Data Science Internship

Final Project: Predicting the persistency of a drug (Healthcare)

Week 8 deliverables

Group name: Gold Standard Team
Name: Mario Rodrigues Peres
Email: mariorope@hotmail.com
Country: Brazil
Specialization: Data Science
Report date: 26th of April 2023
Internship Batch: LISUM19
Project home: GitHub

Contents:

1. Problem Description

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

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

2. Project Timeline

Activity	Section	Deadline
Problem description, project timeline, data intake report and GitHub repository link	Week 7	19 Apr 2023
Understanding the data and checking for problems	Week 8	26 Apr 2023
Data cleansing and transformation	Week 9	2 May 2023
Exploratory data analysis (EDA) and recommendations	Week 10	9 May 2023
EDA presentation and proposed modeling technique	Week 11	16 May 2023
Model selection and model building	Week 12	23 May 2023
Final project report and code	Week 13	30 May 2023

3. Data Understanding

The dataset presents 3424 registers and 69 features. The features include the target, 2 numerical and 66 categorical variables. There is no duplicated register nor missing values.

The 'Persistency_Flag' variable is the target variable and all the others are predictors. This variable presents two categories: Persistent and Non-Persistent. There are more registers for the Non-Persistent category (62.4%).

A chi-square test was applied to check the correlation of categorical variables and the target, and it was observed that some variables are not well correlated with the target, considering a significance level of 0.05.

4. Dataset Problems

4.1. Problems Identification

As mentioned in the previous items, the dataset does not present missing values, but a few problems were identified during this preliminary exploratory analysis as follow:

• Skewness and outliers

Both numerical variables (‘Count_Of_Risks’ and ‘Dexa_Freq_During_Rx’) seems to be strongly correlated with the target, however, these variables are skewed and present outliers that need to be treated before proceeding with the modeling steps.

The 'Dexa_Freq_During_Rx' presents 116 and 693 outliers, whereas the ‘Count_Of_Risks’ presents 4 and 54 outliers considering the Non-Persistent and Persistent groups, respectively.

• High proportion of unknown values

Four variables contain a great percentage of unknown values, but we don’t have very clear information about it and how to overcome the problem for now until further analysis is done.

• Variables representing the same information

A few predictor variables seem to represent the same information, For example, the variables ‘Ntm_Speciality’, ‘Ntm_Specialist_Flag’ and ‘Ntm_Speciality_Bucket’ represent the same information and are highly correlated to each other. Therefore, correlated features will also be treated before the modeling steps.

• Unbalanced dataset

As mentioned before, the target variable (‘Persistency_Flag’) presents more registers for the class ‘Non-Persistent’, and therefore, the data should be balanced prior modeling.

4.2. Problems Mitigation

In order to overcome the identified problems, we will use some different approaches depending on the results of further analysis of the dataset.

In the case of outliers treatment, these registers may be removed from the dataset and the data normalized to overcome the skewness problem. However, further correlation analysis will enable us to decide our next course of action. If categorical predictors are strongly correlated with the numerical variables, the latter ones may be removed from the dataset.

In the case of the high proportion of unknown values (4 variables), these may be also removed if they are well correlated with other predictors. Otherwise, we will have to further investigate their inclusion in the modeling steps.

In order to find out which predictors are correlated to each other, we will perform correlation tests, and remove any variable that represents repeated information. Another approaches to support the features selection will also be used during the modeling process.

Finally, we are planning to use either SMOTE oversampling or Cost-Effective learning(penalizing the training model for wrong predictions) techniques but we have not decided yet. We needed to perform further analysis to select the most adequate method to solve the unbalanced problem.