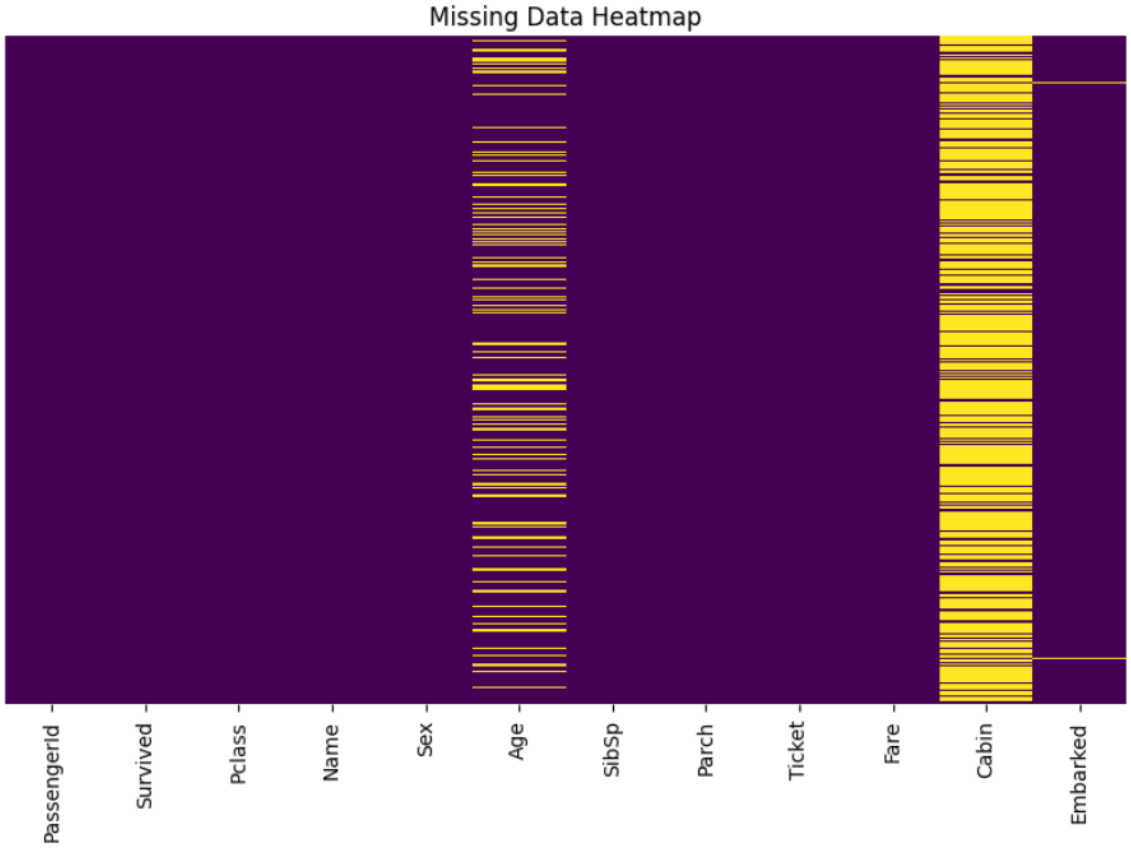
1. Distribution of numerical features (histograms).
2. Bar plots for survival rates by `Pclass`, `Sex`, and `Embarked`.
3. Heatmap for missing data.

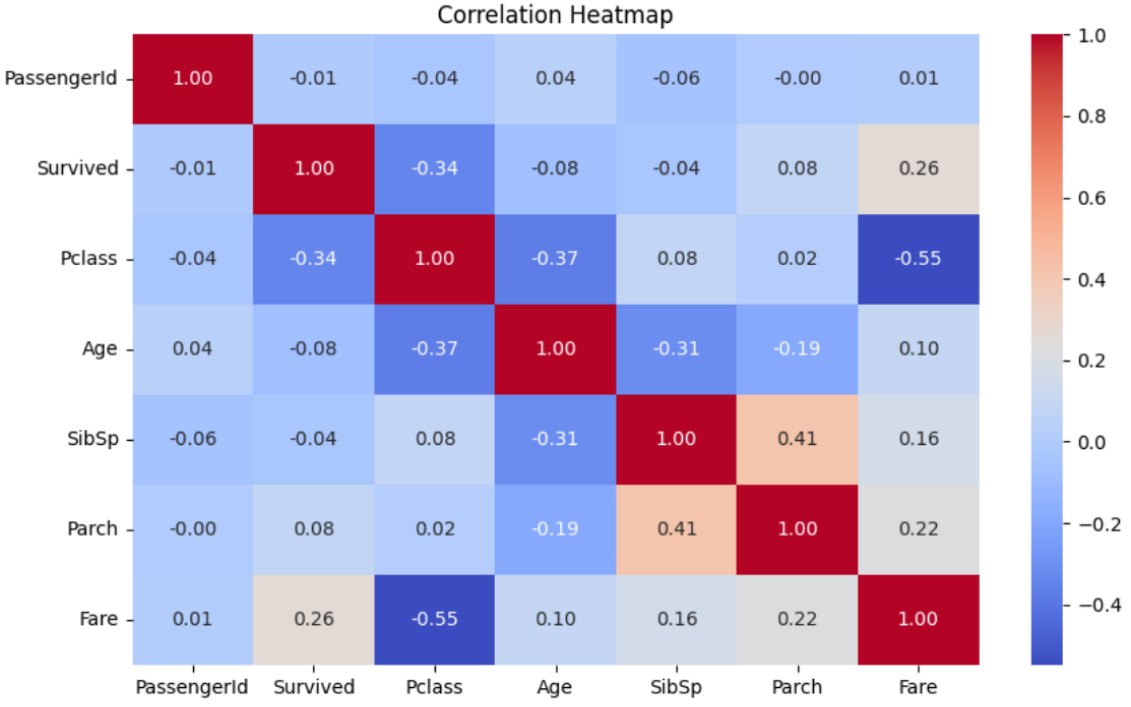
3.3 Handling Missing Data

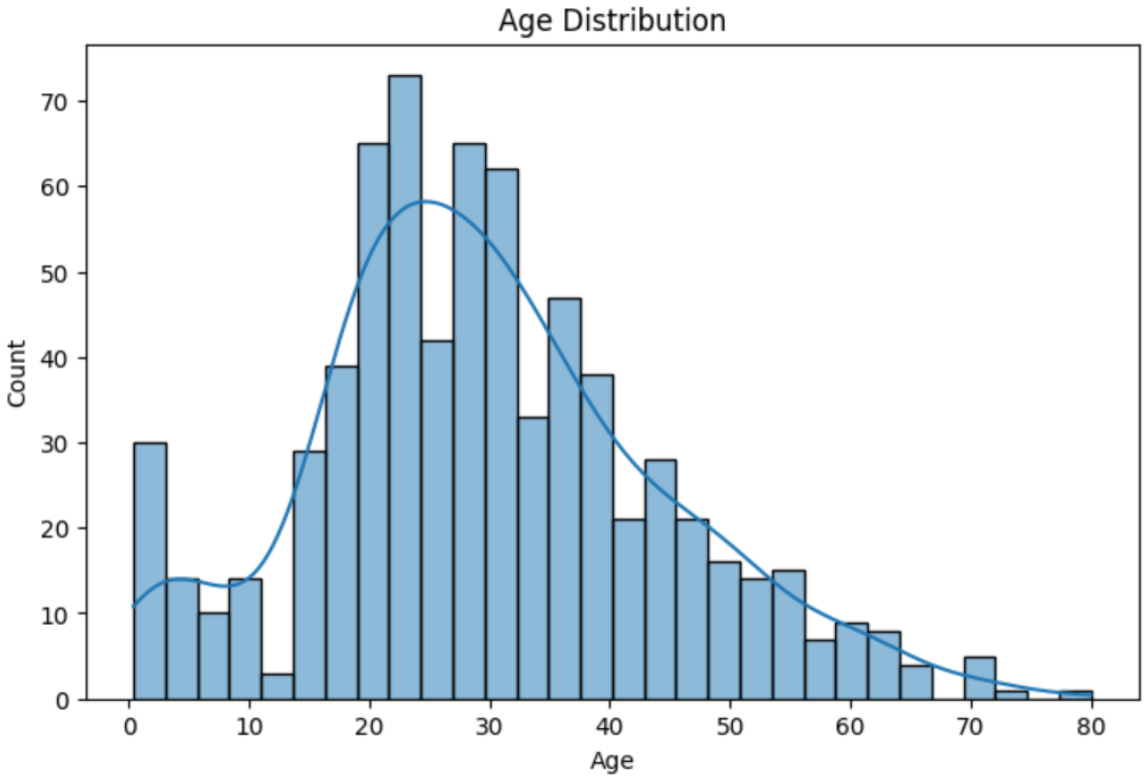
- Impute `Age` based on `Pclass` and `Sex`.
- Fill missing values in `Embarked` with the mode.
- Extract the deck information from `Cabin` and replace missing values with 'Unknown'.

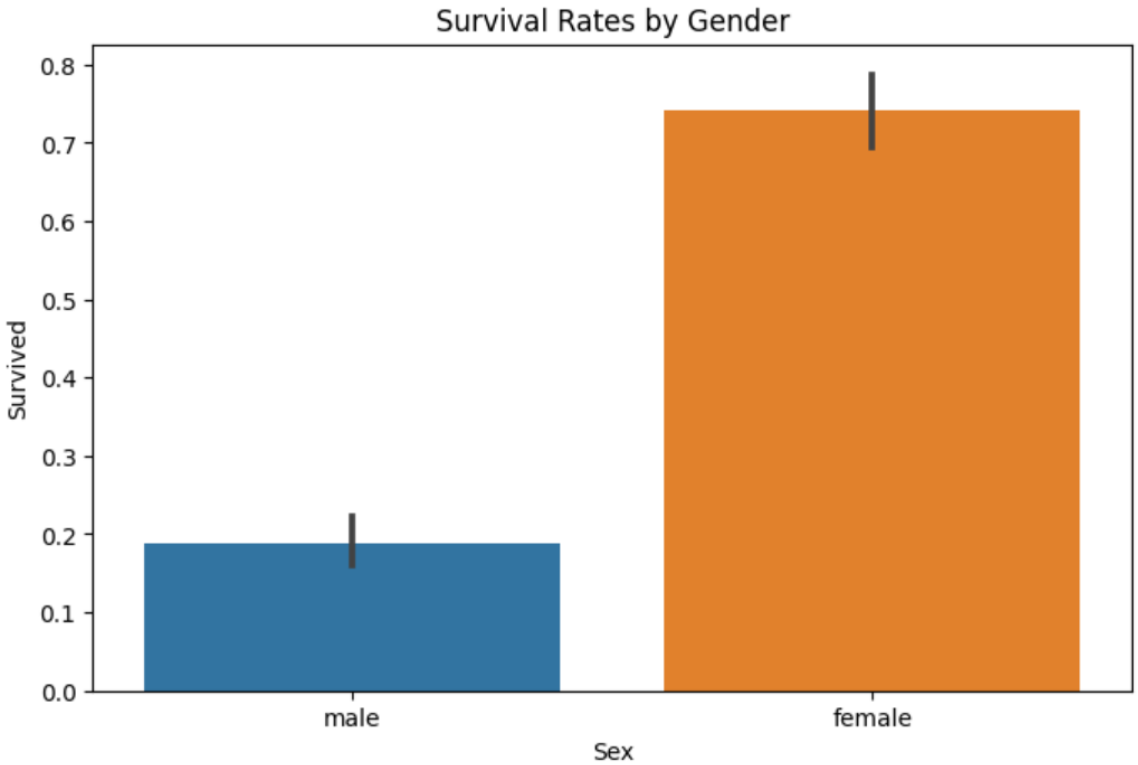
Explanation: Detail the rationale for each imputation strategy.

