Name: Data Science :: Healthcare - Persistency of a drug:: Group Project

Report date: 24th April 2024

Internship Batch: LISUM30:30 Jan24 - 30 Apr 24

Version: 1.0

Data intake by: David Paul Nalumenya

Group name: None

Specialization: Data science

Email: nalumenyad@gmail.com

Company: Makerere University

Data intake reviewer: None

Data storage location:

1.0 Problem description

In this project, we're tackling the challenge of predicting patient drug persistency using machine learning. The data contains 3424 observations and 69 features, with the majority being categorical (67) and stored as object data types. This is important because machine learning models for classification tasks typically can't handle object data directly. To address this, we'll employ a process called feature engineering, which will involve imputing the categorical variables into numerical representations. This is a common practice in data science, with around 80% of data scientists spending more time cleaning and preparing data than actually analyzing it. There are various techniques for encoding categorical features, and the choice of method can significantly impact the model's performance. By carefully transforming our data, we can leverage the power of machine learning to gain valuable insights into patient behavior and medication adherence, a critical area for pharmaceutical companies where studies suggest less than half of patients adhere to their medication regimens.

2.0 Data cleansing and transformation

Data preparation process involved several key steps to ensure the quality and suitability of the data for building a machine learning model that analyzes drug persistency.

1. **Feature Selection**

   We removed the PTID (Patient ID) column because it lacks statistical significance for the specific task of analyzing drug persistency. This decision reflects the concept of **feature importance**, which emphasizes selecting features that are most relevant to the prediction target (drug persistency in this case). Excluding irrelevant features can improve model performance and interpretability.
2. **Missing Value Handling**

   We ensured the absence of missing values (NaN) to avoid introducing bias during model training. Missing values can create uncertainty during model training, potentially leading to inaccurate predictions. Common techniques for handling missing values include imputation (filling in missing values with appropriate strategies) or removal of observations with missing values, depending on the specific context and the amount of missing data. We used the .isnull().sum() function to identify the presence and distribution of missing values.
3. **Outlier Detection**

   Outlier detection is crucial because extreme values can skew model training and potentially lead to unreliable results. We checked for outliers in the integer-type data columns using boxplots. Fortunately, no outliers were identified in this instance. However, depending on the data and modeling goals, various methods can be used to address outliers, such as winsorization (capping extreme values) or removal (if justified).
4. **Feature Engineering**

   Label Encoding

   We addressed the predominantly object-type features (categorical data) as machine learning models typically require numerical features for computations. We employed label encoding, which creates a mapping function to transform categorical levels within each feature (e.g., demographics, risk factors, specialties, comorbidities) into numerical representations. This process preserves some order within the categorical variables (e.g., assigning higher numerical values to categories that might indicate higher risk) while enabling the model to learn relationships between these features and drug persistency. It's important to note that label encoding assumes an inherent order within the categories, which might not always be the case. Alternative encoding techniques like one-hot encoding can be considered if the categories have no intrinsic order.

5. **Data Validation**

   Finally, we validated the imputed numerical variables through inspection and visualization techniques to ensure consistency and identify any potential anomalies before proceeding with model building. Visualizations like histograms and scatter plots can be helpful for exploratory data analysis and identifying potential issues with the transformed data. This step emphasizes the importance of **data quality assurance** before feeding data into a machine learning model.
6. Data normalization and standardization of the outcome variables so that all variables lie on the same scale.