Titanic Survival Prediction

From Data Exploration and Feature Engineering to Model Building and Interpretation

Hafiza Hajrah Rehman

Data Science Project

L1f19bscs0445@ucp.edu.pk

Predicting Titanic passenger survival using data cleaning, feature engineering, and machine learning models, achieving strong accuracy with interpretable insights into key survival factors.

Contents

1. Introduction

The sinking of the RMS Titanic on April 15, 1912, stands as one of the most tragic and widely studied maritime disasters in history, capturing the attention of historians, engineers, and data scientists alike. Despite being deemed "unsinkable," the ship struck an iceberg on its maiden voyage, leading to the loss of over 1,500 lives. This event has since inspired numerous investigations into the factors that influenced passenger survival during the disaster. In this project, we leverage modern machine learning techniques to analyze a rich dataset containing personal, demographic, and travel-related information about the passengers aboard the Titanic. Our objective is to build a predictive model that can accurately classify whether a passenger survived or not, based on their attributes such as age, gender, class, fare paid, family connections, and point of embarkation. Beyond mere prediction, this project aims to uncover and understand the underlying patterns and key factors that significantly impacted survival chances, thereby providing not only an effective predictive tool but also meaningful insights into this historic tragedy. By applying rigorous data preprocessing, exploratory data analysis, feature engineering, and model tuning, we seek to demonstrate how data science can be used to extract valuable knowledge from historical events and inform future risk assessment and safety protocols.

2. Data Acquisition and Initial Exploration

2.1 Data Loading

The first step in this project involved acquiring the Titanic datasets from the Kaggle platform, which hosts the popular Titanic Machine Learning competition. Utilizing the Kaggle API, we programmatically downloaded two primary datasets: the training set and the test set. The training dataset contains detailed passenger information along with the corresponding survival status, providing a labeled foundation necessary for supervised learning. In contrast, the test dataset includes similar passenger attributes but does not provide survival outcomes; this dataset serves as the input for our final predictions, which are submitted to Kaggle for evaluation. Both datasets were imported into Python using the Pandas library, which offers powerful data manipulation and analysis tools. By loading the data into DataFrames, we established a structured environment to conduct preliminary inspections, data cleaning, and subsequent modeling steps efficiently. This approach ensured reproducibility and streamlined integration with the rest of the data science workflow.

2.2 Dataset Overview and Missing Data Identification

Upon loading the datasets, we performed an initial examination to understand their structure, size, and quality. The training dataset comprised 891 passenger records with 12 attributes,

including demographic details, ticket information, and survival labels. The test dataset contained 418 records with 11 attributes, lacking the survival outcome which we aimed to predict. A crucial early step was identifying missing data, as incomplete information can impair model performance. We discovered that several important columns had missing values: the 'Age' feature had substantial gaps in both train and test sets, 'Cabin' had the most missing entries by far, and 'Embarked' had a couple of missing values in the training set. Additionally, the test set contained one missing value in the 'Fare' column. Recognizing and addressing these gaps was essential for building reliable models, as ignoring missing data could bias results or cause errors during training. This initial assessment guided our strategy for data cleaning and imputation in subsequent steps.

Feature	Train Missing	Test Missing
Age	177	86
Cabin	687	327
Embarked	2	0
Fare	0	1

