Data Science Internship

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

Week 10 deliverables

Group name: Gold Standard Team
Name: Harshith Sakala Santhosh
Email:sakala.harshith@gmail.com
Country: United Kingdom
Specialization: Data Science
Report date: 9nd 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. Exploratory Data Analysis

4.1. Feature Description

In the given dataset, there are 69 features out of which 66 are categorical variables. The Description of these variables as follows:

Gender: The great majority of values are from female patients.

Race: The great majority of values are from Caucasian patients.

Ethnicity: The great majority of values are from Not Hispanic patients.

Region: There were observed more patients classified as Non-Persistent from the Midwest, followed by South, West and Northeast, in descending order. However, when considering patients classified as Persistent, most of patients were from the South, followed by Midwest, West and Northeast, in descending order.

Age_Bucket: Most of patients were older than 75 years old, decreasing the number of patients with the decrease of patient's age.

Ntm_Speciality : This variable presents too many classes (36), and about half of the values are classified in the class General Practitioner (1535), followed by Rheumatology (604) and Endocrinology (458). This variable also presents 310 unkown registers.

Ntm_Specialist_Flag : About 50% of patients classified as Persistent were attended by specialists, whereas only 35% of patients classified as Non-Persistent were attended by specialists. This variable might be correlated with the Ntm_Speciality as they bring the same information categorized in two different ways (more specific or broad).

Ntm_Speciality_Bucket : This variable presents another classification of specialities and might also be correlated with the previous two variables.

Gluco_Record_Prior_Ntm : Independently of the target class, most of patients did not present Gluco Record prior Ntm.

Gluco_Record_During_Ntm: The proportion of patients with Gluco record during RX was lower compared to the proportion of patients without Gluco record prior RX for patients classified as Non-Persistent. On the other hand, for patients classified as Persistent, there was a higher proportion of patients with Gluco record during RX compared to Gluco record prior RX.

Dexa_During_Rx : The great majority of patients classified as Non-Persistent did not have Dexa during RX, however for patients classified as Persistent, there were more patients that had Dexa during RX compared to patients that did not have Dexa during RX. This variable may be very important for the modeling steps.

Frag_Frac_Before_Ntm : Independently of the target classes, most of patients did not have Frag_Frac_Before_Ntm.

Frag_Frac_During_Ntm: Similar behaviour of previous variable.

Risk_Segment_Prior_Ntm: There was observed a slightly higher number of patients with VLR-LR compared to patients with HR-VHR for both classes of the target variable.

TScore_Bucket_Prior_Ntm: Similar behaviour of previous variable.

Risk_Segment_During_RX: There were observed very similar counts for patients with VLR-LR or HR-VHR during RX. However, there were a considerable number of unkown values, especially for the class Non-Persistent.

TScore_Bucket_During_Ntm: This variable also presented similar count of patients with <=2.5 or >2.5 for the TScore_bucket_During_Ntm. But, similarly to the previous variable, there were observed a considerable number of unkown values

Change_TScore: This variable presents 4 categories. Considering the patients classified as Non-Persistent, most values are unkown. Within the known values, most patients presented no change in the TScore change, followed by patients that presented worse results and patients that presented some kind of improvement. For the patients classified as Persistent, most values were observed for patients that did not present any change in the TScore, followed by patients with unkown results. There are a lot of unkown results for this variable.

Change_Risk_Segment: Most of the values are unkown for both target classes. Within the known values, most of them were for patients that did not present any change in the Risk Segment, followed by patients that worsen the Change_Risk_Segment. Finally, the category for improved patients was the one that presented less patients.

Note: It is important to note that the variables related to Risk_Segment as well as to TScore might be correlated and present colinearity problems

Adherent_Flag: Most of patients presented Adherent flag. Moreover, it was observed that there was a higher proportion of non-adherent flag among patients that were classified as Persistent.

Idn_Indicator: Most patients from both target classes presented higher positive results. Furthermore, there was observed a higher proportion of positive results among patients classified as Persistent.

Injectable_Experience_During_RX: Most of patients from both taget classes presented a much higher number of positive results.

Comorb_Encounter_For_Screening_For_Malignant_Neoplasms :This variable presents some very interesting results. Among patients classified as Non-Persistent, most of them presented negative results for this variable (approximately double). On the other hand, for patients that were classified as Persistent, there were observed more patients with positive result for this variable (approximately double).

Comorb_Encounter_For_Immunization: A very similar pattern with the previous variable was observed.

Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx: Again, it was observed a higher number of negative results for patients classified as Non-Persistent. But, there were more patients with positive results among the ones classified as Persistent.

Comorb_Vitamin_D_Deficiency: This variable presented higher number of patients with negative results for both target classes. However, when comparing the patients classified as Non-Persistent and Persistent, it was observed a higher proportion of negative results for the Non-Persistent ones.

Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified: Similar behaviour as the previous variable.

28)Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx: Similar behaviour as the previous variable.

Comorb_Long_Term_Current_Drug_Therapy: Similar behaviour as the previous variable.

Comorb_Dorsalgia : Similar behaviour as the previous variable.

