Titanic Survival Analysis

Predict passenger survival on the RMS Titanic using machine learning. This classic dataset demonstra

Jotty SwarmMLComprehensive

February 05, 2026

Contents

Contents

Executive Summary

Predict passenger survival on the RMS Titanic using machine learning. This classic dataset demonstrates fundamental ML concepts including feature engineering, handling missing data, and binary classification.

Key Results

Best Model: Logistic Regression

Performance Metrics:

Metric	Value
Accuracy	0.8324
Precision	0.7910
Recall	0.7681
F1	0.7794
Auc Roc	0.8590

Dataset: 17 features analyzed

Data Profile

Dataset Overview

- **Total Samples:** 891
- **Total Features:** 17

Data Types

Data Type	Count
float64	15

EDA Recommendations

- Age has ~20% missing values - median imputation applied
- Cabin has ~77% missing - dropped from analysis
- Feature engineering added family_size, is_alone, fare_log, is_child

---

Feature Importance Analysis

Feature importance measures how much each feature contributes to the model's predictions. Higher values indicate more influential features.

Top 20 Features

Rank	Feature	Importance
1	age	0.1729
2	fare	0.1324
3	fare_log	0.1252
4	sex	0.1141
5	adult_male	0.1103
6	who	0.0830
7	deck	0.0513
8	pclass	0.0403
9	class	0.0397
10	family_size	0.0356
11	sibsp	0.0242
12	parch	0.0170
13	embarked	0.0163
14	embark_town	0.0155
15	is_child	0.0080
16	is_alone	0.0071
17	alone	0.0071

