**Logistic Regression Analysis: Titanic Survival Prediction**

**Structured Statistical Analysis Report**

**By: Vasukinatha Adiga, 242807003**

**1. Data Description**

# Dataset Overview

**Total observations**: 891 passengers

**Variables**: 11 (including outcome variable 'Survived')

**Study design**: Cross-sectional analysis

**Outcome**: Binary survival status 0=died, 1=survived) **Overall survival rate**: 38.4% 342/891

# Variable Types

|  |  |
| --- | --- |
| Variable Type | Variables |
| Categorical | Sex, Pclass, Embarked, Cabin |
| Continuous | Age, Fare |
| Count | SibSp, Parch |
| Identifier | PassengerId, Ticket |

# Target Variable Distribution

**Died 0** : 549 passengers 61.6%

**Survived 1** : 342 passengers 38.4%

# 2. Preprocessing

## 2.1 Missing Values Analysis

|  |  |  |
| --- | --- | --- |
| Variable | Missing Count | Missing Percentage |
| Cabin | 687 | 77.1% |
| Age | 177 | 19.9% |
| Embarked | 2 | 0.2% |
| Fare | 0 | 0.0% |

**Treatment Applied:**

**Cabin**: High missingness 77.1% - converted to binary indicator Has\_Cabin)

**Age**: Imputed with median 28.0 years)

**Embarked**: Imputed with mode Southampton 'S')

**Justification**: Median imputation for Age preserves central tendency without introducing bias. Mode imputation for Embarked handles minimal missingness effectively. Cabin missingness likely reflects systematic differences (lower class passengers), so converted to binary availability indicator.

## 2.2 Duplicates Analysis

**Total duplicate rows**: 0

**Duplicate PassengerIDs**: 0

## 2.3 Outliers Analysis IQR Method)

|  |  |  |
| --- | --- | --- |
| Variable | Total Outliers | Outlier Percentage |
| Age | 11 | 1.5% |
| SibSp | 46 | 5.2% |
| Parch | 213 | 23.9% |
| Fare | 116 | 13.0% |

**Justification**: Outliers retained for main analysis as they may represent legitimate extreme cases relevant to survival patterns. Sensitivity analysis conducted with outlier removal.

## 2.4 Variable Creation

Created analysis-ready binary/categorical variables:

Female (reference: Male)

Class\_1, Class\_2 (reference: Class\_3 Age groups, Fare quartiles

Family composition variables

Embarked indicators

# 3. Variables Descriptives

## 3.1 Categorical Variables Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Category 1 N% | Category 0 N% | Survival Rate Cat1 | Survival Rate Cat0 |
| Female | 314 35.2% | 577 64.8% | 74.2% | 18.9% |
| Class\_1 | 216 24.2% | 675 75.8% | 63.0% | 30.5% |
| Class\_2 | 184 20.7% | 707 79.3% | 47.3% | 36.1% |
| Alone | 537 60.3% | 354 39.7% | 30.4% | 50.6% |
| Has\_Cabin | 204 22.9% | 687 77.1% | 66.7% | 30.0% |

## 3.2 Continuous Variables Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Mean | Std | Min | Max | Survival Correlation |
| Age | 29.4 | 13.0 | 0.4 | 80.0 | Negative association |
| Fare | 32.2 | 49.7 | 0.0 | 512.3 | Positive association |
| SibSp | 0.5 | 1.1 | 0 | 8 | Mixed association |
| Parch | 0.4 | 0.8 | 0 | 6 | Mixed association |
| Family\_Size | 1.9 | 1.6 | 1 | 11 | U-shaped relationship |

# 4. Assumptions Check

## 4.1 Linearity Assumption Check

Continuous variables vs logit correlations:

|  |  |
| --- | --- |
| Variable | Logit Correlation |
| Age | 0.116 (weak negative) |
| SibSp | 0.089 (weak negative) |
| Variable | Logit Correlation |
| Parch | 0.104 (weak positive) |
| Fare | 0.434 (moderate positive) |

**Assessment**: Linearity assumption reasonably satisfied. Fare shows strongest linear relationship with log-odds.

## 4.2 Multicollinearity Check

**VIF Analysis** - All variables show VIF 5

|  |  |
| --- | --- |
| Variable | VIF |
| Class\_1 | 3.74 |
| Has\_Cabin | 2.54 |
| Alone | 2.27 |
| Fare | 1.77 |

**Assessment**: Multicollinearity assumption satisfied.

## 4.3 Sample Size Adequacy

**Training sample**: 712 observations

**Events (survivors)**: 273

**Predictors**: 11

**Events per variable EPV** : 24.8

**Assessment**: Highly adequate EPV >> 10 recommended minimum)

**4.4 Independence**

Cross-sectional study design ensures independence assumption is met.

# 5. Residual Check

## 5.1 Residual Analysis Summary

**Mean standardized residual**: 0.022 (near zero, appropriate)

**Standard deviation**: 1.08 (close to 1, appropriate)

**Range**: 3.89, 4.70 (reasonable range) **Distribution**: Approximately normal

## 5.2 Outlier Detection

Using Cook's distance and standardized residuals:

**High Cook's distance**: 309 observations 43.4%

**High standardized residuals 2.5** : 38 observations 5.3%

**Combined outliers**: 309 observations

**Assessment**: Large proportion flagged as outliers suggests heterogeneous survival patterns rather than data quality issues.

