D Cube Analytics Case Study

Submitted by- Surya Tripathi
Date: 16 May 2020

Table of Contents

Introduction *3*

Problem Statement: **3**

Framework *4*

Exploratory Data Analysis *5*

NA values Identification and Imputation: **11**

Outlier Detection and Treatment: **11**

Feature transformation: **12**

Model Training and Testing: **12**

Feature Scaling: **12**

Feature Selection *13*

Model Building *14*

Model Evaluation *15*

Introduction

The Purpose of this document is to present the workflow and steps performed while building the model for the D cube analytics case study.

Problem Statement:

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription.

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

Target Variable: Persistency_Flag

Variable description is attached along with the data.

Deliverables:

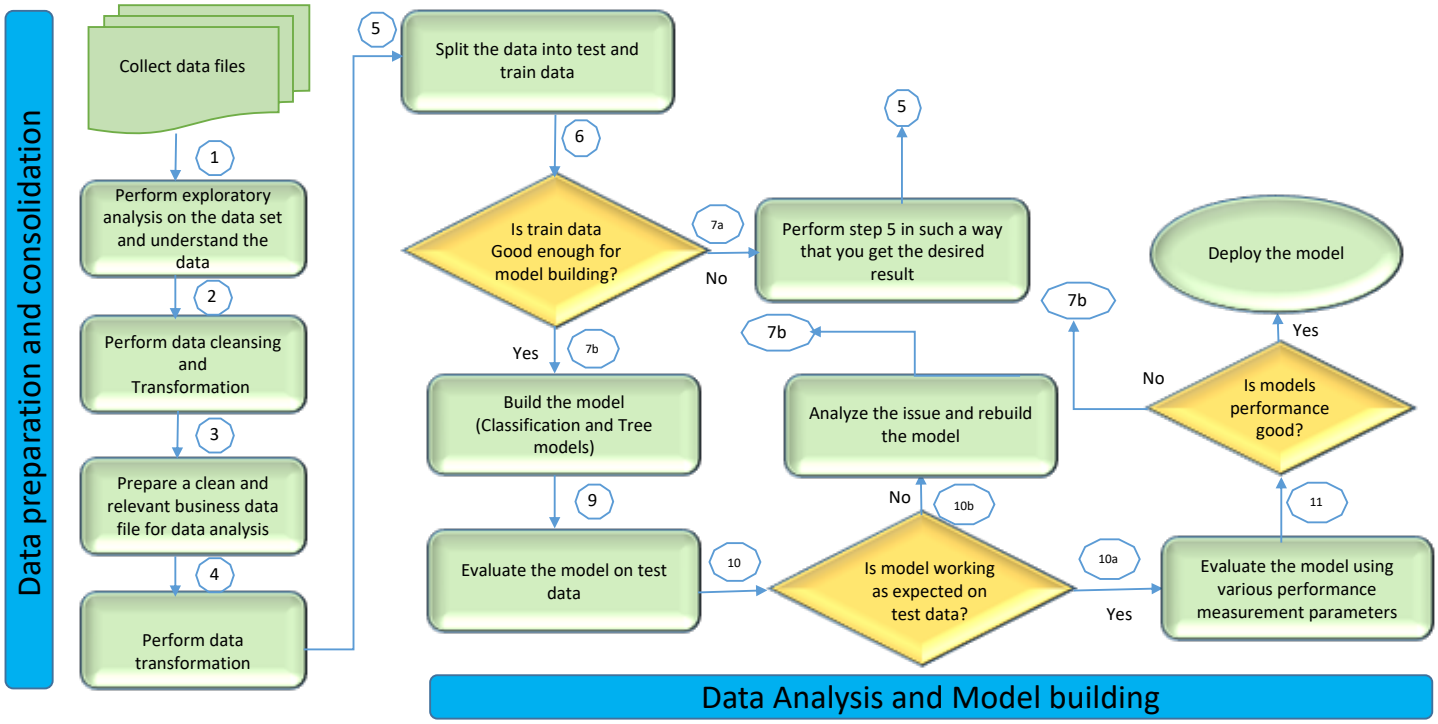
Following are the things we are expecting in the submission:

1. R/Python code (executable) used for the analysis (with proper comments and readability). If it's a Jupyter notebook with all the results in it, that will be best!
2. Model diagnostics to be updated in the attached excel template (Excel File: Analysis Results)
3. Final Analytical (processed) dataset used, which includes the additional derived variables and any other processing applied
4. Attach a document along with brief description of following in the mail:
    a. Changes done in the analytical dataset provided
    b. Any other highlights about the process you followed to ensure a thorough evaluation