Feature Importance Visualization

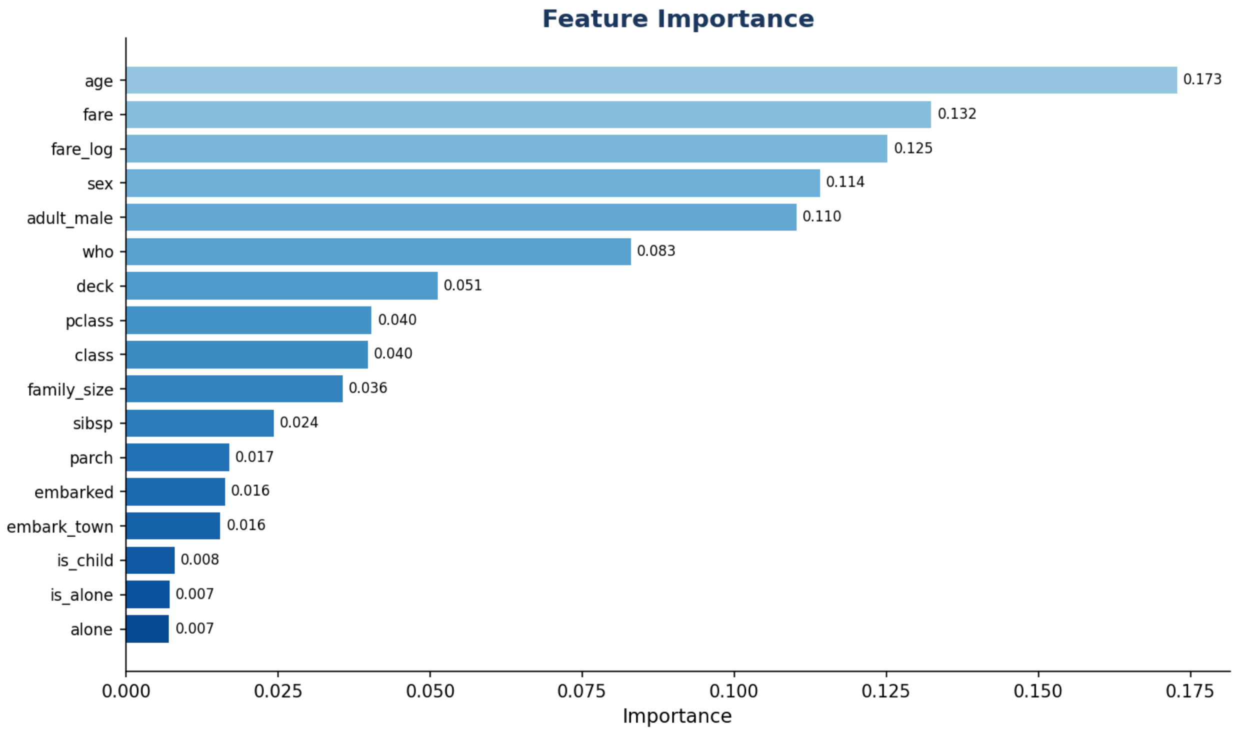

Figure 1: Feature Importance

Model Benchmarking

Multiple machine learning algorithms were evaluated using 5-fold cross-validation. The table below shows the performance of each model.

Model Comparison

Model	CV Score	Std Dev	Test Score	Time (s)
Logistic Regression	0.8105	±0.0203	0.8324	181.08
Gradient Boosting	0.8105	±0.0413	0.8212	0.81
Random Forest	0.7950	±0.0318	0.8156	1.17

Performance Visualization

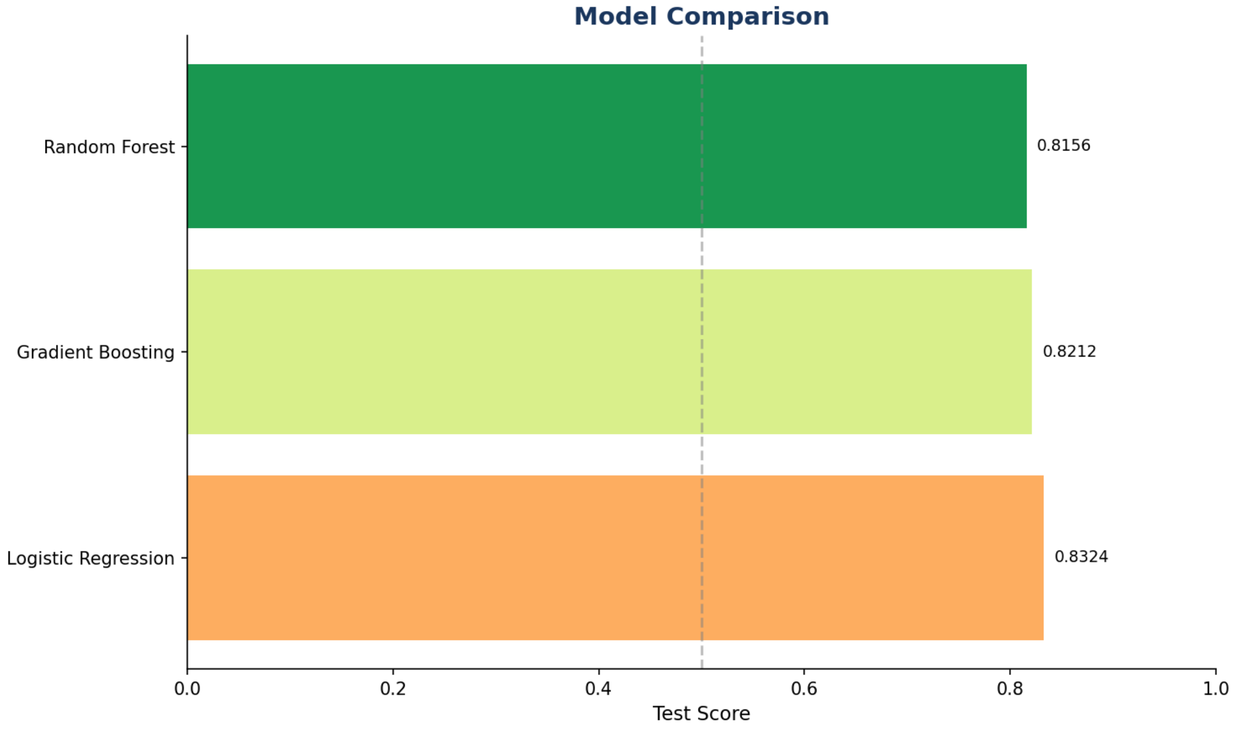

Figure 2: Model Benchmarking

Classification Performance

Classification Report

Class	Precision	Recall	F1-Score	Support
Died	0.857	0.873	0.865	110
Survived	0.791	0.768	0.779	69
Accuracy			**0.832**	