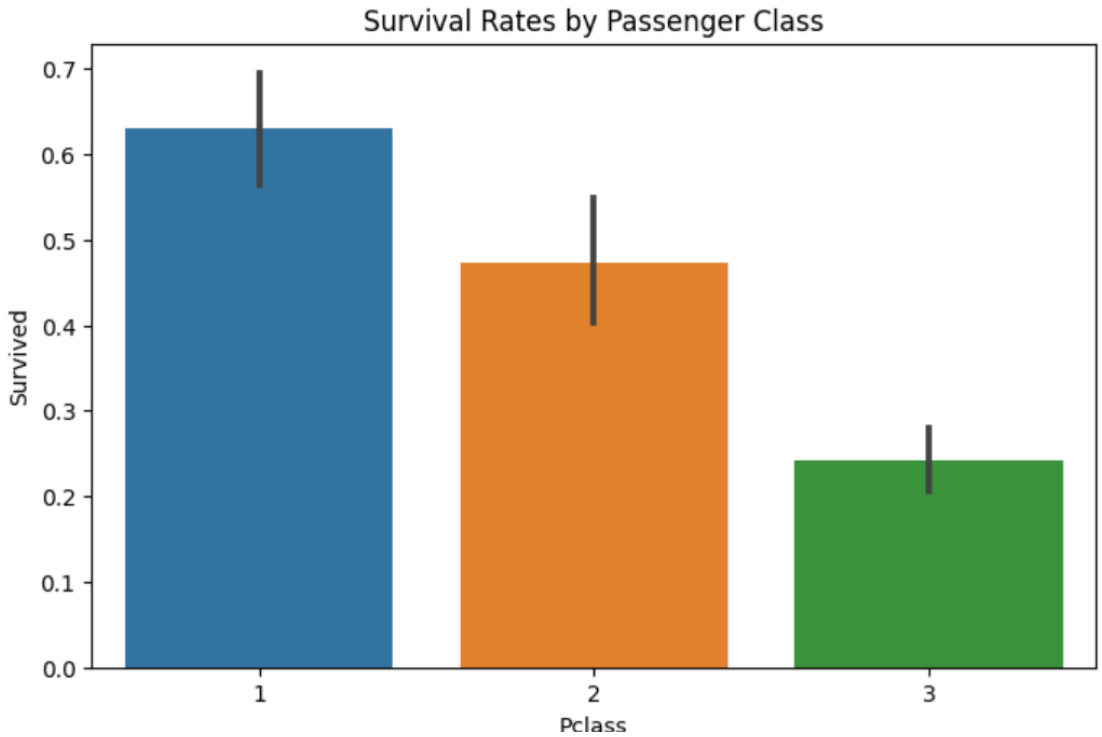


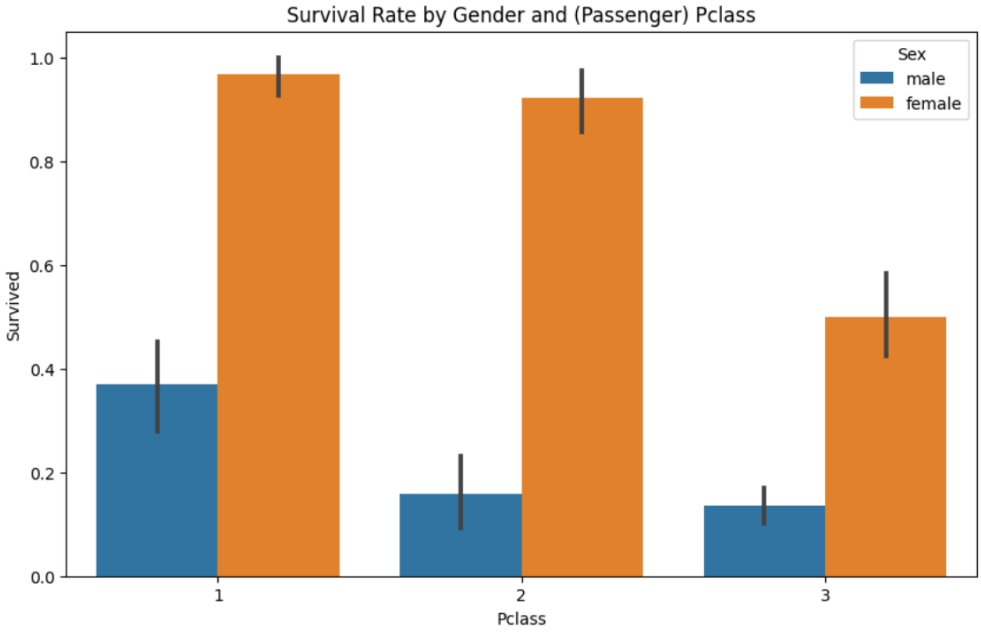

Missing Data Heatmap
PassengerId
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked

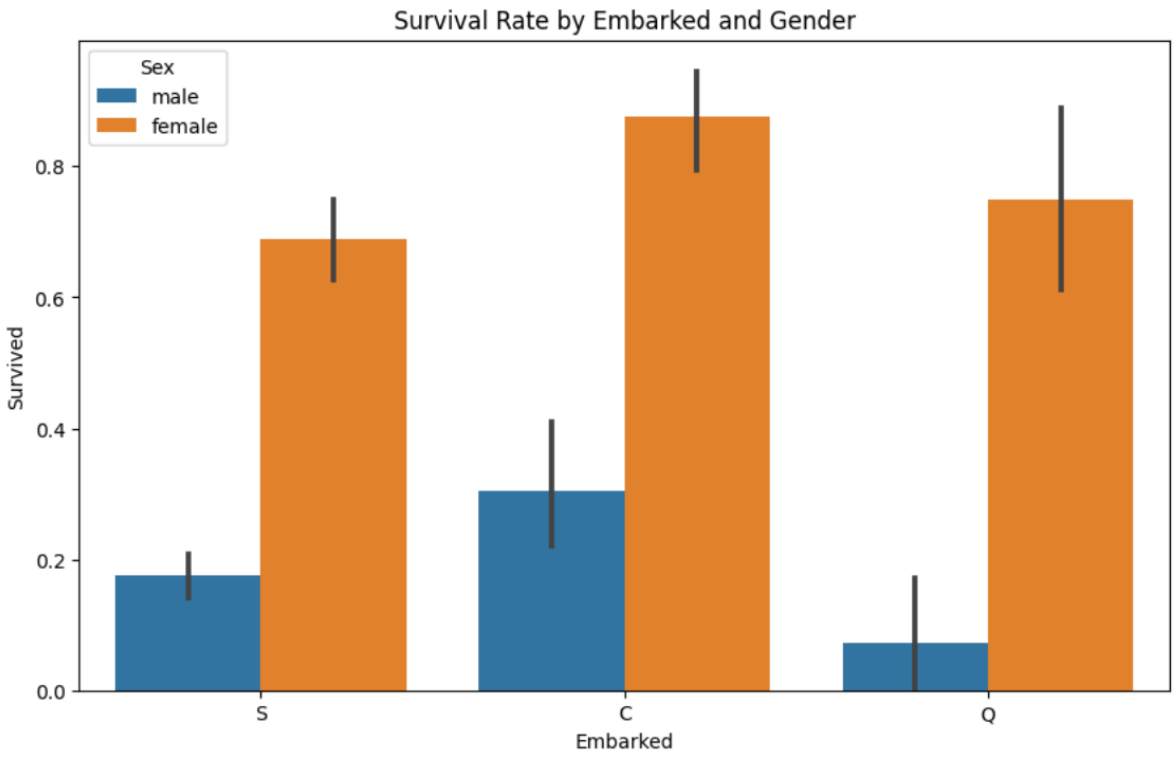

Correlation Heatmap
PassengerId 1.00 -0.01 -0.04 0.04 -0.06 -0.00 0.01
Survived -0.01 1.00 -0.34 -0.08 -0.04 0.08 0.26
Pclass -0.04 -0.34 1.00 -0.37 0.08 0.02 -0.55
Age 0.04 -0.08 -0.37 1.00 -0.31 -0.19 0.10
SibSp -0.06 -0.04 0.08 -0.31 1.00 0.41 0.16
Parch -0.00 0.08 0.02 -0.19 0.41 1.00 0.22
Fare 0.01 0.26 -0.55 0.10 0.16 0.22 1.00
PassengerId Survived Pclass Age SibSp Parch Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

