Hafiz Muhammad Ibraheem
Roll No: [24i-7830]
Course: Data Science Tools and Techniques
Instructor: Dr. Muhammad Mateen Yaqoob

October 6, 2024

1 Part 1: Data Loading & Initial Exploration

In this phase, the dataset was loaded, and the initial records were displayed to understand its structure. Basic descriptive statistics such as mean, median, mode, and standard deviation were calculated for numerical variables like `Age` and `Fare`. These statistics provide insight into the central tendency and variability of the data.

For categorical variables such as `Pclass` and `Sex`, the counts and proportions of each category were summarized. This exploration is crucial as it sets the foundation for the subsequent analysis and highlights the need for further data preprocessing.

Visualizations, including histograms and box plots, were generated to provide graphical representations of the distributions and to identify potential outliers.

2 Part 2: Handling Missing Values

Missing values were identified using the `isnull()` method, revealing specific columns with a significant number of missing entries, particularly `Age`, `Cabin`, and `Embarked`.

For handling these missing values, several strategies were employed:

- For `Age`, mode imputation was chosen due to its categorical nature, providing a reasonable estimate without introducing bias.
- The `Cabin` variable had a high percentage of missing values and was handled using mode imputation to fill in the most common value.
- For `Embarked`, which contained only two missing values, mode imputation was again utilized.

This approach of filling missing values helps maintain the dataset's integrity while preventing a significant loss of data. Because by using removal method the rows reduced from 891 to 183. And that is too much of data loss

Visualizations, such as heatmaps and bar plots, were created to illustrate the distribution of missing values before and after handling them, confirming the effectiveness of the imputation strategies.

3 Part 3: Detecting & Handling Outliers

Outliers were detected using box plots and histograms for numerical features. Statistical methods such as the Z-score and Interquartile Range (IQR) were utilized to identify outliers. For instance, data points that fell beyond the threshold of 1.5 times the IQR were considered outliers and were subsequently capped to limit their influence on the analysis.

The decision to cap rather than remove outliers was based on the need to retain as much data as possible for accurate modeling while mitigating the risk of skewing results.

Visualizations including Z-score plots and distribution plots provided clarity on the outlier handling process, allowing for an understanding of the data distribution.

4 Part 4: Dealing with Duplicates

Duplicate records were checked using the `duplicated()` function. If duplicates were found, they were removed to ensure that each record in the dataset is unique. But no duplicates were found in the dataset.

5 Part 5: Encoding Categorical Variables

For ordinal categorical variables, such as `Sex and Cabin`, Label Encoding was employed. This method is appropriate as it maintains the inherent order among categories, enabling models to interpret the numerical values correctly. Conversely, for nominal categorical variables like`Embarked`, One-Hot Encoding was used to create binary columns for each category, avoiding any misleading implications of ordinal relationships.

The choice of encoding methods was justified based on the variable types, ensuring that the models could effectively interpret the data.

Visualizations were generated to illustrate the distribution of encoded variables, providing clarity on the transformation process.

6 Part 6: Data Distribution & Visualization

Histograms, box plots, and pair plots were plotted to visualize the distributions of numerical data. Bar plots and heatmaps illustrated relationships between categorical and numerical features, allowing for a comprehensive understanding of the data's structure.

Analysis of skewness was performed to comment on the shape of the distributions, and transformations such as log or square root transformations were applied where necessary to normalize the data.

7 Part 7: Feature Scaling

Normalization (Min-Max Scaling) was applied to the dataset to ensure that all features were on the same scale, which is crucial for algorithms that are sensitive to the scale of input data, such as k-means clustering and gradient descent.

Standardization (Z-score Scaling) was also applied where appropriate, particularly for algorithms that assume data is normally distributed. This method is preferred in cases where we want to ensure that the mean of the data is centered at 0 and has a standard deviation of 1.

Visualizations comparing data distributions before and after scaling were included to illustrate the effects of these techniques on the data.

8 Part 8: Final Dataset Overview

A summary of the changes made to the dataset post-preprocessing was provided, highlighting the actions taken to handle missing values, outliers, duplicates, and categorical variable encoding. The first few rows of the final cleaned dataset were displayed, demonstrating its readiness for modeling.