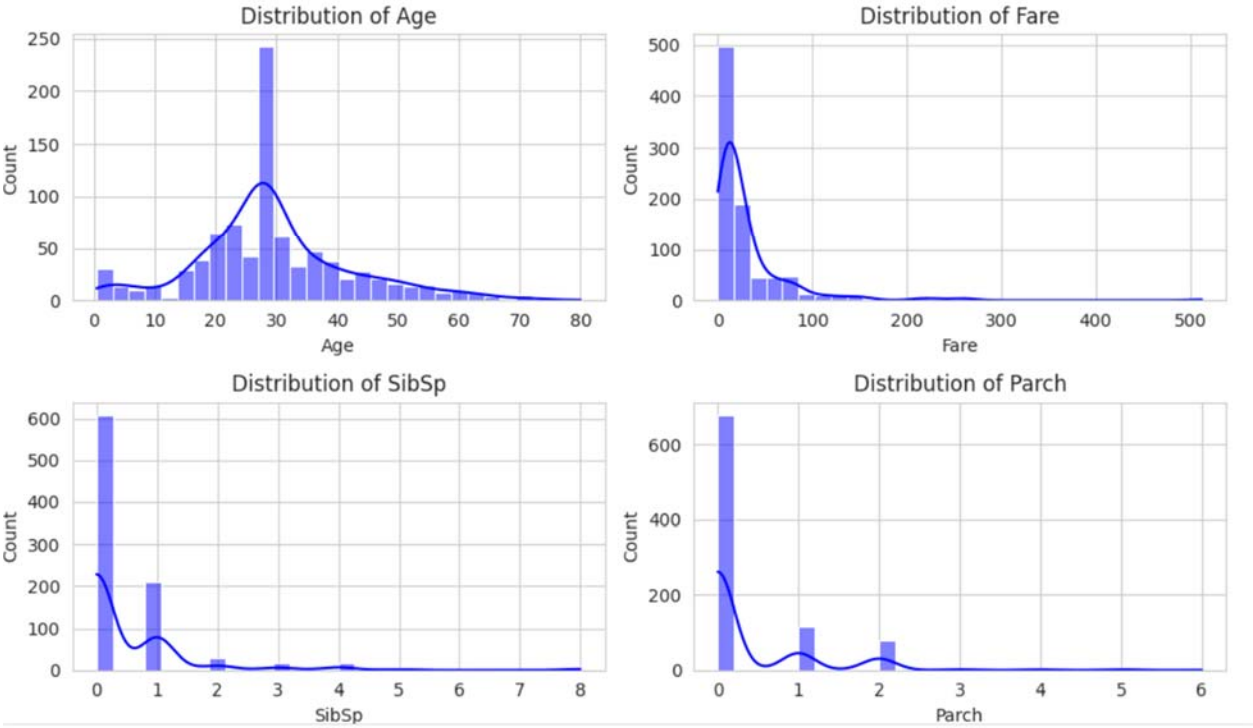
3. Data Cleaning and Imputation

After identifying missing values in the dataset, the next critical phase was to clean and impute these gaps to prepare the data for modeling. Missing values can introduce bias and reduce the effectiveness of predictive models if not handled properly. For the 'Age' feature, which had many missing entries, we opted to fill missing values with the median age rather than the mean, as the median is less sensitive to outliers and better represents the central tendency for skewed data. The 'Embarked' feature, which indicates the port where passengers boarded, had only a few missing entries; these were filled with the mode, the most frequent embarkation point, 'S', ensuring minimal distortion of the dataset. The 'Fare' column in the test set had a single missing value, which we replaced with the median fare to maintain consistency. The 'Cabin' feature presented a unique challenge due to a high proportion of missing data. Instead of discarding the feature, we replaced missing Cabin entries with a placeholder value 'UN' (unknown), preserving the information that the cabin data was missing while allowing the model to potentially learn from this pattern. Additionally, we extracted titles such as 'Mr', 'Mrs', 'Miss', and 'Master' from the passengers' names using regular expressions to capture social status and other demographic cues. Rare or less frequent titles were grouped into a single 'Rare' category to reduce noise and ensure better model generalization. These cleaning and imputation steps were crucial to building a robust dataset that could feed accurate and meaningful information into the predictive models.

