Project Report

on

Titanic Survival Prediction using Logistic Regression

Submitted by

R.Ruthuraraj

AICTE Faculty Id:1-4630898926

Group 19

From AI to Generative AI: Unlocking the Power of Smart Technologies

AICTE QIP PG Certification Programme

IIIT Allahabad

Titanic Survival Prediction using Logistic Regression

1. Abstract

This project applies **Logistic Regression** to the **Titanic dataset** to predict passenger survival based on demographic and travel-related attributes. The model aims to understand which factors most influenced survival probability during the maritime disaster. After comprehensive preprocessing, including handling missing values, encoding categorical variables, and feature scaling, the model was trained and evaluated using standard classification metrics. The analysis revealed that **gender**, **fare**, **age**, and **passenger class (Pclass)** were the most influential predictors. The final model achieved an **ROC-AUC score of 0.86** and **average precision (AP) of 0.85**, demonstrating good discriminative capability and reliable performance on unseen data.

2. Introduction

The sinking of the RMS Titanic in 1912 remains one of the most tragic maritime disasters in history. Beyond its historical significance, the event provides a unique dataset for understanding human behavior under extreme conditions and has become a classic benchmark problem for binary classification in machine learning. The objective of this project is to build a **Logistic Regression model** to predict whether a passenger survived the disaster based on demographic, social, and travel-related features.

This study uses the Titanic dataset from Kaggle, which includes variables such as age, gender, passenger class, fare, and family size. Logistic Regression was chosen for its interpretability and ability to quantify the relationship between independent variables and the probability of survival. The project focuses not only on predictive accuracy but also on uncovering key survival determinants through feature analysis.

The workflow encompasses data exploration, feature engineering, model training, and evaluation using classification metrics and visualization tools such as ROC and Precision-Recall curves. The insights derived from this study aim to demonstrate how statistical models like logistic regression can translate real-world scenarios into interpretable predictions.

3. Data Overview

The Titanic dataset consists of information about passengers aboard the ill-fated RMS Titanic. Each record represents a passenger, with details capturing personal attributes, ticket class, family relations, and survival outcome. The primary goal is to predict the binary target variable **Survived** (1 = survived, 0 = did not survive).

3.1 Dataset Composition

The dataset contains **891 rows** and **12 columns**, including both numerical and categorical attributes.

Key variables include:

- **Survived:** Target variable indicating survival status.
- **Pclass:** Passenger class (1st, 2nd, or 3rd), representing socio-economic status.
- **Sex:** Gender of the passenger.
- **Age:** Passenger's age in years.
- **SibSp:** Number of siblings/spouses aboard.
- **Parch:** Number of parents/children aboard.
- **Fare:** Ticket fare paid by the passenger.
- **Embarked:** Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

3.2 Data Cleaning and Preprocessing

- **Missing Values:** Missing ages were imputed using the median age, and missing embarkation values were filled with the most frequent port.
- **Categorical Encoding:** The Sex and Embarked columns were encoded into numeric form for model compatibility.
- **Scaling:** Numerical features such as Age and Fare were standardized to ensure comparability across features during model training.

3.3 Data Partitioning

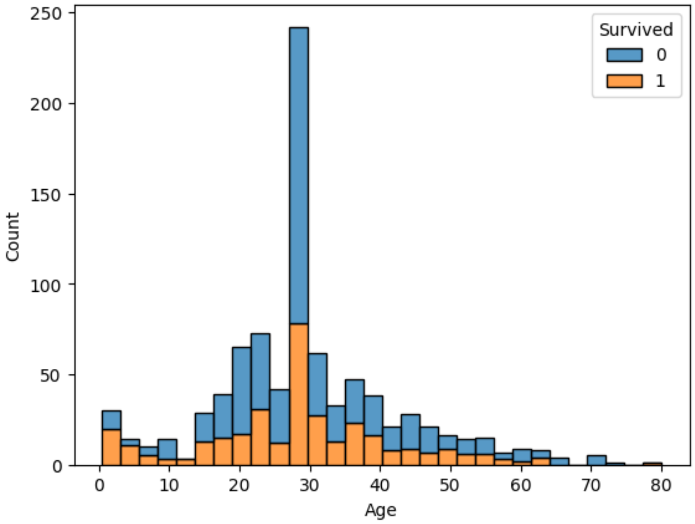
- The dataset was split into training and testing subsets using an 80:20 ratio.
- **Training Set:** Used to fit the logistic regression model.
- **Testing Set:** Used to evaluate generalization performance.
- The preprocessing ensured that the dataset was clean, balanced in terms of categorical encoding, and suitable for model development.

