Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Healthcare – Persistency of a Drug

December 30th, 2022

Group Members

Alireza Samadifardheris

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

Justin Lee

- Email: justindavinlee@gmail.com
- Country: Ontario, Canada
- College: Wilfrid Laurier University, Data Science
- Specialization: Data Science

Data Glacier
Your Deep Learning Partner

Agenda

Executive Summary

Problem Statement

Approach

EDA

EDA Summary

Recommendations

model selection

Executive Summary

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.

With an objective to gather insights on the factors that are impacting the persistence, built a classification model for the given dataset.

Problem Statement

One of the challenges for all Pharmaceutical companies is to understand the persistence of drugs as per the physician's prescription.

To solve this problem, we will perform exploratory data analysis and build classification models to help gather insightful information regarding the topic.

Approach

Our team approached this problem by first performing exploratory data analysis on the given dataset to discover patterns and spot anomalies that may affect the model's predictions.

We will then tidy the dataset by filling or removing the anomalies and perform some feature engineering to have a tidy dataset ready to train our proposed models.

Finally, we will use the dataset to train and test our models, while evaluating them using accuracy and the F1-score to determine the most accurate model.

EDA

We have broken down the exploratory data analysis into multiple steps:

- Find missing values
- Check for skewness
- Look at correlation
- Check for unbalanced data

EDA Summary - Missing Values

There are no missing values present in the given dataset, therefore we have moved on from this step after checking for missing values

EDA Summary - Skewness

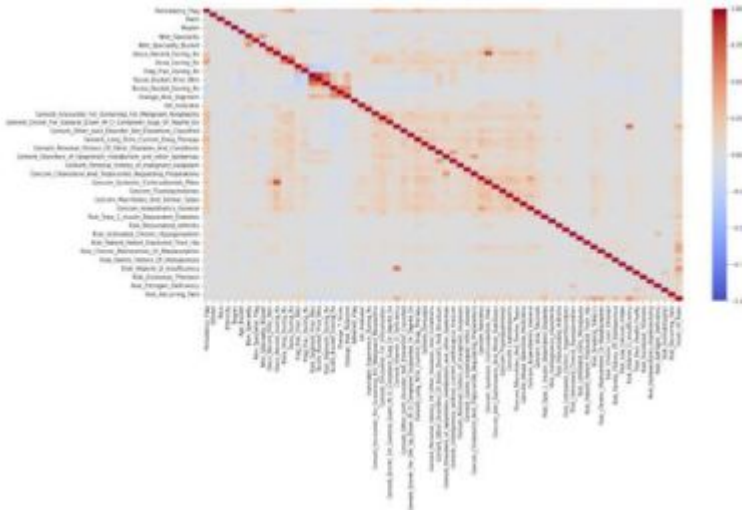
There is skewness present in the data. To solve this issue, we will remove skewness from features with low correlation by standardizing the values.

We 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, we standardize the feature.

We will not remove skewness from features with high correlation, as removing the skew will have a higher impact and could be problematic.

EDA Summary - Correlation

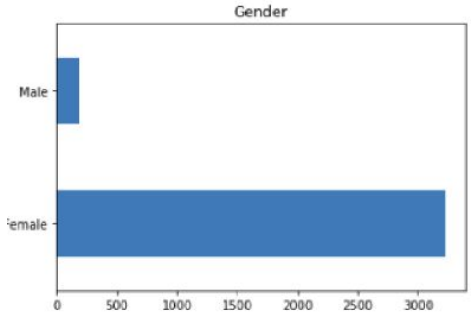
After looking at the correlation of the data, it seems that our target column has correlation with some features.



EDA Summary - Unbalanced Data

The data seems to be heavily unbalanced, with a ratio of up to 1:30 of unbalanced data present in the features.

To deal with this problem of the unbalanced data, we will perform sampling techniques such as the Random Under Sampler and the Synthetic Minority Oversampling Technique (SMOTE) to have a more balanced dataset.



Data Glacier
Your Deep Learning Partner

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.

model selection

The machine learning models that were 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 these machine learning models were 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.

comparing all the models we conclude that it is a good option to proceed with the model created with Support vector machines with SMOTE with the highest score 83.29

```
accuracy: 83.29
F1-Score: 76.92
Confusion matrix:
 [[418  58]
 [ 68 210]]
```

Data Glacier
Your Deep Learning Partner

Thank You