Anderson Eguasa

Report on Titanic Data set

Exploratory Data Analysis Report: Titanic Train Dataset

1. Executive Summary

This report provides an in-depth exploration of the Titanic training dataset. The analysis focuses on identifying key features that may influence survival outcomes aboard the Titanic. By examining variables such as passenger class, sex, age, and family connections, we aim to highlight trends and insights that can support predictive modeling and further investigations into the disaster's human impact.

2. Introduction

The Titanic dataset is one of the most popular datasets used for introductory data science and machine learning tasks. It contains details about the passengers aboard the Titanic, including demographics, ticket information, and survival status. This report outlines the exploratory data analysis (EDA) process and summarizes the key findings regarding the survival factors of passengers.

3. Data Overview

- **Total Observations:** 891 passengers (as provided in the train dataset)
- **Key Variables:**
  - **Survived:** Indicator of survival (0 = No, 1 = Yes)
  - **Pclass:** Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
  - **Name:** Name of the passenger
  - **Sex:** Gender of the passenger
  - **Age:** Age in years (some missing values)
  - **SibSp:** Number of siblings/spouses aboard
  - **Parch:** Number of parents/children aboard
  - **Ticket:** Ticket number
  - **Fare:** Passenger fare
  - **Cabin:** Cabin number (many missing values)
  - **Embarked:** Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

4. Data Preparation and Cleaning

- **Missing Values:**
  - *Age* has missing entries, which were imputed using median values.
  - *Cabin* has a high percentage of missing values; thus, it was excluded from some analyses.

    - *Embarked* contains a few missing entries, which were imputed with the most frequent value.
- **Feature Engineering:**
    - A new feature, **FamilySize**, was created by combining *SibSp* and *Parch*.
    - Titles were extracted from passenger names to provide additional insights into social status and age.

5. Exploratory Analysis and Key Findings

5.1 Survival Rate

- **Overall Survival Rate:** Approximately 38% of passengers survived.
- A survival imbalance is apparent, prompting further analysis into the underlying factors.

5.2 Gender Analysis

- **Female vs. Male Survival:**
    - Females had a significantly higher survival rate compared to males.
    - The “women and children first” policy appears evident, with female passengers more likely to survive.

5.3 Class Analysis (Pclass)

- **Survival by Passenger Class:**
    - 1st class passengers had the highest survival rates.
    - 3rd class passengers experienced significantly lower survival rates.
    - This suggests that socio-economic status played a key role in survival.

5.4 Age Analysis

- **Survival by Age:**
    - Children (age < 16) showed higher survival percentages.
    - A survival drop is observed for middle-aged passengers.
    - The age distribution helps reinforce the narrative that younger passengers and children were given priority during evacuation.

5.5 Family Size Analysis

- **Effect of Family Size on Survival:**
    - Passengers traveling with a small family (or alone) had different survival odds compared to those in larger family groups.
    - Both very small (singleton) and very large families tended to have lower survival rates, potentially due to difficulties in coordinating group evacuations.

5.6 Fare Analysis

- **Fare and Survival:**
  - Higher fares (often associated with 1st class) correlated with higher survival rates.
  - Fare distribution supports the class-based survival discrepancies.

5.7 Port of Embarkation

- **Embarked Variable:**
  - The majority of passengers boarded at Southampton (S), with noticeable survival differences among embarkation points.
  - Variations in survival across ports could be linked to differences in passenger demographics or ticket class distribution.

6. Conclusions and Recommendations

- **Key Insights:**
  - **Gender and Class:** These were the most decisive factors in survival. Female and higher-class passengers had better odds.
  - **Age:** Younger passengers, particularly children, had higher survival rates.
  - **Socio-Economic Factors:** Evident through the relationship between fare, class, and survival.
- **Recommendations for Future Analysis:**
  - Further investigation into cabin information could be valuable if more complete data becomes available.
  - Advanced modeling techniques (e.g., decision trees or ensemble methods) may capture interactions among features better.
  - Incorporating feature engineering (like title extraction) can enhance predictive performance in classification tasks.
- **Limitations:**
  - The dataset has inherent biases (e.g., missing data in age and cabin variables).
  - The retrospective nature of the analysis limits causal conclusions.

7. Final Thoughts

The Titanic dataset continues to be a rich resource for demonstrating data analysis techniques. This report underscores the significance of data preprocessing and exploratory analysis in uncovering patterns, which, when applied to predictive modeling, can lead to robust insights and better model performance.