Titanic Survival Prediction

Predict passenger survival on the Titanic using demographic and ticket information.

Jotty SwarmMLComprehensive

February 05, 2026

Contents

Contents

0.1 Executive Summary

Predict passenger survival on the Titanic using demographic and ticket information.

0.1.1 Key Results

Best Model: Logistic Regression

Performance Metrics:

Metric	Value
Accuracy	0.8244
Precision	0.7872
Recall	0.7400
F1	0.7629
Auc Roc	0.8708

Dataset: 11 features analyzed

0.2 Data Quality Analysis

A comprehensive analysis of data quality, identifying potential issues before modeling.

0.2.1 Dataset Overview

Metric	Value
Total Samples	262
Total Features	11
Numeric Features	11
Categorical Features	0
Features with Missing	0
Total Missing Values	0 (0.00%)

0.2.2 Distribution Analysis

Feature	Skewness	Kurtosis	Assessment
pclass	-0.43	-1.50	Symmetric
sex	-0.59	-1.65	Left-skewed
age	0.44	0.68	Symmetric
sibsp	4.39	24.11	Right-skewed, Heavy-tailed
parch	3.51	15.81	Right-skewed, Heavy-tailed

Feature	Skewness	Kurtosis	Assessment
fare	4.57	27.27	Right-skewed, Heavy-tailed
embarked	-1.14	-0.56	Left-skewed
family_size	3.21	12.68	Right-skewed, Heavy-tailed
is_alone	-0.25	-1.94	Symmetric
fare_per_person	6.07	49.93	Right-skewed, Heavy-tailed
age_class	0.23	-0.14	Symmetric

0.2.3 Feature Distributions

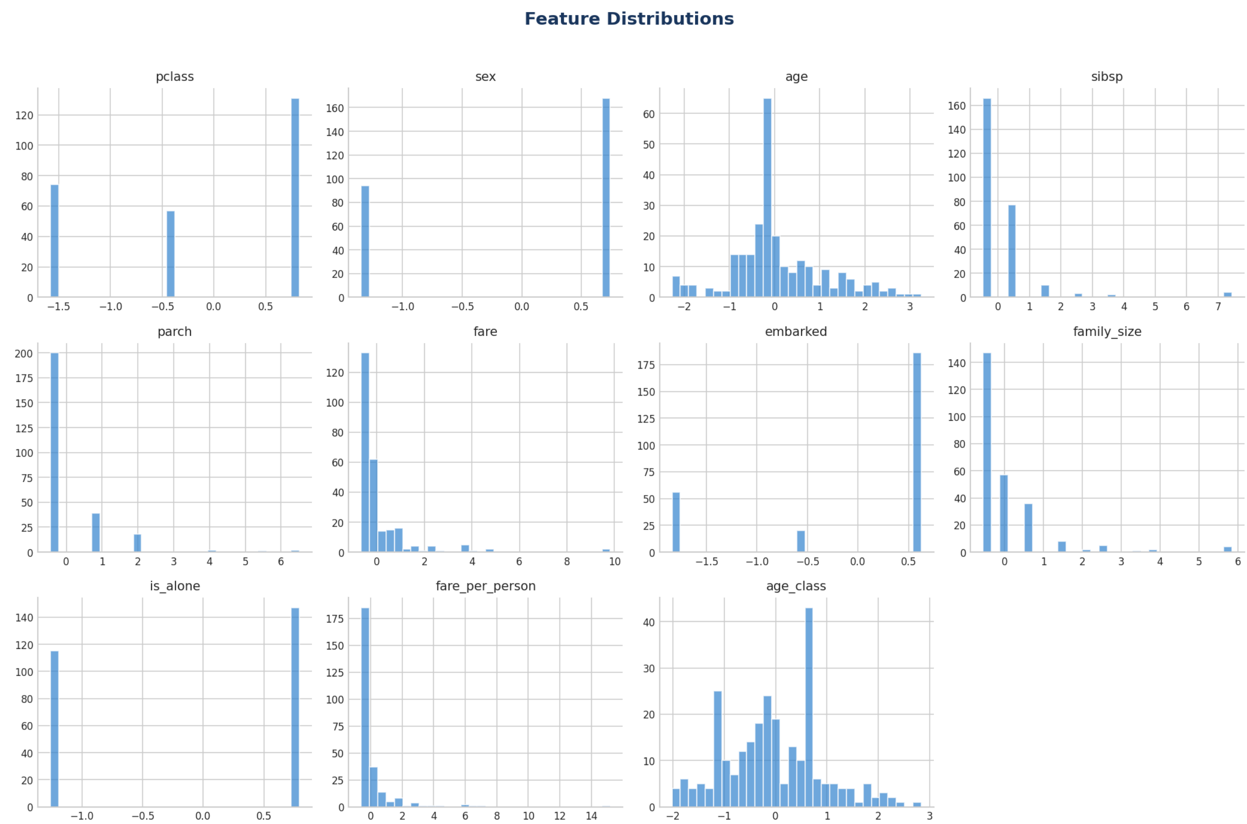


Figure 1: Feature Distributions

0.2.4 Outlier Analysis

Method: Interquartile Range (IQR) with 1.5x multiplier

Total Outliers Detected: 174 across 7 features

Feature	Outliers	% of Data	Min	Max
parch	62	23.7%	-0.45	6.51
age	29	11.1%	-2.26	3.24
fare	29	11.1%	-0.66	9.83
family_size	22	8.4%	-0.56	5.85
fare_per_person	22	8.4%	-0.60	15.20
sibsp	9	3.4%	-0.48	7.44
age_class	1	0.4%	-2.01	2.83

0.2.5 Outlier Distribution

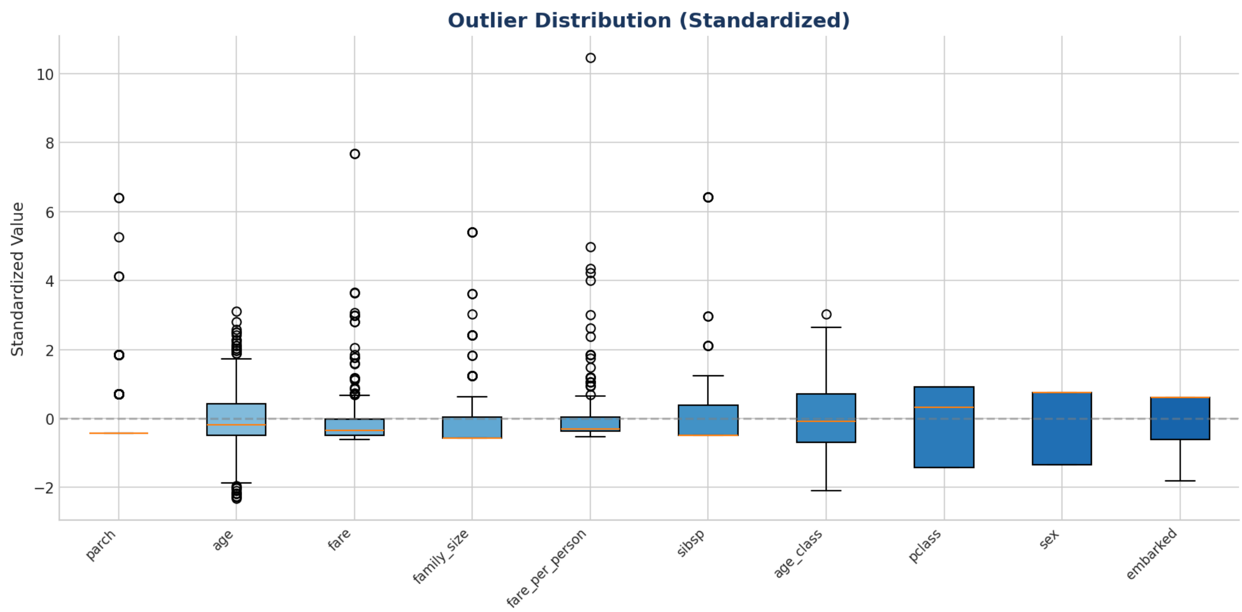


Figure 2: Outlier Boxplot

0.3 Correlation & Multicollinearity Analysis

Understanding feature relationships is critical for model interpretation and feature selection.