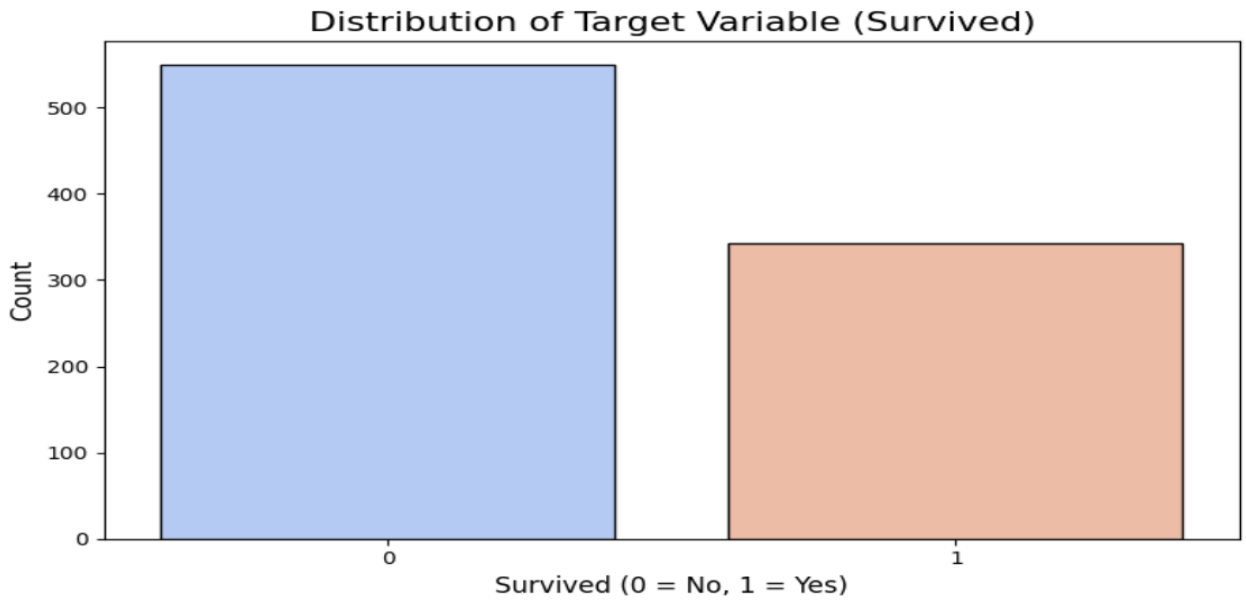

Age Distribution
Count
Age
0
10
20
30
40
50
60
70
80

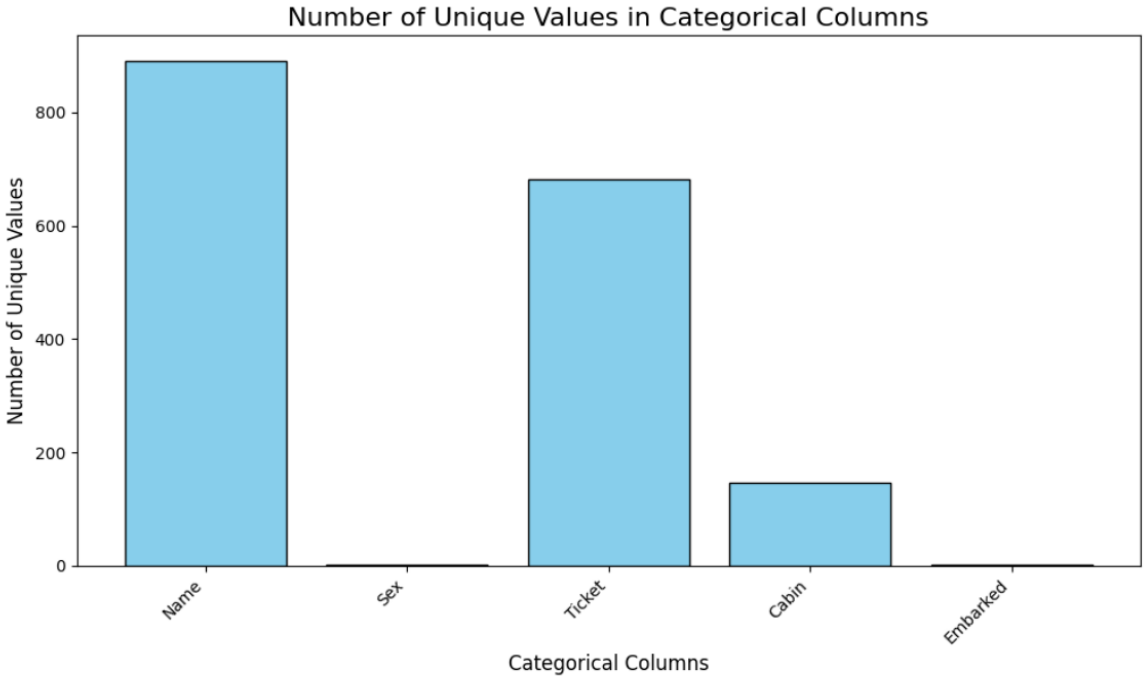

Survival Rates by Gender
Survived
Sex
male
female
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

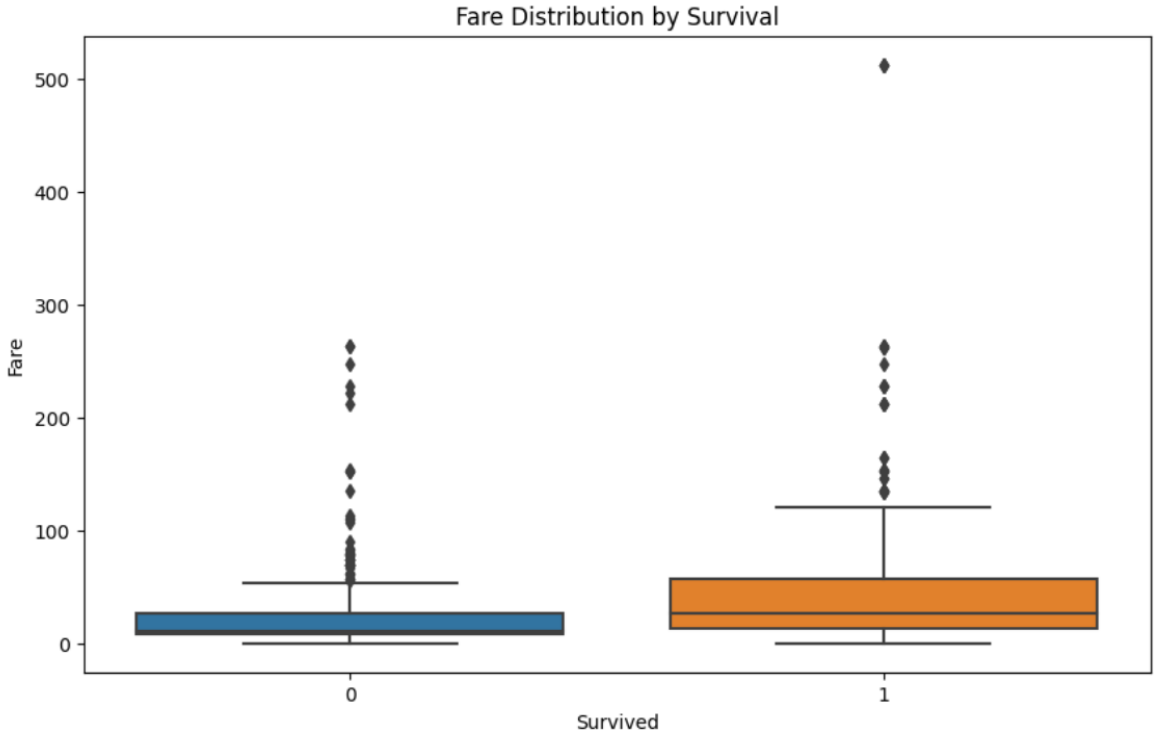

Survival Rates by Passenger Class
Survived
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Pclass

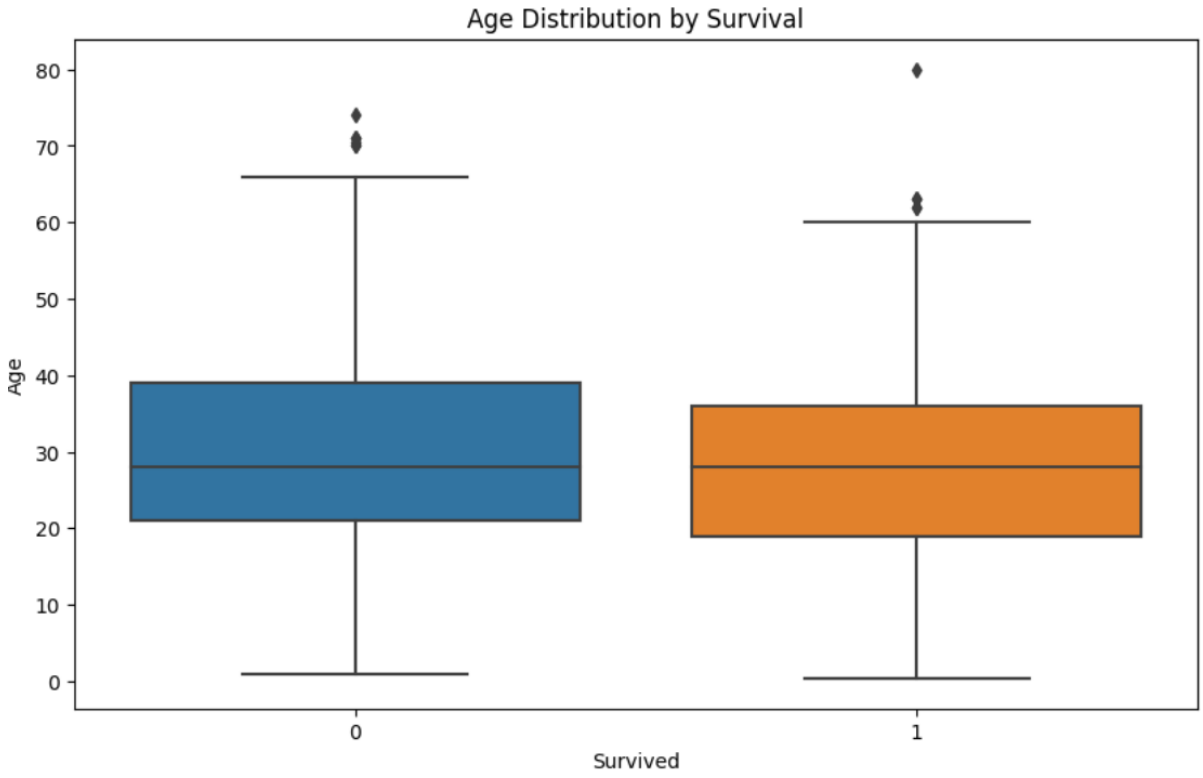

Survival Rate by Gender and (Passenger) Pclass
Sex
male
female
Survived
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
Pclass