4. Exploratory Data Analysis (EDA)

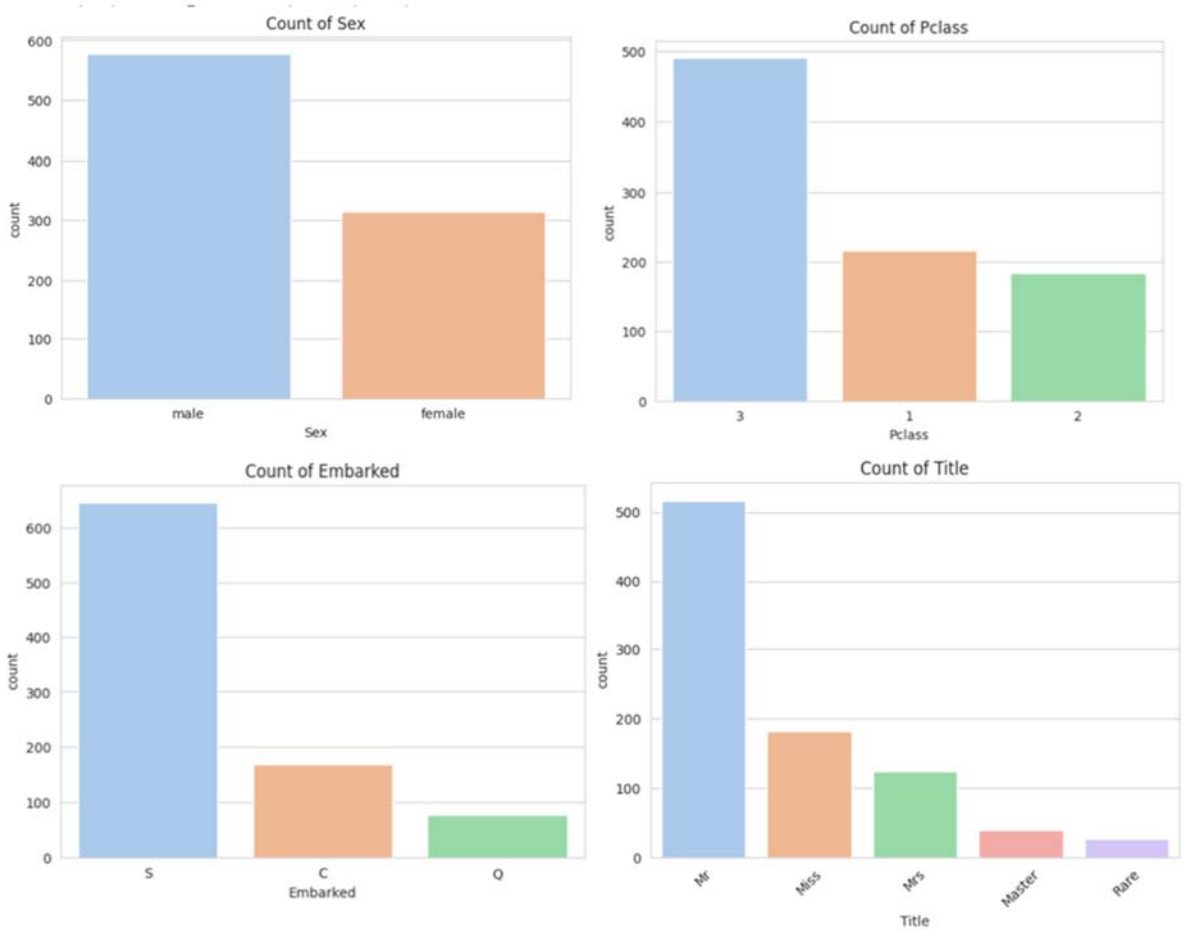
4.1 Univariate Analysis

Exploratory Data Analysis (EDA) began with univariate analysis to understand the distribution and characteristics of individual features within the dataset. For numerical variables such as Age, Fare, SibSp (number of siblings or spouses aboard), and Parch (number of parents or children aboard), histograms were plotted to visualize their frequency distributions. The Age feature showed an approximately normal distribution with a slight right skew, reflecting a passenger population that was mostly adults but included younger children and some elderly passengers. The Fare feature exhibited a heavily right-skewed distribution with a small number of passengers paying very high fares, indicating variability in ticket prices largely driven by passenger class and accommodations. Family-related variables SibSp and Parch revealed that most passengers traveled alone or with a small family group, with family sizes predominantly ranging from one to three members.



Turning to categorical variables, bar plots for Sex, Passenger Class (Pclass), Embarked port, and Title highlighted important demographic patterns. The dataset consisted mainly of male passengers, consistent with historical records. The majority of passengers were traveling in third class, representing the largest group on board. Regarding embarkation points, Southampton ('S') was the most common port of boarding, followed by Cherbourg and Queenstown. Titles extracted from passenger names were dominated by 'Mr' and 'Miss', reflecting the prevalent social groups