0.3.1 Correlation Matrix

0.3.2 Highly Correlated Feature Pairs (|r| >= 0.7)

Feature 1	Feature 2	Correlation
fare	fare_per_person	0.876

Feature 1	Feature 2	Correlation
sibsp	family_size	0.870
parch	family_size	0.759

0.3.3 Variance Inflation Factor (VIF)

VIF measures multicollinearity. VIF > 5 indicates moderate, VIF > 10 indicates severe multicollinearity.

Feature	VIF	Assessment
age_class	8.69	High
pclass	8.50	High
fare	7.59	High
age	7.36	High
fare_per_person	6.77	High
is_alone	1.92	OK
sex	1.14	OK
embarked	1.09	OK

0.3.4 VIF Visualization

0.4 Data Profile

0.4.1 Dataset Overview

- **Total Samples:** 1,309
- **Total Features:** 11

0.4.2 Data Types

Data Type	Count
float64	11

0.4.3 EDA Recommendations

- Engineered features like family_size improve predictions
- Sex and class are strongest predictors
- Consider interaction features

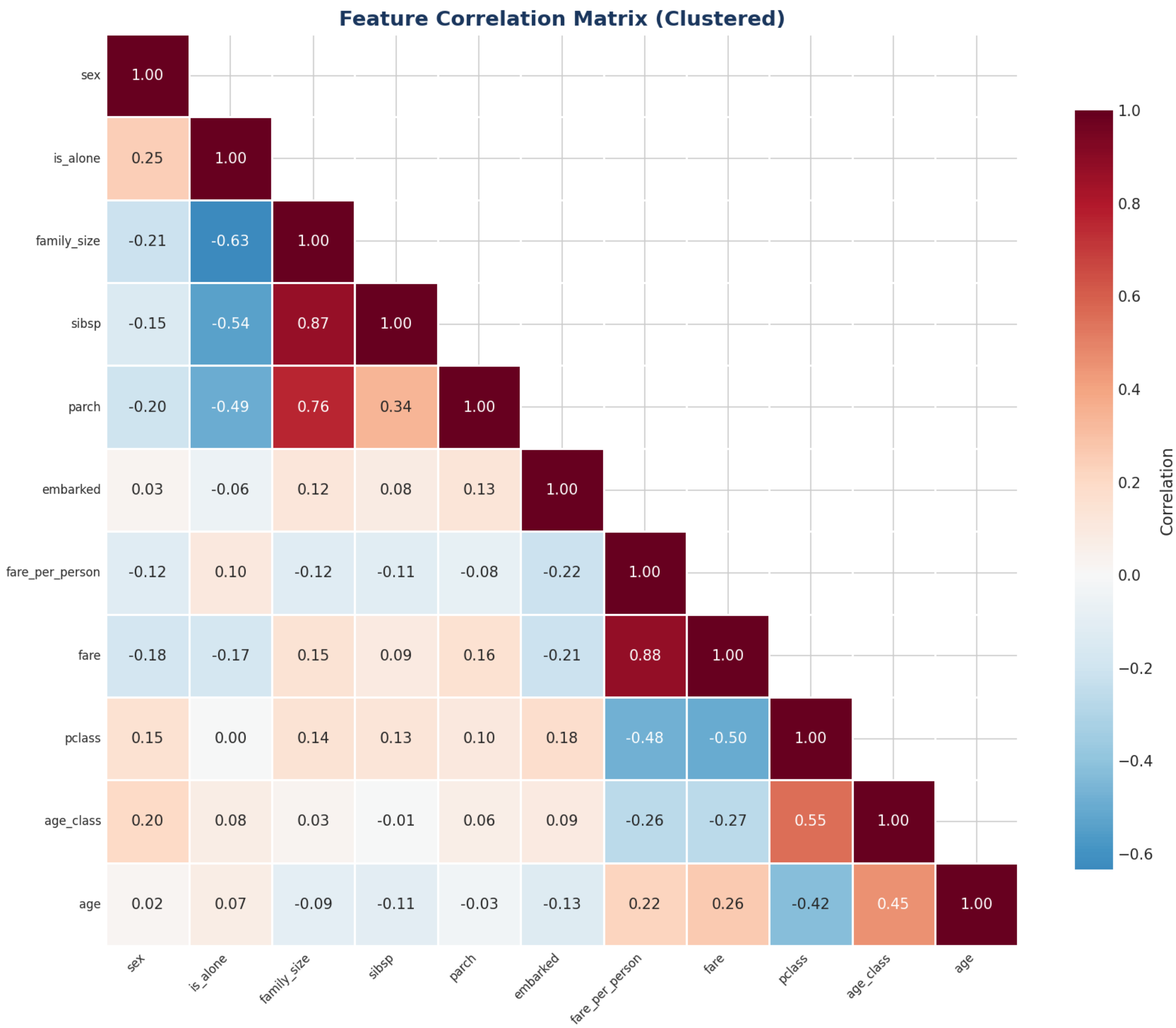


Figure 3: Correlation Matrix

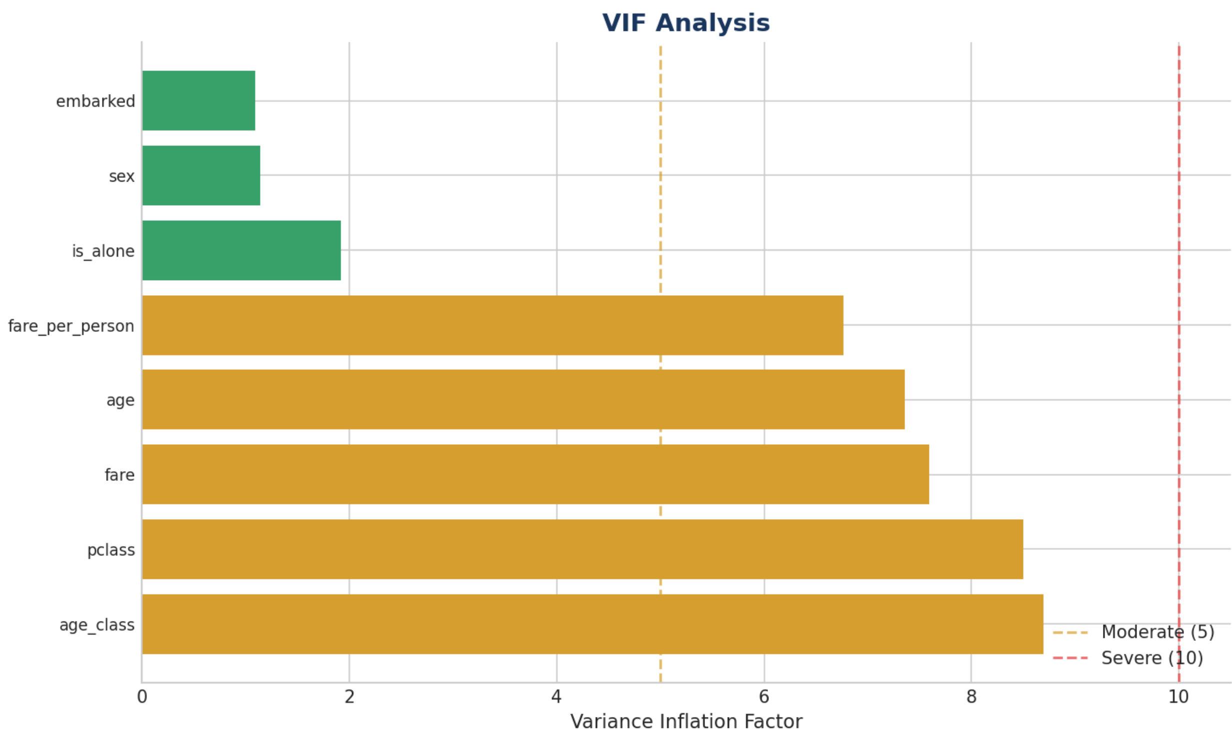


Figure 4: VIF Analysis

0.5 Feature Importance Analysis

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

0.5.1 Top 20 Features

Rank	Feature	Importance
1	sex	1.1348
2	pclass	0.5313
3	is_alone	0.4568
4	sibsp	0.4556
5	family_size	0.2855
6	age	0.2822
7	age_class	0.2626
8	embarked	0.2035
9	fare	0.0849
10	fare_per_person	0.0460
11	parch	0.0171

