Report: Exploratory Data Analysis of the Titanic Dataset

By Ritvik Mittal
For Course: Exploratory Data Analysis for Machine Learning.
Coursera

1. Introduction

In this report, I present an exploratory data analysis of the Titanic dataset, which contains information about the passengers aboard the ill-fated RMS Titanic. Our goal is to gain insights into the factors that influenced survival and perform an initial analysis to understand the patterns in the data.

2. Data Description

The Titanic dataset consists of several attributes, including:
- PassengerId: Unique identifier for each passenger
- Survived: Whether the passenger survived (0 = No, 1 = Yes)
- Pclass: Ticket class (1 = 1st class, 2 = 2nd class, 3 = 3rd class)
- Name: Passenger's name
- Sex: Gender of the passenger (0 = Male, 1 = Female)
- Age: Age of the passenger (some missing values were filled with the mean)
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number
- Fare: Passenger fare
- Cabin: Cabin number (a significant number of missing values)
- Embarked: Port of embarkation (0 = Cherbourg, 1 = Queenstown, 2 = Southampton; some missing values were filled with the mode)

3. Data Exploration

I began my analysis by exploring the dataset containing 891 rows and 12 columns. The 'Survived' column is our target variable; the rest are potential predictor variables. I computed summary statistics and visualised the distribution of numerical variables like 'Age' to gain initial insights into the data.

4. Key Findings and Insights

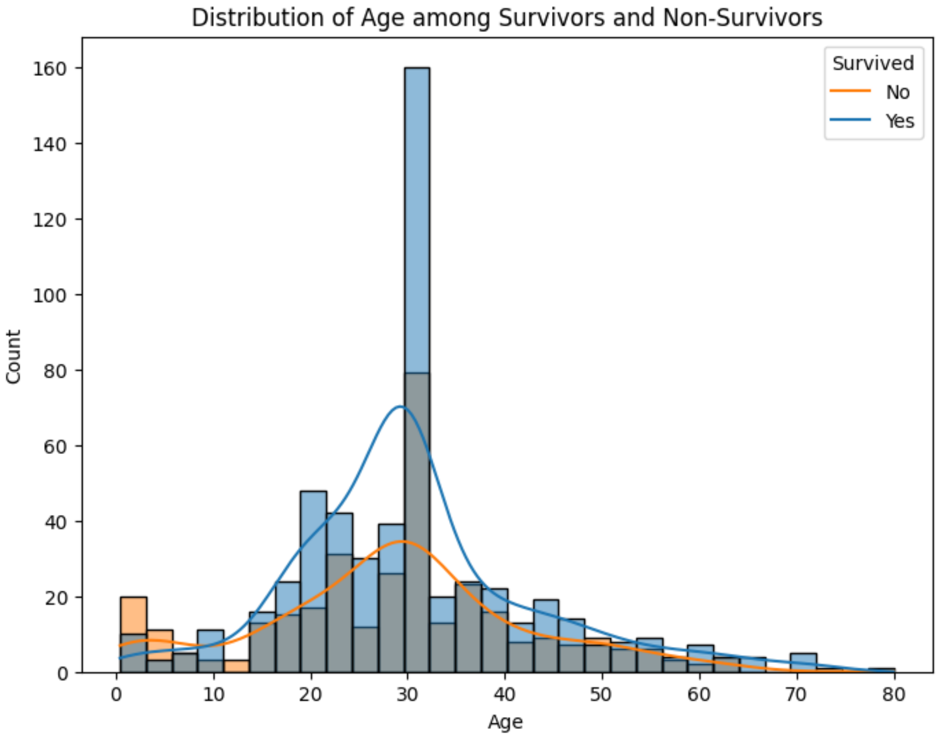
I found several exciting insights during this exploratory data analysis:
- The overall survival rate is approximately 38.4%.
- Female passengers had a significantly higher survival rate (74.2%) than male passengers (18.9%).
- Passengers in the 1st class had a higher chance of survival (62.6%) compared to those in the 2nd class (47.3%) and 3rd class (24.2%).
- The age distribution among survivors and non-survivors showed that children and young adults had a relatively higher chance of survival.

5. Hypotheses

I formulated three hypotheses to explore the data further:
1. Hypothesis: Passengers in higher classes (Pclass = 1) had a higher chance of survival.
2. Hypothesis: Female passengers had a higher chance of survival than male passengers.
3. Hypothesis: Passengers who boarded from Cherbourg (Embarked = 0) had a higher chance of survival.

6. Significance Test

To test one of our hypotheses, I conducted a significance test to examine the relationship between sex and survival. The results showed that the relationship between sex and survival is statistically significant (p-value < 0.05), confirming that female passengers indeed had a higher chance of survival.



7. Suggestions for Next Steps

To deepen the analysis and gain more insights, I propose the following next steps:
1. Feature Engineering: Create new features such as family size (combining 'SibSp' and 'Parch') and cabin deck from the 'Cabin' column to capture additional patterns.
2. Predictive Modeling: Build machine learning models to predict survival based on different features. Use Logistic Regression, Random Forest, or Gradient Boosting for predictions.
3. Feature Importance: Identify the most important features contributing to the prediction of survival, which can provide additional insights.

8. Data Quality Summary and Request for Additional Data

The Titanic dataset is relatively clean, with some missing values handled in the 'Age' and 'Embarked' columns. However, the 'Cabin' column contains many missing values and could be further explored or excluded from the analysis.

I request additional data related to the following:
1. Passenger's Social and Economic Background: Information such as occupation, income, or socioeconomic status could provide further insights into survival patterns.
2. Lifeboat Assignment: Knowing which lifeboat each passenger was assigned to could help understand the rescue process.

3. Crew Information: Including data about the crew members could provide a more comprehensive analysis of survival rates for different roles on the ship.

9. Conclusion

In conclusion, this exploratory data analysis of the Titanic dataset revealed valuable insights into the factors influencing passenger survival. Female passengers and those in the 1st class had a higher chance of survival. Further analysis, including predictive modelling and feature importance, could provide more accurate predictions and additional insights into this tragic event.

10. References

- Kaggle Titanic: Machine Learning from Disaster: https://www.kaggle.com/c/titanic