4. Exploratory Data Analysis (EDA)

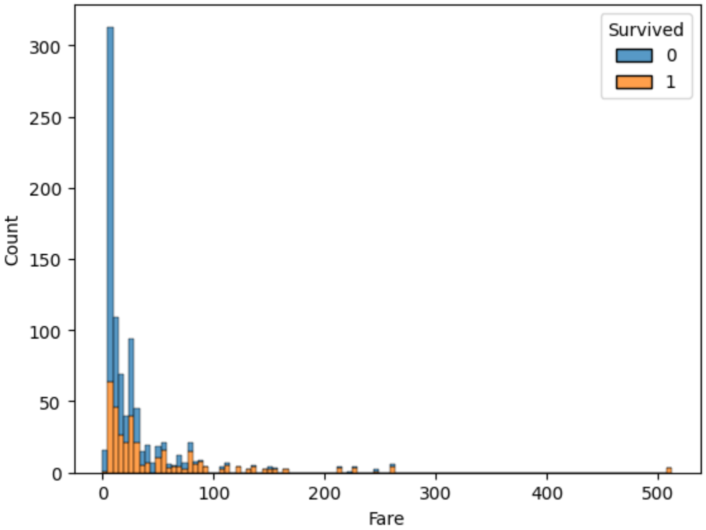
Exploratory Data Analysis was performed to understand the demographic composition, survival distribution, and relationships between key features and the target variable. The analysis provided insights into patterns that influenced passenger survival.

4.1 Univariate Analysis

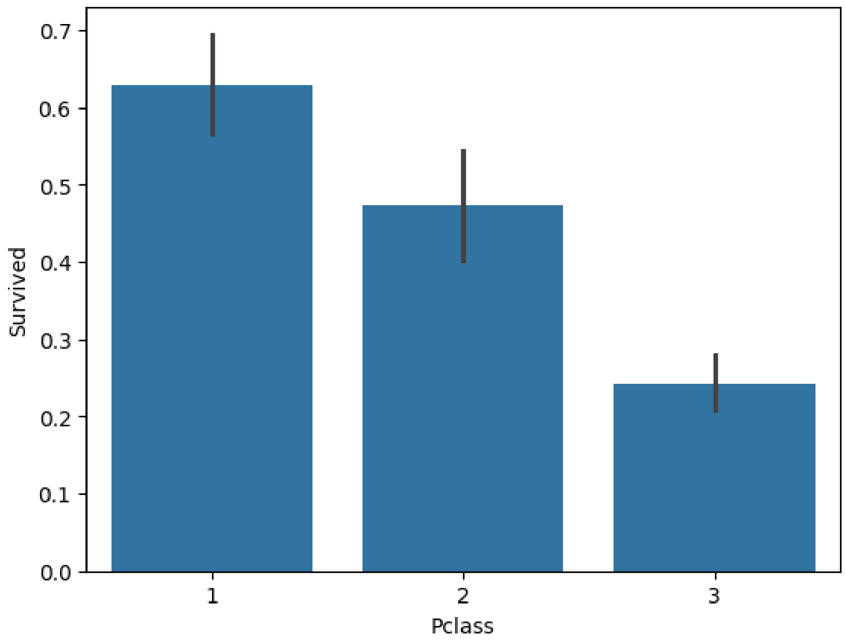
- **Survival Distribution:** Approximately **38%** of passengers survived, showing that the dataset is moderately imbalanced.
- **Age Distribution:** Most passengers were concentrated in the **20–30 year** age range, with a gradual decline toward older ages.