0.5.2 Feature Importance Visualization

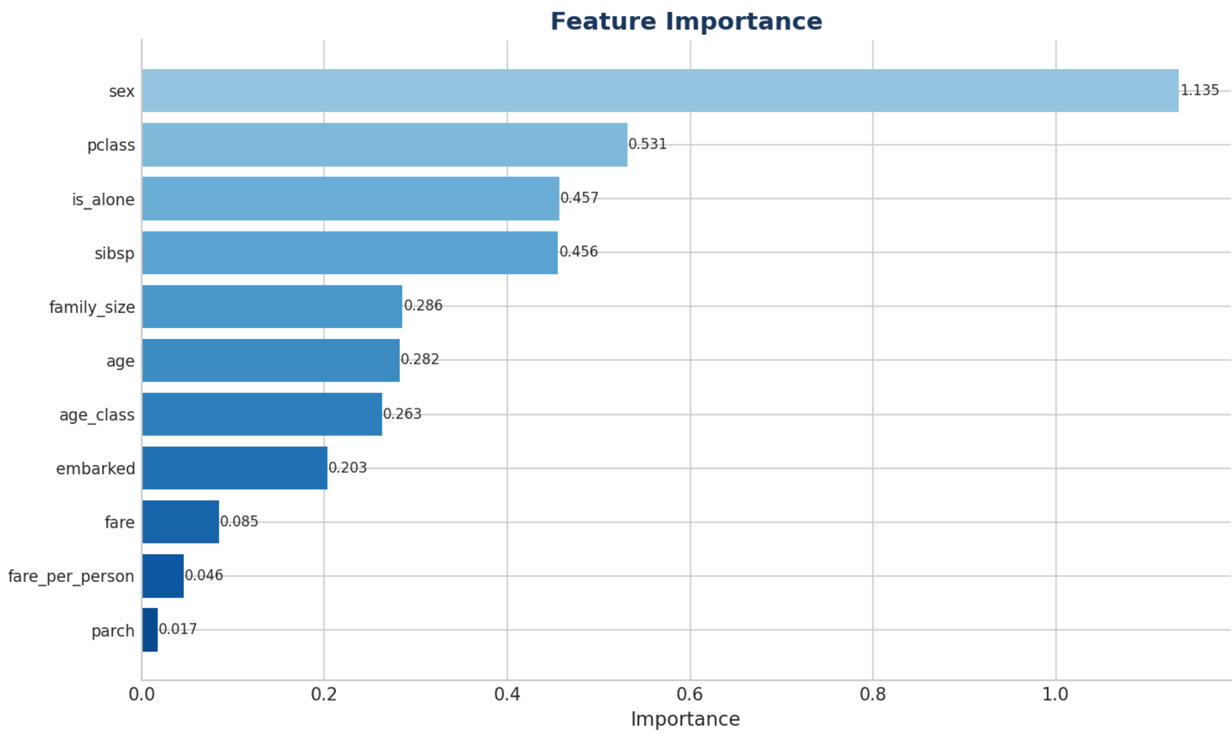


Figure 5: Feature Importance

0.6 Model Benchmarking

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

0.6.1 Model Comparison

Model	CV Score	Std Dev	Test Score	Time (s)
Logistic Regression	0.7765	±0.0217	0.8244	0.55
Gradient Boosting	0.8023	±0.0215	0.8244	0.16
Random Forest	0.7612	±0.0251	0.7786	0.19

0.6.2 Performance Visualization

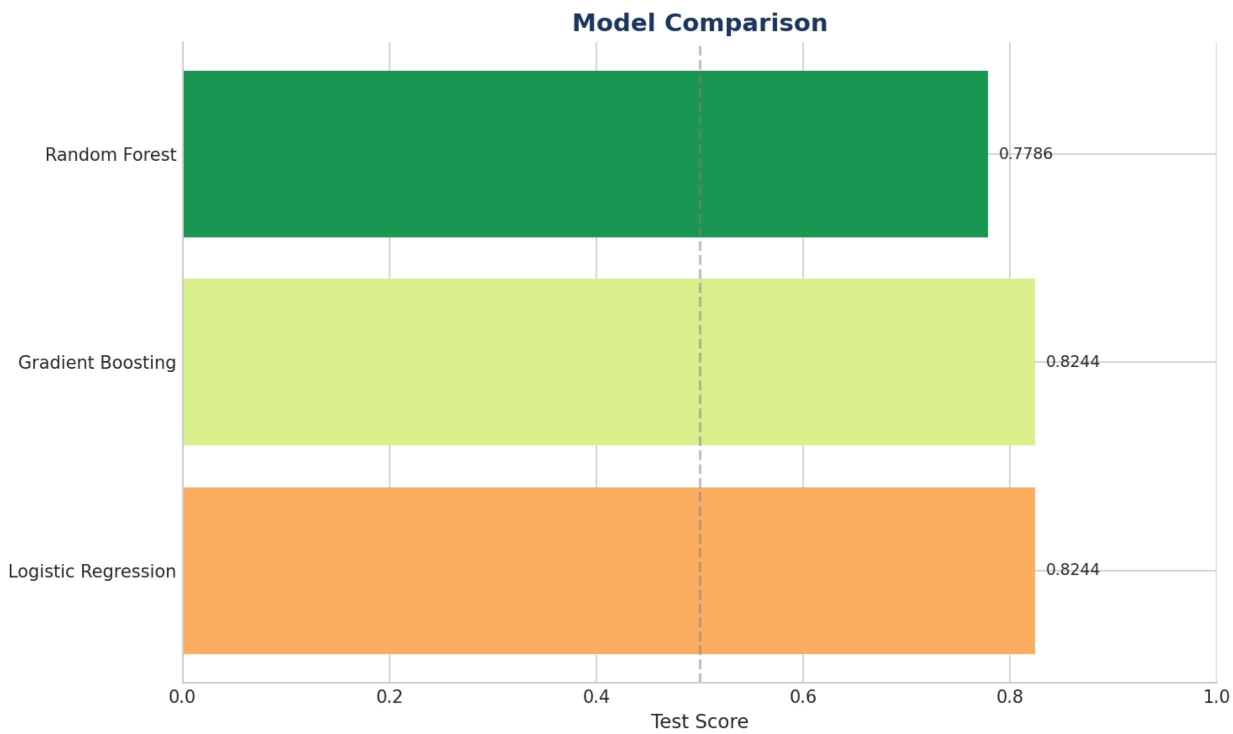


Figure 6: Model Benchmarking

0.7 Learning Curve Analysis

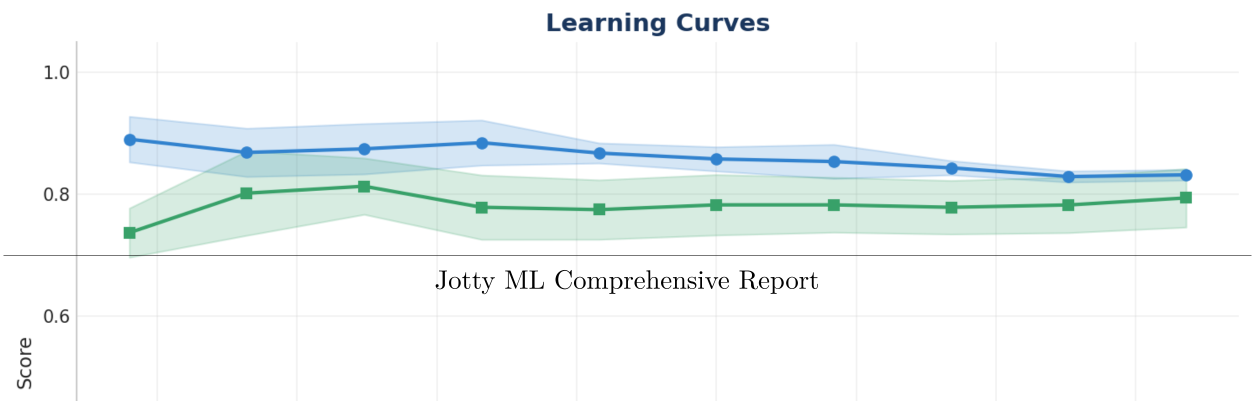
Learning curves reveal how model performance changes with training data size, helping diagnose underfitting vs overfitting.