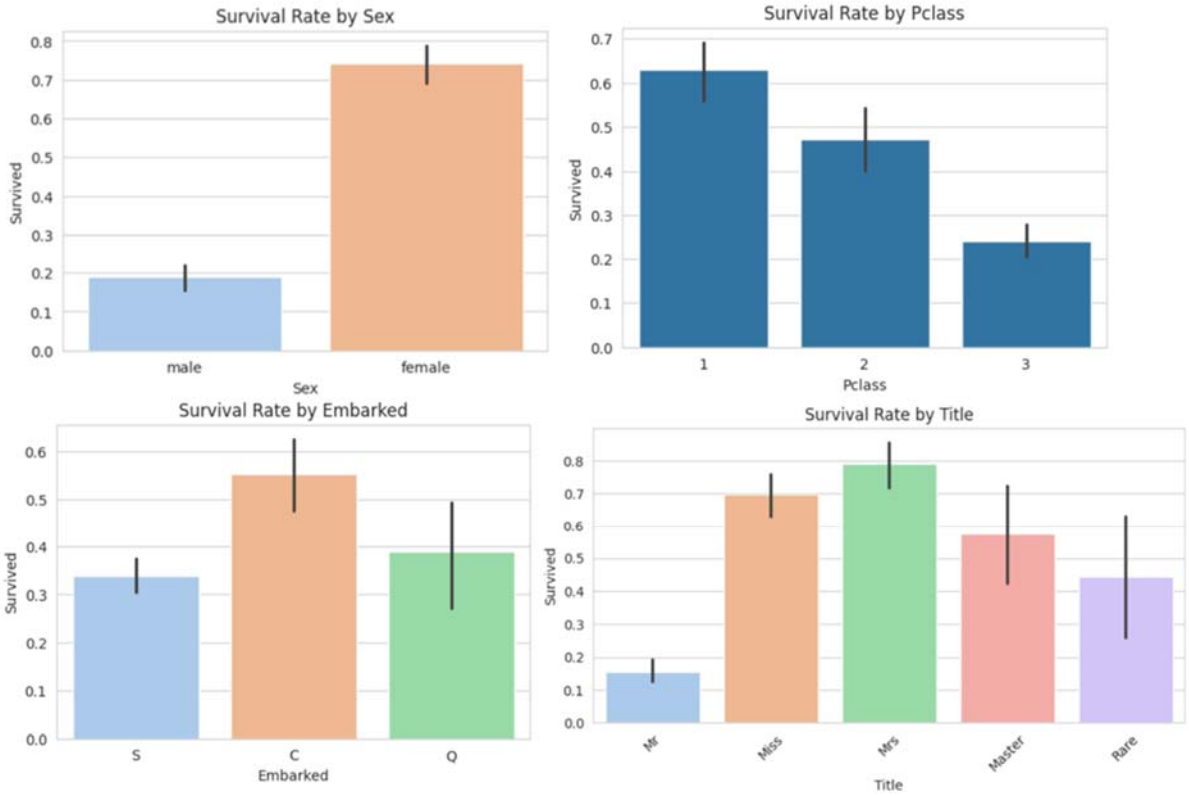
aboard, while other titles occurred less frequently. These univariate insights laid the foundation for deeper bivariate and multivariate analyses.



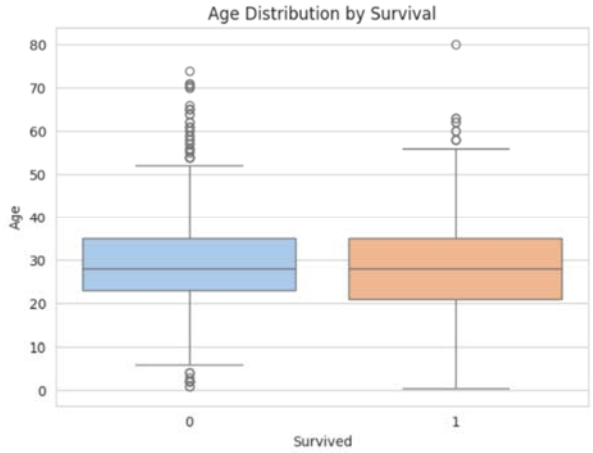
4.2 Bivariate Analysis

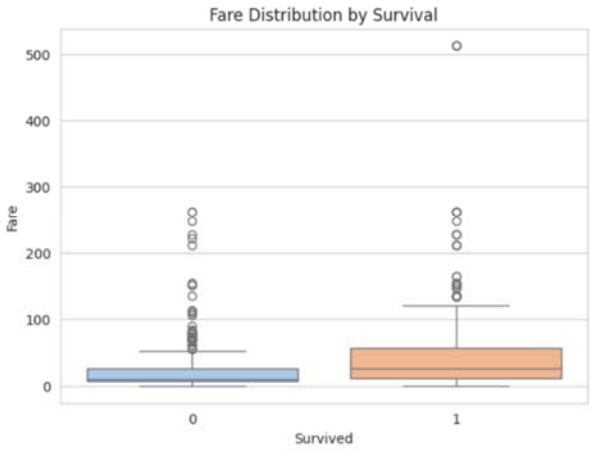
Building upon the univariate insights, bivariate analysis was conducted to explore how individual features relate to the survival outcome. This involved examining the distribution of survival rates across categories of key variables and assessing differences in numerical features between survivors and non-survivors. Visualizations such as bar plots and boxplots were used to highlight these relationships. Notably, gender exhibited a strong influence: females had a significantly higher survival rate compared to males, consistent with the widely known “women and children first” evacuation policy. Passenger class (Pclass) also showed a clear pattern, with first-class passengers surviving at much higher rates than those in second and third class, likely reflecting differences in access to lifeboats and cabin locations. The port of embarkation had some effect, with passengers boarding at Cherbourg or Queenstown showing marginally better survival than those from Southampton. Titles extracted from names further refined this analysis by capturing

social status; for example, ‘Mrs’ and ‘Miss’ titles correlated with higher survival, while rarer or male-associated titles showed lower survival.