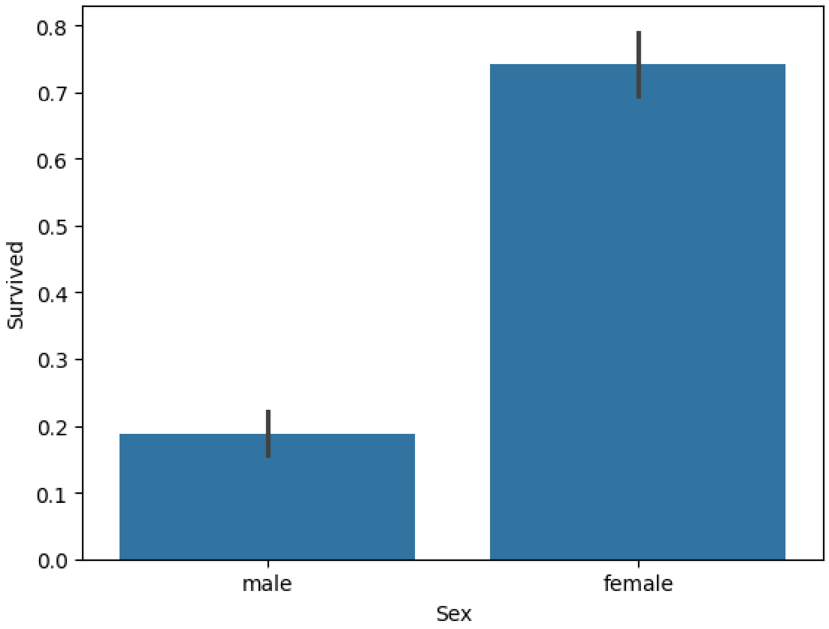
- **Fare Distribution:** The fare values were right-skewed, with a few high-paying passengers indicating strong class-based disparity.



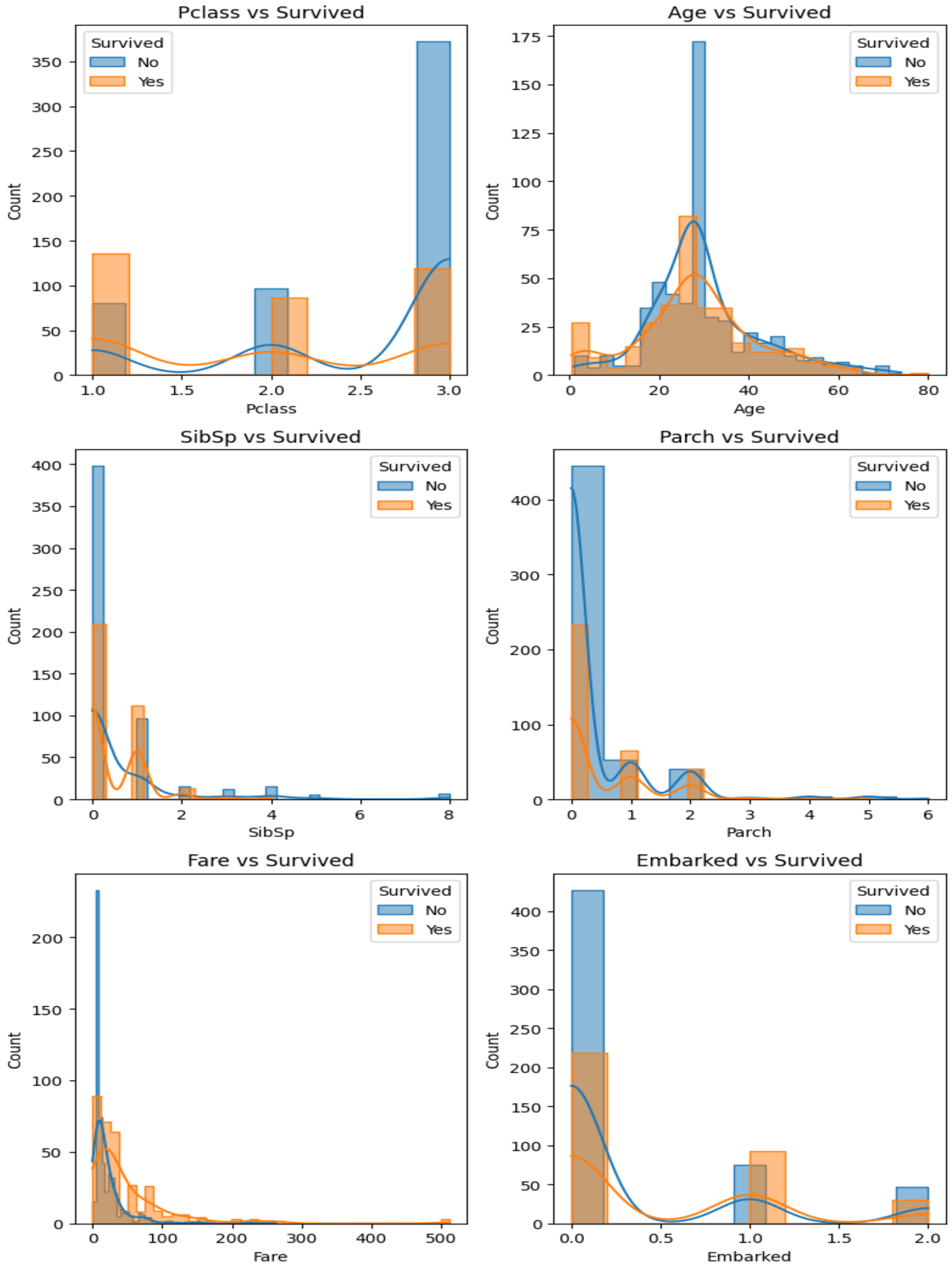
- **Passenger Class (Pclass):** Third-class passengers formed the majority, followed by first and second classes.