Confusion Matrix

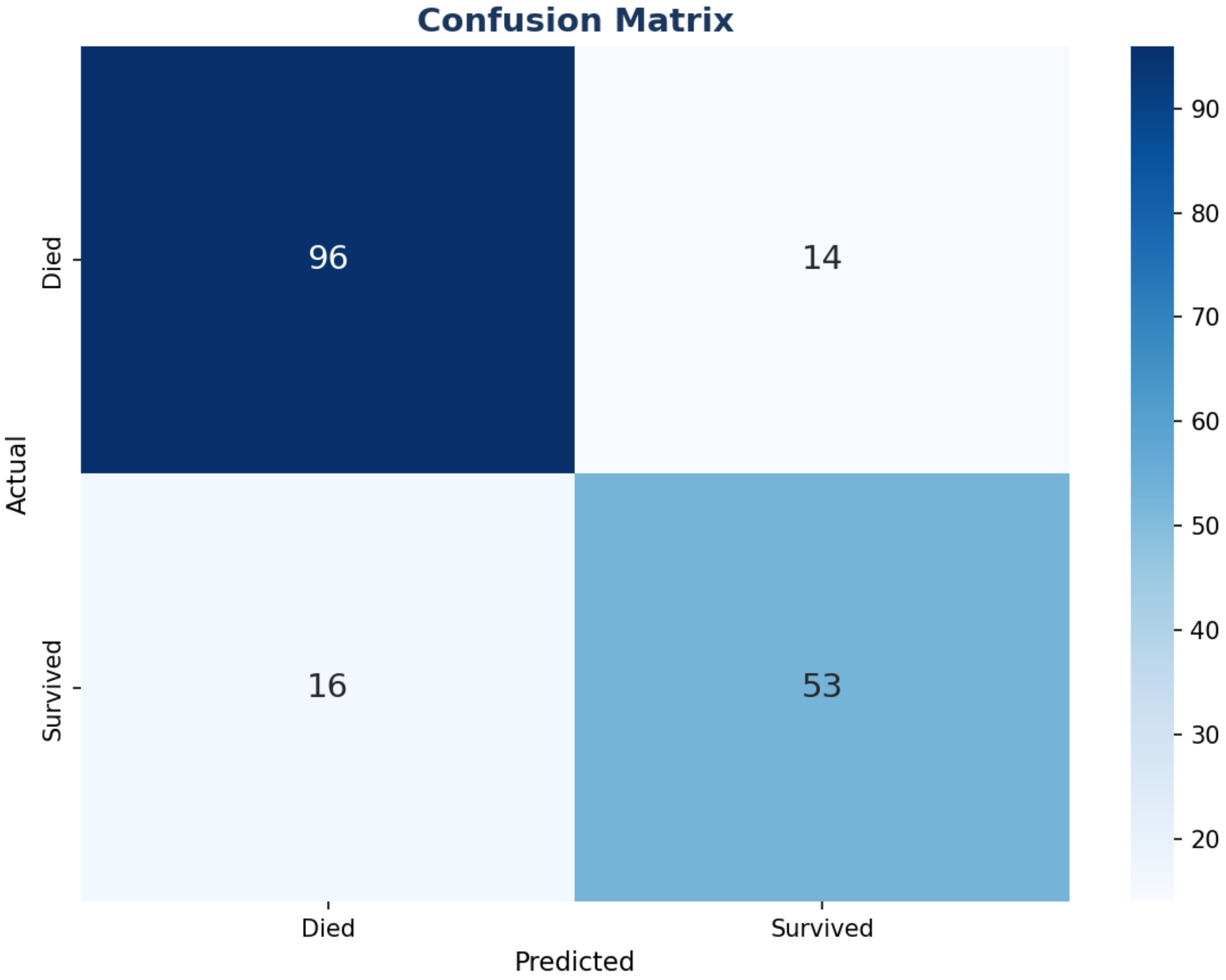

Figure 3: Confusion Matrix

ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve shows the trade-off between true positive rate and false positive rate at various classification thresholds.

