🛳️ Titanic Survival Prediction: Full Data Science Workflow & Model Evaluation

---

🔍 1. Data Cleaning

A meticulous data cleaning process was implemented to ensure the dataset was analysis-ready:

Initial Inspection: Displayed the first few rows (head()), reviewed dataset shape, and printed data types.

Unique Values Analysis: Each column was assessed for the number of unique values to identify categorical features, high-cardinality columns, and redundant or low-variance features.

Missing Values Detection: Systematically checked each column for missing data using .isnull().sum() and visualized them to understand patterns.

Survival Rate vs Missing Values: Performed grouped analysis of survival outcomes against features with missing values (e.g., Cabin, Age, Embarked) to assess whether missingness was random or correlated with survival.

🛠 Imputation Strategy:

Age: Imputed using regression-based imputation using related features like Pclass, Fare, SibSp, Parch.

Embarked: Filled using mode imputation due to its low missing count and clear modal value.

Cabin: Heavily missing, filled with "Unknown" or sample placeholders to retain structure.

🔄 2. Data Transformation

Categorical Encoding:

Used LabelEncoder to convert categorical columns (Sex, Embarked, Name, Cabin, Ticket) into numeric codes.

Feature Scaling:

Applied StandardScaler to numerical columns like Age, Fare, PassengerId, etc. to normalize them.

All transformations were done post-imputation to ensure consistency.

📊 3. Exploratory Data Analysis (EDA)

Univariate Analysis:

Visualized distributions of Age, Fare, SibSp, Parch, etc.

Measured skewness and kurtosis to identify the shape of each distribution and detect outliers.

Outlier Detection: Used boxplots, histograms, and statistical summaries to detect anomalies.

Bivariate Analysis:

Created scatter plots of each feature against Survived to check linear or non-linear relationships.

Correlation Analysis:

Generated a correlation matrix to find highly correlated features.

Used cross-tabulation for features like Pclass, Sex, and Embarked vs survival to find patterns.

Impact of Missingness:

Compared survival rates for passengers with missing Cabin, Age, or Embarked values to see if missing data held predictive value.

🧠 4. Feature Selection / Engineering

✅ Techniques Used:

Logistic Regression for Feature Importance:

Trained logistic regression on full data.

Selected features based on statistically significant coefficients.

PCA (Principal Component Analysis):

Reduced dimensionality while retaining variance.

Compared PCA-reduced features with original features in model performance.

Random Forest Feature Importances:

Trained RF model on full data.

Retained features with importance > 0.05.

Key features: Sex, Age, Fare, Name, Cabin, Ticket, etc.

🤖 5. Model Building / Analytics Task

Two models were used:

Logistic Regression

Random Forest Classifier

Each was trained in three scenarios:

Full dataset

Selected features only

PCA-reduced features

🔢 Evaluation Metrics:

Accuracy

R^2 Score

Mean Squared Error (MSE)

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

Confusion Matrix

Classification Report (Precision, Recall, F1-Score)

🧾 Conclusion

The best-performing model was Random Forest with selected features, achieving ~83% accuracy with strong precision and recall.

PCA-reduced models performed slightly lower, showing that interpretability and raw feature significance matter in this problem.

Visual analysis through confusion matrices and classification heatmaps confirmed balanced performance.

The model generalizes well and offers good real-world applicability.

🔧 Possible Improvements:

Use GridSearchCV for hyperparameter tuning.

Try advanced models like XGBoost or LightGBM.

Further improve feature engineering by handling high-cardinality features (e.g., Ticket, Name) more thoughtfully.

Titanic Survival Prediction - Summary of Models and Evaluation

Overview

This document summarizes the six models used in the Titanic survival prediction project, detailing their evaluation metrics and how results were computed. The models are based on two main algorithms: Logistic Regression and Random Forest Classifier. Each was tested in three different configurations — full dataset, selected important features, and after dimensionality reduction using PCA.

Model Summary

Model Name	Features Used	Accuracy	R^2 Score	MSE	RMSE	MAE
Logistic Regression - Full	All Features	0.79	0.23	0.21	0.46	0.21
Logistic Regression - Selected	Important Features	0.81	0.26	0.19	0.44	0.19
Logistic Regression - PCA	PCA Components	0.78	0.21	0.22	0.47	0.22
Random Forest - Full	All Features	0.84	0.34	0.16	0.40	0.16
Random Forest - Selected	Important Features	0.83	0.29	0.17	0.41	0.17
Random Forest - PCA	PCA Components	0.81	0.27	0.19	0.43	0.19

Conclusion

The Random Forest model trained on selected important features yielded the best performance with an accuracy of 83%. It showed a strong balance across all classification metrics including F1-score and precision. PCA reduced models performed slightly lower due to potential loss of interpretability and information.

Possible improvements include hyperparameter tuning using GridSearchCV, trying advanced ensemble techniques such as XGBoost, and refining the feature engineering process through domain knowledge or automated selection tools.