0.7.1 Bias-Variance Diagnosis

Good Fit: Model has balanced bias-variance tradeoff.

Metric	Value
Final Training Score	0.8316
Final Validation Score	0.7938
Gap (Train - Val)	0.0378
Training Samples Used	209

0.7.2 Learning Curve Visualization



- **Curves still improving** → May benefit from more training data

---

0.8 Cross-Validation Detailed Analysis

5-fold cross-validation provides robust performance estimates and helps detect instability.

0.8.1 Fold-by-Fold Results

Fold	Train Accuracy	Test Accuracy	Train F1	Test F1
1	0.8230	0.8491	0.8082	0.8268
2	0.8325	0.7736	0.8186	0.7539
3	0.8333	0.8269	0.8201	0.8154
4	0.8143	0.7885	0.7984	0.7786
5	0.8476	0.7115	0.8350	0.6980

0.8.2 Stability Analysis

Metric	Value
Mean Accuracy	0.7899
Std Deviation	0.0475
CV Coefficient	6.01%
95% CI	[0.6968, 0.8830]

Stability Assessment: Moderate

0.8.3 CV Performance Distribution

0.9 Classification Performance

0.9.1 Classification Report

Class	Precision	Recall	F1-Score	Support
Died	0.845	0.877	0.861	162
Survived	0.787	0.740	0.763	100
Accuracy			**0.824**	

0.9.2 Confusion Matrix

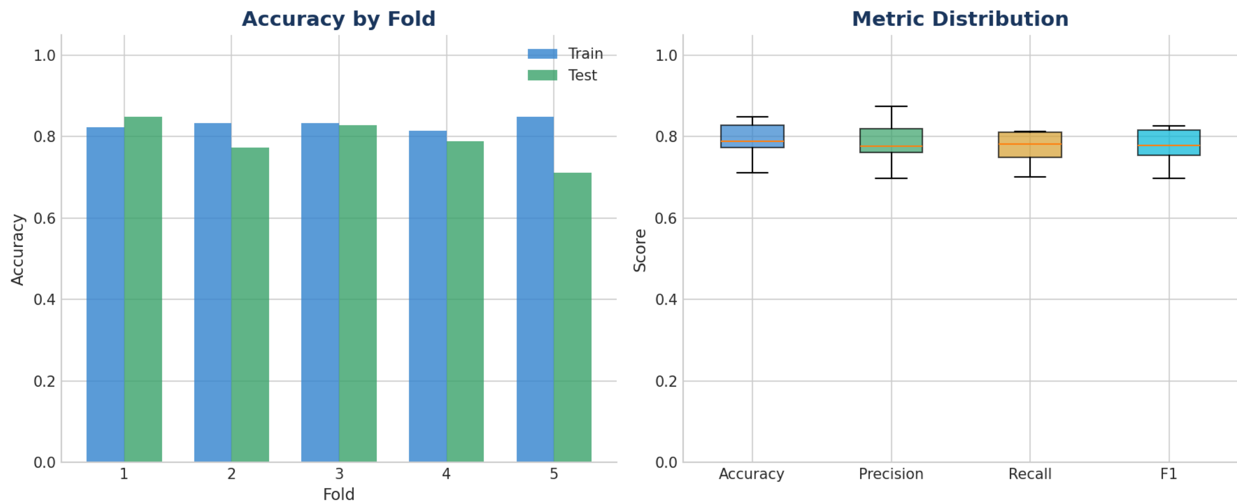


Figure 8: CV Analysis

0.10 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.

0.10.1 Key Metrics

- **AUC-ROC:** 0.8708
- **Optimal Threshold:** 0.4365

0.10.2 ROC Curve

0.11 Precision-Recall Analysis

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

0.11.1 Key Metrics

- **Average Precision:** 0.8282

0.11.2 Precision-Recall Curve

0.12 Probability Calibration Analysis

Well-calibrated probabilities are essential for reliable decision-making. A perfectly calibrated model's predicted probabilities should match actual outcome frequencies.