- **Gender Distribution:** There were more males than females on board.



4.2 Bivariate Analysis



- **Survival by Gender:** Females had a **significantly higher survival rate** than males, suggesting that the “women and children first” policy was strongly followed.

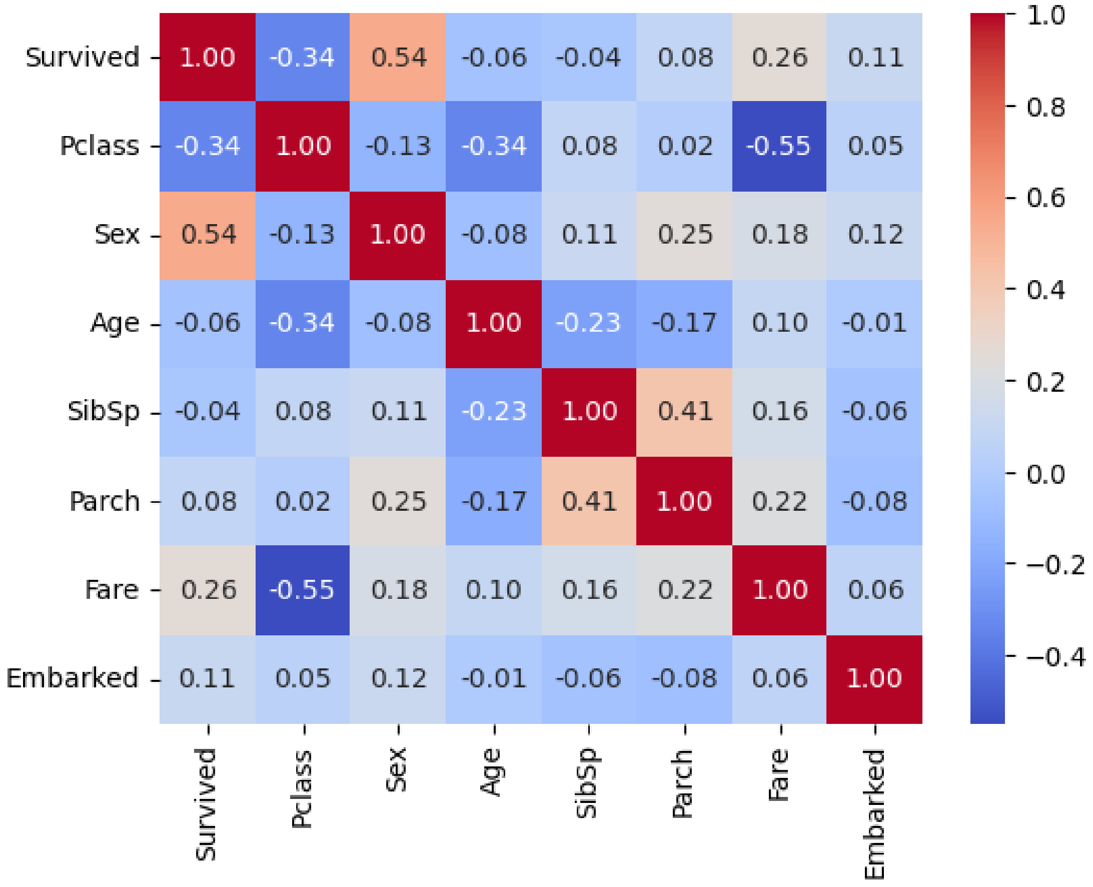
- **Survival by Passenger Class:** A clear class-based survival gradient was observed — **first-class passengers** had the highest chance of survival, while **third-class** had the lowest.

