**Report - Logistic Regression Model**

**Data description:**

* **This report analyses the Titanic dataset to build a model predicting passenger survival.**
  + **Response Variable (Y): Survived (a binary variable where 1 = Survived, 0 = Did Not Survive).**
  + **Predictor Variables (X): Pclass (Passenger Class: 1, 2, or 3), Sex (male or female), Age (age in years), and SibSp (number of siblings/spouses aboard).**

**Shapiro-Wilk Test**

* **Variables not normally distributed: Age, Fare (both p-value < 0.001)**

| **Variable** | **Median** | **IQR (Q1, Q3)** |
| --- | --- | --- |
| **Age** | **28** | **(20.1, 38.0)** |
| **Fare** | **14.45** | **(7.91, 31.00)** |

* **Descriptive Statistics (Categorical Variables)**

|  |  |
| --- | --- |
| **Variable** | **n (%)** |
| **Survived = 0** | **549 (61.6%)** |
| **Survived = 1** | **342 (38.4%)** |
| **Sex = Male** | **577 (64.8%)** |
| **Sex = Female** | **314 (35.2%)** |
| **Pclass = 1** | **216 (24.2%)** |
| **Pclass = 2** | **184 (20.7%)** |
| **Pclass = 3** | **491 (55.1%)** |

**Data Preprocessing and Imputation:**

**The initial dataset contained 891 observations. Significant preprocessing was required before modelling.**

* **Missing Data: The Cabin column was removed due to excessive missingness (687 missing values). The Age and Embarked columns had 177 and 2 missing values, respectively .**
* **Missing Data Analysis: A test for data Missing Completely at Random (MCAR) was rejected (p < 0.05), indicating that the missing data was not random. Further Chi-square tests revealed that missingness in the Age variable was significantly associated with other variables like Pclass, SibSp, and Embarked. This suggests the data is**

**Missing at Random (MAR), which justifies using imputation.**

* **Imputation Strategy: A simple imputation method was used instead of regression imputation since the variables were skewed. Missing Age values were filled with the median age of 28, and the two missing Embarked values were filled with the mode.**
* **Influential Points: After fitting an initial model, Cook's distance was used to identify highly influential data points that could disproportionately affect the model's coefficients. A total of 67 influential points were identified .**

**A graph of a number of objects

AI-generated content may be incorrect.**

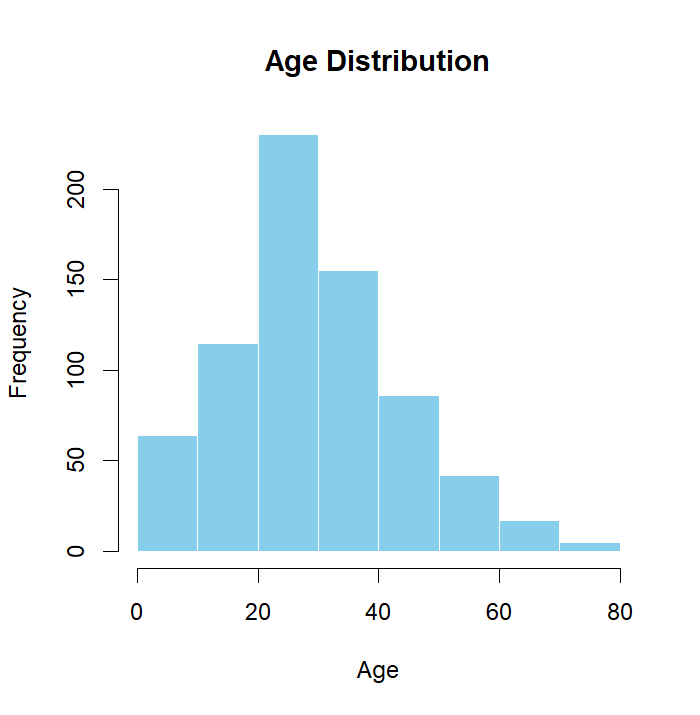
* **The data had no duplicated row. Additionally, the dataset was checked for any illogical values and none were found. Thus, the dataset was deemed suitable for immediate analysis.**