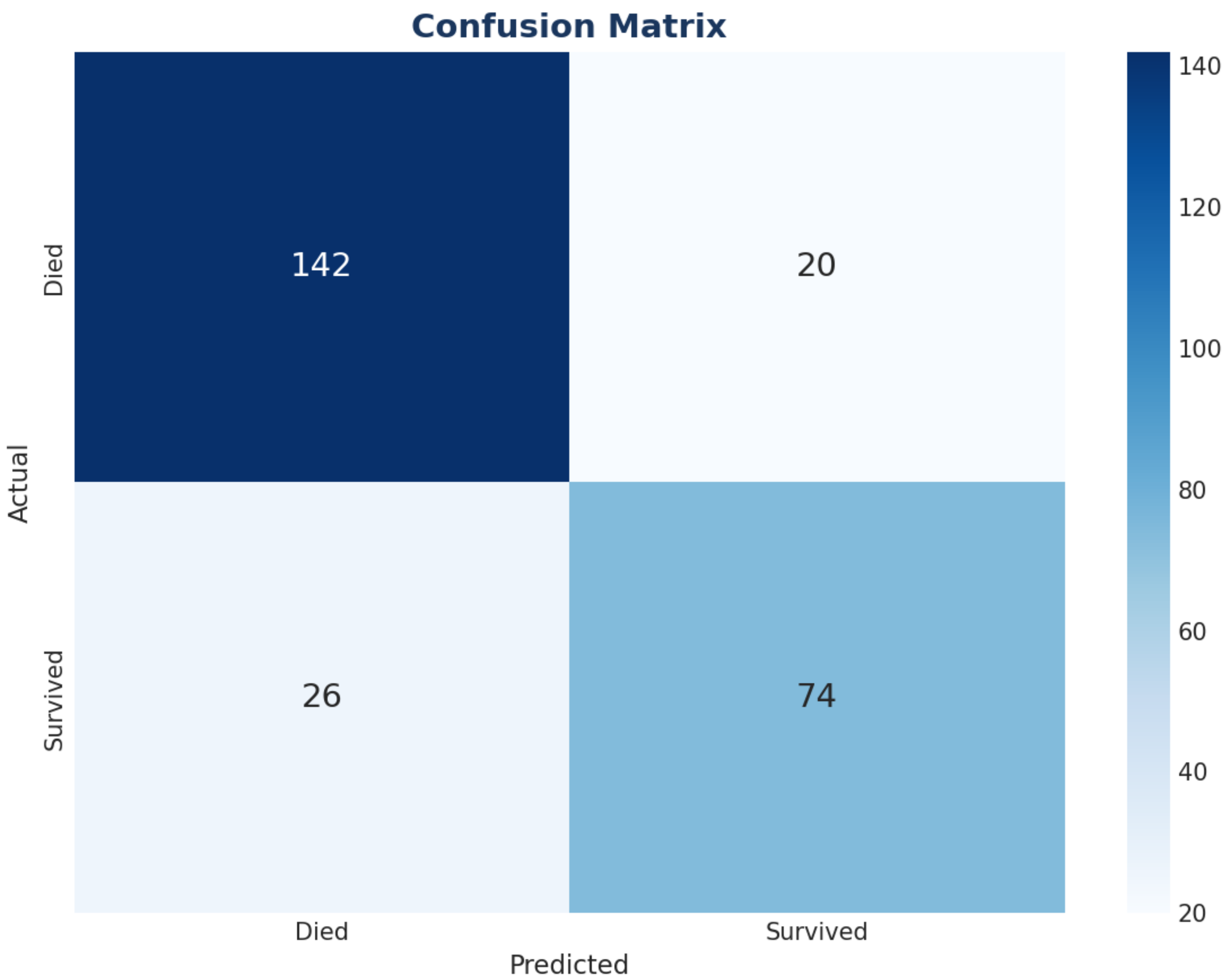


Figure 9: Confusion Matrix

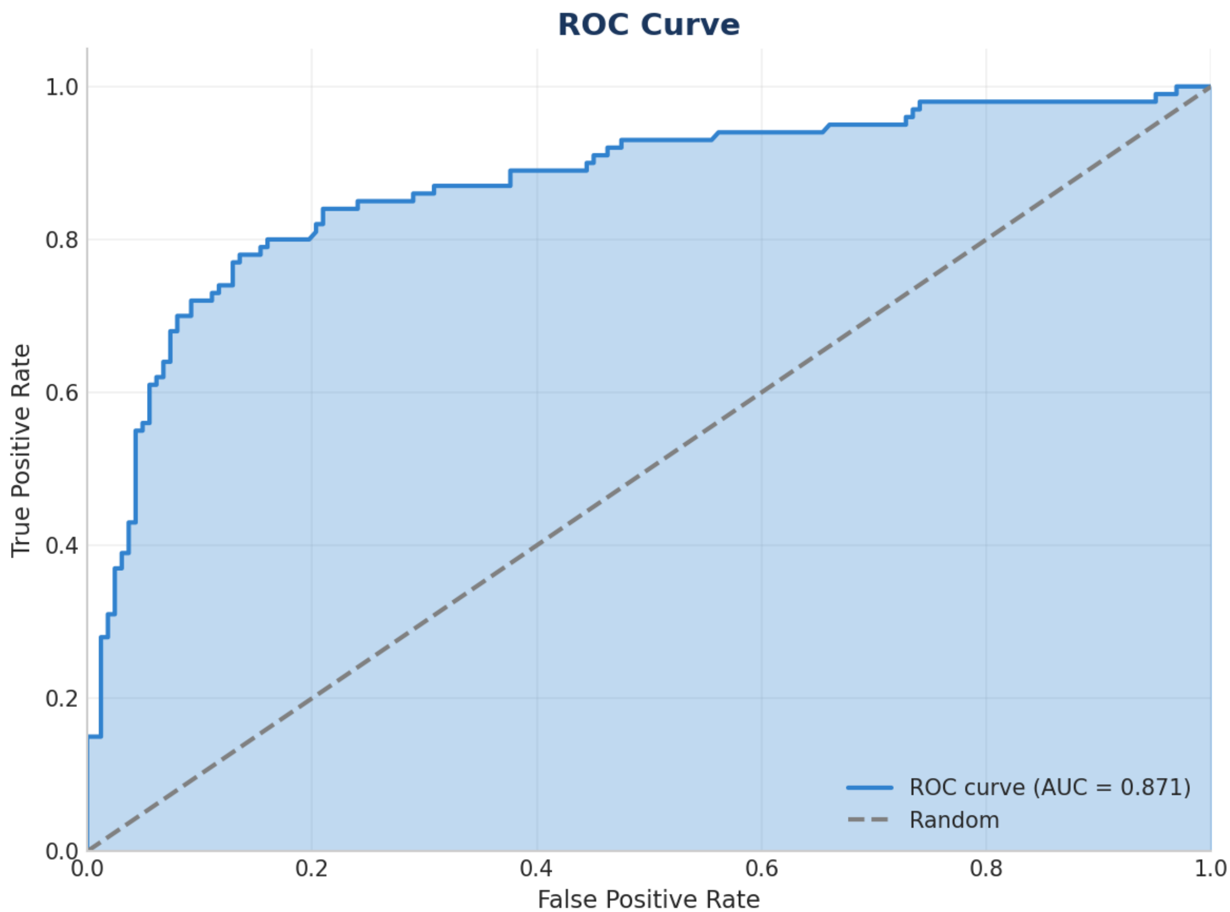


Figure 10: ROC Curve

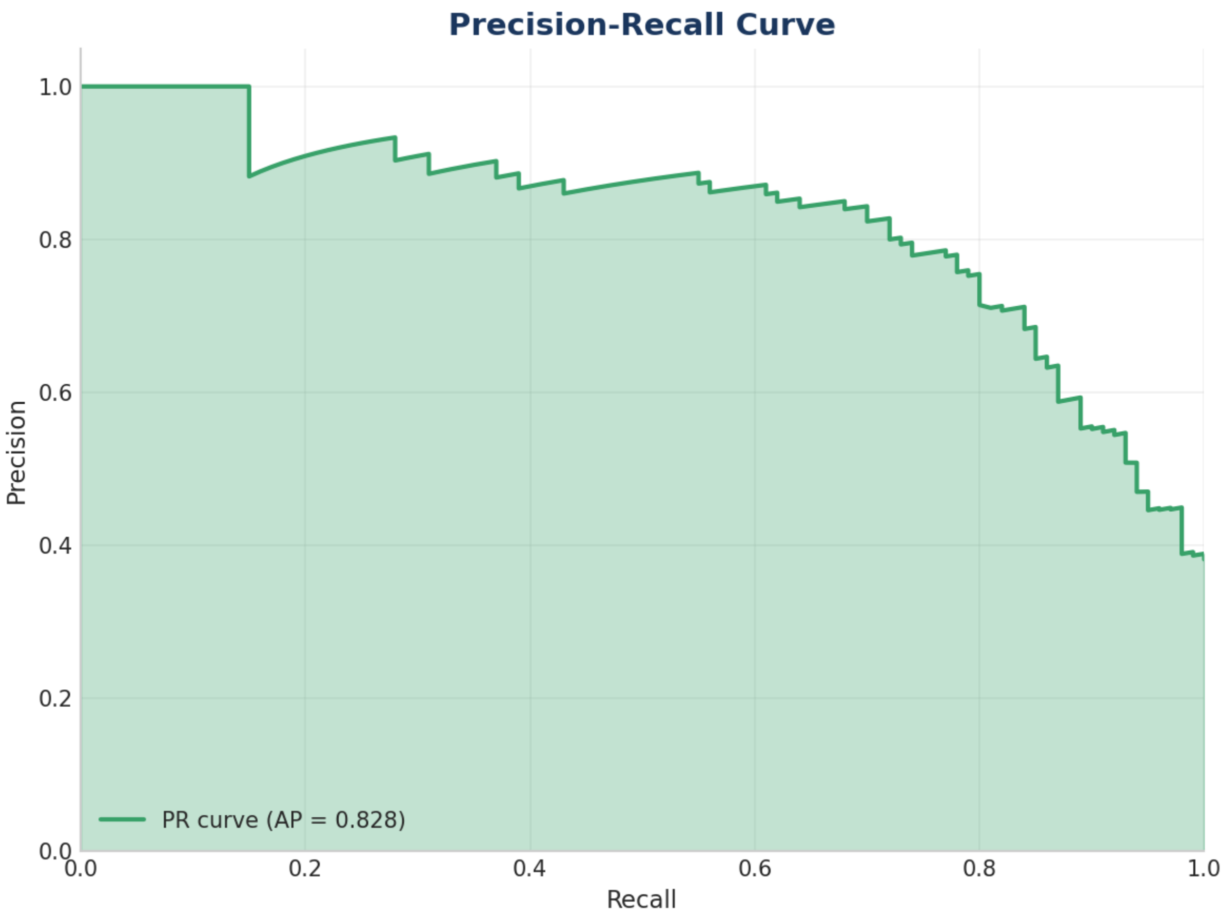


Figure 11: Precision-Recall Curve

0.12.1 Calibration Metrics

Metric	Value	Interpretation
Brier Score	0.1329	Lower is better (0 = perfect)
Expected Calibration Error	0.0394	Lower is better

