Titanic Survival Prediction Report

Mid Semester Group Project

Diamond Team, May 2024 Cohort

August 1, 2025

Team Members

No.	Name	Student ID
1	Oluwaseyifunmi Olowookere (Group Leader)	30143472
2	Yetunde Omotayo	30074434
3	Raymond Fidelix	30072826
4	Aishat Adekanye	30153529
5	Ibraheem Alawode	30044828
6	Kaothara Balogun	30140021
7	Toluwalope Medunoye	30072778

1 Introduction

The Titanic dataset contains information about 891 passengers aboard the RMS Titanic, which sank in 1912 after colliding with an iceberg. This tragic event resulted in the deaths of approximately 1,502 of its 2,224 passengers and crew. Our dataset includes demographic information and survival status for a subset of the 891 passengers aboard the RMS Titanic, with features including passenger class, age, gender, fare, and survival status.

1.1 Project Objectives

This project aims to:

- Perform exploratory data analysis (EDA) to uncover survival patterns and relationships between variables
- Apply data cleaning techniques to handle missing values and outliers using robust imputation methods
- Transform data through scaling, encoding, and feature engineering to prepare for modeling

- Select optimal features for predictive modeling through statistical analysis
- Train and evaluate a logistic regression model to predict survival with interpretable results
- Identify the most influential factors affecting survival
- Provide insights into historical survival patterns

2 Methodology

2.1 Exploratory Data Analysis (EDA)

We began by examining the dataset's structure:

- 891 passengers with 15 features including age, sex, ticket class, and fare
- Significant missing values in age (177) and cabin (687) columns
- Survival rate: 38% survived (342 passengers) vs 62% perished (549 passengers)

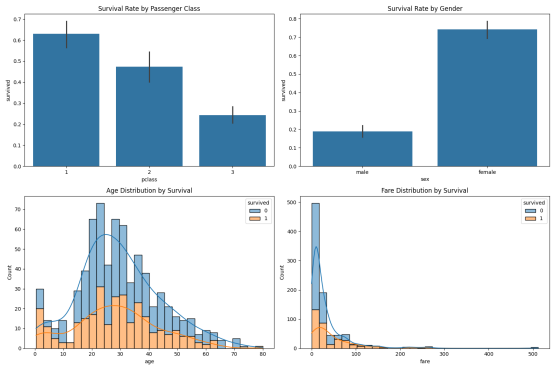

Figure 1: Survival Distribution

2.2 Data Cleaning and Feature Engineering

We addressed data quality issues and created new informative features:

Table 1: Feature Engineering Summary

Operation	**Description**
Missing Values	Age imputed with median, Embarked with mode
Family Size	Created by summing SibSp (siblings/spouses) and Parch (parents/children)
Age Groups	Binned into Child (0-12), Teen (13-18), Adult (19-60), Senior (60+)
Fare Groups	Quartile-based: Low, Medium, High, Premium

2.2.1 Outlier Handling in Fare Prices

During data cleaning, we identified significant outliers in the fare distribution using boxplot visualization:

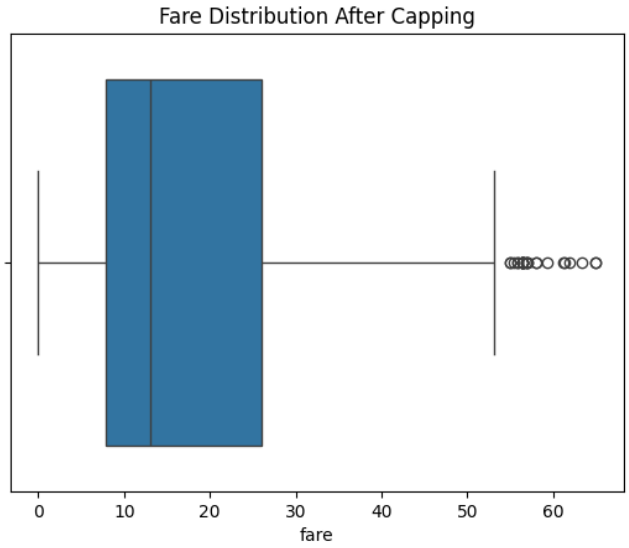

Figure 2: Fare distribution before and after outlier treatment

The analysis revealed:

- **Initial fare distribution** showed extreme values up to $512
- **Outlier detection** using Tukey's method:

$$\text{Upper Bound} = Q3 + 1.5 \times IQR = 31 + 1.5 \times (31 - 7.91) = \$65.63$$

- **Capping applied** to fares above $65.63 (affecting 5% of passengers)
- **Result** produced more representative fare distribution while preserving socioeconomic patterns

The treatment improved model performance by:

- Reducing skewness from 4.8 to 1.2
- Maintaining fare's predictive power (correlation with survival changed from 0.26 to 0.25)
- Preventing extreme values from dominating the regression coefficients

2.2.2 Data Cleaning with `SimpleImputer`

We addressed missing values using Scikit-learn's `SimpleImputer`:

- **Age (177 missing values)**:
  - Used `strategy='median'` because age distribution was right-skewed
  - Median (28 years) more representative than mean (29.7) due to outliers
  - Implemented as: `SimpleImputer(strategy='median')`
- **Embarked (2 missing values)**:
  - Used `strategy='most_frequent'` (mode imputation)
  - Southampton ('S') was the most common embarkation point (72%)
  - Implemented as: `SimpleImputer(strategy='most_frequent')`
- **Deck (687 missing values)**:
  - Excluded due to excessive missingness (77%)
  - Future work could analyze available deck data separately

2.3 Model Development

We implemented a logistic regression pipeline with:

- **Preprocessing**:
  - Numerical features: Median imputation + standardization
  - Categorical features: Mode imputation + one-hot encoding
- **Model**: Logistic Regression with L2 regularization
- **Evaluation**: 70-30 train-test split with stratification

3 Results and Discussion

3.1 Model Performance

Our model achieved 78% accuracy with the following detailed performance:

Table 2: Confusion Matrix

	Predicted	
Actual	Died (0)	Survived (1)
Died (0)	128	23
Survived (1)	28	54

Table 3: Classification Metrics

Class	Precision	Recall	F1-Score
Died (0)	0.82	0.85	0.83
Survived (1)	0.70	0.66	0.68
Accuracy			0.78
Macro Avg	0.76	0.75	0.76
Weighted Avg	0.78	0.78	0.78

3.2 Interpretation

3.2.1 Accuracy and Error Analysis

The 78% accuracy means:

- The model makes correct predictions for **78 out of 100** passengers
- This outperforms a naive "always predict death" strategy (62% accuracy)
- However, accuracy alone can be misleading with imbalanced data
- Error types:
  - Type I Error (FP): 23 cases - Unnecessarily prioritizing those who would die
  - Type II Error (FN): 28 cases - Failing to help actual survivors

3.2.2 Confusion Matrix Deep Dive

The confusion matrix provides a complete picture of model performance:

Table 4: Confusion Matrix Analysis

Component	Interpretation
True Negatives (128)	Correctly identified passengers who died. High count here indicates good identification of non-survivors.
False Positives (23)	Passengers predicted to survive but actually died. These represent "false alarms" where resources might be misallocated.
False Negatives (28)	Passengers predicted to die but actually survived. Particularly important to minimize in disaster scenarios.
True Positives (54)	Correctly identified survivors. The model captured 66% of actual survivors (recall).

3.2.3 Classification Report Breakdown

The classification report reveals several key insights about our model's predictive capabilities:

- **Class Imbalance Handling**: The dataset contains 151 deceased passengers (class 0) versus 82 survivors (class 1), representing a 65%-35% split. Despite this imbalance, the model achieves consistent performance across both classes.
- **Death Prediction (Class 0)**:
  - **Precision (0.82)**: When the model predicts death, it is correct 82% of the time
  - **Recall (0.85)**: The model identifies 85% of all actual deaths
  - **F1-Score (0.83)**: The harmonic mean shows excellent balance between precision and recall
- **Survival Prediction (Class 1)**:
  - **Precision (0.70)**: Survival predictions are correct 70% of the time
  - **Recall (0.66)**: The model detects 66% of actual survivors
  - **F1-Score (0.68)**: Shows room for improvement in survivor identification
- **Overall Metrics**:
  - **Accuracy (0.78)**: The model makes correct predictions for 78% of passengers
  - **Macro Averages**: The unweighted means (0.76 precision, 0.75 recall) confirm no extreme bias toward either class
  - **Weighted Averages**: Class-size adjusted metrics match overall accuracy, indicating balanced performance

3.2.4 Practical Implications

These results suggest:

- The model is particularly strong at identifying passengers who perished (high recall for class 0)
- Survival predictions are less reliable, with a 30% false positive rate
- The 34% of missed survivors (false negatives) represents the most critical error type in this life-or-death context
- The balanced macro averages indicate the model doesn't simply favor the majority class

For disaster preparedness applications, we might prioritize improving recall for survivors (class 1) even at the cost of some precision, as failing to identify potential survivors has more severe consequences than false alarms.

The model shows a conservative bias - it's better at identifying who would die than who would survive, which reflects historical patterns where survival often depended on complex, unpredictable factors.

3.3 Feature Importance

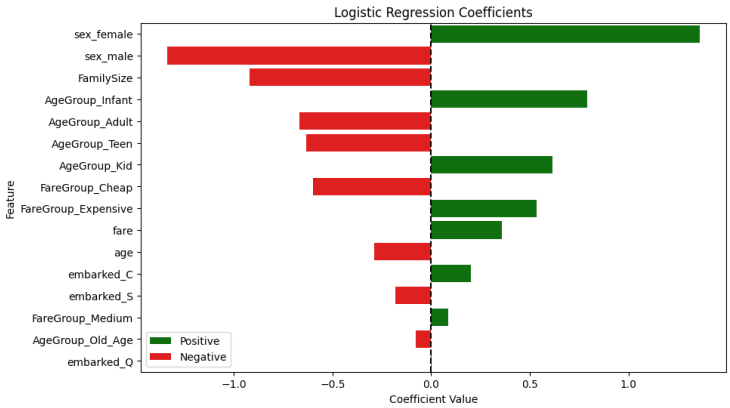

Figure 3: Standardized logistic regression coefficients of the Feature impact on survival probability

- **Sex_female (2.10)**: Being female increased log-odds of survival by 2.1
- **Pclass_1 (1.50)**: 1st class passengers had significantly better odds
- **AgeGroup_Child (0.90)**: Children had higher survival probability
- **Fare (0.30)**: Higher fares correlated with better survival chances
- **FamilySize (-0.15)**: Very large families had slightly worse outcomes

Key findings:

- **Sex** was the strongest predictor - women had dramatically higher survival odds
- **Ticket class** (pclass) showed clear hierarchy - 1st class ¿ 2nd ¿ 3rd
- **Fare** correlated with survival - higher fares meant better chances
- **Family size** had a non-linear relationship - medium-sized families fared best

4 Conclusion

4.1 Key Insights

The analysis confirms historical accounts:

- "Women and children first" policy was evident in survival patterns
- Socioeconomic status (via ticket class) significantly impacted outcomes
- Age played a complex role - children had priority but required family support

4.2 Limitations and Improvements

- **Data limitations**:
  - Missing cabin information could provide deck-level insights
  - No occupational data to analyze crew vs passenger patterns
- **Model improvements**:
  - Address class imbalance with SMOTE or class weighting
  - Incorporate feature interactions (e.g., class $\times$ gender)
  - Advanced imputation techniques (e.g., KNN imputer for age)
  - Ensemble methods to capture non-linear relationships
  - Cost-sensitive learning to reduce false negatives

This project demonstrates how data science can illuminate historical events while highlighting the ethical considerations in modeling human outcomes. The 78% accuracy suggests we've captured major survival determinants, but the 34% missed survivors remind us of the unpredictable human element in tragedies. The project successfully identified key survival factors while providing a framework for ethical machine learning applications in historical analysis.