**Univariate Analysis with Graphs**

**Histogram:**

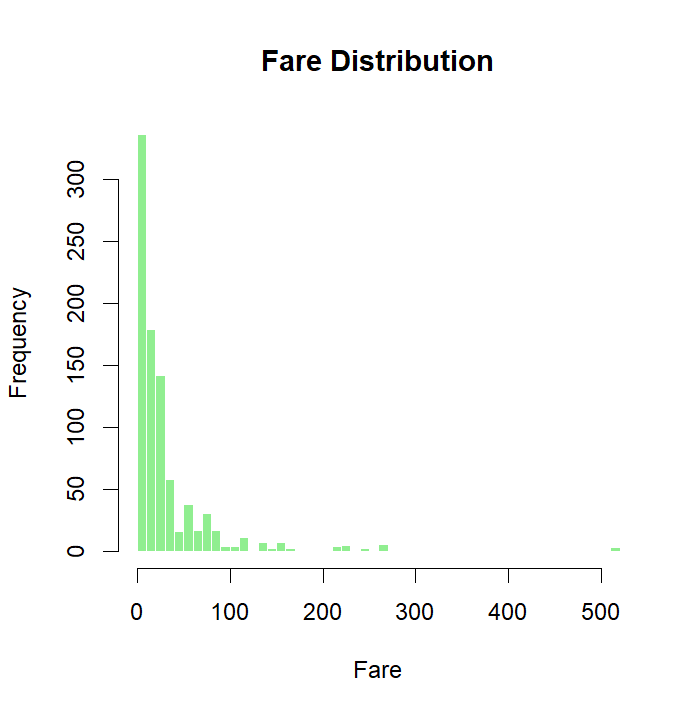
**Age**

**Displays a unimodal but right-skewed distribution, with most passengers between 20–40 years.**

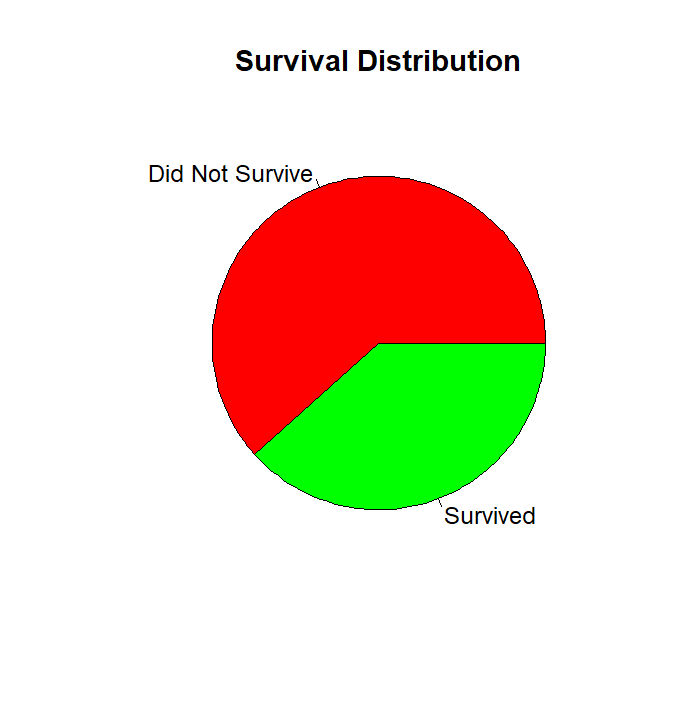
****

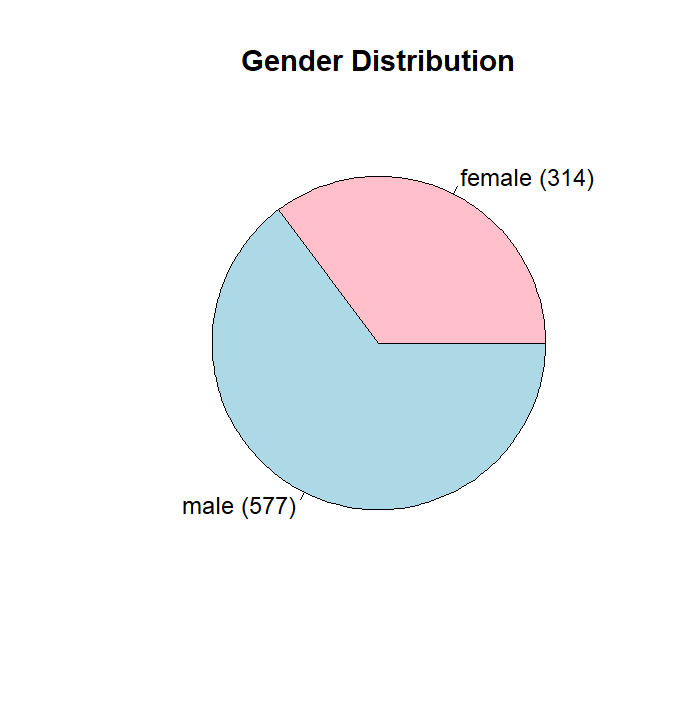
**Fare**

**Most fares are clustered below 50, with a few extreme outliers (up to 512).**

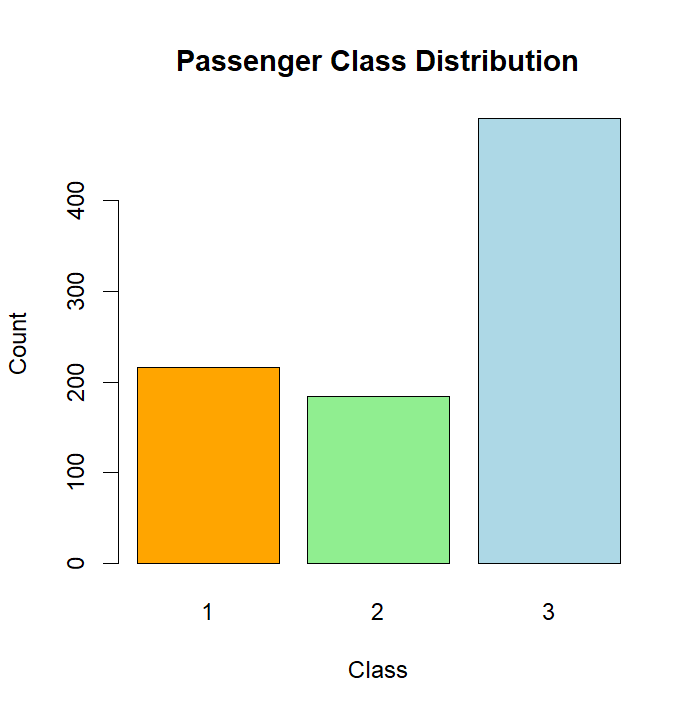
****

**Pie Charts:**

****

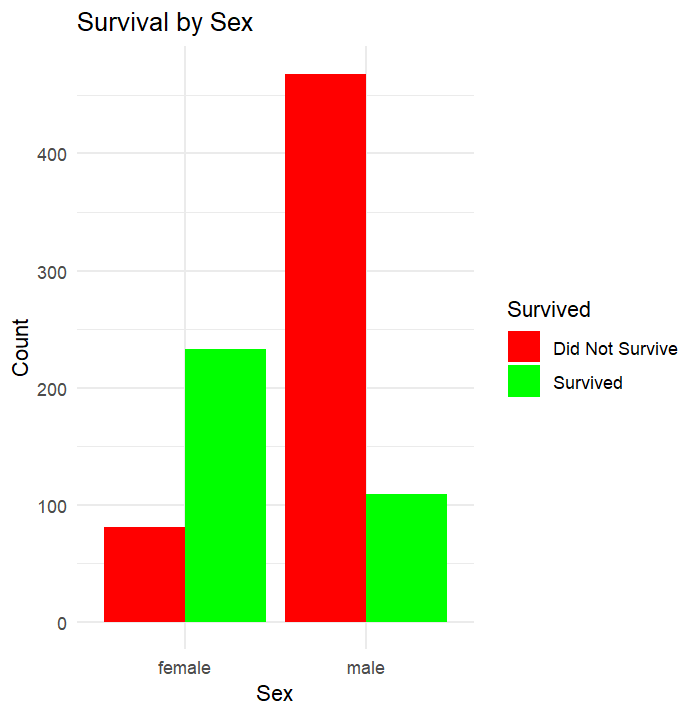
****

**Bar plots:**

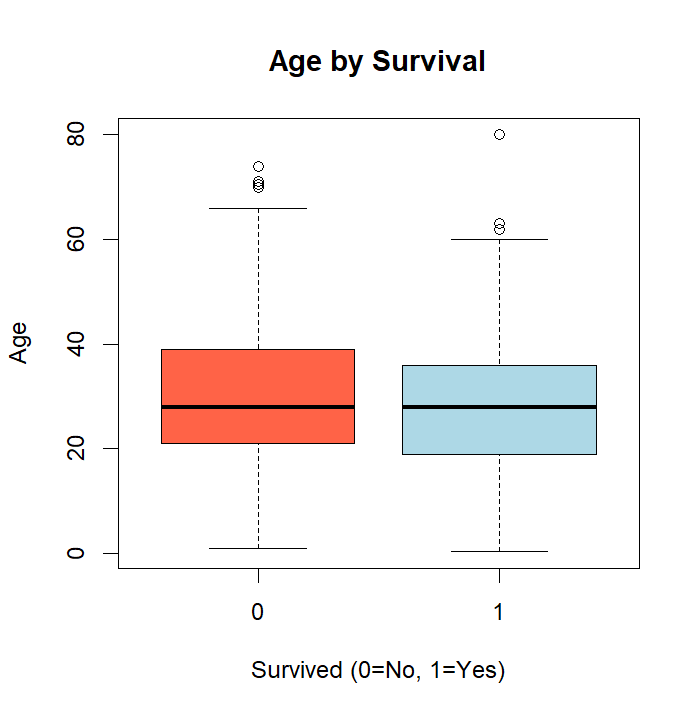
****

**A graph of a bar graph

AI-generated content may be incorrect.**

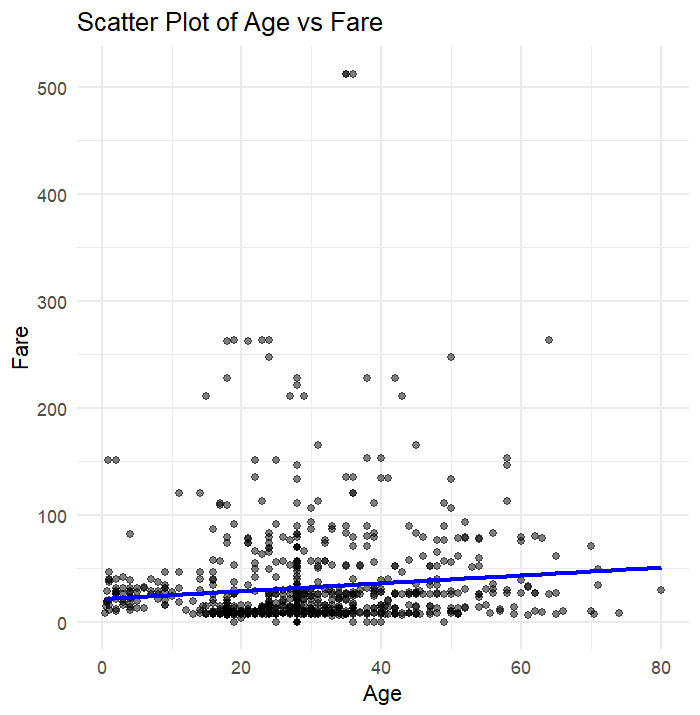
****

**Box plot**

****

**Correlation**

**Correlation of Age and Fare was 0.10**

****

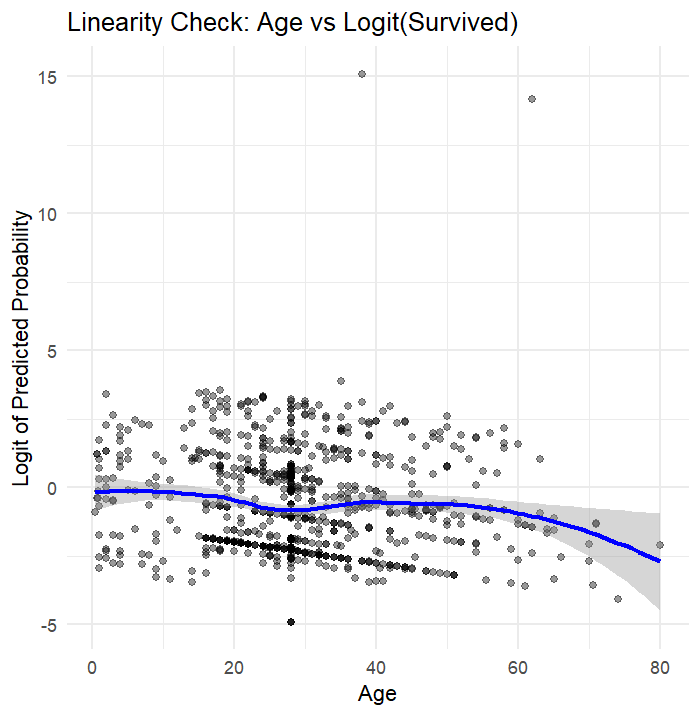
**Model form:**

**We wish to fit a logistic regression model of the form:**

**log(P/(1-P)) = β0​+β1​x1​+β2​x2​+...+ϵ**

**where Y is the binary outcome Survived, and the x variables are the predictors. The model predicts the log-odds of survival.