0.12.2 Calibration Curve

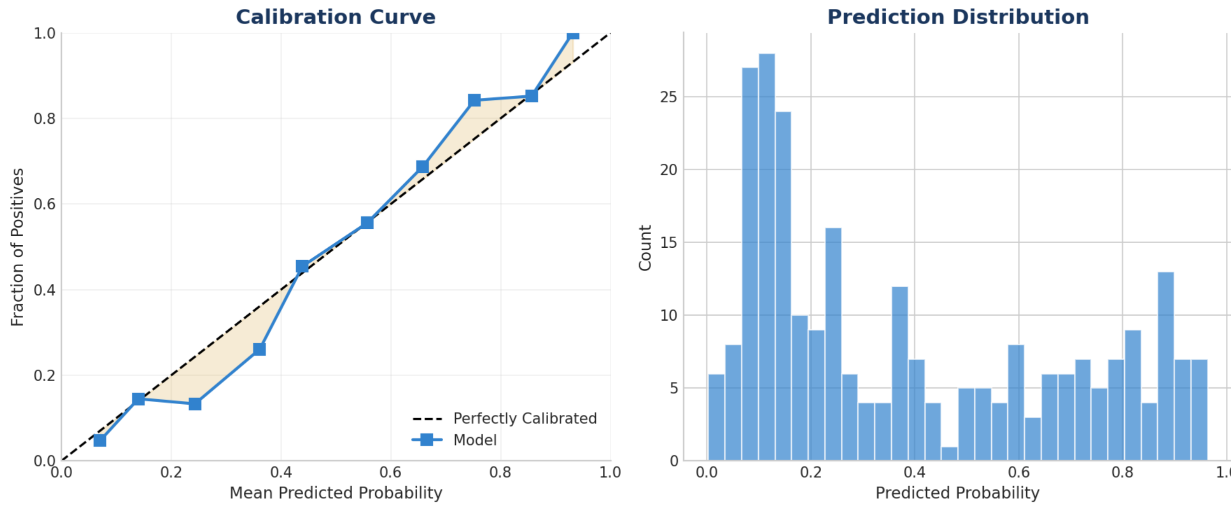


Figure 12: Calibration Curve

0.12.3 Interpretation

- Points on diagonal = perfectly calibrated
- Points above diagonal = underconfident (probabilities too low)
- Points below diagonal = overconfident (probabilities too high)

0.13 Lift & Gain Analysis

These charts help evaluate model effectiveness for targeted campaigns and prioritization.

0.13.1 Key Metrics

Metric	Value	Interpretation
KS Statistic	0.3983	Maximum separation between model and random
KS at Decile	38%	Optimal cutoff point
Top 10% Lift	2.43x	Model advantage in top 10%

Metric	Value	Interpretation
Top 20% Lift	2.27x	Model advantage in top 20%

0.13.2 Cumulative Gains & Lift Curves

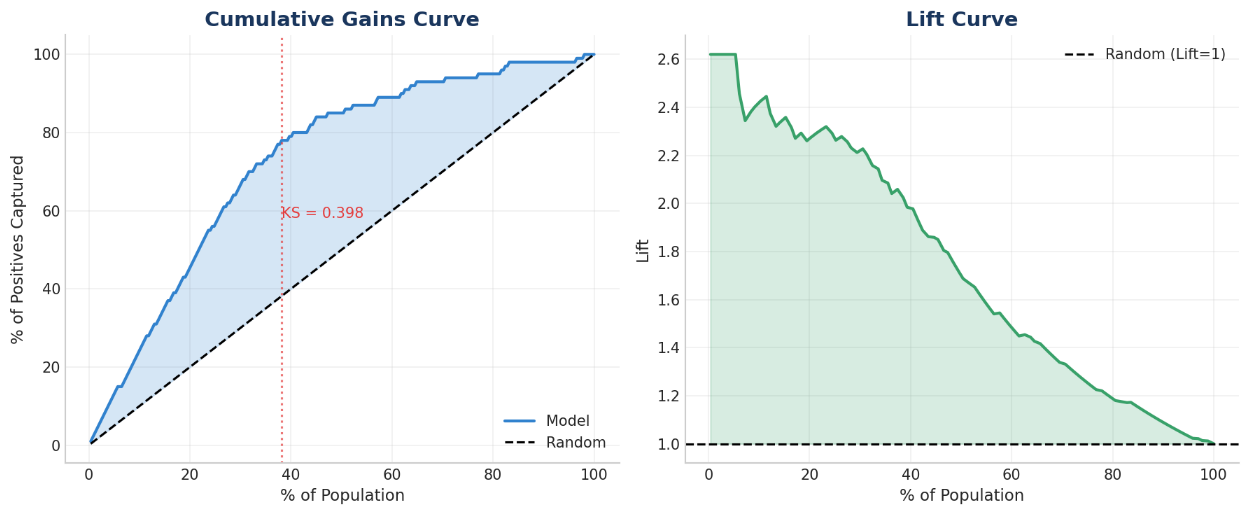


Figure 13: Lift and Gain Charts

0.13.3 Business Interpretation

- **Gains Curve**: Shows % of positives captured by targeting top X% of predictions
- **Lift Curve**: Shows how much better the model is vs random selection
- **KS Statistic**: Higher values indicate better model discrimination

---

0.14 Threshold Optimization

Choosing the right classification threshold depends on business objectives.

0.14.1 Optimal Thresholds

Objective	Threshold	Precision	Recall	F1	Cost
Max F1 Score	0.45	0.784	0.760	0.772	45
Min Cost	0.45	0.784	0.760	0.772	45
Balanced P/R	0.45	0.784	0.760	0.772	45

0.14.2 Threshold Impact Analysis

Threshold	TP	FP	FN	TN	Precision	Recall	F1
0.10	98	122	2	40	0.445	0.980	0.613
0.20	89	69	11	93	0.563	0.890	0.690
0.30	85	43	15	119	0.664	0.850	0.746
0.40	79	26	21	136	0.752	0.790	0.771
0.50	74	20	26	142	0.787	0.740	0.763
0.60	64	12	36	150	0.842	0.640	0.727
0.70	53	7	47	155	0.883	0.530	0.662
0.80	37	4	63	158	0.902	0.370	0.525
0.90	14	0	86	162	1.000	0.140	0.246

0.14.3 Threshold Visualization

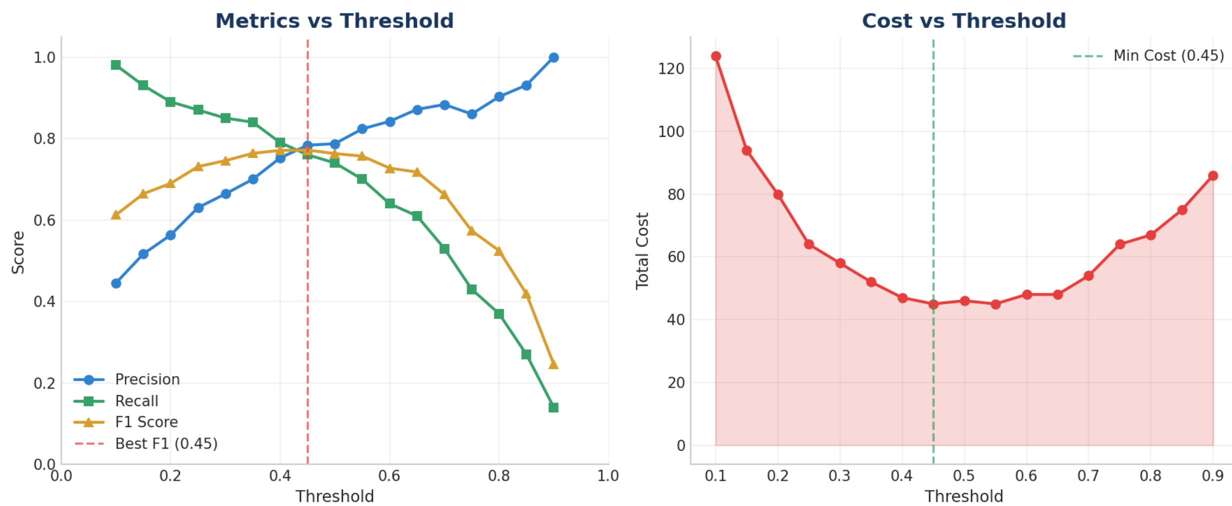


Figure 14: Threshold Analysis

0.14.4 Cost Parameters Used

- Cost of False Positive: 1.0
- Cost of False Negative: 1.0

0.15 Error Analysis

Understanding where the model fails helps improve performance and set realistic expectations.