Framework

We have followed CRISP-DM framework for building this model. Following are the phases of CRISP-DM framework:

- Business Understanding
- Data understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment (Not applicable for this case study)



Exploratory Data Analysis

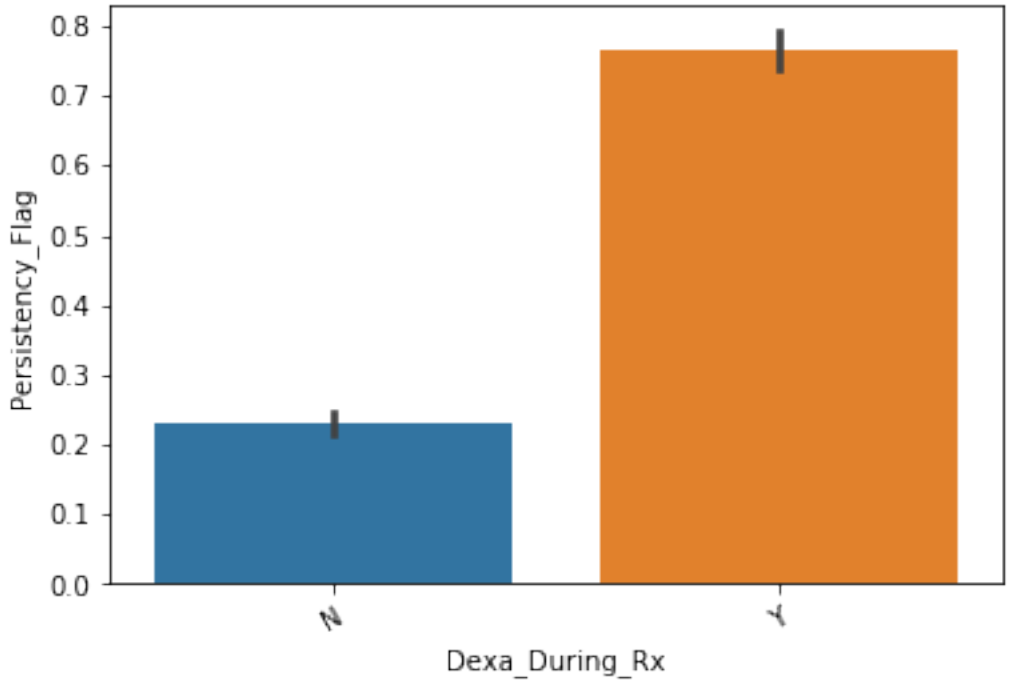
We have performed exploratory data analysis on the dataset:

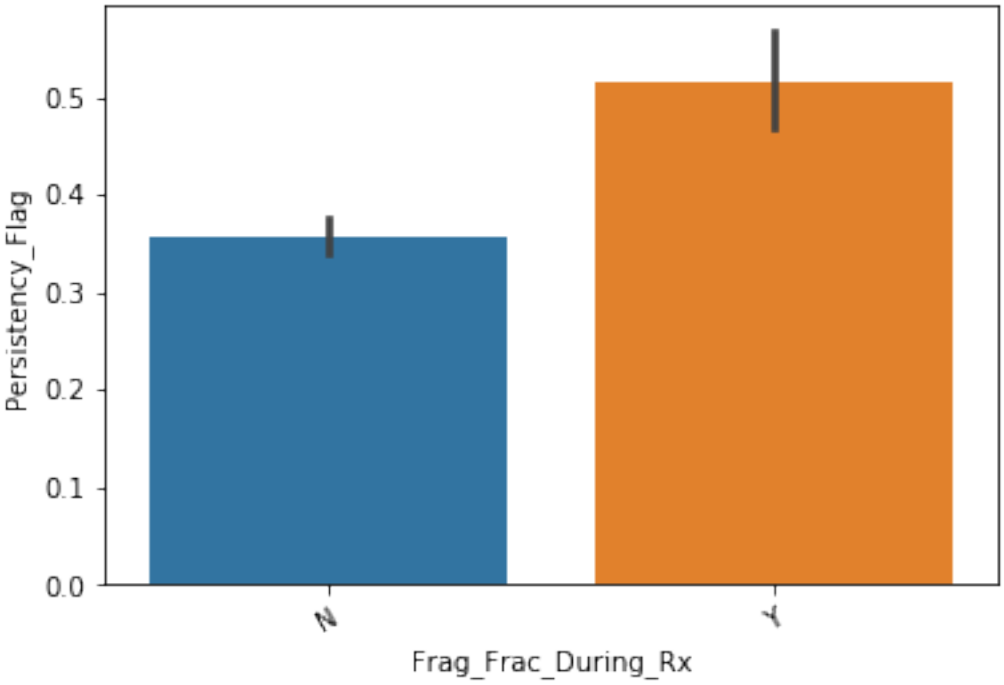
Data Points: 3424

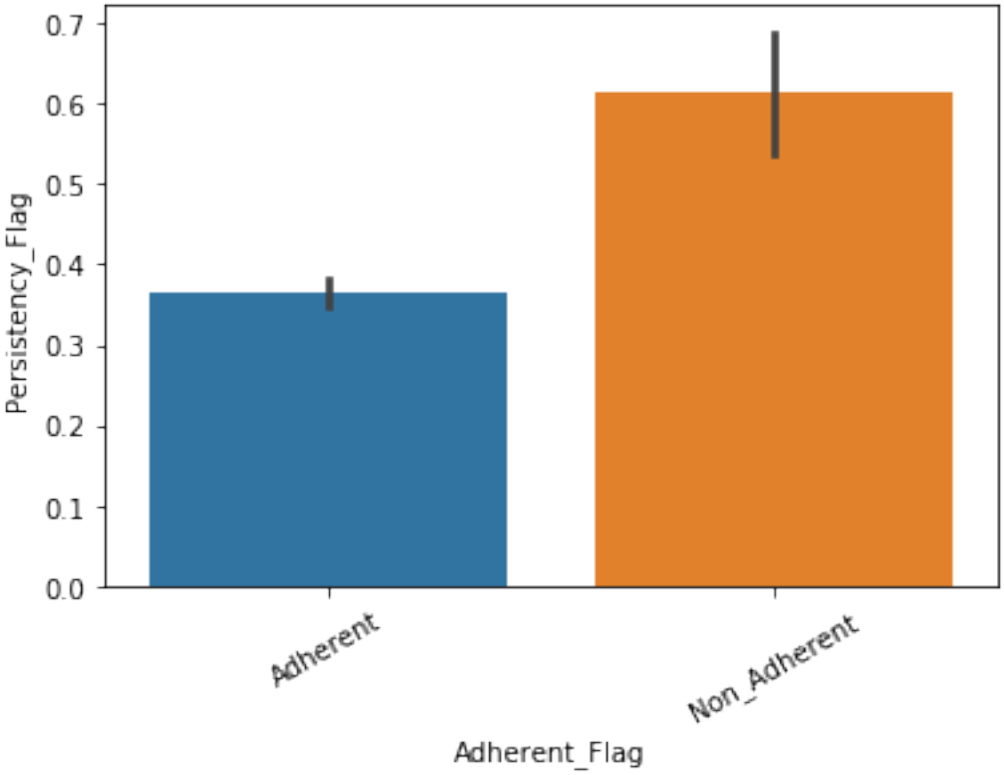
Features: 69

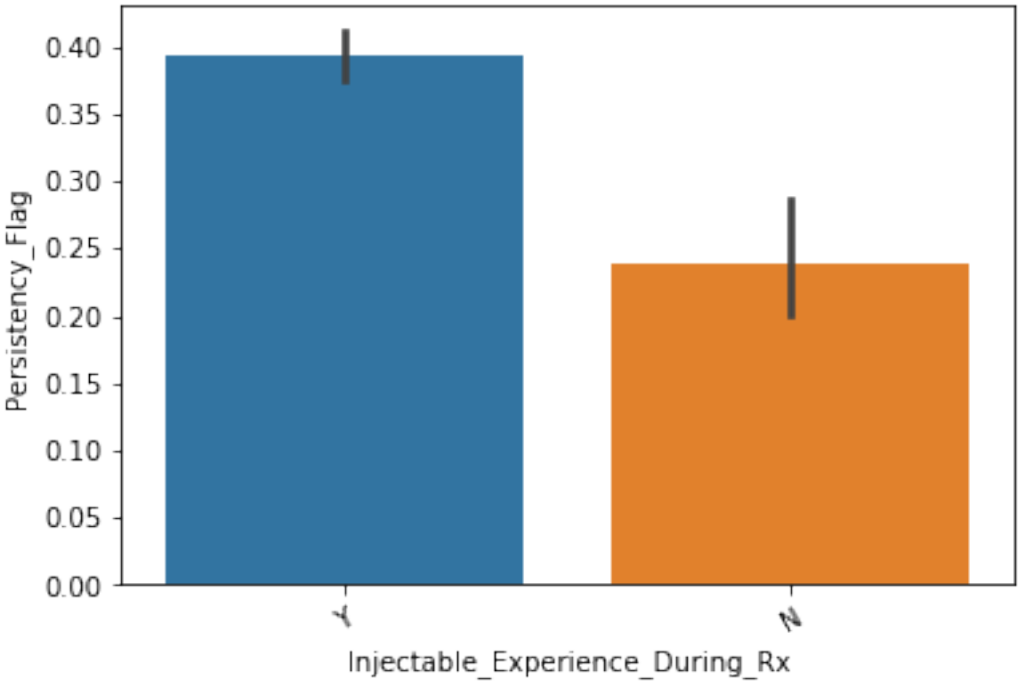
We have found that Counts_of_Risks and Dexa_Frq_during_Rx are numeric feature and all other are categorical feature.