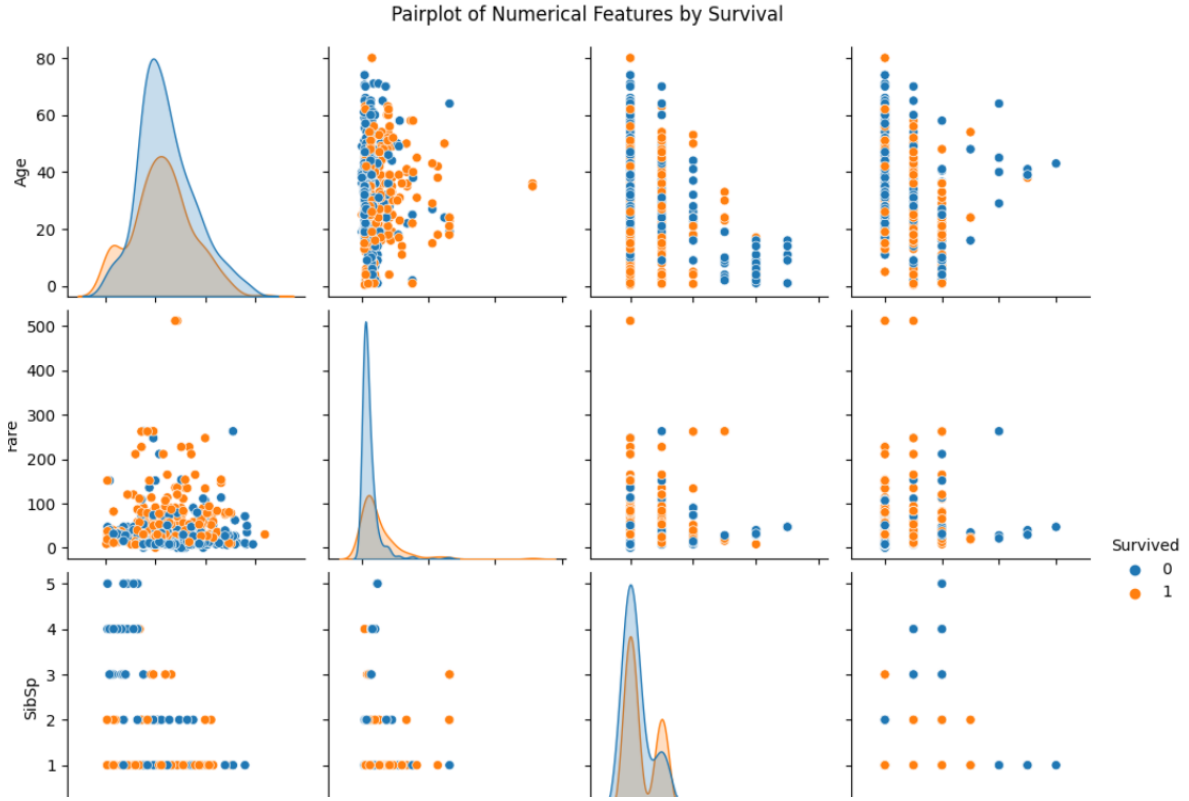

Survival Rate by Embarked and Gender
Sex
male
female
Survived
0.0
0.2
0.4
0.6
0.8
S
C
Q
Embarked


Distribution of Target Variable (Survived)
Count
0
100
200
300
400
500
0
1
Survived (0 = No, 1 = Yes)


Number of Unique Values in Categorical Columns
Number of Unique Values
0
200
400
600
800
Name
Sex
Ticket
Cabin
Embarked
Categorical Columns


Fare Distribution by Survival
Fare
0
100
200
300
400
500
0
1
Survived

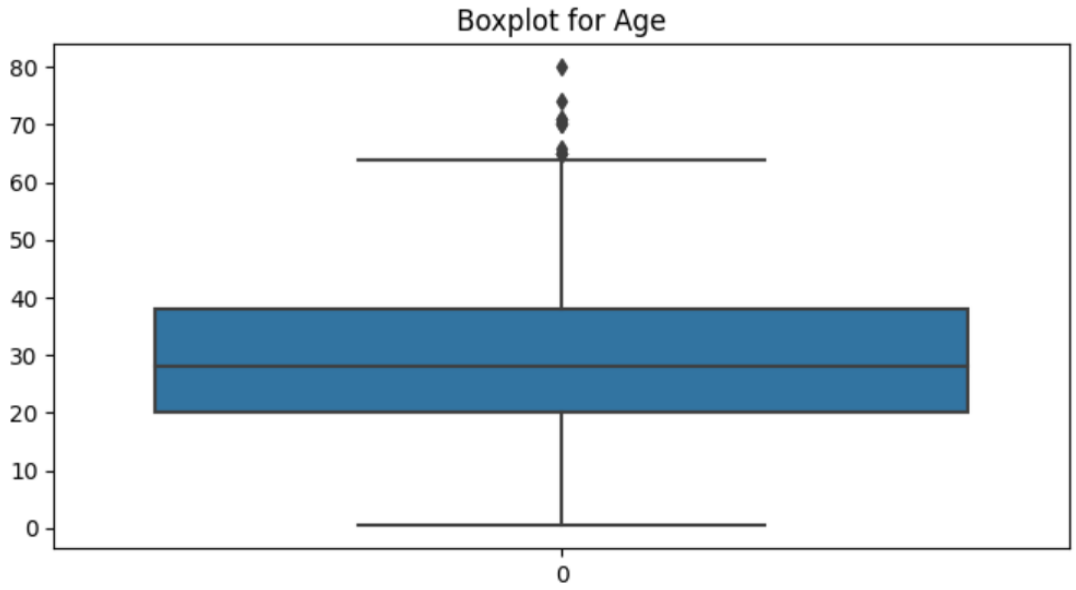

Age Distribution by Survival
Age
Survived
0
1
0
10
20
30
40
50
60
70
80

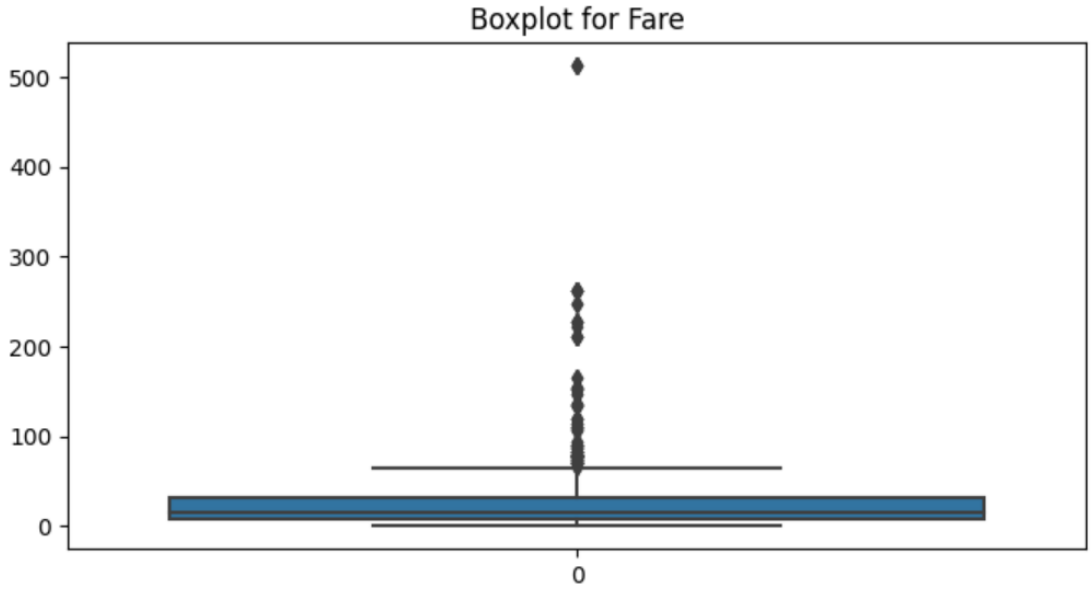

Pairplot of Numerical Features by Survival
Age
Fare
SibSp
Survived
0
1

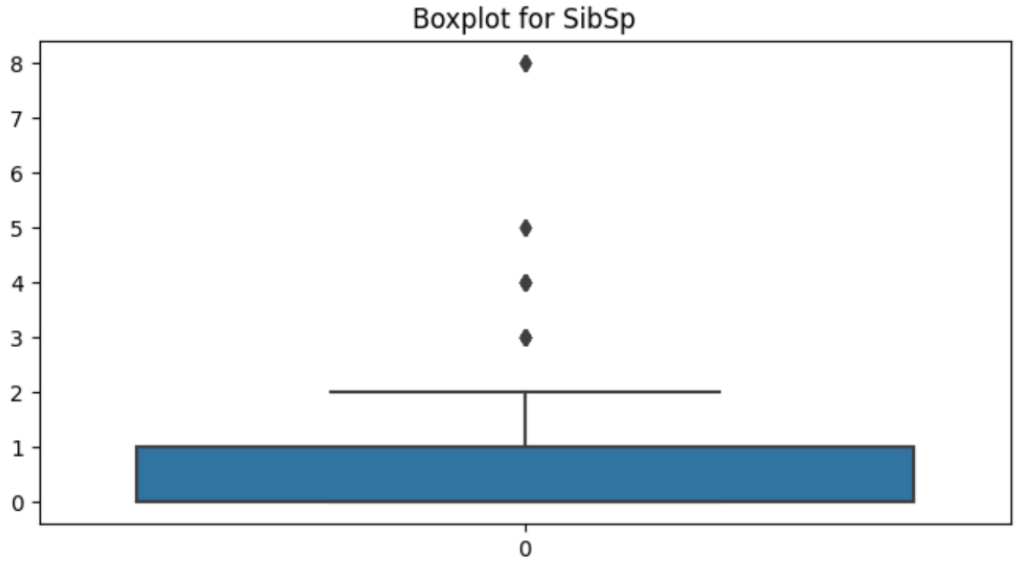
3.5 Outlier Detection and Handling

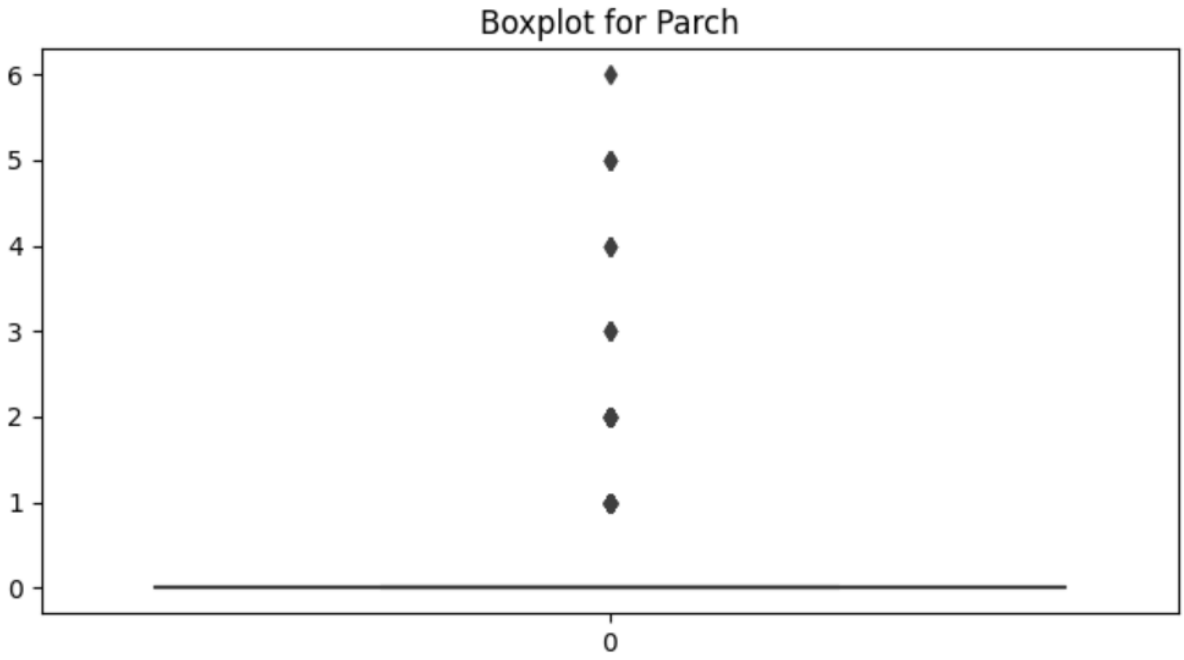
- Identify outliers using box plots and Z-scores.
- Cap `Fare` at the 95th percentile to reduce the impact of extreme values.