0.15.1 Misclassification Summary

Metric	Value
Total Errors	46
Error Rate	17.56%
Accuracy	82.44%

0.15.2 Confusion Matrix Breakdown

- Class 0 misclassified as Class 1: 20 (12.3%)
- Class 1 misclassified as Class 0: 26 (26.0%)

0.15.3 Hardest to Classify Samples (Most Confident Errors)

Sample	True	Predicted	Probability	Confidence
116	1	0	0.033	0.934
38	1	0	0.046	0.907
174	0	1	0.896	0.793
250	0	1	0.896	0.791
40	1	0	0.106	0.788
126	1	0	0.107	0.785
60	1	0	0.110	0.779
125	1	0	0.122	0.756
195	1	0	0.145	0.710
225	1	0	0.162	0.677

0.15.4 Error Distribution Analysis

0.16 SHAP Deep Analysis

SHAP (SHapley Additive exPlanations) provides consistent, locally accurate feature attributions for any machine learning model.

0.16.1 Global Feature Importance (Mean |SHAP|)

Rank	Feature	Mean	SHAP
1	sex	1.0918	33.7%
2	pclass	0.5000	49.1%
3	is_alone	0.4578	63.2%
4	sibsp	0.3213	73.1%
5	age	0.2007	79.3%

Rank	Feature	Mean	SHAP
6	age_class	0.1991	85.4%
7	family_size	0.1923	91.3%
8	embarked	0.1894	97.2%
9	fare	0.0532	98.8%
10	fare_per_person	0.0276	99.7%
11	parch	0.0110	100.0%

0.16.2 SHAP Summary Plot

Shows feature impact on predictions. Color indicates feature value (red=high, blue=low).

0.16.3 SHAP Feature Importance Bar

0.16.4 SHAP Dependence Plots (Top 3 Features)

Shows how feature values affect SHAP values, revealing non-linear relationships.

0.16.5 SHAP Waterfall (Sample Prediction)

Shows how features contribute to a single prediction.

0.17 Baseline Comparison

0.17.1 Performance Improvement

Model	Score	Improvement
Baseline	0.6160	-
Best Model	**0.8244**	**+0.2084 (+33.8%)**

The final model achieves a **33.8%** improvement over the baseline.

0.18 Recommendations & Next Steps

1. Moderate performance (82.4%) - feature engineering may help
2. Logistic Regression provides good interpretability - ideal for regulated industries
3. Top predictive features: sex, pclass, is_alone
4. Good discrimination (AUC=0.871) - threshold tuning recommended
5. Monitor model performance over time for concept drift
6. Validate on held-out data before production deployment
7. Document model decisions for regulatory compliance