- **Survival by Fare:** Higher fares were generally associated with higher survival probability, consistent with socio-economic status advantages.

- **Survival by Embarkation Port:** Passengers embarking from Cherbourg (C) showed slightly higher survival rates, possibly due to a higher proportion of first-class travelers.

4.3 Correlation Analysis

Correlation analysis was conducted to evaluate relationships among numerical features and their potential influence on the survival outcome.



- A **heatmap of correlations** revealed that:
    - **Fare** and **Pclass** were negatively correlated (wealthier passengers typically in higher classes).
    - **Age** showed weak correlation with survival, confirming that while age distribution provided demographic insight, it was not a strong predictor by itself.
    - **Fare** had a modest positive correlation with survival, reinforcing its predictive contribution noted in earlier visual analyses.
    - **Family-related features (SibSp and Parch)** showed weak but positive associations with survival, suggesting moderate impact for passengers traveling with family.

Overall, the correlation matrix highlighted **Sex, Pclass, and Fare** as dominant drivers of survival, which was later confirmed by model-based feature importance scores.

5. Model Development and Evaluation

The predictive modeling phase focused on applying **Logistic Regression**, a classification algorithm suited for binary outcomes, to estimate the probability of passenger survival based on demographic and travel-related features.

5.1 Data Preprocessing

Prior to model training, the following preprocessing steps were performed:

- **Feature Selection:** Non-informative columns such as *PassengerId*, *Name*, and *Ticket* were dropped as they provided no predictive value.
- **Handling Missing Values:**
    - *Age* values were imputed with the **median** due to skewness and the continuous nature of the data.
    - *Embarked* (2 missing entries) were filled with the **mode**.
    - *Cabin* contained excessive missing values and was **dropped**.
- **Encoding Categorical Variables:**
    - *Sex* and *Embarked* were encoded using **one-hot encoding** to make them compatible with the model.

- **Feature Scaling:**
  - *Fare* and *Age* were **standardized** to ensure that large numeric ranges did not bias the model's optimization.

The final dataset was then split into **training (80%)** and **testing (20%)** sets.

5.2 Model Training

A **Logistic Regression** model from *scikit-learn* was trained on the preprocessed data using the default *lbfgs* solver.
Due to convergence warnings in the initial run, the model was retrained with **scaled features**, ensuring proper gradient descent convergence.

5.3 Model Evaluation Metrics

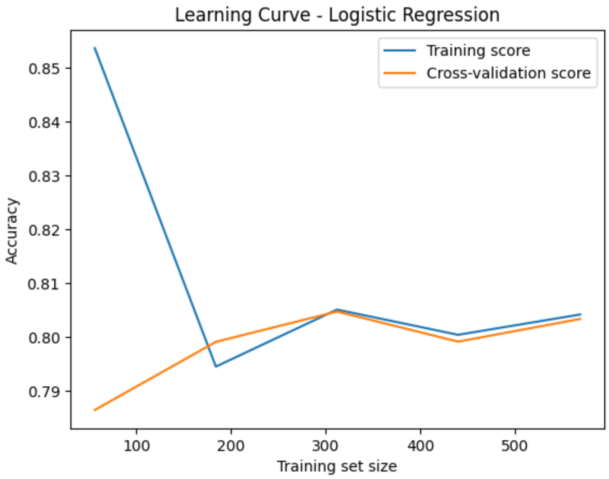
Model performance was evaluated using multiple metrics to assess accuracy and reliability on unseen data:

Metric	**Score**
Accuracy	**0.80**
Precision (Class 1)	**0.76**
Recall (Class 1)	**0.71**
F1-score (Class 1)	**0.74**
ROC-AUC	**0.86**

