Project: Healthcare - Persistency of a drug

Week 9: Deliverables

Name: Krishna Ratna Deepika Haripuram

Email: haripuramdeepika@gmail.com

Country: Canada

Batch Code: LISUM26

Specialization: Data Science

Submission Date: Nov 29th, 2023

Submitted to: Data Glacier

Table of Contents

1. Problem Description:

This project focuses on analyzing the persistency of drug usage as prescribed by physicians in a healthcare dataset. The main goal is to identify factors influencing drug persistency and build a predictive model to classify patients based on their medication adherence (Persistency_Flag). This analysis is critical for pharmaceutical companies to enhance patient care and optimize treatment strategies.

2. Data Cleaning and Transformation:

a. **Handling Missing Values:**
   - Applied various techniques like mean, median, and mode imputation to handle missing values in different columns.
   - Experimented with K-Nearest Neighbors (KNN) imputation as a model-based approach for a comprehensive understanding of handling missing data in mixed-type datasets.

b. **Outlier Detection:**
   - Identified and handled outliers using methods like the Interquartile Range (IQR) for 'Dexa_Freq_During_Rx' and Z-score for 'Count_Of_Risks'.
   - These techniques helped in normalizing the data distribution and removing anomalies that could potentially skew the analysis.

c. **Weight of Evidence (WoE) Calculation:**
   - Computed WoE for categorical variables to transform them into a continuous scale and gain insights into the predictive power of each category.

3. NLP Featurization and Data Cleaning:

- Utilized regex for cleaning text data in columns such as 'Ntm_Speciality' and 'Risk_Segment_Prior_Ntm', removing non-alphabetic characters and standardizing the text.
- Applied TF-IDF Vectorization on the 'Ntm_Speciality' column to convert the cleaned text data into a numerical format. This transformation is crucial for feeding textual data into machine learning models.
- The sparse nature of the TF-IDF matrix (mostly zeros) is typical and reflects the uniqueness of terms in the documents.

4. Review and Reflections:

In this project, I navigated challenges such as choosing suitable data cleaning methods and handling the sparsity in the TF-IDF matrix. These experiences underscored the importance of understanding data at a deep level and maintaining data integrity through careful documentation. This project reinforced the value of adaptability and critical analysis in data science, lessons that I will apply in my future endeavors.

Code Repository and Documentation:

- All code and detailed documentation are maintained in the provided GitHub repository. Regular updates and comprehensive commenting in the code have been a priority to ensure clarity and reproducibility of the analysis.

Github Repo Link
https://github.com/krishnaharipuram/Data-Glacier/tree/main/Week%209