**

**Linearity for logistic regression:**

****

* **If the line is roughly straight (linear), this suggests that the assumption of linearity in the logit (required for logistic regression for continuous predictors like Age) is reasonably satisfied.**
* **If the line is curved (concave/convex), it indicates non-linearity, meaning Age might not have a strictly linear relationship with the logit of survival.**

**Deviance residual plot:**

**A graph with black dots and red line

AI-generated content may be incorrect.**

**Pearson residual plot:**

**A graph with black dots and numbers

AI-generated content may be incorrect.**

**Model With All Data (Before Removing Influential Points):**

**Model Results (Final Model After Backward Selection)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor** | **Estimate (β)** | **Std. Error** | **z value** | **p-value** | **Odds Ratio (exp(β))** |
| **Intercept** | **4.85** | **0.66** | **7.36** | **<0.001** | **–** |
| **Pclass2** | **-1.22** | **0.36** | **-3.37** | **<0.001** | **0.29** |
| **Pclass3** | **-2.35** | **0.34** | **-6.94** | **<0.001** | **0.10** |
| **Sex (Male)** | **-2.63** | **0.30** | **-8.79** | **<0.001** | **0.07** |
| **Age** | **-0.04** | **0.01** | **-4.14** | **<0.001** | **0.96** |
| **SibSp (1)** | **-0.34** | **0.34** | **-0.99** | **0.322** | **0.71** |
| **SibSp (2)** | **-1.23** | **0.68** | **-1.81** | **0.070** | **0.29** |
| **SibSp (3)** | **-2.15** | **0.84** | **-2.57** | **0.011** | **0.12** |
| **SibSp (4)** | **-1.05** | **1.19** | **-0.88** | **0.380** | **0.35** |
| **SibSp (5)** | **-1.00** | **1.24** | **-0.81** | **0.418** | **0.37** |
| **SibSp (8)** | **-3.99** | **1.41** | **-2.83** | **0.005** | **0.02** |

**Interpretation of Coefficients**

***Model With Influential Points:***

* **Sex (Male): Strong negative association with survival (p < 0.001). Being male decreases the odds of survival by ~93%.**
* **Pclass: Both 2nd and 3rd class passengers had significantly lower odds of survival compared to 1st class (OR ≈ 0.29 for Pclass2, OR ≈ 0.095 for Pclass3).**
* **Age: Each additional year of age reduces the odds of survival by ~4% (p < 0.001).**
* **SibSp: Having 3 siblings/spouses aboard was significantly associated with lower odds of survival (p = 0.011). Other SibSp levels were not statistically significant.**

**Model Fit Statistics:**

* **Null deviance: 876.75 (657 df)**
* **Residual deviance: 567.26 (652 df)**
* **AIC: 579.26**
* **McFadden’s Pseudo R²: 0.353**

**Goodness-of-Fit & Diagnostics:**

* **ROC AUC: 0.8538 which means Excellent discrimination**
* **Hosmer–Lemeshow Test: χ² = 26.56, df = 8, p = 0.00084 → Some lack of calibration**
* **VIF: All predictors < 2 which means no multicollinearity issues**

**Classification Performance (Test Data)/Confusion matrix**

|  | **Actual = 0** | **Actual = 1** |
| --- | --- | --- |
| **Predicted = 0** | **89** | **24** |
| **Predicted = 1** | **13** | **40** |