Visualizations: Box plots for numerical features to highlight outliers.









3.6 Data Preprocessing

- Encode categorical features using label encoding and one-hot encoding.
- Standardize numerical features to have a mean of 0 and a standard deviation of 1.

	Survived	Age	SibSp	Parch	Fare	FamilySize	IsAlone	Pclass_2	\
0	0	22.0	1	0	7.2500	2	0	0.0	
1	1	38.0	1	0	71.2833	2	0	0.0	
2	1	26.0	0	0	7.9250	1	1	0.0	
3	1	35.0	1	0	53.1000	2	0	0.0	
4	0	35.0	0	0	8.0500	1	1	0.0	

	Pclass_3	Sex_1	...	FareBin_2	FareBin_3	AgeBin_1	AgeBin_2	AgeBin_3	\
0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	
2	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	
4	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0	

	AgeBin_4	SibSpBin_1-2	SibSpBin_3+	ParchBin_1-2	ParchBin_3+
0	1.0	1.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0
2	1.0	0.0	0.0	0.0	0.0
3	1.0	1.0	0.0	0.0	0.0
4	1.0	0.0	0.0	0.0	0.0

3.7 Model Training and Testing

- Split the data into training and testing sets (80%-20%).
- Train multiple models:
  - Logistic Regression
  - Random Forest
  - XGBoost
  - Stacking Ensemble

3.8 Model Evaluation

- Evaluate models using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.
- Visualize confusion matrices and ROC curves for each model.

Visualizations:

1. Confusion matrices for each model.
2. ROC curves comparing model performance.

3.9 Model Comparison

- Compile accuracy and AUC-ROC scores for all models.
- Visualize model performance using bar charts.

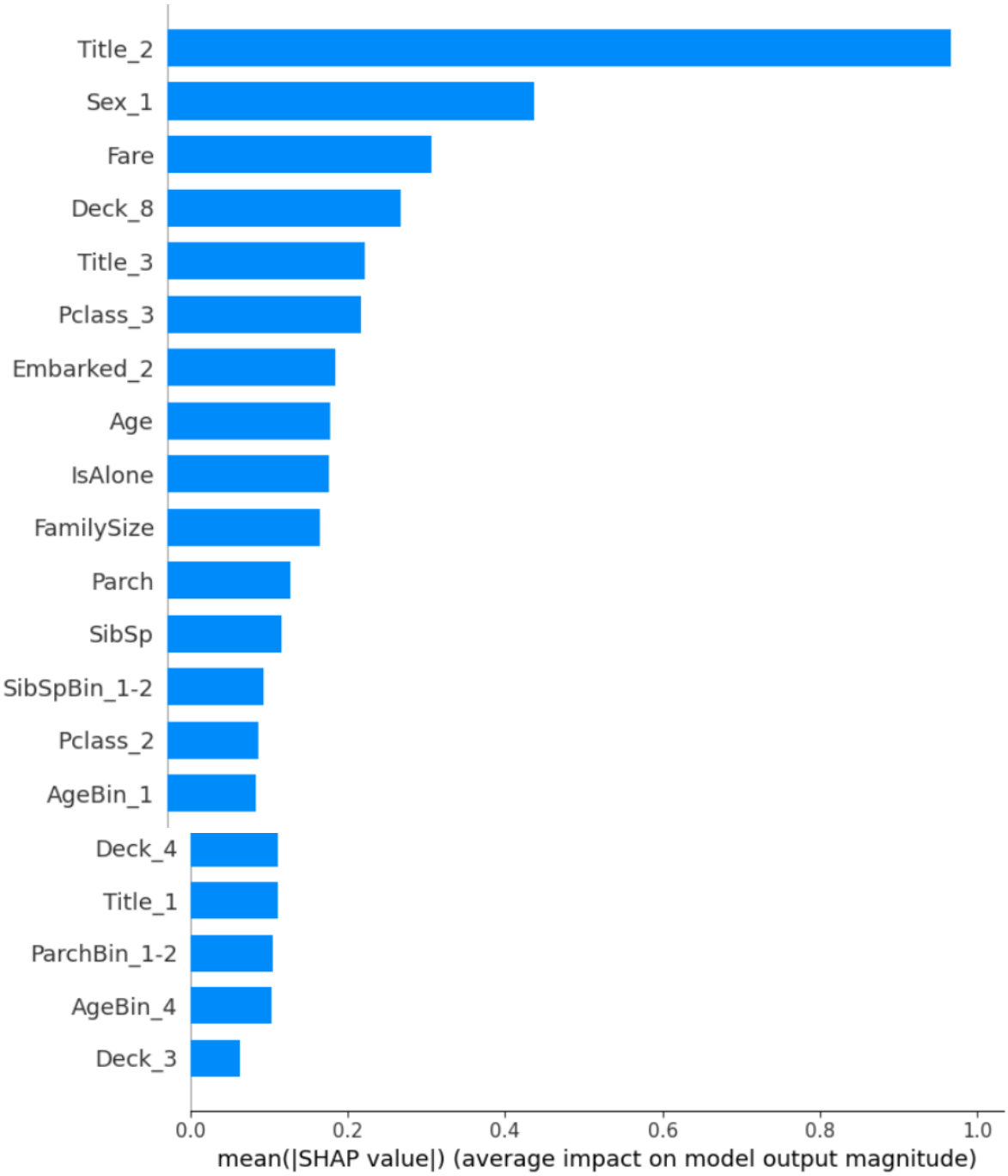
Visualization: Bar chart comparing accuracy and AUC-ROC for all models.

3.10 Explainability Analysis

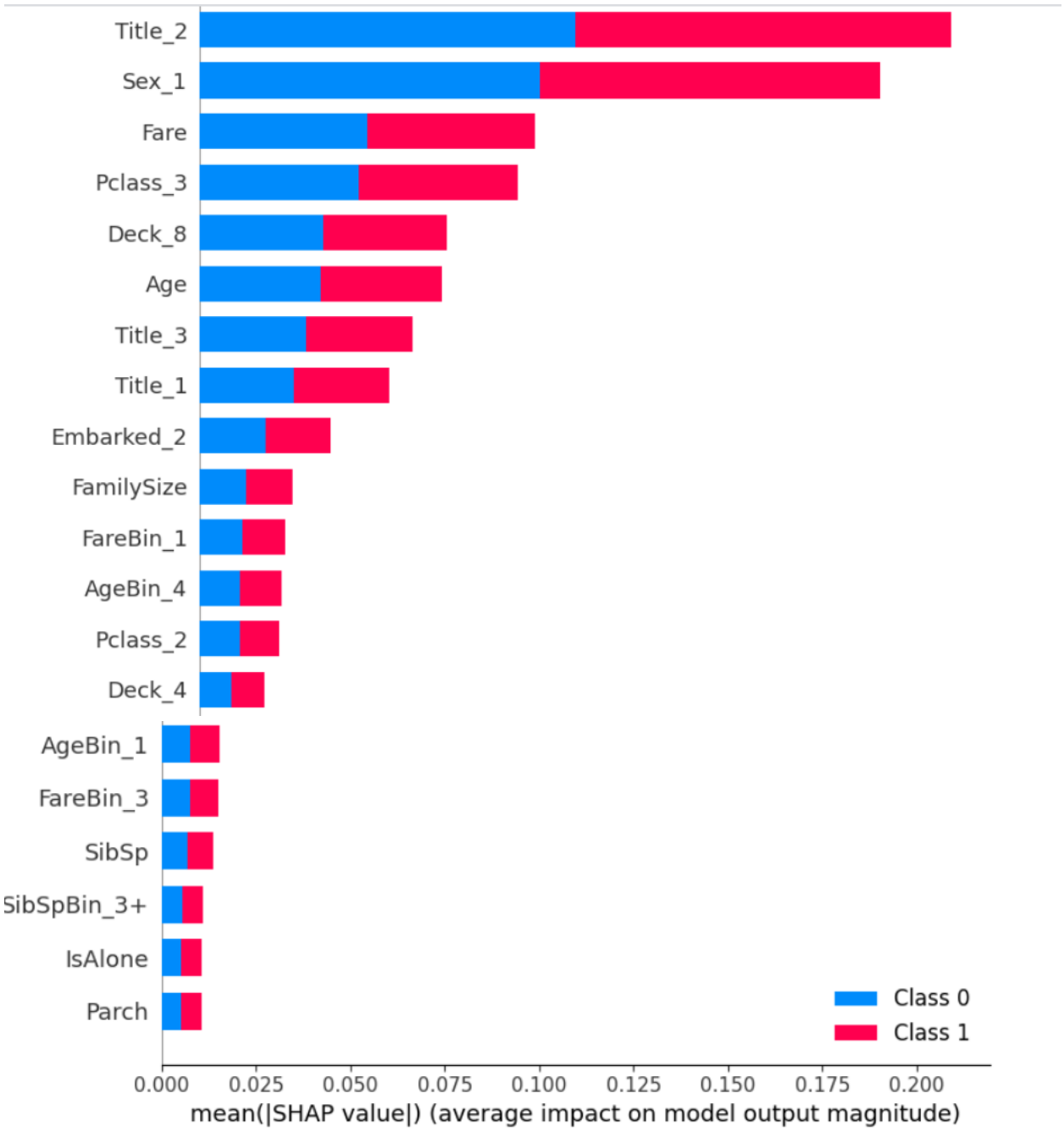
- Use SHAP to interpret predictions for:
    - Logistic Regression
    - Random Forest
    - XGBoost
- Visualize feature importance using SHAP summary plots and beeswarm plots.