Key Metrics

- **AUC-ROC:** 0.8590
- **Optimal Threshold:** 0.4976

ROC Curve

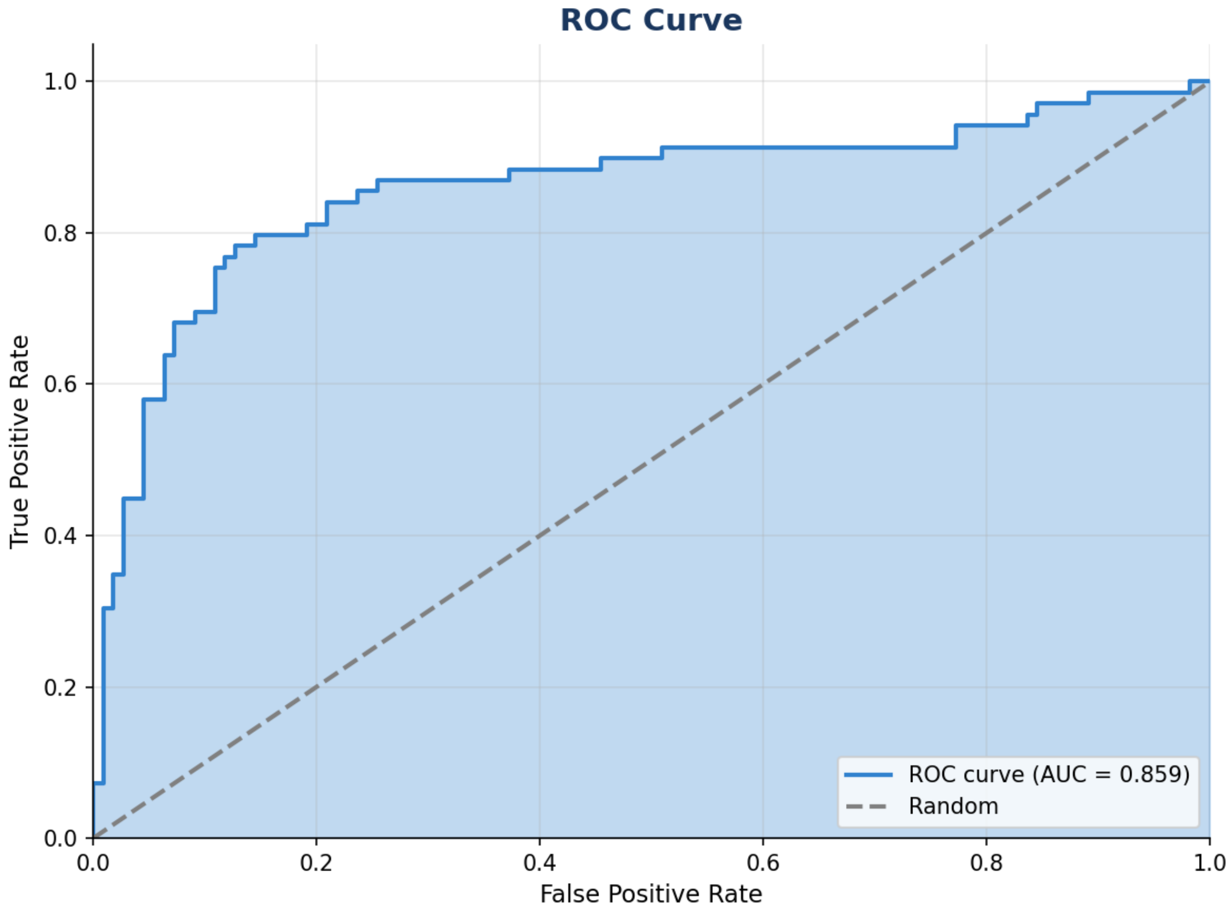

Figure 4: ROC Curve

Precision-Recall Analysis

The Precision-Recall curve is especially useful for imbalanced datasets, showing the trade-off between precision and recall.