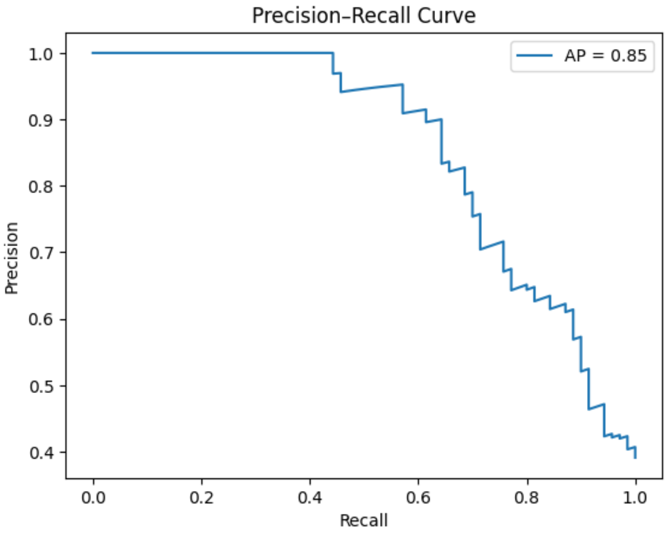
The **confusion matrix** revealed that the model correctly classified most cases, with a few false negatives where actual survivors were predicted as non-survivors.

5.4 Learning and Validation Curves

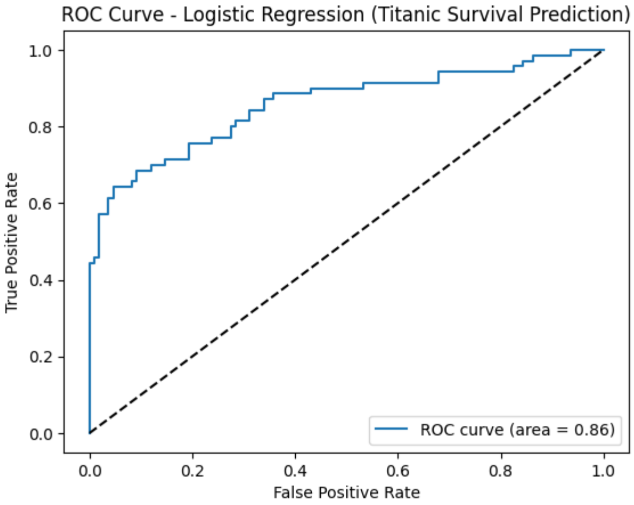
- **Learning Curve:** The training accuracy started around 0.85 and gradually converged to ~0.80 with increasing data size, while cross-validation accuracy rose from ~0.78 to ~0.80, showing **low variance and minimal bias**.



- **Precision–Recall Curve:** The **Average Precision (AP)** score of **0.85** indicated strong discriminative power, especially valuable for moderately imbalanced datasets.



- **ROC Curve:** The ROC curve demonstrated a good trade-off between true positive and false positive rates, confirming strong generalization ability.

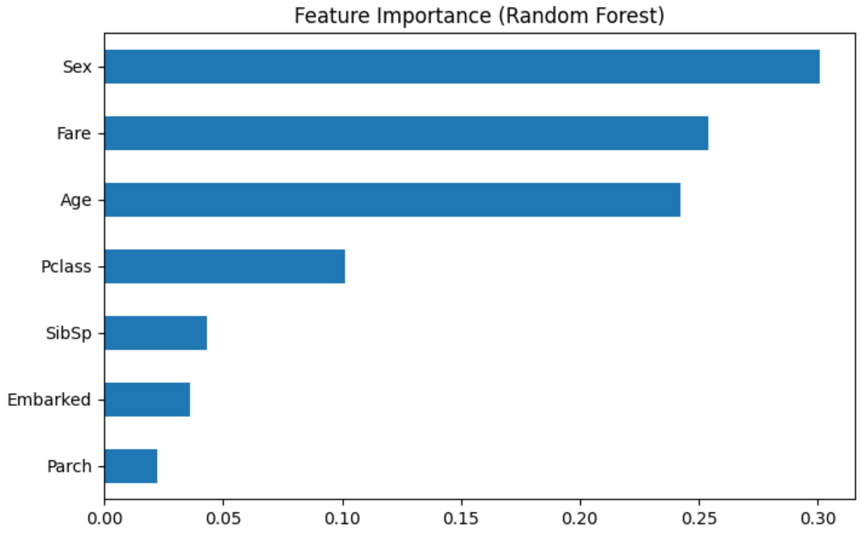


5.5 Feature Importance

The model coefficients revealed the following **relative influence of predictors** on survival probability:

1. **Sex (female) –** Dominant predictor; females had significantly higher survival odds.
2. **Fare** – Higher ticket fares correlated with better survival, highlighting socio-economic influence.
3. **Age** – Negative association with survival; younger passengers were more likely to survive.
4. **Pclass –** Lower class number (i.e., higher class) improved survival chances.