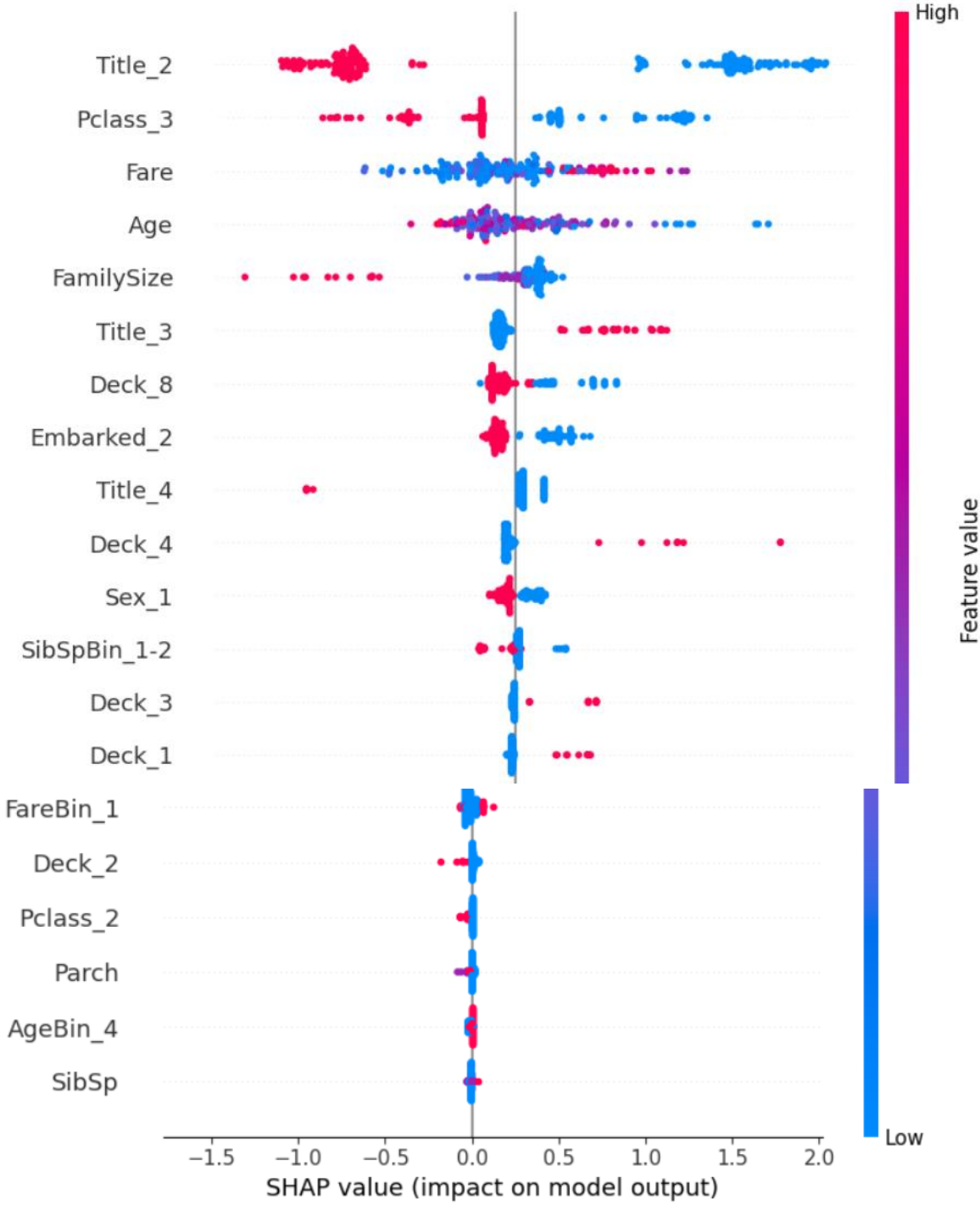
Visualizations:

1. SHAP summary plots (bar and beeswarm).
2. Force plots for individual predictions.


Title_2
Sex_1
Fare
Deck_8
Title_3
Pclass_3
Embarked_2
Age
IsAlone
FamilySize
Parch
SibSp
SibSpBin_1-2
Pclass_2
AgeBin_1
Deck_4
Title_1
ParchBin_1-2
AgeBin_4
Deck_3
0.0
0.2
0.4
0.6
0.8
1.0
mean(|SHAP value|) (average impact on model output magnitude)


Title_2
Sex_1
Fare
Deck_8
Title_3
Pclass_3
Embarked_2
Age
IsAlone
FamilySize
Parch
SibSp
SibSpBin_1-2
Pclass_2
AgeBin_1
Deck_4
Title_1
ParchBin_1-2
AgeBin_4
Deck_3
−1.0
−0.5
0.0
0.5
1.0
SHAP value (impact on model output)
High
Feature value
Low


Title_2
Sex_1
Fare
Pclass_3
Deck_8
Age
Title_3
Title_1
Embarked_2
FamilySize
FareBin_1
AgeBin_4
Pclass_2
Deck_4
AgeBin_1
FareBin_3
SibSp
SibSpBin_3+
IsAlone
Parch
Class 0
Class 1
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
mean(|SHAP value|) (average impact on model output magnitude)

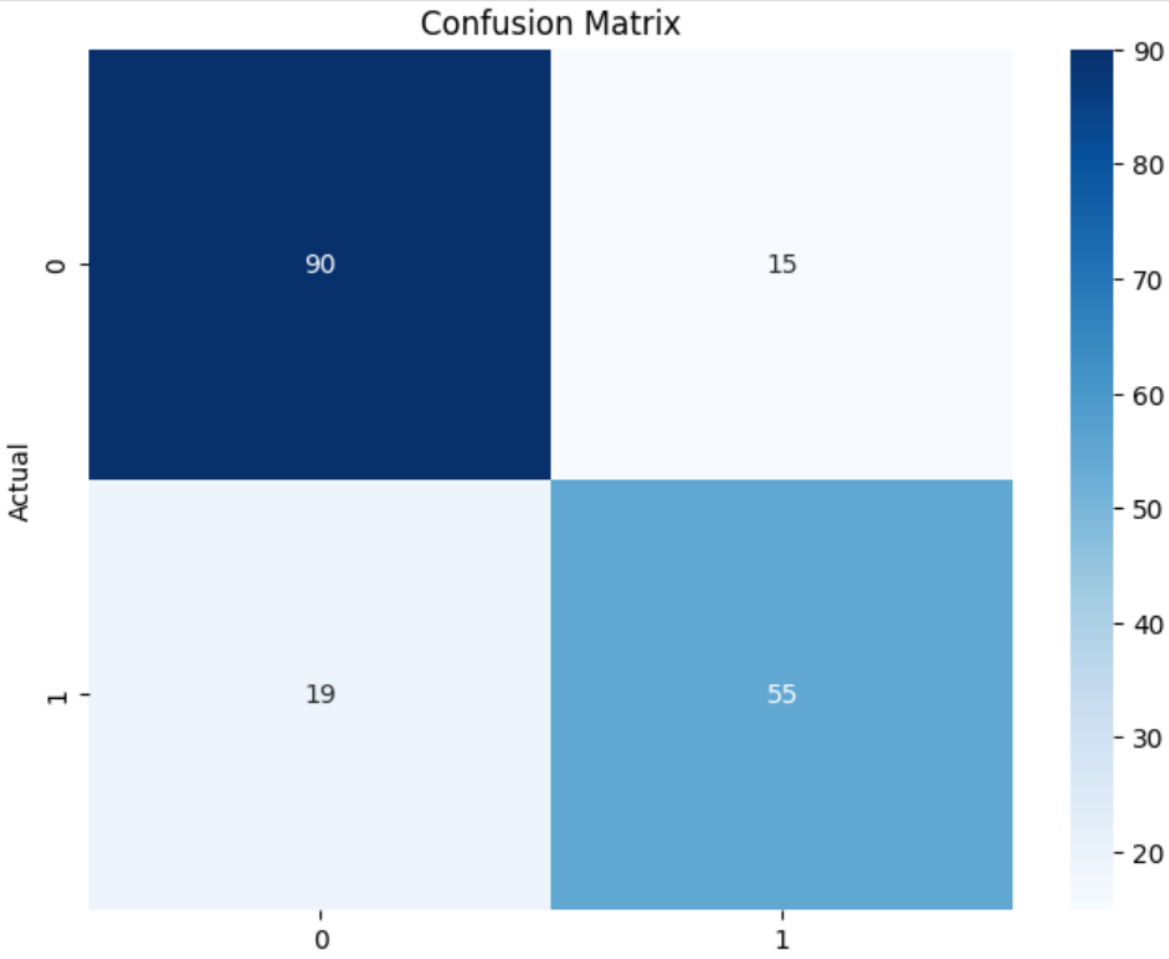

Title_2
Pclass_3
Fare
Age
FamilySize
Title_3
Deck_8
Embarked_2
Title_4
Deck_4
Sex_1
SibSpBin_1-2
Deck_3
Deck_1
FareBin_1
Deck_2
Pclass_2
Parch
AgeBin_4
SibSp
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
SHAP value (impact on model output)
High
Feature value
Low

4. Results and Conclusion

4.1 Model Performance

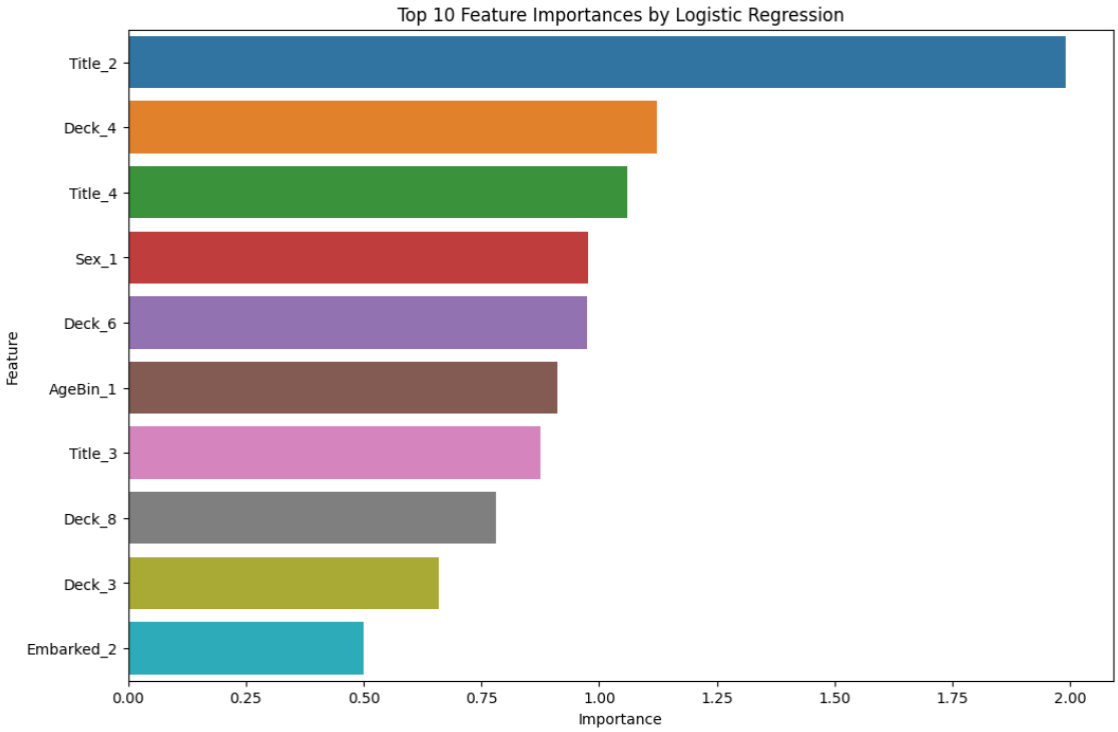
Logistic Regression: Accuracy = 81.01%, AUC-ROC = 0.874.

```
Accuracy: 0.8100558659217877

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.86      0.84       105
           1       0.79      0.74      0.76        74

    accuracy                           0.81       179
   macro avg       0.81      0.80      0.80       179
weighted avg       0.81      0.81      0.81       179
```