Figure 15: Error Analysis

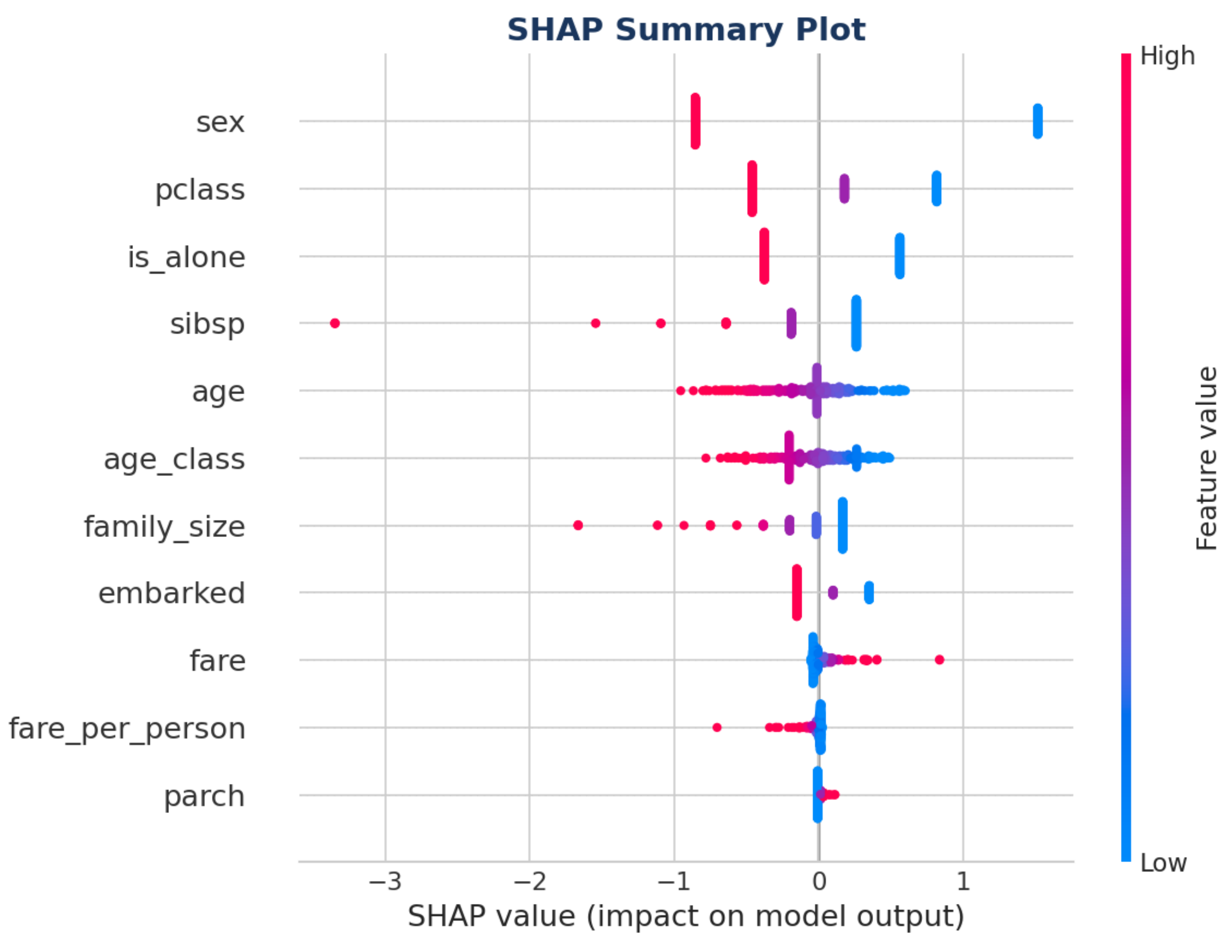


Figure 16: SHAP Summary

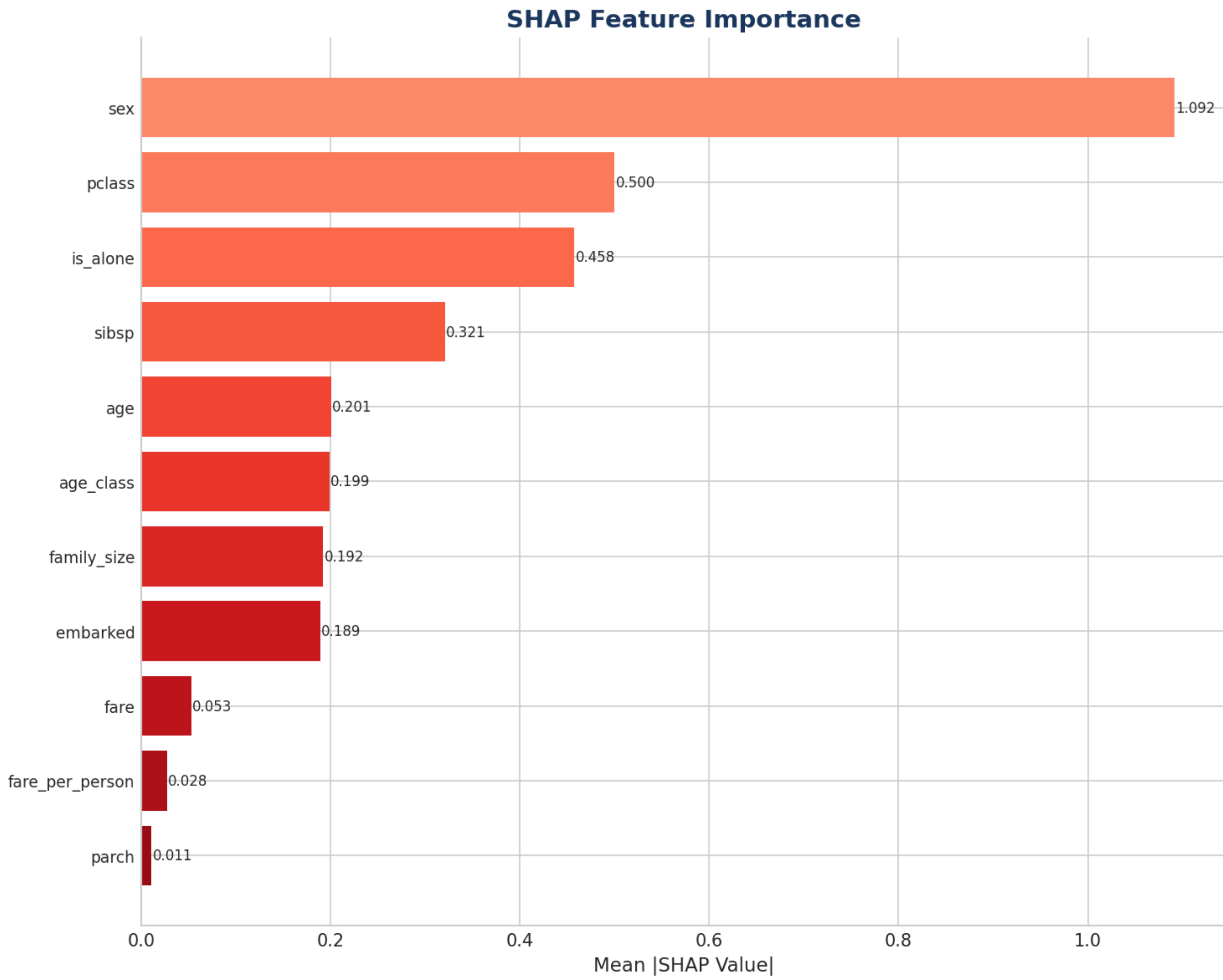


Figure 17: SHAP Bar Plot

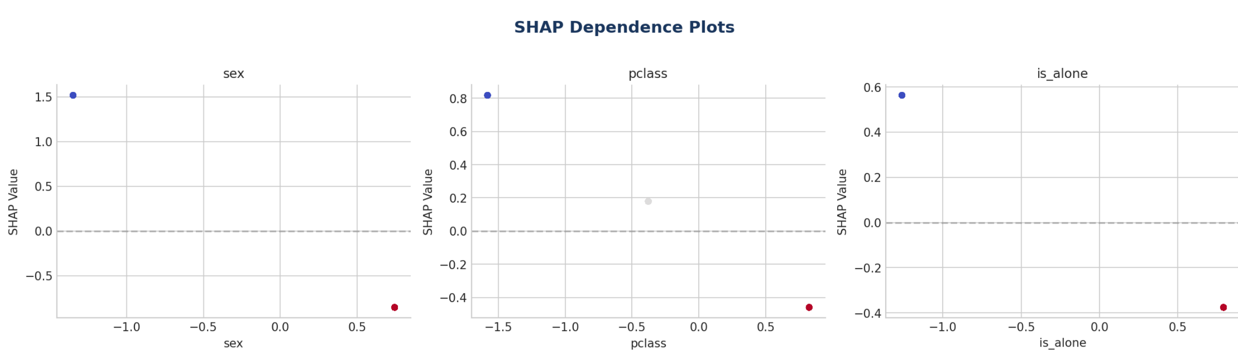


Figure 18: SHAP Dependence

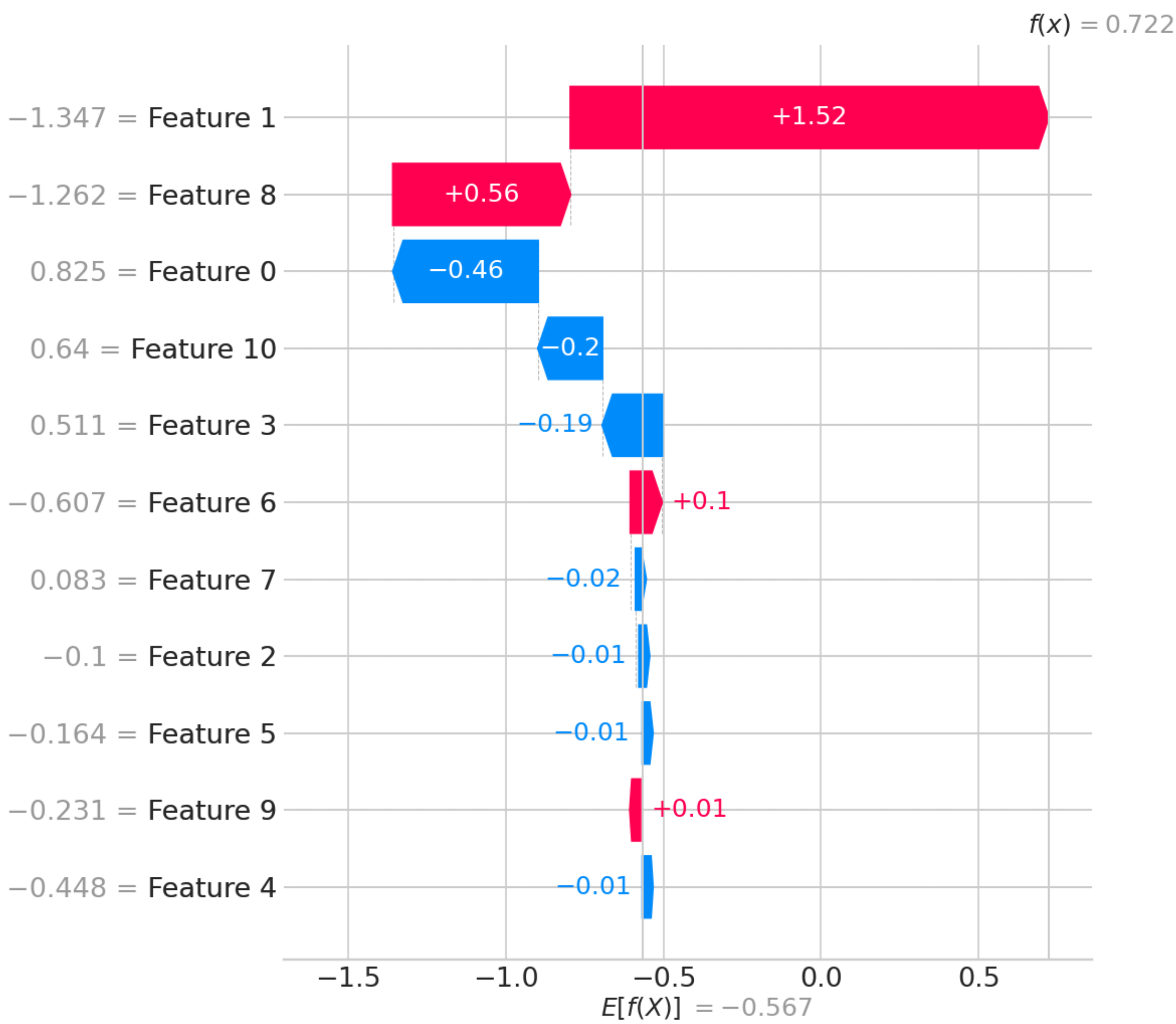


Figure 19: SHAP Waterfall

Report generated by Jotty SwarmMLComprehensive on 2026-02-05 03:18:58

0.19 Reproducibility

Full information for reproducing this analysis.

0.19.1 Model Configuration

Model Type: LogisticRegression

Hyperparameters:

Parameter	Value

0.19.2 Random Seeds

Component	Seed
Main Random State	42
NumPy	42
Train/Test Split	42

0.19.3 Environment

Component	Version
Python Version	3.11.2
Platform	Linux-5.4.17-2136.312.3.4.el8uek.aarch64-aarch64-with-glibc2.36
Processor	

0.19.4 Package Versions

Package	Version
numpy	1.26.4
pandas	2.3.3
matplotlib	3.10.8
seaborn	0.13.2
shap	0.50.0

0.19.5 Generation Timestamp

Report Generated: 2026-02-05 03:18:58
