Data Understanding

Team Members: Apurwa Bhausaheb Sontakke, Vedant Wagh

Group Name - HealthData Innovators

Name – Apurwa Bhausaheb Sontakke

Email – sontakke.ap@northeastern.edu

Country - USA

College– Northeastern University

Specialization – Data Science

Github Link - https://github.com/apurwasontakke/Week-8-Data-Science-Healthcare-Project-Data-Understanding

Problem Description

Pharmaceutical companies face a significant challenge in understanding why patients continue or discontinue their prescribed medications. To address this, ABC Pharma has sought the help of an analytics company to automate the identification process of factors influencing drug persistency. The aim is to develop a classification model that predicts whether a patient will persist with a prescribed drug (Persistency_Flag).

Data Understanding

Data Type

The dataset consists of healthcare-related data containing 3424 entries and 69 columns. The data includes both numerical and categorical variables. Numerical variables are primarily related to frequency counts and risk scores, while categorical variables describe patient demographics, medical history, and treatment details.

Data Problems

1. Missing Values - Upon initial inspection and verification, the dataset was found to contain no missing values. All columns had complete entries for each of the 3424 records.
2. Outliers - Outliers were detected in some of the numerical columns, particularly in the Dexa_Freq_During_Rx column. These outliers could potentially skew the results and affect the overall analysis.
3. Skewness - The dataset exhibited skewness in certain numerical variables. For instance, the Dexa_Freq_During_Rx column showed a right-skewed distribution with a long tail on the right, indicating the presence of extreme values.

Approaches to Overcome Data Problems

- Handling Outliers

1. Z-score Method:

   Purpose: To identify and handle outliers by standardizing the data and identifying values that fall beyond a certain number of standard deviations from the mean.

   Method: Values with Z-scores greater than 3 were identified as outliers and were replaced with the mean value of the column.

   Rationale: This method is effective for normally distributed data and ensures that extreme values do not disproportionately affect the analysis.

2. IQR Method:

   Purpose: To identify and handle outliers by using the interquartile range, which is less sensitive to extreme values.

   Method: Values falling below Q1 - 1.5IQR or above Q3 + 1.5IQR were identified as outliers and replaced with the median value of the column.

   Rationale: This method is robust to skewed distributions and helps to minimize the impact of extreme values on the analysis.

- Reducing Skewness

  Log Transformation:

  Purpose: To reduce skewness in the data and make the distribution more symmetric.

  Method: A log transformation was applied to the Dexa_Freq_During_Rx column.

  Rationale: Log transformation reduces the influence of large values and helps in normalizing the distribution, making it more suitable for statistical analysis.

Conclusion

The data analysis process involved a thorough examination of the dataset to identify and address potential issues such as outliers and skewness. By applying the Z-score and IQR methods, outliers were effectively managed, and the log transformation was used to normalize skewed data. These preprocessing steps ensured that the dataset was cleaned and transformed appropriately, making it suitable for subsequent analysis and modeling efforts.