* **Accuracy: 78.4%**
* **Sensitivity: 62.5%**
* **Specificity: 87.3%**

**Model After Removing Influential Points (Cook’s Distance > 4/n)**

**A total of 67 influential points were removed to improve model stability.**

**Model Results (Refit After Backward Selection)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Predictor** | **Estimate (β)** | **Std. Error** | **z value** | **p-value** | **Odds Ratio (exp(β))** |
| **Intercept** | **3.894** | **0.412** | **9.45** | **<0.001** | **–** |
| **Pclass2** | **-1.196** | **0.270** | **-4.43** | **<0.001** | **0.30** |
| **Pclass3** | **-2.206** | **0.250** | **-8.82** | **<0.001** | **0.11** |
| **Sex (Male)** | **-2.766** | **0.199** | **-13.90** | **<0.001** | **0.06** |
| **Age** | **-0.042** | **0.008** | **-5.12** | **<0.001** | **0.96** |
| **SibSp1** | **0.172** | **0.215** | **0.80** | **0.424** | **1.19** |
| **SibSp2** | **-0.337** | **0.540** | **-0.62** | **0.532** | **0.71** |
| **SibSp3** | **-2.104** | **0.683** | **-3.08** | **0.002** | **0.12** |
| **SibSp4** | **-1.450** | **0.772** | **-1.88** | **0.060** | **0.23** |
| **SibSp5** | **-16.17** | **1050.00** | **-0.02** | **0.988** | **~0.00** |
| **SibSp8** | **-15.93** | **748.90** | **-0.02** | **0.983** | **~0.00** |

**Interpretation of Coefficients:**

* **Pclass:**
  + **Passengers in 2nd class had about 70% lower odds of survival compared to 1st class (OR ≈ 0.30).**
  + **Passengers in 3rd class had nearly 90% lower odds of survival compared to 1st class (OR ≈ 0.11).**
* **Sex:**
  + **Males had about 94% lower odds of survival compared to females (OR ≈ 0.06), holding other variables constant.**
* **Age:**
  + **Each additional year of age was associated with about a 4% decrease in the odds of survival (OR ≈ 0.96), keeping other variables constant.**
* **SibSp (Number of siblings/spouses aboard):**
  + **Having 1 sibling/spouse slightly increased the odds of survival (OR ≈ 1.19), though this was not statistically significant.**
  + **Having 2 or more siblings/spouses showed progressively lower odds of survival, with 3 siblings/spouses significantly reducing the odds (OR ≈ 0.12).**
  + **Categories with very rare counts (SibSp 5, 8) had unreliable estimates and should be interpreted with caution.**

**Model Fit Statistics:**

* **Null deviance: 1098.10 (823 df)**
* **Residual deviance: 730.75 (818 df)**
* **AIC: 742.75**
* **McFadden’s Pseudo R²: 0.375 (↑ from 0.353)**

**Goodness-of-Fit & Diagnostics:**

* **ROC AUC: 0.8604 (no loss in discrimination)**
* **Hosmer–Lemeshow Test: p < 0.01 which means slight lack of fit persists, but improved calibration visually**
* **VIF: All predictors < 2 which means no multicollinearity**

**Classification Performance (Clean Test Data) /Confusion matrix:**

|  | **Actual = 0** | **Actual = 1** |
| --- | --- | --- |
| **Predicted = 0** | **95** | **16** |
| **Predicted = 1** | **15** | **53** |

* **Accuracy: 82.6%**
* **Sensitivity: 76.8%**
* **Specificity: 86.4%**

**Conclusion:**

**Both models (before and after removing influential points) identified the same key predictors of survival: Pclass, Sex, Age, and SibSp. The direction and magnitude of effects remained consistent, with Sex (male) and Pclass3 showing the strongest negative impact on survival probability.**

**Although the Hosmer–Lemeshow test (p < 0.01) indicates a slight lack of fit, the model shows excellent discrimination (ROC AUC = 0.8604), stable coefficients, and no multicollinearity, making it robust and suitable for predicting survival.**

**After removing influential points:**

* **Model fit improved (AIC = 742.75, McFadden’s pseudo R² = 0.375).**
* **Predictive performance improved, with accuracy increasing from 78.4% to 82.6% and sensitivity increasing from 62.5% to 76.8%; specificity remained high at 86.4%.**
* **Odds ratios remained similar, confirming the robustness of the model.**

**Overall, the final model (after removing influential points) provides a more stable and better-fitting model for predicting Titanic survival while retaining the same key insights about risk factors.**