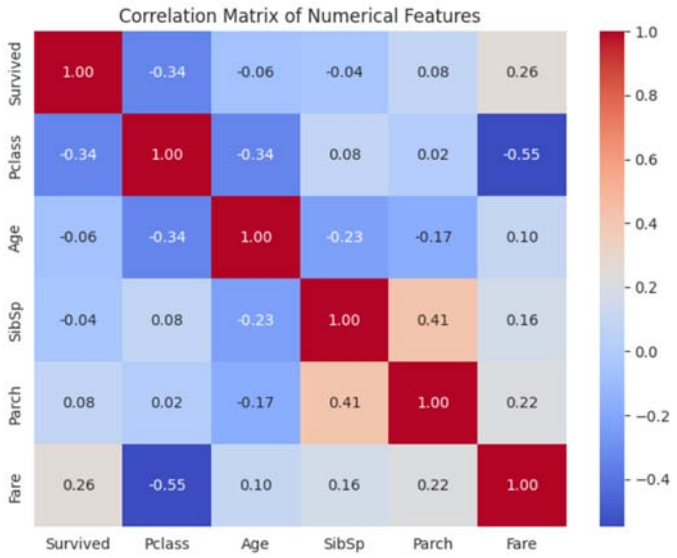
Age distributions revealed that younger passengers, including children, generally had higher survival chances, while older passengers tended to have lower survival rates. Fare amounts paid by passengers also correlated positively with survival, as those paying higher fares were more likely to be in first or second class. These bivariate insights confirmed the importance of demographic and socioeconomic factors in survival and guided feature selection and engineering for modeling.





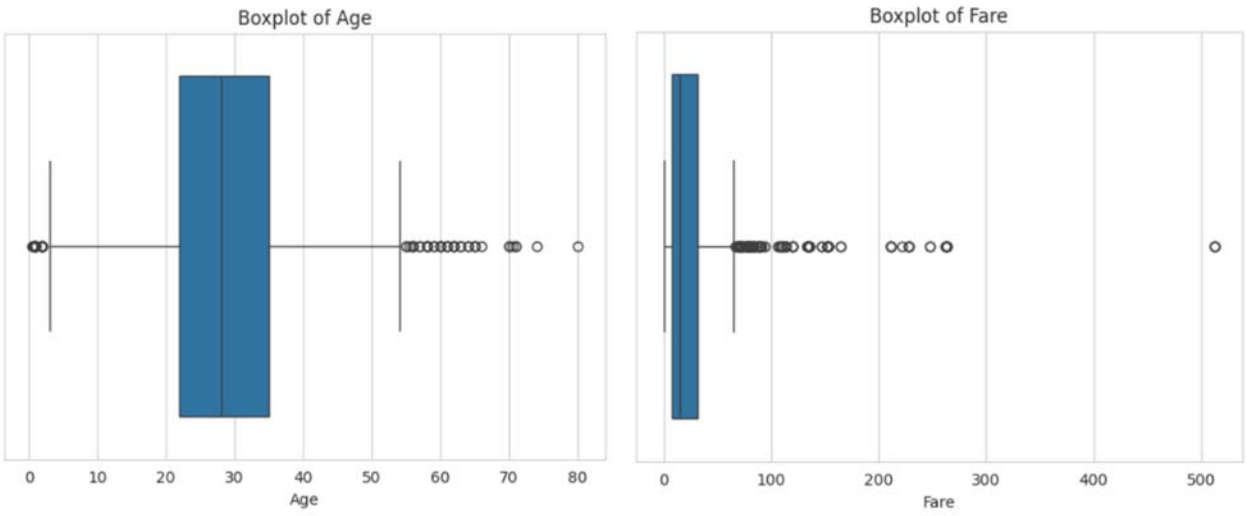
4.3 Correlation Analysis

To further quantify the relationships between numerical features and the target variable, a correlation matrix was computed. The matrix measures the strength and direction of linear associations between pairs of variables, with values ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation). The survival outcome showed a moderate negative correlation with Passenger Class (Pclass) at approximately -0.34, indicating that passengers in higher classes (lower Pclass number) had better survival chances. Fare exhibited a moderate positive correlation of about 0.26 with survival, reflecting that those who paid higher fares were more likely to survive, which aligns with the socioeconomic stratification on board. Age showed a weak negative correlation (-0.065) with survival, suggesting that younger passengers had slightly higher survival odds, although this effect was not as strong as other factors. Family-related variables SibSp (siblings/spouses aboard) and Parch (parents/children aboard) demonstrated very weak correlations with survival (-0.035 and 0.082 respectively), indicating that family size alone was not a strong predictor of survival without further feature engineering. Additionally, the correlations among features highlighted that Pclass and Fare were strongly negatively correlated (-0.55), as higher class passengers typically paid more for their tickets. SibSp and Parch were moderately positively correlated (0.41), reflecting the natural relationship between different family members aboard. These insights emphasized the importance of class and fare as primary predictors while guiding the incorporation of family-related features into the modeling process.