Key Metrics

- **Average Precision:** 0.8243

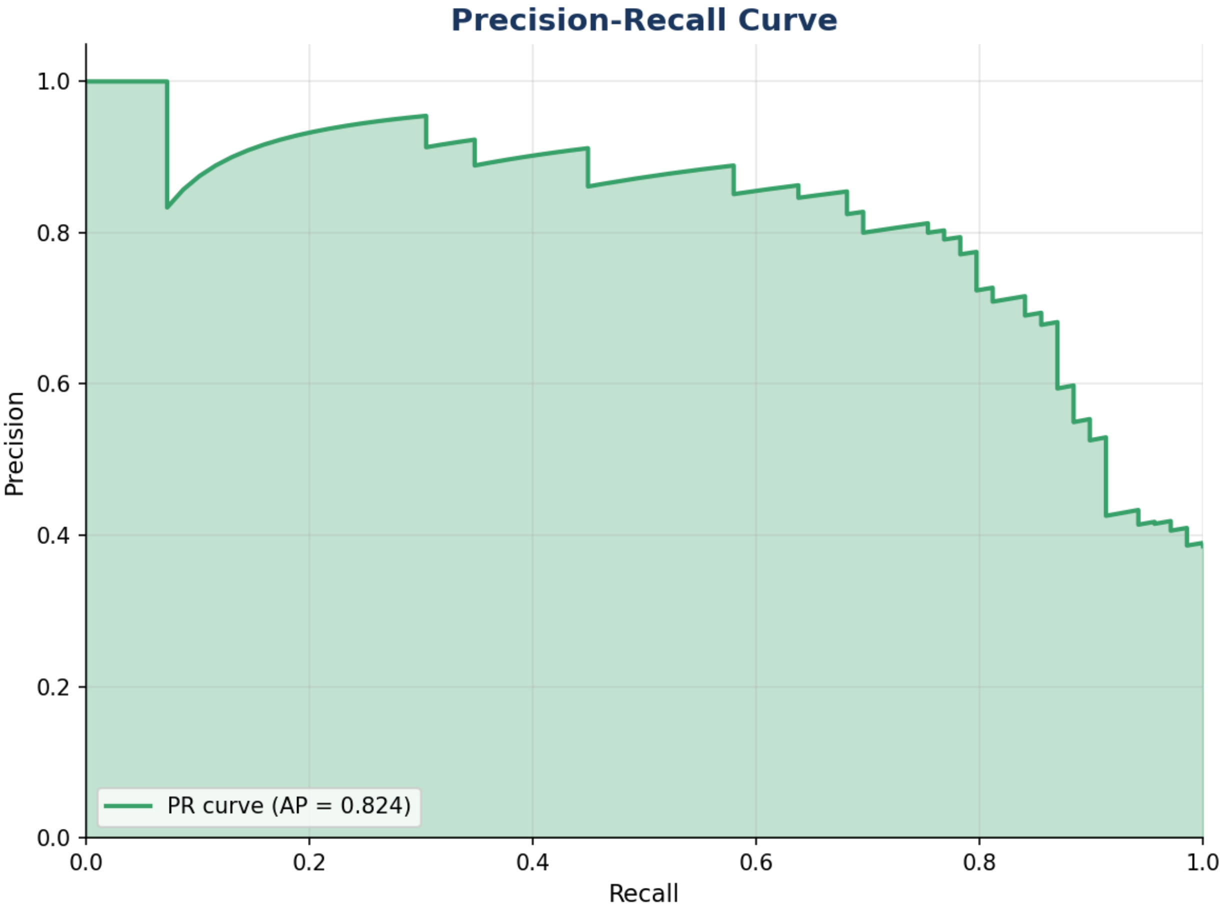

Figure 5: Precision-Recall Curve

Precision-Recall Curve

SHAP Feature Analysis

SHAP (SHapley Additive exPlanations) values provide model-agnostic explanations showing how each feature contributes to individual predictions.

