Team Name: Data Minders

Team Members:
Jacob Farrington; jtsfarrington@gmail.com , USA, University of North Texas, Data Science
Xiaoyan Zhang; xiaoyanhouston@gmail.com , USA, University of Texas at Dallas, Data Science
Marina Tsvetkova; marinatsv07@gmail.com , Poland, N/A , Data Science
Pamela S. D. Martey; pmlmartey@gmail.com, USA, Clark Atlanta University, Data Science

Problem description:

ABC is a pharmaceutical company keen on understanding persistency of a drug based on physician prescriptions for patients. To tackle this, ABC has engaged an analytics company to
automate the identification of persistency. This analytics firm has tasked Team Data Minders with creating an automated solution to evaluate and enhance the persistency of a drug for ABC.

Data Understanding:

The healthcare dataset comprises 69 columns with data types including 'object' and 'int64'. The
target variable, "Persistency_Flag," is binary.
Most of the columns contain binary data labeled "Y" and "N."
Upon analysis, it was determined that only a few columns are numerical, while the rest are string
data stored as 'object' type.

Exploratory Data Analysis:
No missing values ("N/A") were found in the dataset. Outliers identified in certain columns were
removed, decreasing the dataset size from 3,424 to 2,956 entries.
The dataset primarily consists of binary data, with entries marked as "Y" or "N."
A correlation analysis using the phik matrix exposed multicollinearity among some columns, which will require attention during subsequent model training stages.
What type of data you have got for analysis:

- Patient Information:
  - Ptid: Patient ID
  - Gender: Gender of the patient
  - Race: Race of the patient
  - Ethnicity: Ethnicity of the patient
  - Region: Geographic region
- Medical Data:
  - Persistency_Flag: Indicates whether the patient is persistent or non-persistent

- Age Bucket: Age range of the patient
- Ntm_Speciality: Specialty of the treating physician
- Ntm_Specialist_Flag: Indicates if the specialist is an NTM specialist
- Ntm_Speciality_Bucket: Bucketed specialty information
- Idn_Indicator: Indicator for IDN involvement
- Injectable_Experience_During_Rx: Experience with injectables during treatment

The dataset contains 69 columns, mainly with binary data labeled "Y" or "N," and data types include 'object' and 'int64'. The target variable, "Persistency_Flag," is binary.

What are the problems in the data (number of NA values, outliers , skewed etc):

Outliers:
• Summary statistics such as minimum, maximum, mean, and standard deviation show a potential presence of outliers in our data. Outliers will be further examined visually using box plots and histograms.
• Outliers were found and removed from the dataset, specifically in columns such as 'dexa_freq_during_rx' and 'count_of_risks', reducing the dataset size from 3,424 to 2,956 entries.
Skewness:
• Notable skewness in columns like Dexa_Freq_During_Rx and Count_Of_Risks.
NA Values: There are no NA values in the dataset.
What aproaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?:
We will approach skewness using log or square root transformations, and we will approach outliers via capping in order to maintain data integrity. Outliers in the columns 'dexa_freq_during_rx' and 'count_of_risks' were identified and removed to improve the reliability
of the data for accurate analysis and modeling.
Data Handling: With no NA values present, there was no need for strategies related to missing
data.

Data cleaning and transformation done on the data.

In the healthcare dataset, numerical columns were standardized using StandardScaler to normalize the data distribution, which is crucial for models sensitive to the scale of input features. Binary data represented as 'Y' and 'N' were converted to 0 and 1, and the target variable 'persistency_flag' was encoded from 'Non-Persistent'/'Persistent' to 0/1. Categorical columns were transformed using OneHotEncoder with the drop='first' option to prevent multicollinearity by creating binary columns for each category. Outliers in 'dexa_freq_during_rx' and 'count_of_risks' were removed based on their quantiles to eliminate extreme values that could skew the analysis. Duplicate entries were identified and removed to ensure data uniqueness. Risk-related columns were converted from 'Yes'/'No' to 1/0, and an aggregate risk score was calculated by summing these values, providing a summarized risk level for each record. The dataset was also grouped by various categories

and analyzed to determine the counts of persistent and non-persistent cases per category, offering deeper insights into the data. This series of transformations enhances the dataset's integrity and usability for accurate modeling.

Final Recommendation

Based on the exploratory data analysis (EDA) results and the insights from the Phik matrix, it is planned to develop robust predictive models with special attention to addressing multicollinearity and class imbalance. Multicollinearity, identified using the Phik matrix, is intended to be mitigated through methods such as Principal Component Analysis (PCA) and regularization (Lasso and Ridge), which will help adjust coefficients and enhance model stability. To address class imbalance, techniques such as SMOTE will be applied to increase the representation of the minority class or adjust class weights algorithmically within the model. Algorithms suitable for binary classification, such as logistic regression, decision trees, and ensemble methods like random forests and gradient boosting machines, are planned due to their reduced sensitivity to multicollinearity and effectiveness in handling class imbalance. To fine-tune model parameters and assess their effectiveness, cross-validation, and performance metrics such as precision, recall, and AUC-ROC are intended to be used. These measures will ensure that the predictive modeling efforts are based on sound statistical principles and best practices.

Github Link:

https://github.com/S0n0f1saac/LISUM33-GroupProject/Week_10