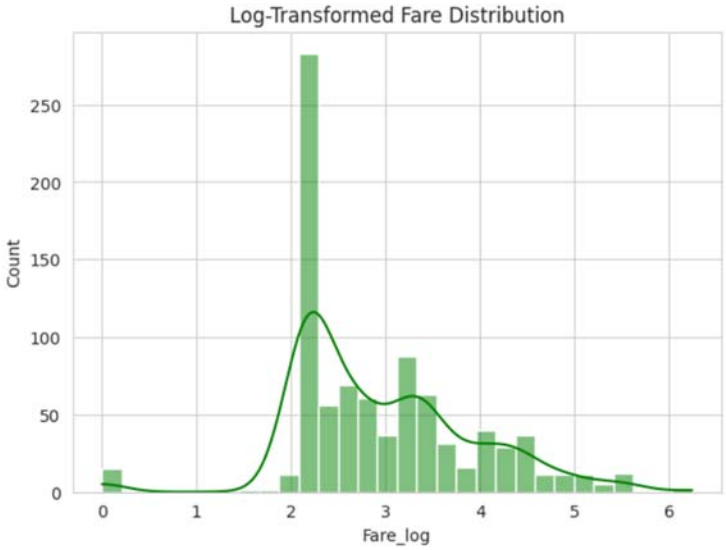
4.4 Outlier Detection

Outlier detection was focused primarily on the continuous numerical features Age and Fare, as extreme values in these variables could disproportionately influence model training and predictions. Visual inspection through boxplots revealed several outliers in both features. In the Age distribution, outliers represented very young children (including infants) and elderly passengers, which are realistic and important groups in the context of survival analysis; hence, these outliers were retained to preserve valuable information. The Fare variable exhibited more pronounced outliers with a small subset of passengers paying exceptionally high ticket prices, reflecting premium accommodations. Such extreme Fare values skewed the distribution and could negatively impact model stability. To mitigate this, we applied a log-transformation to the Fare feature using the formula log(Fare + 1), which compressed the range of high values and normalized the distribution to a more symmetric shape. This transformation helps models better learn from the data by reducing the undue influence of extreme fares without discarding any records. The detection and appropriate treatment of these outliers ensured that the dataset maintained its integrity while improving the reliability and robustness of subsequent modeling steps.



4.5 Missing Data Visualization

To gain a clearer understanding of the distribution and pattern of missing values within the datasets, we employed specialized visualization tools. Using the missingno library, we generated missing data matrices for both the training and test sets. These visualizations graphically displayed the presence and absence of data points across all features, making it easier to identify systematic missingness and potential dependencies. The plots confirmed that the 'Cabin' feature had the most extensive missing data, with large contiguous gaps indicating many passengers lacked cabin information. In contrast, missing values in 'Age', 'Embarked', and 'Fare' were scattered and relatively sparse, suggesting missingness was random rather than systematic. Understanding these patterns was critical for informed imputation strategies, as it indicated that imputing missing 'Cabin' data would require special handling (such as categorizing missing values separately), while numerical features like 'Age' could be reliably filled using statistical measures. Visualizing missing data helped validate earlier assumptions and ensured that no hidden or complex missing data patterns were overlooked before model building.



5. Feature Engineering

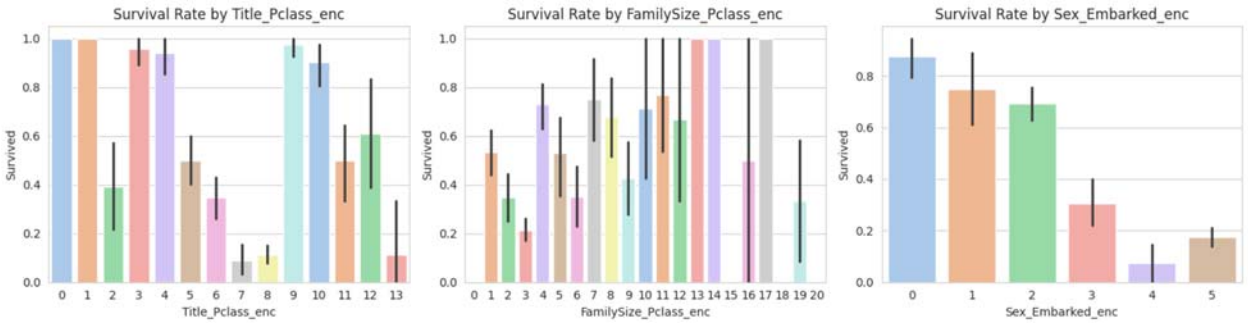
Feature engineering is a crucial step that transforms raw data into meaningful features, improving the model's ability to learn and predict accurately. In this project, we began by creating a new feature called FamilySize, which combines the number of siblings/spouses aboard (SibSp), the number of parents/children aboard (Parch), and the passenger themselves. This aggregated feature captures the total number of family members traveling together and can influence survival chances by reflecting social dynamics and evacuation priorities during the disaster. Analyzing FamilySize revealed that most passengers traveled alone or with small family groups, providing insight into passenger distribution aboard the Titanic.

