Ali-Just Team Members

Name: Alireza Samadifardheris
Email: alirezasamadii71@gmail.com
Country: Rome, Italy
College: Sapienza University of Rome, Computer Engineering
Specialization: Data Science

Name: Justin Lee
Email: justindavinlee@gmail.com
Country: Ontario, Canada
College: Wilfrid Laurier University, Data Science Concentration Big Data
Specialization: Data Science

Link to GitHub Repository:
https://github.com/justindavin/DataGlacierInternship/tree/main/Week9

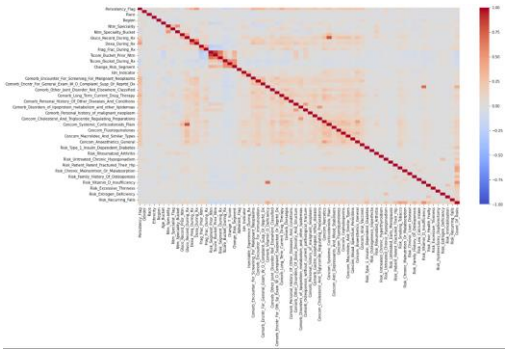
Problem Description

One of the challenges for all Pharmaceutical companies is to understand the persistence of drugs as per the physician's prescription. the persistency of a drug may be defined as "the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen." Medication persistence refers to the act of continuing the treatment for the prescribed duration.

Exploratory Data Analysis

After performing extensive exploratory data analysis on this dataset, we have found some interesting information about the data.

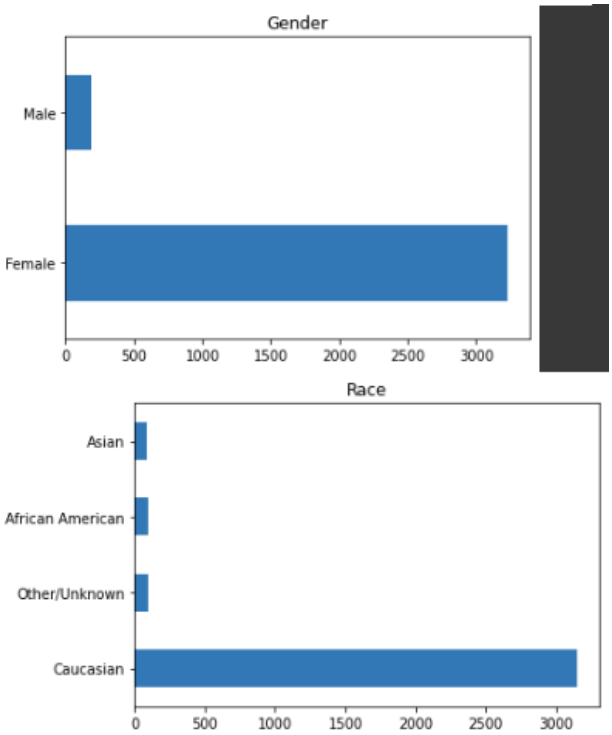
After looking at the correlation of the data, it seems that our target column has correlation with some features.



We have also discovered that there is skewness in the data. To solve the issue of skewness, we will remove skewness from features with low correlation, as removing skew from data with

higher correlation will have a higher impact and could be problematic. To remove the skewness of features with low correlation, we have first calculated the skewness of each feature. Then, we compared the values to a certain threshold of 0.01. If the correlation of the feature is below the threshold, then we remove the skewness from the features by square rooting the skewed column.

We have also discovered that the dataset is unbalanced. There seems to have a 1:15 to 1:30 ratio of unbalanced data in many of the features included within the dataset.



To deal with the problem of the unbalanced data, we will perform sampling techniques such as: RandomUnderSampler and SMOTE (Synthetic Minority Oversampling Technique) which will be performed in the model building portion of the project.

Final Recommendation

After performing extensive exploratory data analysis, data cleaning, and data transformations on the dataset, we have found that the dataset is unbalanced, and highly skewed. While there are many features included within the dataset, we will first test out models using all the included features to test out the base performance of the model. We will then fine tune the models my performing some sampling techniques to test the model on a more balanced dataset. To finalize the model, only features that are important for the model's predictive power will be used to test if the performance of the model will improve. Some machine learning models that will be used in this project are the GaussianNB, Logistic Regression, Support Vector Machines (SVM), Linear Support Vector Classification (LSVC), Perceptron, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier, Stochastic Gradient Descent (SGD) Classifier, as well as Gradient Boosting Classifier. All of these machine learning models will be evaluated using the accuracy metric, the F1-Score, as well as the confusion matrix. With the confusion matrix, the precision and recall can also be calculated.

Data Intake Report

Name: Healthcare Industry
Report date: December 2nd, 2022
Internship Batch: LISUM14
Version: 0.3
Data intake by: Alireza Samadifardheris and Justin Lee
Data intake reviewer: Alireza Samadifardheris and Justin Lee
Data storage location:
https://docs.google.com/spreadsheets/d/1P_oMc6gOBlhw6dY5PxaqxV2swdHMUooK/edit#gid=2047360270

Tabular data details:

Total number of observations	3424
Total number of files	
Total number of features	69
Base format of the file	.xlsx
Size of the data	900 KB

Note: Convert this doc in pdf and provide the link of pdf file in your dashboard.
Please do not forget to remove this section while converting the file into pdf.