SHAP Feature Importance

Feature	Mean
adult_male	0.0913
sex	0.0882
who	0.0533
fare	0.0440
fare_log	0.0412
age	0.0377
deck	0.0362
pclass	0.0337
class	0.0333
embarked	0.0142
embark_town	0.0138
family_size	0.0130
sibsp	0.0094
is_alone	0.0048
is_child	0.0048

SHAP Summary Plot

Baseline Comparison

Performance Improvement

Model	Score	Improvement
Baseline	0.6100	-
Best Model	**0.8324**	**+0.2224 (+36.5%)**

The final model achieves a **36.5%** improvement over the baseline.

Recommendations & Next Steps

1. Best model Logistic Regression achieves 83.2% survival prediction accuracy

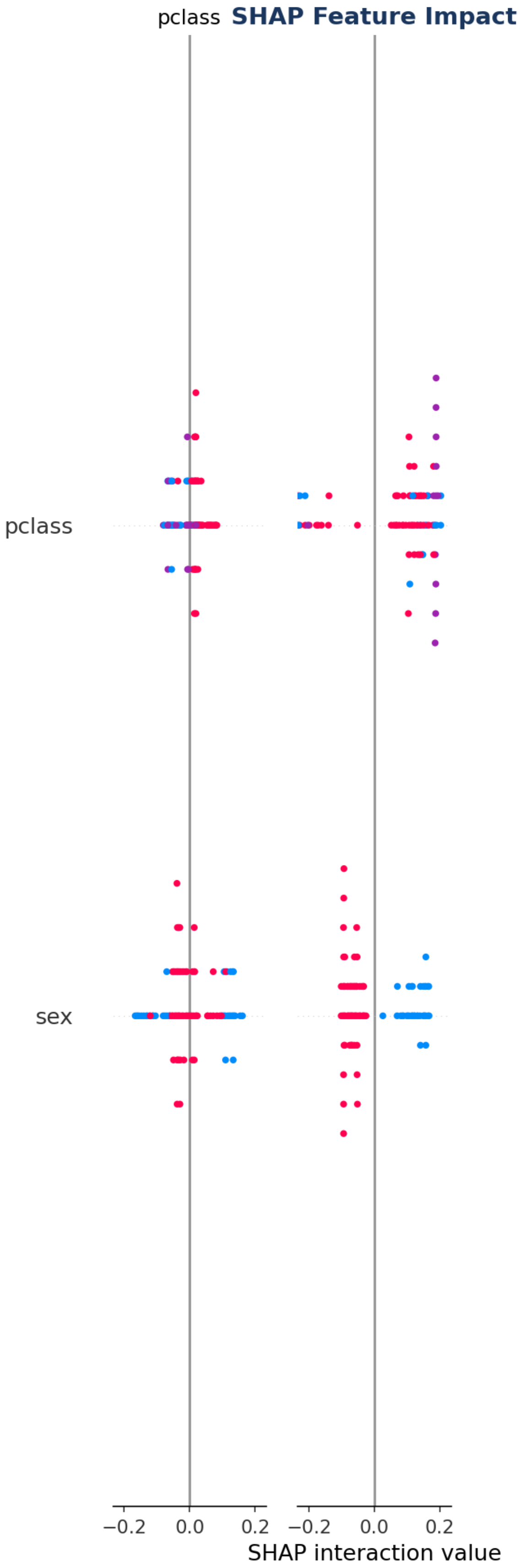

Figure 6: SHAP Analysis

2. AUC-ROC of 0.859 indicates good discrimination ability
3. Key predictive features: sex, fare, class, age
4. Women and children had higher survival rates (‘women and children first’)
5. First class passengers had significantly higher survival rates
6. Consider ensemble methods for production deployment

Report generated by Jotty SwarmMLComprehensive on 2026-02-05 03:01:48