Group Project Deliverable Week 10

Project Name: Healthcare - Persistency of a drug
Report date: 26.09.2023
Internship Batch: LISUM24
Group: HealthData Collective
Data storage location: https://github.com/venusflytrapfairy/HealthDataCollective-Group

- **Team member's details**

Group: HealthData Collective
Group members:

Name	Email	Country	College/Company	Specialization
Selena	selena.elensto@gmail.com	Macedonia	-/-	Data Science
Anusha Asim	anushaasim21@yahoo.com	UAE	Regent Middle East	Data Science
Sukurat	salamsukurat@gmail.com	United Kingdom	Teesside University	Data Science
Khushboo	Khushboomasih193@gmail.com	United Kingdom	University of Aberdeen	Data Science

- **Problem description**

The problem to be solved is related to pharmaceutical companies' need to understand the persistence of drug usage among patients based on physician prescriptions. Specifically, the challenge is to automate the process of identifying whether a patient is persistent in following their prescribed therapy regimen. The primary problem is to build a classification model using machine learning to determine the persistence of patients based on various features and factors.

- **EDA performed on the data**

Our exploratory data analysis covered several key steps including data loading, data profiling, data cleaning, and feature engineering.

Here's a summary of the steps and tasks performed:

- *Data Loading:* We loaded the dataset from an Excel file using the Pandas library.

- *Data Profiling:* The explanations for each feature in the dataset was provided to help understand the meaning and relevance of each variable.

- *Data Cleaning:* We identified and handled the presence of "Unknown" values in the dataset, which can be crucial for data preprocessing.

- *Deduplication:* The potential duplicate records in the dataset were identified and handled.

- *Data Visualization:* We generated a profile report using the Pandas Profiling library to gain insights into the dataset's statistics and distributions.

- *Label Encoding:* The non-numerical features were encoded, including categorical and boolean features, in preparation for modeling.

- *Model Development:* We performed basic linear regression modeling and generated a summary using the statsmodels library.

- *Train-Test Split:* We split the dataset into training and testing sets.

- *Dummy Classifier:* A Dummy Classifier to create a baseline model and evaluate its performance on both the original and encoded datasets.

- *Oversampling:* We applied oversampling techniques such as SMOTE, ADASYN, SMOTETomek, and SMOTEENN to address the issue of imbalanced target classes.

- *Results Summary:* The results of each oversampling technique were collected and summarized, including the shape of the oversampled data and the accuracy score of a Dummy Classifier.

- *Comparison with Original Dataset:* We've included a row in the summary to compare results with the original dataset without oversampling.

This EDA serves as the foundation for subsequent modeling and decision-making in the context of medication persistence.

- **Final Recommendations**

In our pursuit of understanding medication persistence based on physician prescriptions, we recommend the following machine learning models tailored to the dataset. Our dataset encompasses a comprehensive set of features, including patient demographics, clinical factors, provider attributes, disease/treatment factors, and more. The target variable of interest is 'Persistency_Flag,' indicating whether a patient is persistent with their prescribed therapy.

- *Random Forest:* Random Forest is a versatile ensemble learning method that can be particularly advantageous for our problem. It excels in handling datasets with a mix of numerical and categorical variables, making it a strong candidate for our dataset, which includes patient attributes like age, gender, race, and physician specialty. Random Forest can be utilized to uncover patterns and interactions among various patient attributes and their effect on medication persistence. By examining feature importances, we can identify the most influential factors, aiding in a better understanding of persistence determinants.

- *Support Vector Machine (SVM):* Support Vector Machine is a robust classification algorithm known for its effectiveness in handling binary classification tasks. Given that

our primary focus is on distinguishing between persistent and non-persistent patients, SVM can be a suitable choice. SVM can be applied to establish an optimal decision boundary that segregates patients into these two categories. Its ability to handle both linear and nonlinear data can help capture complex relationships within our dataset, such as the impact of physician specialty and clinical factors on medication persistence.

- *Naive Bayes:* Naive Bayes, a probabilistic algorithm, is particularly useful when dealing with categorical features and feature independence assumptions. In our dataset, features like race, ethnicity, and physician specialty are categorical, aligning with Naive Bayes' strengths. Naive Bayes can be employed to model the likelihood of a patient being non-persistent given their attributes. By considering the independence assumption, it helps uncover relationships between categorical variables and medication persistence, enabling us to draw insights from the data.

- *XG Boost (Extreme Gradient Boosting):* XG Boost is a high-performance gradient boosting algorithm designed to handle complex datasets efficiently. Our dataset, with a mix of categorical and numerical variables, can benefit from XG Boost's capabilities. XG Boost can model intricate interactions between patient attributes and medication persistence, offering valuable insights. It is particularly adept at dealing with imbalanced datasets, a common scenario in medical datasets. Through feature importance analysis, we can identify the most influential factors and understand their impact on persistence.

- *Decision Trees:* Decision Trees are simple yet effective models that create a flowchart-like structure to make decisions. They can be used to understand the hierarchy of factors affecting medication persistence. Decision Trees can reveal the most critical features and their thresholds that influence patient persistence. They provide a visual representation of the decision-making process, aiding in the interpretability of the model.

These models represent a selection tailored to our dataset's characteristics and the specifics of our problem statement. The next step is to determine which one best aligns with the objective of understanding and predicting medication persistence.