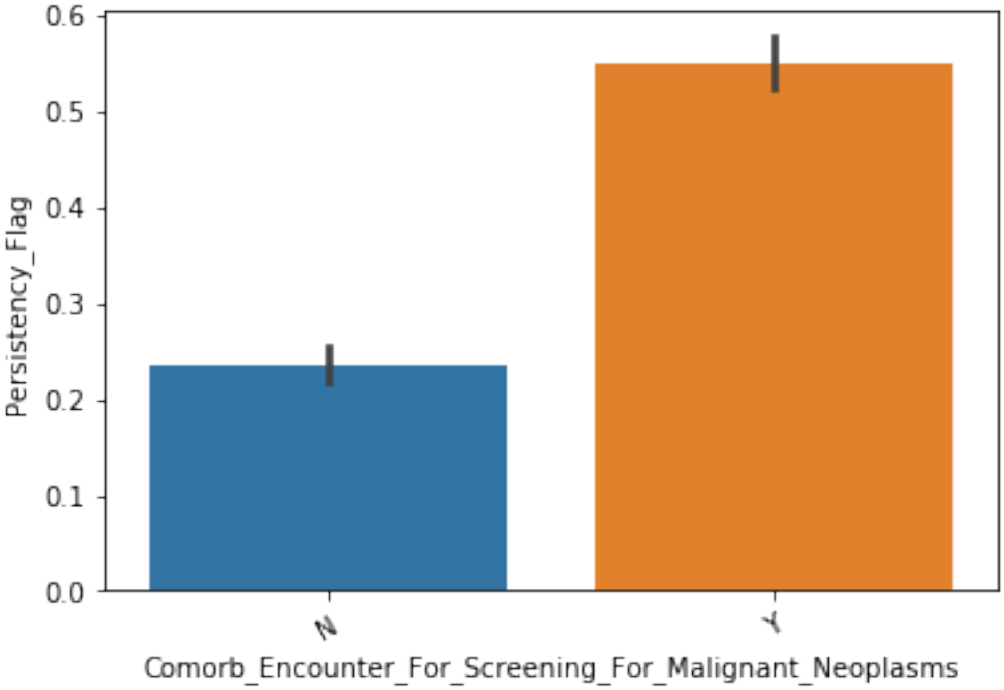
Following are the snaps of EDA:

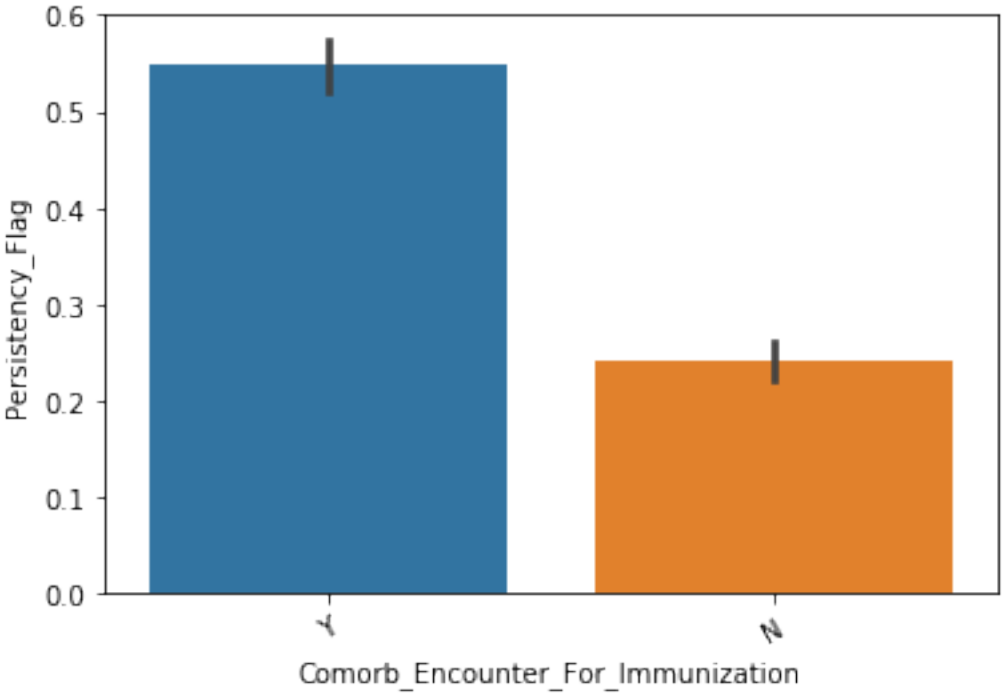


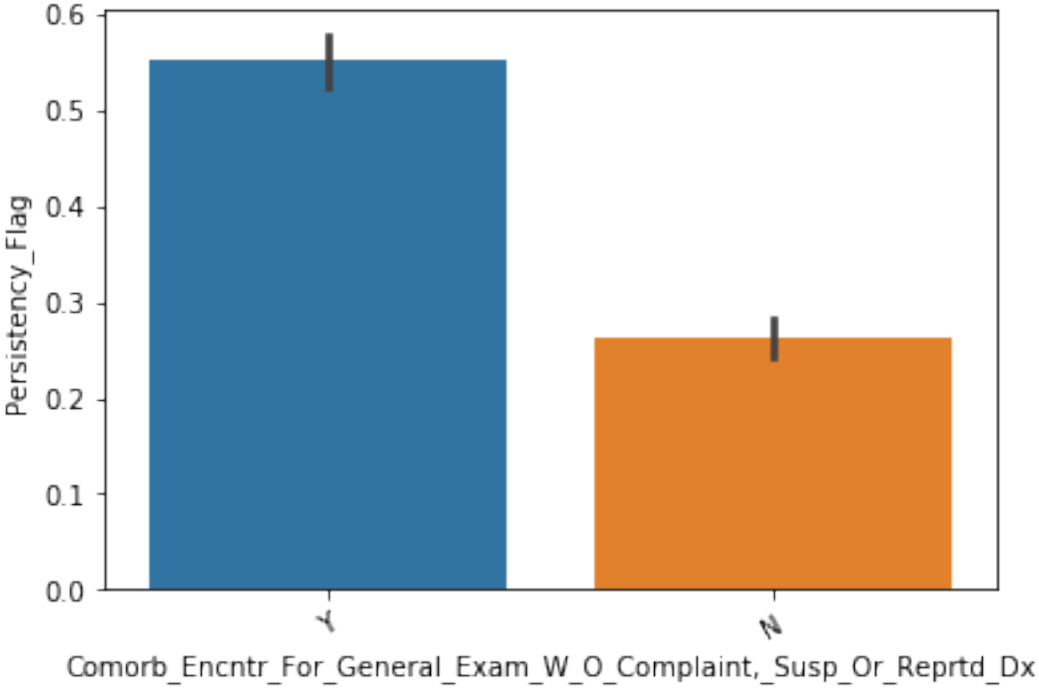

Persistency_Flag
0.0
0.1
0.2
0.3
0.4
0.5
N
Y
Frag_Frac_During_Rx

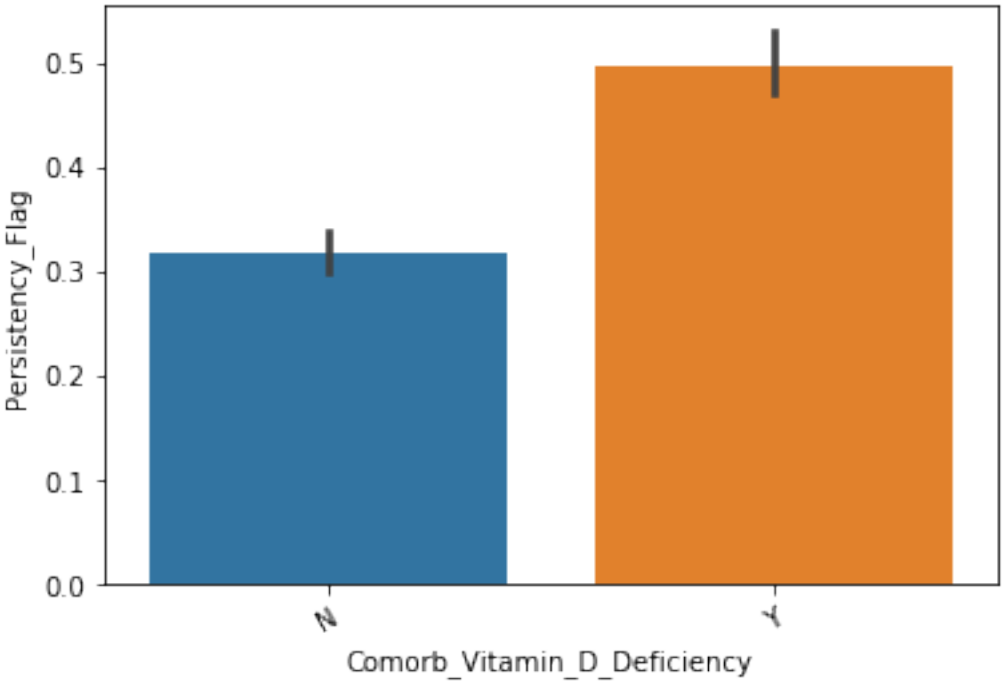

Persistency_Flag