Comorb_Personal_History_Of_Other_Diseases_And_Conditions: Similar behaviour as the previous variable.

Comorb_Other_Disorders_Of_Bone_Density_And_Structure: Similar behaviour as the previous variable.

Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias: This variable presented higher number of negative values for the Non-Persistent patients but higher number of positive values for the Persistent patients. This variable may be important for the modeling steps.

Comorb_Osteoporosis_without_current_pathological_fracture: Similar behaviour as the variable Comorb_Other_Disorders_Of_Bone_Density_And_Structure.

Comorb_Personal_history_of_malignant_neoplasm: Similar behaviour as the previous variable.

Comorb_Gastro_esophageal_reflux_disease: Similar behaviour as the previous variable.

Concom_Cholesterol_And_Triglyceride_Regulating_Preparations: There are more negative values for both target classes, however, there is a higher proportion of negative values for the Non-Persistent class.

Concom_Narcotics: Similar behaviour as the previous variable.

Concom_Systemic_Corticosteroids_Plain: Similar behaviour as the previous variable.

Concom_Anti_Depressants_And_Mood_Stabilisers: Similar behaviour as the previous variable.

Concom_Fluoroquinolones : Similar behaviour as the previous variable.

Concom_Cephalosporins: Similar behaviour as the previous variable.

Concom_Macrolides_And_Similar_Types: Similar behaviour as the previous variable.

Concom_Broad_Spectrum_Penicillins: Similar behaviour as the previous variable.

Concom_Anaesthetics_General: Similar behaviour as the previous variable.

Concom_Viral_Vaccines: Similar behaviour as the previous variable, however with more extreme proportions.

Risk_Type_1_Insulin_Dependent_Diabetes :There were much more negative results for both target classes, and the proportion does not seems to be very discrepant.

Risk_Osteogenesis_Imperfecta: This variable presents only 3 positive results and will be removed for the modeling steps.

Risk_Rheumatoid_Arthritis : There were higher number of negative results for both target classes. There was a slight higher number of patients with negative results in the class Non-Persistents.

Risk_Untreated_Chronic_Hyperthyroidism :This variable presents only 2 positive results and will be removed for the modeling steps.

Risk_Untreated_Chronic_Hypogonadism: Similar behaviour as the variable Risk_Rheumatoid_Arthritis

Risk_Untreated_Early_Menopause :This variable presents only 12 positive results and will be removed for the modeling steps.

Risk_Patient_Parent_Fractured_Their_Hip: This variable presented much more negative results for both target classes and the proportion seems to very similar.

Risk_Smoking_Tobacco: There were observed more negative results for both target classes and it seems that there were a slightly higher proportion of negative results for the Non-Persistent patients.

Risk_Chronic_Malnutrition_Or_Malabsorption: Similar behaviour as the previous variable.

Risk_Chronic_Liver_Disease: This variable presents only 18 positive results, which are spread over the Non-Persistent and Persistent patients, and will be removed for the modeling steps.

Risk_Family_History_Of_Osteoporosis A higher number of patients with negative results were observed for both target classes. The proportion of negative results does not seems to be considerably different for patients classified as Persistent and Non-Persistent.

Risk_Low_Calcium_Intake: This variable also presents a very low number of values (42) and will also be removed before the modeling steps.

Risk_Vitamin_D_Insufficiency: There was a slight higher number of negative results among the patients classified as Non-Persistent, but there was a slight higher number of positive results among the Persistent patients. This variable may also be important for the model.

Risk_Poor_Health_Frailty :This variable also presents much more negative results for both target classes and the proportion of negative results seems to be very similar.

Risk_Excessive_Thinness : Similar behaviour with the previous variable, however there are just a few positive results in total.

Risk_Hysterectomy_Oophorectomy: Similar behaviour with the previous variable, however there are just a few positive results in total.

Risk_Estrogen_Deficiency: This variable presents only 11 positive results and will be removed for the modeling steps.

Risk_Immobilization: This variable presents only 14 positive results and will be removed for the modeling steps.

Risk_Recurring_Falls: Similar behaviour with the variable Risk_Hysterectomy_Oophorectomy

4.2. Analysing Correlation between Input and Output Features

The Correlation value of features is calculated using two statistical Tests. They are as follows:

1) Z-test and T-test (Between Numerical Variables and Categorical Output)
2) Chi-Square Analysis(Between Categorical input and output)

Note: The significance level for all tests is set at 0.05

4.2.1 Correlation Analysis of Numerical Variables

There are two numerical variables('Dexa_Frequency_During_Rx', 'Count_of_Risks') and these both contains outliers and to treat outliers four approaches were employed:

Approach 1: Replace outlier values by the median value(Approach by: Harshith Sakala Santhosh)

In the first approach, 'Dexa_Frequency_During_Rx' seems to be having good correlation with 'Persistency_Flag' but 'Count_of_Risks' has very weak correlation as p-value goes more than significance value.

Approach 2: Replace outlier values by the upper limit threshold(Approach by: Mario Rodrigues Peres)

In this approach, both variables seem to be showing good correlation with 'Persistency_Flag' as both have p-values less than significance level.