# 6. Univariate Models Crude Odds Ratios)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Crude OR | 95% CI | P-value | Interpretation |
| Female | 3.345 | 3.11 3.60 | 0.001 | Strong protective effect |
| Fare | 2.248 | 2.09 2.42 | 0.001 | Higher fare increases survival |
| Has\_Cabin | 2.052 | 1.91 2.21 | 0.001 | Cabin availability protective |
| Class\_1 | 1.863 | 1.73 2.00 | 0.001 | First class advantage |
| Embarked\_C | 1.430 | 1.33 1.54 | 0.001 | Cherbourg embarkation advantage |
| Class\_2 | 1.145 | 1.06 1.23 | 0.001 | Second class mild advantage |
| Age | 0.862 | 0.80 0.93 | 0.001 | Older age decreases survival |
| Alone | 0.644 | 0.60 0.69 | 0.001 | Being alone decreases survival |

**Significant predictors (p<0.05)**: 9 out of 11 variables

# 7. Multivariate Models

## 7.1 Model Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | With Outliers | Without Outliers | Difference |
| Test Accuracy | 0.810 | 0.816 | 0.006 |
| Test AUC | 0.844 | 0.841 | 0.003 |

**Decision**: Model WITHOUT outliers selected based on superior accuracy.

Training sample reduced from 712 to 403 observations after outlier removal.

## 7.2 Justification for Outlier Removal

While outliers improved AUC slightly 0.003 , the accuracy improvement 0.6%) without outliers indicates better overall classification performance. Outlier removal reduces model complexity and potential overfitting.

# 8. Crude vs Adjusted Odds Ratios

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Crude OR | Adjusted OR | % Change | Confounding Assessment |
| Female | 3.345 | 21.777 | 551.0% | Severe positive confounding |
| Fare | 2.248 | 1.492 | 33.6% | Moderate negative confounding |
| Has\_Cabin | 2.052 | 2.511 | 22.4% | Mild positive confounding |
| Class\_1 | 1.863 | 2.454 | 31.8% | Moderate positive confounding |
| Class\_2 | 1.145 | 1.997 | 74.5% | Severe positive confounding |
| Age | 0.862 | 0.439 | 49.0% | Severe negative confounding |
| Alone | 0.644 | 0.556 | 13.7% | Mild negative confounding |

## 8.1 Adjusted Odds Ratios Interpretation

**Strongest Predictors Final Model)**:

**Female**: OR 21.78 2078% increased odds) - Dominant protective factor

**Has\_Cabin**: OR 2.51 151% increased odds) - Strong protective effect

**Class\_1**: OR 2.45 145% increased odds) - Strong socioeconomic advantage

**Class\_2**: OR 2.00 100% increased odds) - Strong protective effect

**Age**: OR 0.44 56% decreased odds) - Strong age penalty

**Alone**: OR 0.56 44% decreased odds) - Isolation penalty

# 9. Confusion Matrix

## 9.1 Confusion Matrix Final Model - Without Outliers)

|  |  |  |
| --- | --- | --- |
|  | Predicted |  |
| **Actual** | **Dead** | **Survived** |
| **Dead** | 98 | 12 |
| **Survived** | 21 | 48 |

**9.2 Performance Metrics Primary Metrics**:

**Overall Accuracy**: 81.56%

**Sensitivity Recall** : 69.57% Ability to identify survivors

**Specificity**: 89.09% Ability to identify non-survivors

**Precision**: 80.00% Accuracy of survival predictions

**F1 Score**: 74.42% Harmonic mean of precision and recall

**AUC**: 84.10% Excellent discrimination ability **Classification Quality**:

**True Negatives**: 98 (correctly identified deaths)

**True Positives**: 48 (correctly identified survivors)

**False Positives**: 12 (incorrectly predicted survival) **False Negatives**: 21 (missed survivors)

## 9.3 Model Assessment

**Strengths**:

High overall accuracy 81.6%

Excellent specificity 89.1% - low false positive rate

Good AUC 84.1% - strong discriminative ability

Robust after outlier removal **Areas for Improvement**:

Moderate sensitivity 69.6% - missing some survivors

21 false negatives represent missed life-saving predictions

**Conclusions**

# 1. Methodological Validation

All logistic regression assumptions adequately satisfied

Appropriate sample size .

Minimal multicollinearity concerns

Model without outliers selected for optimal performance

# 2. Key Findings

Gender is the dominant survival predictor OR21.78

Socioeconomic factors (class, fare, cabin) strongly predictive

Age and family composition significantly impact survival

Substantial confounding present - multivariate analysis essential

# 3. Clinical/Practical Significance

Model achieves 81.6% accuracy with excellent specificity

Strong predictive capability for emergency response planning

Clear socioeconomic and demographic survival disparities identified

The analysis demonstrates that rigorous statistical methodology, combined with comprehensive feature engineering, produces reliable and interpretable predictive models for complex survival scenarios.