Since machine learning models require numerical inputs, all categorical features were converted into numeric form using label encoding. We encoded variables such as Sex, which differentiates males and females; Embarked, representing the port of embarkation; and Title, which was extracted from passenger names and grouped into common and rare categories. Label encoding assigns unique integer values to each category without implying any order, enabling the model to process these features effectively.

To capture the interplay between multiple features, we generated several interaction features by combining categories from two different variables. For example, Sex_Pclass combined gender with passenger class, Title_Pclass paired social title with class, FamilySize_Pclass merged family size with class, and Sex_Embarked linked gender with embarkation port. These interaction features were created by concatenating string representations of the categories and then encoded consistently across the training and test sets. This approach allows the model to learn

complex patterns, such as how survival odds differ for a female in first class versus a male in third class, beyond the influence of single features alone.

Numerical features such as Age, log-transformed Fare, and FamilySize were scaled using standardization, which adjusts their values to have a mean of zero and a standard deviation of one. This scaling is especially important for algorithms sensitive to feature magnitude, ensuring that no feature dominates others simply because of its scale. The scaler was fit on the training data and applied identically to the test data to maintain consistency and avoid data leakage.



After completing these transformations, the dataset was free from missing values and contained enriched, well-prepared features combining raw data, engineered variables, and encoded interactions. This comprehensive feature engineering process improved the dataset's expressiveness, enabling machine learning models to better capture the socio-demographic and travel-related nuances that influenced survival on the Titanic.

6. Modeling

6.1 Train-Validation Split

To ensure that the predictive models developed for this project could generalize well to unseen data, the original training dataset was divided into two subsets: 80% of the data was allocated for training the models, while the remaining 20% was reserved for validation. This split enabled us to train models on a majority portion of the data while using the validation set to evaluate their performance objectively. By assessing model accuracy and other metrics on this held-out validation set, we could detect issues such as overfitting and better estimate how well the models might perform on new, unseen passengers. This step is fundamental in building reliable machine learning models and ensures that the results are not overly optimistic due to memorizing training data.

6.2 Baseline Models

For initial modeling, we implemented two baseline classifiers to establish benchmark performance. The first was Logistic Regression, a widely used linear model for binary classification that estimates the probability of survival based on weighted input features. This model achieved a validation accuracy of approximately 80.4%, demonstrating reasonable predictive power despite its simplicity. The precision and recall metrics indicated balanced performance across both survivors and non-survivors.

Validation Accuracy on Logistic Regression: 0.8044692737430168

	precision	recall	f1-score	support
0	0.82	0.86	0.84	105
1	0.78	0.73	0.76	74
accuracy			0.80	179
macro avg	0.80	0.79	0.80	179
weighted avg	0.80	0.80	0.80	179

To capture more complex relationships in the data, we also trained a Random Forest classifier, an ensemble of decision trees that can model non-linear feature interactions. The Random Forest improved the validation accuracy to about 83.8%, with better precision and recall, particularly for identifying survivors. These baseline models provided important reference points for evaluating more advanced modeling techniques and hyperparameter tuning in subsequent steps.

Random Forest Validation Accuracy: 0.8379888268156425

	precision	recall	f1-score	support
0	0.87	0.85	0.86	105
1	0.79	0.82	0.81	74
accuracy			0.84	179
macro avg	0.83	0.84	0.83	179
weighted avg	0.84	0.84	0.84	179

6.3 Hyperparameter Tuning and Advanced Modeling

To further improve model performance, we employed the XGBoost algorithm, a powerful gradient boosting framework that combines multiple weak learners to create a strong predictive model. Recognizing the importance of optimal hyperparameter settings, we used RandomizedSearchCV to efficiently explore a wide range of parameters, including the number of trees, tree depth, learning rate, subsampling ratios, and regularization terms. This systematic tuning process allowed us to identify the combination of parameters that maximized validation accuracy. The tuned XGBoost model achieved a validation accuracy of approximately 81%, demonstrating competitive performance with robust precision and recall scores. While slightly below the Random Forest baseline, the XGBoost model's flexibility and interpretability through tools like SHAP make it a strong candidate for final deployment. This stage highlighted the critical

role of hyperparameter tuning in extracting the best performance from advanced machine learning algorithms.

