Data Science Internship

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

Week 12 deliverables

Group name: Gold Standard Team
Names:
1) Harshith Sakala Santhosh
2) Mario Rodriques Peres
3) Alexis Michael-Igbokwe
Email: mariorope@hotmail.com
Country: Brazil
Specialization: Data Science
Report date: 2nd of May 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%).

4. Feature Selection and Modeling

4.1. Feature Selection

Several techniques were used for feature selection including the correlation analysis, checking the importance of variables based on Random Forest and the tool SelectKBest from the Sklearn package. A few different datasets were tested using the Random Forest algorithm in order to support the selection of the final features used for the development of the model. The final predictors included in the model with higher performance (presented in the following section) are as follow:

- ‘Ntm_Speciality_Bucket'
- ‘Dexa_During_Rx'
- ‘Tscore_Bucket_During_Rx'
- ‘Risk_Segment_During_Rx'
- ‘Gluco_Record_During_Rx'
- ‘Frag_Frac_Prior_Ntm','Adherent_Flag'
- ‘Risk_Rheumatoid_Arthritis'
- ‘Risk_Untreated_Chronic_Hypogonadism'
- ‘Risk_Smoking_Tobacco'
- ‘Risk_Excessive_Thinness'
- ‘Risk_Immobilization'
- ‘Comorb_Long_Term_Current_Drug_Therapy'
- ‘Concom_Systemic_Corticosteroids_Plain'

4.2. Model Development and Selection

A few models were tested during this stage. There were tested the Linear Regression, Random Forest, SVM, ADABoost, XGBoost and Dense Neural Network. Each member of the team tried different approaches and a single Jupiter file was prepared, which included all the Jupiter notebooks produced for each member of the team.

A few models presented a good performance, such as ADABoost (accuracy of 79%) and XGBoost (accuracy of 80%), However, the model that presented best performance was the Dense Neural Network, which achieve an accuracy of about 80% in training and 82% in test. The model presented a precision score of 76.7%, recall of 75.5% and F1-score of 76.1%.