Survival Analysis of Titanic Passengers

Author: Samiksha Atul Hujare

Internship Provider: ShadowFox Data Science Internship

Project Type: Jupyter Notebook Data Analysis

---

1. Introduction

The **Titanic Disaster of 1912** is one of the most well-known maritime tragedies in history. The ship, carrying **2,224 passengers and crew**, collided with an iceberg and resulted in the loss of more than **1,500 lives**. This project aims to analyze the **Titanic dataset** from Kaggle to determine the factors that influenced survival rates.

The key objectives of this study are:
- To explore the dataset and identify missing values.
- To analyze **passenger demographics**, **class distribution**, and **survival rates**.
- To perform **feature engineering** and handle missing data.
- To apply a **Logistic Regression model** to predict survival.
- To evaluate the model's performance and summarize findings.

---

2. Data Exploration and Cleaning

2.1 Loading the Dataset

The dataset was obtained from **Kaggle's Titanic competition** and contains information about passengers, including their age, gender, ticket class, fare, and survival status.

2.2 Understanding the Data

The dataset consists of the following key columns:
- **PassengerId** – Unique identifier for each passenger.
- **Survived** – Target variable (0 = No, 1 = Yes).
- **Pclass** – Ticket class (1st, 2nd, 3rd).
- **Name** – Passenger's name.
- **Sex** – Gender of the passenger.
- **Age** – Age of the passenger.
- **SibSp** – Number of siblings/spouses aboard.
- **Parch** – Number of parents/children aboard.

- **Ticket** – Ticket number.
- **Fare** – Ticket price.
- **Cabin** – Cabin number (many missing values).
- **Embarked** – Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

2.3 Handling Missing Values

Several columns had missing values:
- **Age**: Missing values were replaced with the median age.
- **Cabin**: Dropped due to excessive missing values.
- **Embarked**: Missing values were replaced with the most common port (‘S’).

---

3. Exploratory Data Analysis (EDA)

3.1 Survival Distribution

The survival rate among passengers was analyzed:
- **38.38% (342 passengers) survived.**
- **61.62% (549 passengers) did not survive.**

3.2 Gender and Survival

- **Females had a much higher survival rate (74%) than males (19%).**
- This confirms the “women and children first” evacuation policy followed on the Titanic.

3.3 Passenger Class and Survival

- **First-class passengers had a higher survival rate** (62%), compared to **second-class (47%) and third-class (24%)**.
- The wealthier passengers had **better access to lifeboats**, which contributed to higher survival rates.

3.4 Age Distribution and Survival

- **Children (under 10 years) had higher survival rates** than adults.
- **Elderly passengers (above 60) had lower survival chances.**

3.5 Fare and Survival

- Higher ticket fares correlated with a **higher chance of survival**.
- First-class passengers, who paid higher fares, had a better chance of surviving.

---

4. Feature Engineering

4.1 Encoding Categorical Variables

- **Sex:** Converted ‘male’ to 0 and ‘female’ to 1.
- **Embarked:** Converted ‘S’ to 0, ‘C’ to 1, and ‘Q’ to 2.

4.2 Handling Numerical Data

- Missing **Age** values were filled with the median age.
- **Fare** values were log-transformed to handle skewness.

---

5. Machine Learning Model: Logistic Regression

A **Logistic Regression model** was used to predict survival.

5.1 Model Training

- The dataset was split into **training (80%) and testing (20%) sets**.
- Selected **features**: `Pclass`, `Sex`, `Age`, `SibSp`, `Parch`, `Fare`, and `Embarked`.

5.2 Model Performance

- **Accuracy:** 81.34%
- **Precision:** 78.21%
- **Recall:** 74.56%
- **F1-Score:** 76.32%

5.3 Confusion Matrix Results

- The model correctly identified **most survivors and non-survivors**, but some misclassifications occurred.

---

6. Insights and Conclusion

6.1 Key Findings

1. **Women had a significantly higher survival rate (74%) compared to men (19%).**
2. **First-class passengers were more likely to survive** compared to second and third-class passengers.
3. **Children had better survival chances than adults and elderly passengers.**
4. **Higher ticket fare correlated with higher survival rates.**

5. **The Logistic Regression model achieved an accuracy of 81.34%**, which is reasonable for predicting survival outcomes.

6.2 Limitations and Future Improvements

- The dataset is **limited in scope** and does not include all possible survival factors.
- The model can be improved by using **advanced machine learning techniques** such as **Random Forest or Neural Networks**.

---

7. References

- Kaggle Titanic Dataset: https://www.kaggle.com/c/titanic
- Scikit-learn Documentation: https://scikit-learn.org
- Matplotlib & Seaborn Documentation: https://matplotlib.org/

End of Report