Predicting Persistency of a Drug (Group Project

DATA-SCIENCE INTERNSHIP

DATA GLACIER LISUM 19

Table of Contents

1. Problem Description
2. Data Understanding
3. Dataset Problems

    a. Problem Identification

    b. Problem Mitigation

4. Project life-cycle along with deadline

GROUP NAME: GOLD STANDARD TEAM

NAME: Alexis Dymphina Michael-Igbokwe

EMAIL: alexis.phina@gmail.com

COUNTRY: Ireland

SPECIALIZATION: DATA-SCIENCE

INTERNSHIP BATCH: LISUM19

REPORT DATE: 26TH APR 2023

PROBLEM DESCRIPTION

Pharmaceutical companies are going through the challenge of understanding the persistency of drug as per the physician prescription. ABC company wants us to automate the process of identification for them.

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.

DATASET PROBLEM

PROBLEM 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

1. 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.

2. 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.

3. 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.

4. 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.

PROBLEM 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(penaulysing the training model for wrong predictions) techniques but still not yet decided and needed to do some more analysis to optthe method to solve unbalanced dataset problem.

PROJECT LIFE-CYCLE

ACTIVITY	WEEK	TIME FRAME
Problem description, data intake report, project timeline and Github link	WEEK 7	19 APR 2023
Understanding the data and checking for problems	WEEK 8	26 APR 2023
Data cleansing and transformation	WEEK 9	2ND MAY 2023
Exploratory data analysis (EDA) and recommendations	WEEK 10	9TH MAY 2023
EDA presentation and proposed modelling technique	WEEK 11	16TH MAY 2023
Model selection and model building	WEEK 12	23RD MAY 2023
Final project report and code	WEEK 13	30TH MAY 2023