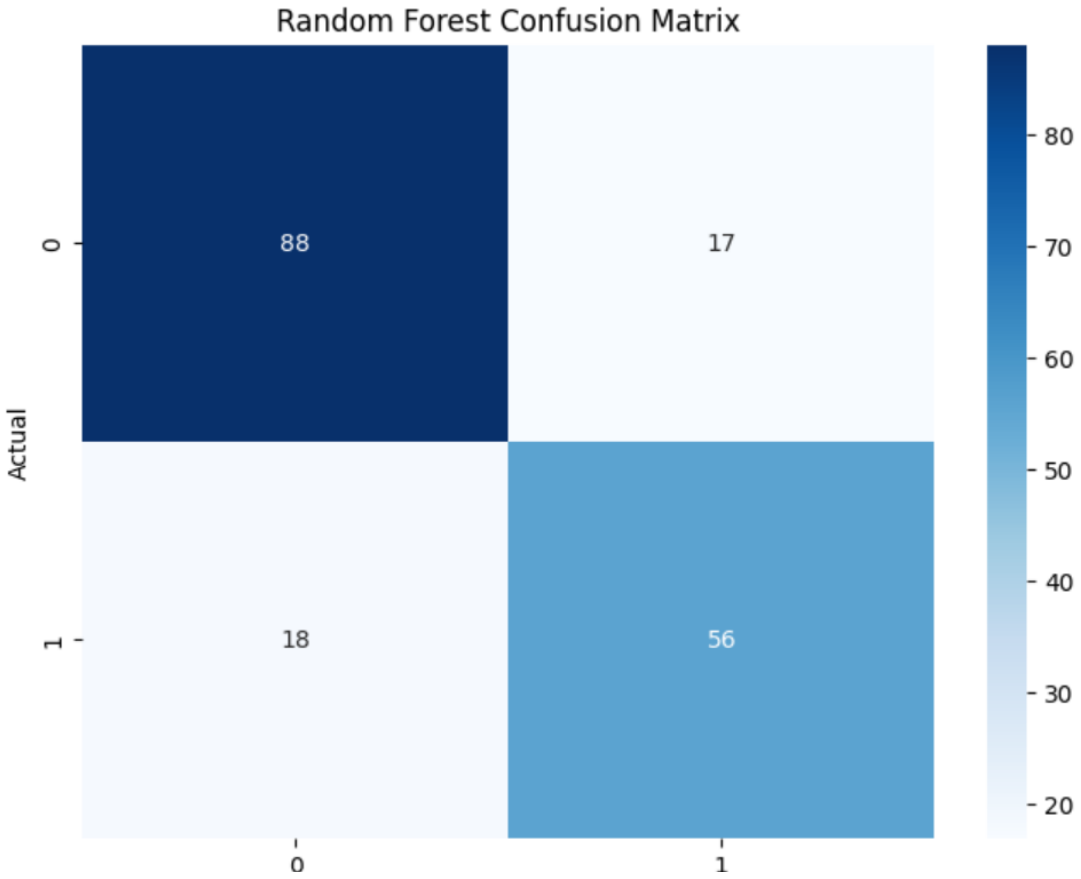
Top Features by Logistic Regression:

	Feature	Coefficient	Importance
12	Title_2	-1.990977	1.990977
18	Deck_4	1.122063	1.122063
14	Title_4	-1.059864	1.059864
8	Sex_1	-0.976020	0.976020
20	Deck_6	-0.974838	0.974838
26	AgeBin_1	0.911873	0.911873
13	Title_3	0.876208	0.876208
22	Deck_8	-0.780276	0.780276
17	Deck_3	0.659725	0.659725
10	Embarked_2	-0.499972	0.499972

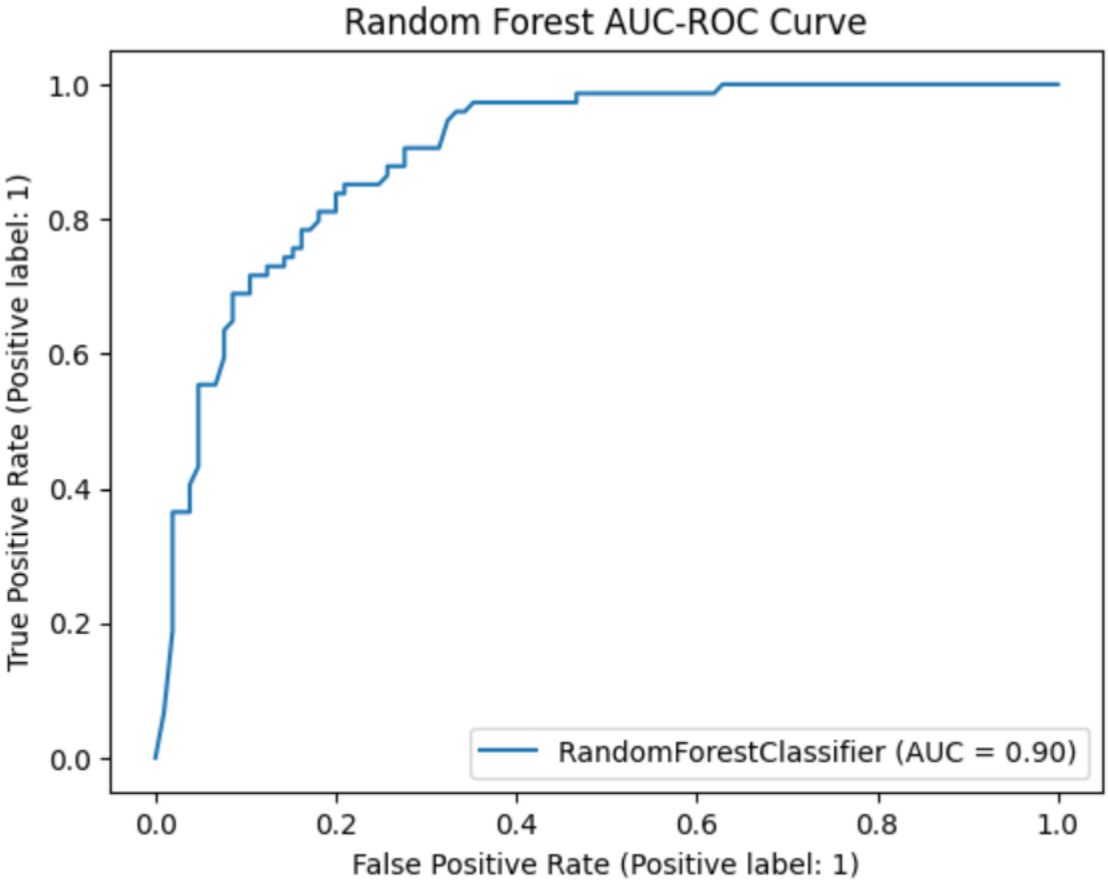


Random Forest: Accuracy = 80.45%, AUC-ROC = 0.898.

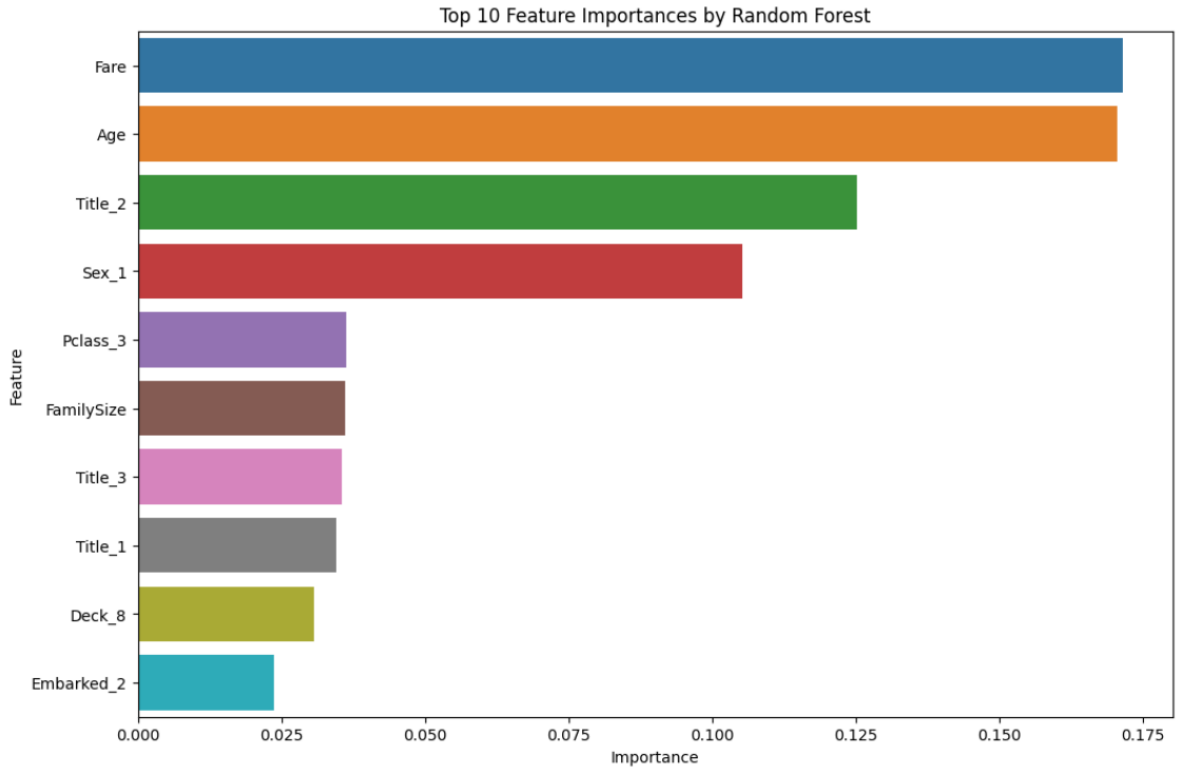
```
Random Forest Accuracy: 0.8044692737430168

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.84      0.83       105
           1       0.77      0.76      0.76        74

    accuracy                           0.80       179
   macro avg       0.80      0.80      0.80       179
weighted avg       0.80      0.80      0.80       179
```