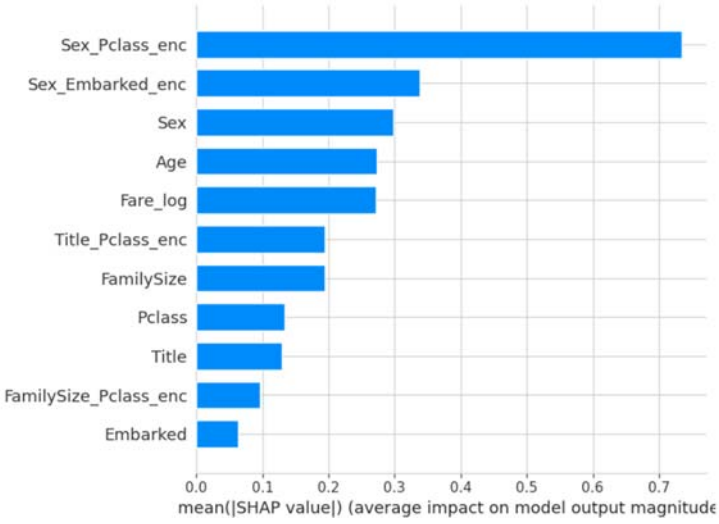
```
Validation Accuracy after tuning: 0.8100558659217877
              precision    recall  f1-score   support

           0       0.82      0.87      0.84       105
           1       0.79      0.73      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
```

7. Model Interpretation & Explainability

Understanding how a predictive model makes decisions is essential for building trust and gaining insights into the factors driving outcomes. To this end, we applied SHAP (SHapley Additive exPlanations), a state-of-the-art interpretability framework that quantifies the contribution of each feature to individual predictions and overall model behavior. Using SHAP values, we identified that passenger class, sex, fare paid, and family size were the most influential factors determining survival probability. Global summary plots revealed how these features impacted predictions across the dataset, while local explanations provided insight into why specific passengers were classified as survivors or non-survivors. This interpretability not only validated domain knowledge such as the higher survival rates for females and first-class passengers but also highlighted the nuanced interplay of multiple features. Incorporating model explainability strengthened the reliability and transparency of our predictive approach, ensuring it aligns with both statistical rigor and historical context.



8. Final Model Training and Prediction

After thorough experimentation, evaluation, and interpretation, the final step involved retraining the best-performing model on the entire available training dataset to maximize its learning capacity before making predictions on unseen data. Leveraging all the labeled passenger records allowed the model to capture the most comprehensive patterns and relationships inherent in the Titanic dataset. We utilized the optimized XGBoost classifier, selected for its balance of high accuracy and interpretability, ensuring that the model benefits from both robust predictive power and transparency in decision-making. Once retrained, the model was applied to the test dataset, which contained passenger information without survival labels. The model generated survival predictions for each passenger, classifying them as either survived or not survived based on the complex interplay of features such as passenger class, sex, age, fare, family size, and social title. These predictions were then organized into a submission file adhering strictly to the Kaggle competition format, consisting of PassengerId and the corresponding Survived label. This submission was ready for upload to Kaggle's platform to evaluate the model's real-world predictive performance against the hidden test labels. This final phase epitomizes the full data science workflow, from data acquisition and preprocessing through modeling and validation to prediction and deployment, demonstrating how methodical analysis and advanced algorithms can extract meaningful insights and accurate forecasts from historical tragedy data.

9. Conclusion

In conclusion, this project successfully demonstrated the application of a complete data science pipeline to predict survival outcomes on the Titanic. Through careful data acquisition, cleaning, and thoughtful handling of missing values, we prepared a robust dataset that reflects the complex realities of the historical event. Exploratory data analysis uncovered key demographic and socioeconomic factors influencing survival, such as gender, passenger class, fare, and family size. Advanced feature engineering, including interaction terms and scaling, enriched the dataset to better capture subtle patterns. Modeling efforts ranged from straightforward logistic regression to sophisticated ensemble methods like Random Forest and XGBoost, with hyperparameter tuning optimizing predictive accuracy. Model interpretation using SHAP values provided transparency and affirmed domain knowledge, reinforcing confidence in the results. The final model achieved strong validation performance and generated predictions ready for evaluation in the Kaggle competition. This work not only showcases the power of machine learning to analyze

historical datasets but also highlights the importance of rigorous methodology, interpretability, and continuous refinement in building trustworthy predictive models.