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Adherent
Non_Adherent
Adherent_Flag

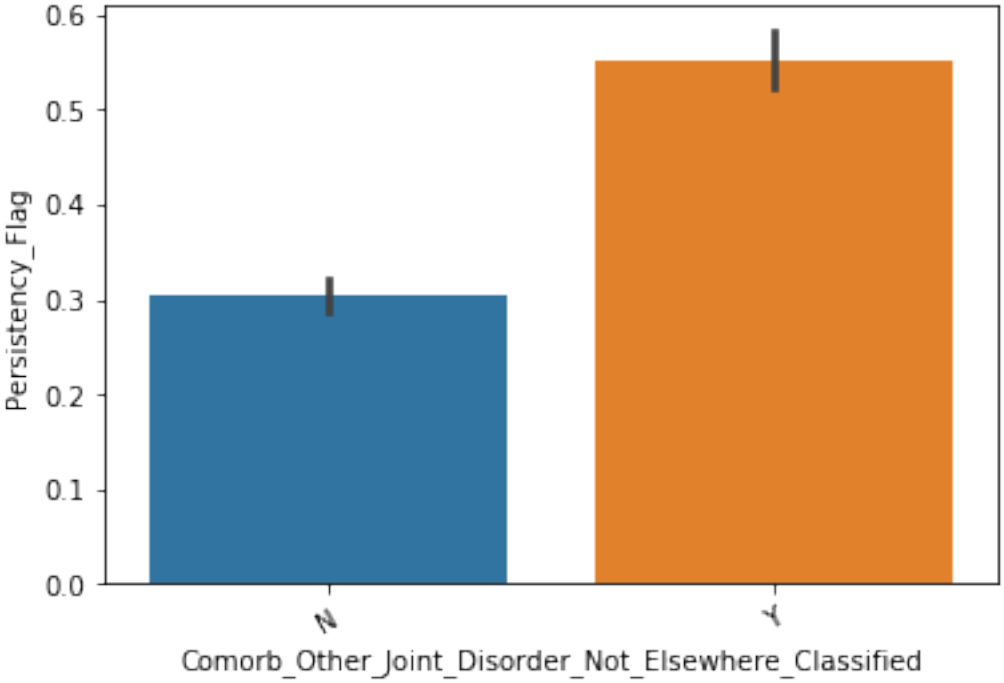

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Persistency_Flag
Y
N
Injectable_Experience_During_Rx