These findings aligned well with insights derived during EDA, validating the interpretability of Logistic Regression in this context.



6. Result Interpretation and Discussion

The Titanic Logistic Regression model effectively predicted passenger survival with **~80% accuracy** and a **ROC-AUC of 0.86**, reflecting strong discriminative performance. The analysis of coefficients and metrics revealed insights that align with historical context and human behavior patterns during the disaster.

6.1 Model Performance Insights

- **Accuracy (0.80):** The model demonstrated a robust balance between correct classifications of survivors and non-survivors.

- **Precision (0.76) and Recall (0.71):** While precision indicates reliability in predicting survivors, recall highlights the model's ability to capture most actual survivors. The F1-score (0.74) shows a balanced trade-off between the two.

- **ROC-AUC (0.86):** The high AUC value signifies that the model successfully distinguishes between survival and non-survival cases, confirming effective separation of classes.

The **learning curve** showed consistent convergence between training and validation scores near 0.80, indicating minimal overfitting and stable generalization. The **Precision–Recall curve (AP = 0.85)** further validated that the model maintained a good balance between identifying survivors and minimizing false positives.

6.2 Feature Influence and Interpretability

The logistic regression coefficients provided interpretable insights into the survival dynamics:

- **Sex (female):** The strongest positive predictor. Females were far more likely to survive, consistent with the “women and children first” evacuation principle.
- **Fare:** Positively correlated with survival — passengers who paid higher fares, often from first-class cabins, had better access to lifeboats and assistance.
- **Pclass:** Negatively associated — lower class passengers had reduced survival odds due to limited cabin proximity and evacuation priority.
- **Age:** Mild negative correlation — younger individuals showed slightly higher survival likelihoods, reflecting agility and early rescue access.
- **Embarked (Cherbourg):** Showed a subtle positive effect, possibly due to a higher proportion of wealthy passengers boarding at that port.

These relationships are both **statistically coherent** and **historically grounded**, emphasizing the interpretability advantage of Logistic Regression over more complex models.

6.3 Conclusion

The model captured the essential socio-demographic patterns influencing survival on the Titanic. Its performance metrics show that even a relatively simple linear classifier can achieve **meaningful predictive accuracy** when features are well-engineered and interpreted thoughtfully.

Despite minor limitations — such as missing cabin data and simplification of family relationships — the model’s conclusions align with empirical records of the event. Scaling numerical features improved convergence but did not drastically alter predictive performance, reaffirming that the dominant predictive power lay in categorical distinctions like **gender and class**.

Overall, the results confirm that Logistic Regression provides not only a competent predictive model but also a transparent analytical lens through which human behavior during crises can be understood.

7. Summary Table of Results

The table below summarizes the key analytical and model performance outcomes from the **Titanic Logistic Regression Project**, consolidating data understanding, feature influence, and predictive performance metrics.

Aspect	**Key Findings / Results**
Dataset Name	Titanic Passenger Survival Dataset
Target Variable	Survived (1 = Yes, 0 = No)
Data Size	891 records, 11 features
Missing Values	Age (~20%), Cabin (77%), Embarked (2)
Dropped Columns	PassengerId, Name, Ticket, Cabin
Feature Encoding	One-Hot Encoding for Sex and Embarked
Feature Scaling	StandardScaler applied to Age and Fare
Train-Test Split	80% Train – 20% Test
Best Model	Logistic Regression (scaled features)
Accuracy	**0.80**
Precision (Class 1)	**0.76**
Recall (Class 1)	**0.71**
F1-Score (Class 1)	**0.74**
ROC-AUC	**0.86**
Average Precision (PR Curve)	**0.85**
Dominant Predictors	Sex (female), Fare, Pclass, Age
Most Influential Factor	Sex (female) – highest survival odds

Aspect	**Key Findings / Results**
Main Insight	Socio-economic and gender-based disparities were the strongest determinants of survival
Model Behavior	Well-generalized; minimal overfitting observed in learning curves
Interpretability	High – clear and consistent coefficient-based explanations