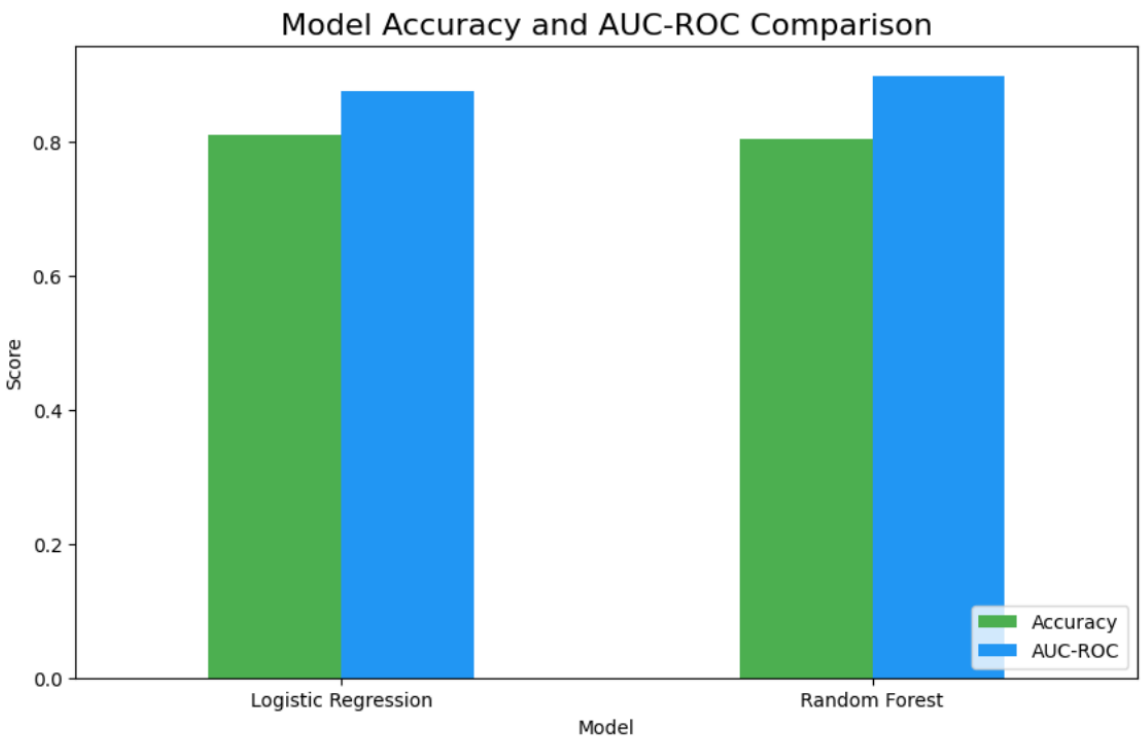

Random Forest Confusion Matrix
Actual
0
1
88
17
18
56
0
1
80
70
60
50
40
30
20

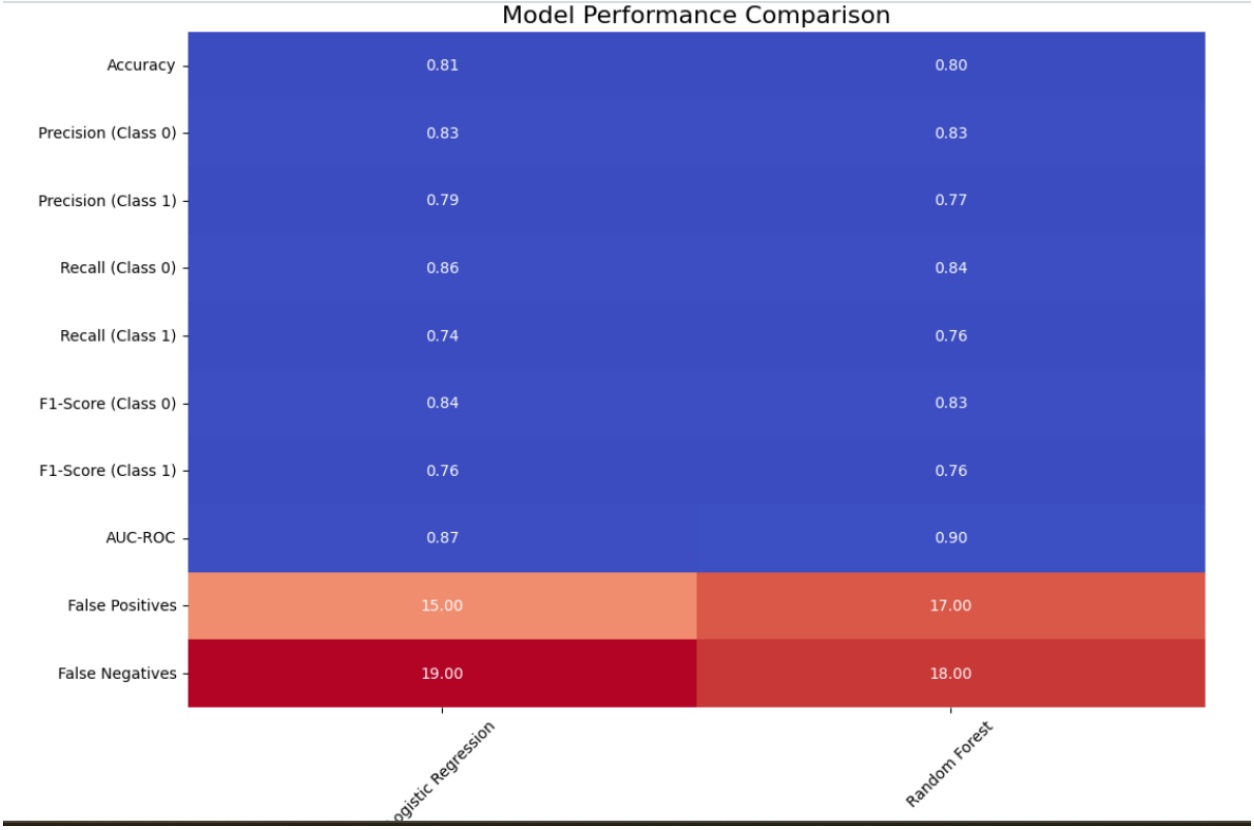
```
Random Forest AUC-ROC Score: 0.8984555984555984
```



```
Top Features by Importance:
        Feature  Importance
3          Fare    0.171571
0           Age    0.170557
12      Title_2    0.125272
8         Sex_1    0.105216
7      Pclass_3    0.036258
4    FamilySize    0.036078
13      Title_3    0.035508
11      Title_1    0.034657
22       Deck_8    0.030619
10   Embarked_2    0.023719
```







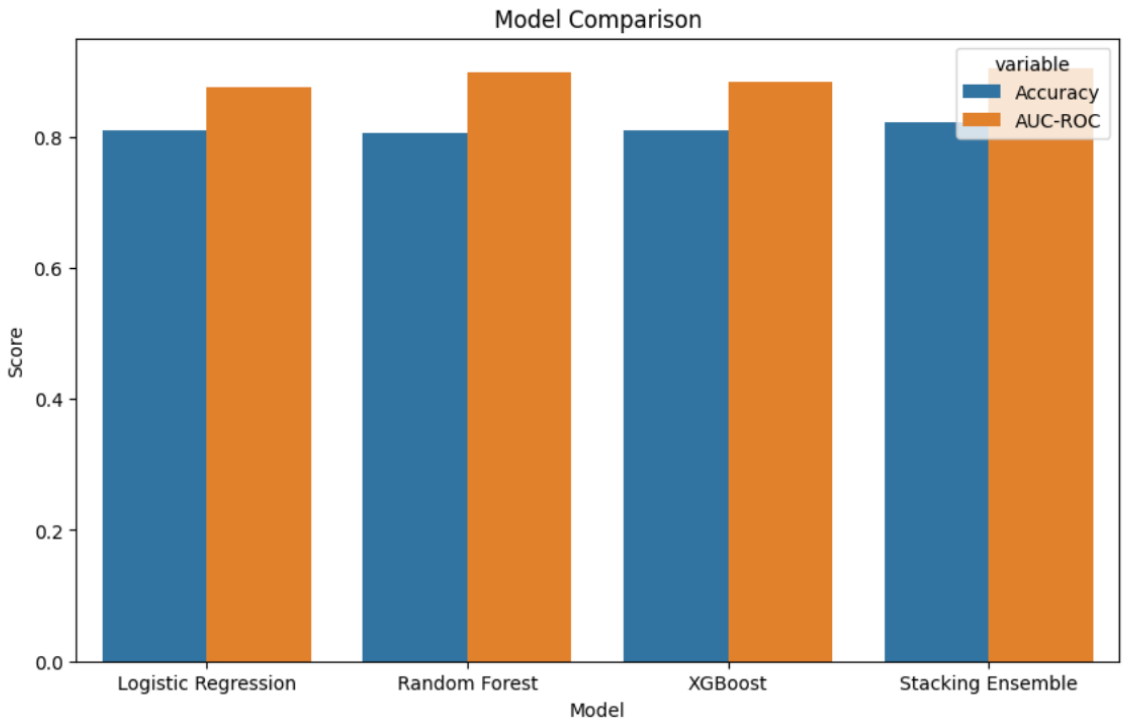
- **XGBoost**: Accuracy = 81.01%, AUC-ROC = 0.884.

```
XGBoost Accuracy: 0.8100558659217877

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.85      0.84       105
           1       0.78      0.76      0.77        74

    accuracy                           0.81       179
   macro avg       0.80      0.80      0.80       179
weighted avg       0.81      0.81      0.81       179


XGBoost AUC-ROC: 0.8842985842985842
```

- **Stacking Ensemble**: Accuracy = 82.12%, AUC-ROC = 0.905.

```
Stacking Accuracy: 0.8212290502793296

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.86      0.85       105
           1       0.79      0.77      0.78        74

    accuracy                           0.82       179
   macro avg       0.82      0.81      0.81       179
weighted avg       0.82      0.82      0.82       179


Stacking AUC-ROC: 0.9046332046332046
```

	Model	Accuracy	AUC-ROC
0	Logistic Regression	0.810056	0.874517
1	Random Forest	0.804469	0.898456
2	XGBoost	0.810056	0.884299
3	Stacking Ensemble	0.821229	0.904633



The **Stacking Ensemble** model achieved the best performance, leveraging the strengths of multiple models to optimize predictions.

The **Stacking Ensemble** model achieved the best performance, leveraging the strengths of multiple models to optimize predictions.

4.2 Key Insights

- Features like `Fare`, `Pclass`, `Sex`, and `Title` significantly influenced survival predictions.
- Feature engineering and handling missing data improved model performance.
- SHAP analysis provided transparency and explainability for model decisions.