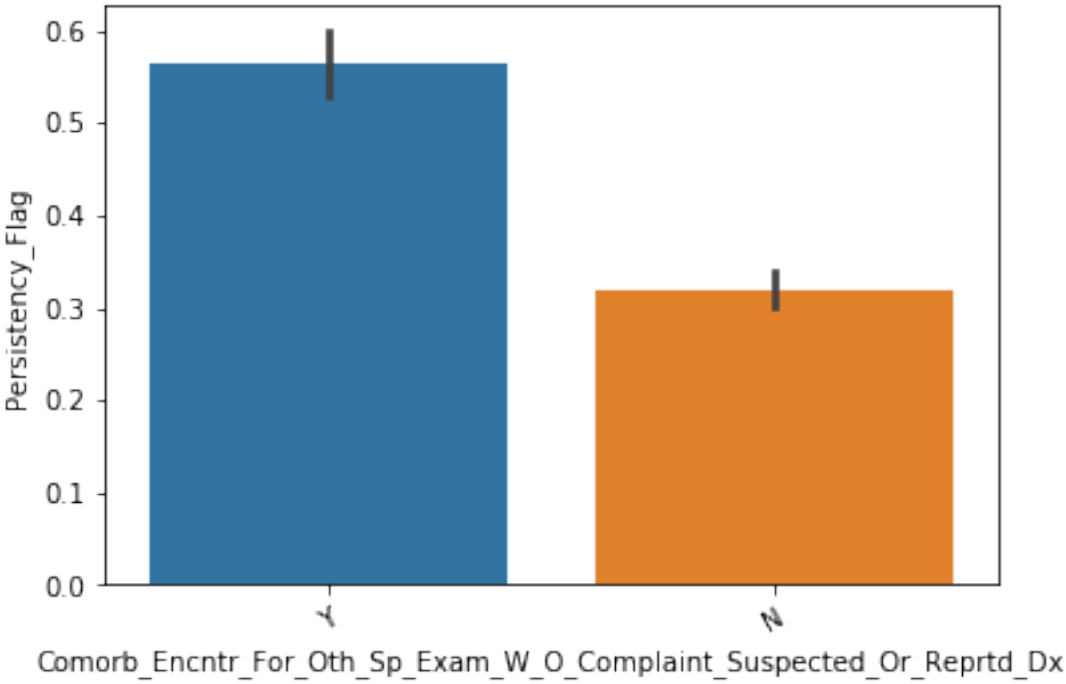

0.6
0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
N
Y
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms

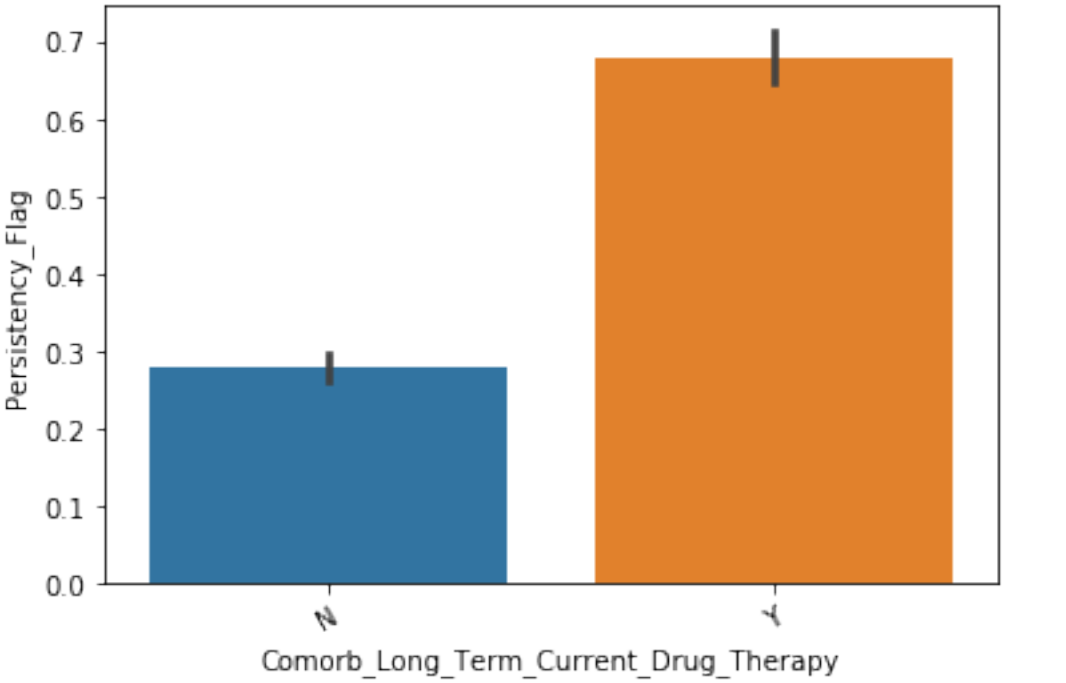

0.6
0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
Y
N
Comorb_Encounter_For_Immunization

