Techconative Assignment

Data Analysis Report

Question 2: Classification Model

Introduction

In this document, I present the exploratory data analysis (EDA) results derived from the Titanic dataset - https://www.kaggle.com/c/titanic/overview . My analysis focuses on understanding the relationships and trends among various features. This report contains only the data observations I gathered. **For a detailed implementation, including the code and visualizations that led to these results, please check out the main code file, I have attached in the GitHub Repo link.**

Data Overview

The dataset contains various features about passengers, some of which have missing values:

- **Features with Missing Values**:
  - Age
  - Cabin
  - Embarked

Types of Features

- **Categorical Features**: Sex, Embarked
- **Ordinal Feature**: Pclass
- **Continuous Features**: Age, Fare
- **Discrete Features**: SibSp, Parch

Survival Rate

The survival of passengers is encoded as follows:

- 0 = No
- 1 = Yes

Approximately **38.4%** of the passengers in the training set survived.

Feature Analysis

1. Sex

- Indicates the sex of the passenger (Male or Female).
- **Observations**:

  - More men than women were on the ship.
  - Women saved: nearly twice the number of men saved.
  - Women's survival rate: ~75%.
  - Men's survival rate: ~18-19%.

2. Embarked

- Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).
- **Observations**:
  - Passengers from port C had the highest survival rate (around 0.55), while those from port S had the lowest.
  - Maximum passengers boarded from S; majority from Pclass 3.
  - High survival rates for women in Pclass 1 and Pclass 2.
  - Port Q showed extremely low survival for men, with almost all passengers from Pclass 3.

3. Pclass

- A proxy for socio-economic status (SES): 1st = Upper, 2nd = Middle, 3rd = Lower.
- **Observations**:
  - Passengers in Pclass 1 had a high priority during rescue (survival rate ~75%).
  - Women in Pclass 1 had a survival rate of about 95-96%, while men had significantly lower survival rates.

4. Age

- Age is recorded as a continuous feature, with the oldest passenger at 80 years and the youngest at 0.42 years.
- **Observations**:
  - Children (age < 5) were saved in large numbers due to the "Women and Children First" policy.

    - Maximum deaths occurred in the age group of 30-40.
    - Survival chances for passengers aged 20-50 in Pclass 1 are high, especially for women.

5. Fare

- Passenger fare is another continuous feature.
- **Observations**:
    - Highest fare recorded: 512.33; lowest: 0.00; average fare: 32.20.
    - Significant distribution of fares among Pclass 1 passengers, decreasing with lower classes.
    - Consideration for binning to convert this continuous variable into discrete values.

6. SibSp

- Represents family relations (Siblings and Spouses).
- **Observations**:
    - Passengers without siblings had a survival rate of 35%.
    - Survival rate decreases with an increase in siblings.
    - Families with 5-8 members had a 0% survival rate.

7. Parch

- Represents family relations (Parents and Children).
- **Observations**:
    - Higher chances of survival for passengers with parents onboard (1-3 present).
    - Survival chances decrease for those with more than 4 parents.

Correlation Analysis

The heatmap analysis indicates that the features are not significantly correlated. The highest correlation (0.41) is between SibSp and Parch, suggesting some relationship, but not enough to confirm the removal of either feature. Hence, all features can be included in the model without concerns of multicollinearity.

Conclusion

From this analysis, I learned more about my data and how it is organized. I discovered how different factors, like gender, social class, and family connections, affected whether passengers survived. These insights showed me the variety in the dataset. For more details and visual examples of these findings, please refer to the main code.