Explanation Outcome

Persistency of a Drug

Name: Jorge Alberto Jaramillo Bermúdez
Email: jorge_jb1@hotmail.com
Country: México
Specialization: Health Data Scientist
18 de enero de 2024

Dataset Overview:

The dataset comprises 2 discrete value features, 1 continuous value feature, and 66 categorical value columns, including the target variable, which is the consistency flag.

Data Quality Issues:

The dataset exhibits missing values, particularly in the following features:

- **'Race'**
- **'Ethnicity'**
- **'Region'**
- **'Ntm_Speciality'**
- **'Risk_Segment_During_Rx'**
- **'Tscore_Bucket_During_Rx'**
- **'Change_T_Score'**
- **'Change_Risk_Segment'**

Handling Missing Values:

Addressing missing values is crucial for robust analysis. Several strategies can be employed, considering the nature of the dataset:

- Drop Missing Values:
  - While straightforward, this approach may result in significant data loss and is not always the most efficient.
- Missing Value as a Label:
  - Assigning missing values a distinct label can be considered, although this might not be the optimal strategy.
- Fill with Most Popular Label:

- Imputing missing data with the most frequently occurring label in a given feature is a simple strategy but may not capture the inherent complexity of the data.

- Machine Learning-Based Imputation:
  - Leveraging machine learning algorithms to predict and fill missing values can offer a more sophisticated approach.

Adopted Approach:

In my analysis, the data visualization underscores a distinct pattern among patients labeled as "Persistent" in the ['Persistency Flag'] column, predominantly identifying as females. Notably, a significant proportion of these female patients falls within the age group exceeding 75 years, accounting for nearly 1250 cases. Subsequently, approximately 1000 cases fall within the age range between 65 and 75 years. In the 55 to 65 age group, there are around 700 cases, while the count drops to less than 200 in the age group younger than 55.

Delving into race-related insights, the data reveals that the Caucasian race exhibits the highest persistence, showcasing a substantial number of persistent cases within this demographic. Furthermore, the ethnicity with the highest persistence is non-Hispanic, suggesting a heightened propensity for persistence in this population.

Geographical analysis indicates a notable prevalence of persistent cases in the Midwest region, signifying a higher incidence of persistence compared to other regions.

Analysis of High Cardinality in 'Ntm_Speciality':

The elevated cardinality in 'Ntm_Speciality' points to a diverse range of medical specialties among those prescribing the medication. Addressing this diversity is paramount for effective analysis and modeling in the context of drug persistence.

Upon examination, it becomes evident that general practitioners encounter this medication most frequently. Following closely is the specialty of rheumatology, where specialists focus on diseases of the musculoskeletal system and connective tissue, some of which can impact bone health. Rheumatologists' interest in T-score and bone densitometry stems from their association with osteoporosis, characterized by a decrease in bone density.

Persistency_Flag in Binary Format for Model Training:

In the data preparation process for our machine learning model, we opted to transform the 'Persistency_Flag' variable into a binary format. This common practice in classification problems enhances model interpretation and training efficiency.

The original 'Persistency_Flag' variable with two categories, "Persistent" and "Non-Persistent," has been assigned binary numeric values as follows:

- Persistent: 1
- Non-Persistent: 0

This approach simplifies model interpretation and aids in calculating performance metrics like accuracy, recall, and area under the ROC curve (ROC-AUC).

Moreover, a consistent encoding has been applied to other variables, such as Ptid, using:

- Y: 1
- N: 0

This uniform encoding standardizes the handling of binary categorical variables in our dataset, contributing to a more consistent and interpretable representation for the machine learning model.

These transformations are integral to our data preparation efforts, establishing a robust foundation for the training and evaluation of classification models in our project.