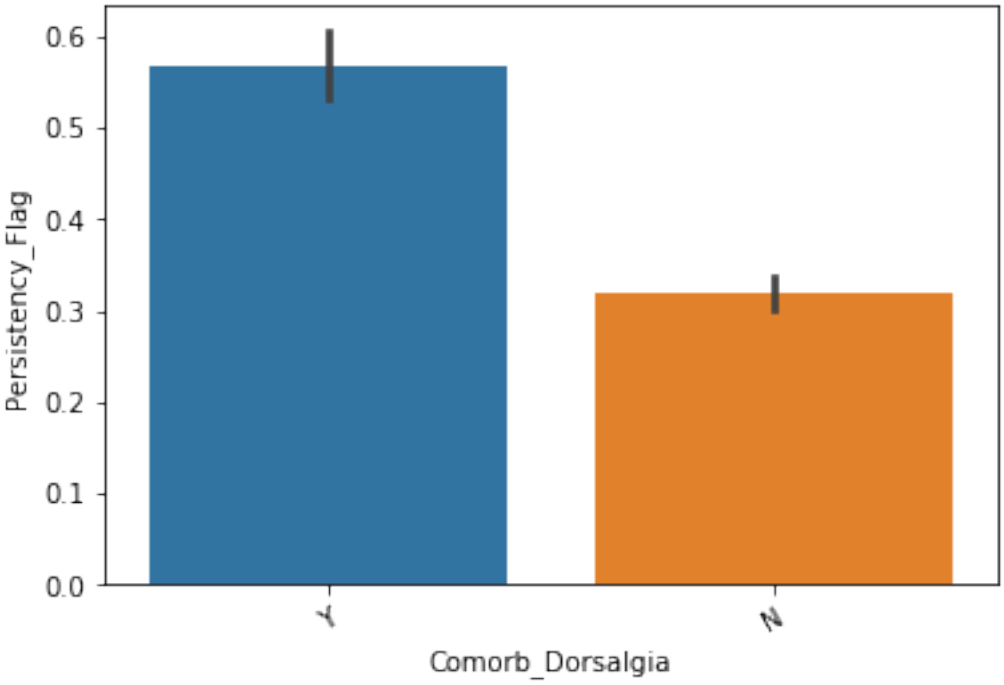

0.6
0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
Y
N
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx


0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
N
Y
Comorb_Vitamin_D_Deficiency


0.6
0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
N
Y
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified


0.6
0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
Y
N
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Persistency_Flag
N
Y
Comorb_Long_Term_Current_Drug_Therapy



NA values Identification and Imputation:

We have not encountered any NA values in the data set.

Outlier Detection and Treatment:

There are two numeric variables and Count_Of_Risks and Dexa_Freq_During_Rx. Both has outlier in the dataset and we have used IQR to impute the outliers.

We have taken 25 percentile and 75 percentile of the dataset and subtract that to get IQR.

Then We have imputed that values which are less than Q1(25 percentile)- 1.5 IQR

Or greater than Q3+1.5 IQR.

We have imputed these values with 86 and 99 percentiles respectively (as we had only outlier which had value greater than Q3+1.5IQR).

Feature transformation:

We have encoded with 0 and 1 for features which had only 2 levels. For categorical features which had more than 2 levels, we had used the dummy variables.

Model Training and Testing:

We have used 70% data for model training and 30% data for model testing.

Feature Scaling:

We have used Z value to scale the numeric features and have not performed scaling on other features as that features were already in the range of 0 to 1.

Feature Selection

We have used stats model api and RFE (SK learn) for feature elimination.

- We have passed all the features in **RFE** and pulled only 15 most significant features.
- Post the above step, we have used stats model and **VIF** to perform backward elimination of the features. Using stats model API we have utilized **p-value** to understand the feature significance while VIF we have used to understand the multicollinearity in the data set.

Model Building

We have used following models in this case study:

1. Logistics Regression
2. Random Forest

We have used stats model API to build the logistics regression because it provides better statistical summary and we have used SK learn for remaining two models.

Model Evaluation

We have used following metrics for model evaluation:

- Accuracy
- Specificity
- Sensitivity
- Precision
- Recall
- F1 Score
- AUC_ROC Curve

We have found that logistic regression performed better on this data set as compared to Random forest.

Following are the final result:

Precision	0.71
Recall/Sensitivity	0.74
Specificity	0.8
Accuracy	0.78
F1 Score	0.72
AUC	0.87