Approach 3: Remove outliers(Approach by: Alexis Michael-Igbokwe)
This approach the outliers were reduced by much and seems to be showing good correlation with output label.

Note: It is observed that in 'count_of_risks' the most concentrated values are '1,2,3' but there are values more than these and their appearance in the dataset was very less and to counter this issue all values >=3 is replaced with value '3' making it a complete and pure categorical variable. It is achieved using buckenisation and this made the feature more flexible with other categorical features.

4.2.2 Correlation Analysis of Categorical Variables

As mentioned previously, the correlation of categorical variables and 'Persistency_Flag' is calculated using chi-square analysis. Out of 66 for now top ten features will be set aside and in future based on feature selection and model training performances these number may change.

The top 10 categorical features are as follows:

1) **Dexa_During_Rx**
2) **Comorb_Long_Term_Current_Drug_Therapy**
3) **Comorb_Encounter_For_Screening_For_Malignant_Neoplasms**
4) **Comorb_Encounter_For_Immunization ,**
5) **Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx**
6) **Comorb_Other_Disorders_Of_Bone_Density_And_Structure ,**
7) **Concom_Systemic_Corticosteroids_Plain**
8) **Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified**
9) **Concom_Anaesthetics_General**
10) **Concom_Viral_Vaccines**

4.3. Elimination of unwanted and Similar correlated Variables

It is always better to remove features that has weak correlation with output label and removing few features when there are multiple features highly correlated to each other as this increases speed of training and most importantly it increases accuracy of the model. There are features that constitute high correlation with others. They are as follows:

1) **NTM_Speciality Analysis(3 Variables):**
   It is found out all NTM_Speciality variables are highly correlated to each other .Based on the correlation analysis, the variable Ntm_Speciality_Bucket was the one that presented the highest correlation with the target variable, and therefore this variable will be kept, and the other two variables will be removed. When testing the models, we can also try to substitute the Ntm_Speciality_Bucket by the Ntm_Speciality_Flag to see how it goes. The latter variable is interesting because it is also well correlated with the target and presents only two categories instead of three. Moreover, the values are more well distributed among the categories.

2) **Dexa_Frequency_During_Rx and Dexa_During_Rx Analysis:**
   As similar to previous case these variables possess good correlation. These variables presented a significant correlation, and therefore, one of them might also be removed during the following steps. As the Dexa_During_Rx presented a more significant correlation with the target variable, it seems to be a good idea to remove the Dexa_Freq_During_Rx, which was the variable that presented outliers that were treated by 3 different approaches.

3) Segment Variables Analysis:
   It was observed that ‘Risk_Segment_Prior_Ntm’ has weak correlation with target variable. It is seen that; nearly half of the values are unknown for both variables and correspondents. There are a few more unknown values for the variable “Change_Risk_Segment”. Moreover, these variables are correlated, and one of them should also be removed. As the variable Risk_Segment_During_Rx presented a more significant correlation with the target, the Change_T_Score variable may be removed. Finally, the Risk_Segment_During_Rx also presents the advantage of having less categories with the values more equally distributed.

4) TScore Variables Analysis:
   Like segment group of variables, the variable Tscore_Bucket_Prior_Ntm did not present a significant correlation with the target and may be excluded in the following steps before modeling. Similarly, to the previous analysis, the variables Tscore_Bucket_During_Rx and Change_T_Score are correlated and one of them must be removed. The Change_T_Score presented higher correlation with the target variable, however the 4 categories of this variable are not evenly distributed, whereas the variable Tscore_Bucket_During_Rx presented 3 categories with the values better distributed.

5) Gluco_Record_Prior_Ntm and Gluco_Record_During_Rx:
   These two variables are very correlated, and therefore, one of them should be removed before the modeling steps. Moreover, the Gluco_Record_Prior_Ntm is not correlated with the target variable. Therefore, we will keep the Gluco_Record_During_Rx for the next steps of the modeling.

6) Frag_Frac_Prior_Ntm and Frag_Frac_During_Rx:
   These two variables are also very correlated, and one of them should be also removed before proceeding with the modeling steps. Similarly,to the Gluco variables, the

Frag_Frac_Prior_Ntm did not present a correlation with the target variable and will be removed before modeling. Therefore, we will keep the Frag_Frac_Prior_Ntm.

7) Comorb Analysis:
   All variables of Comorb are highly correlated to target variable except 'Comorb_Other_Disorders_Of_Bone_Density_and_Structur, Comorb_Vitamin_Deficiency, 'Comorb_Vitamin_D_Deficiency.' Most of these variables might be removed for model training and maybe only of these will be used. The variable 'comorb_long_term_current_drug' is one of the highly correlated variable among comorb variables.
8) Concom Analysis:
   All the Concom variables are correlated to each other, So, lets choose the one with more significant correlation with the target. In this case, It should be the Concom_Systemic_Corticosteroids_Plain.

   Note: There are still some feature analysis that was done but not included here to keep the length of the document less than 2500 words.