Healthcare - Persistency of a drug
Data Science Final Project Report

Group Name: Data Science Warrior
Name: Ugur Selim Ozen
Email: ugur_ozen58@hotmail.com
Country: Turkey
Collage/Company : Yıldız Technical University
Specialization : Data Science
Submission date: 12.18.2021
Github Repo Link:
https://github.com/UGURSELIMOZEN/Data_Glacier_DS_Internship/tree/main/DataScience_Healthcare_Final_Project

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, build a classification for the given dataset.

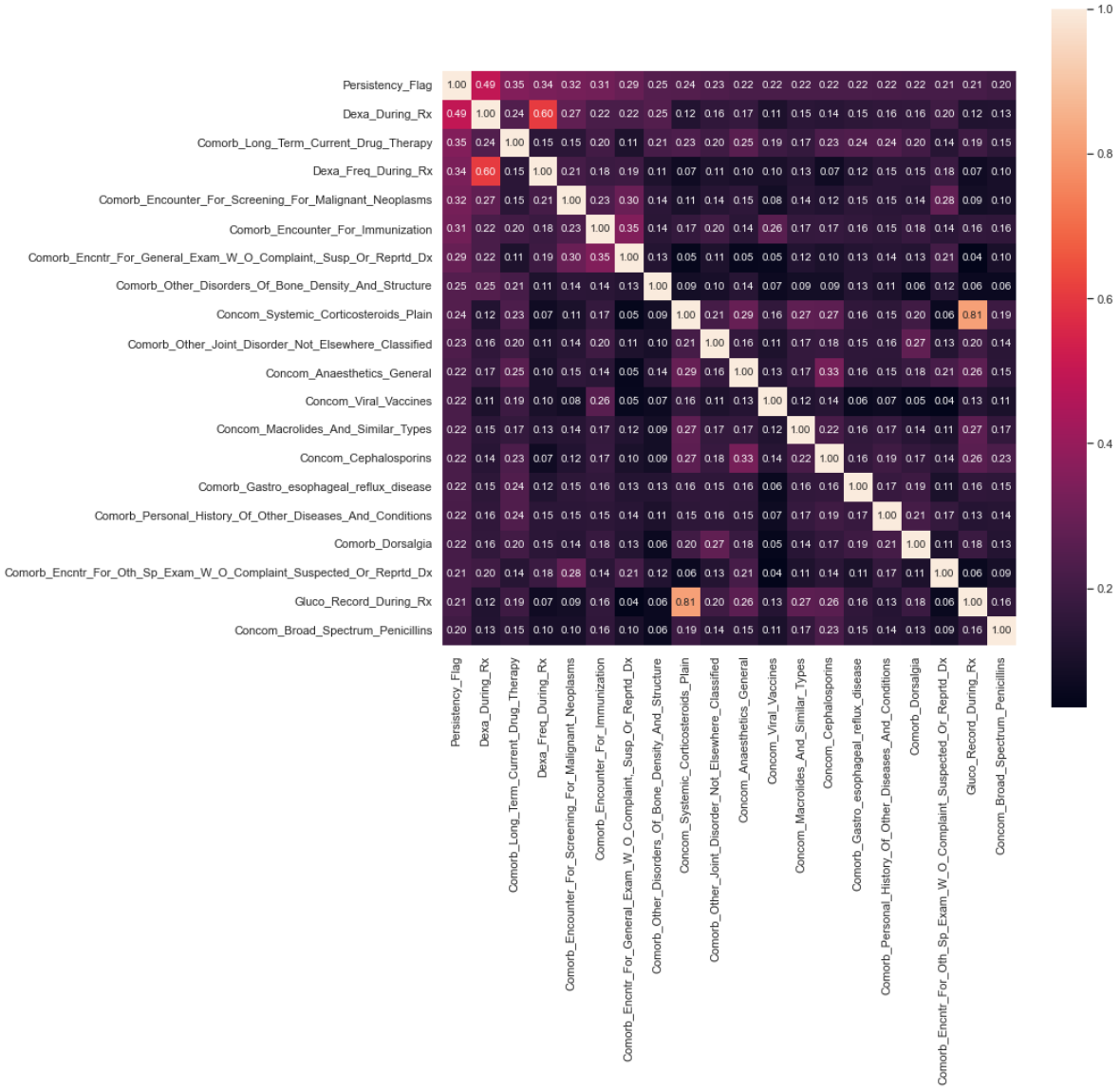
2.Data Understanding

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	.xlsx
Size of the data	1.8 MB
Null/NA Values	0

3. Exploratory Data Analysis and Feature Engineering

3.a. Correlation Heatmap Results

As seen from correlation heatmap given below ; Our target variable's **highest correlation** is **0.49** with **Dexa_During_Rx** and the others are lower than **0.35**.Concom_Systemic_Corticosteroids_Plain and Gluco_Record_During_Rx are extremely correlated each other with **0.81** so one of them can be dropped in model building process.Dexa_During_Rx and Dexa_Freq_During_Rx are highly correlated each other with **0.60** so may one of them can be dropped in model building process.But other features correlation is lower than **0.35** with each other.



3.b. Feature Transformation on Some Columns

By utilizing the ‘persistency_ratio’ function, I printed persistency ratio of some column's unique values.Using this function , I can **make grouping operation on some columns** by merging values that have same persisteny ratio to reduce column's unique value number.I will make this transformation **on 3 columns** which are ; **Count_Of_Risks** , **Ntm_Speciality** , **Dexa_Freq_During_Rx** .

3.c. Final Recommendations

Following features are certainly (%100) has PERSISTENT value so if your case has following values you have caught some wanted cases ;

• Ntm_Speciality = 9

• Dexa_Freq_During_Rx = 12

Following features are very likely (%80-%100) has PERSISTENT value so if your case has following values you may caught some wanted cases ;

• Dexa_Freq_During_Rx = 10

• Dexa_Freq_During_Rx = 11

Following features are likely (%60-%80) has PERSISTENT value so if your case has following values it is possible that catching some wanted cases ;

• Ntm_Speciality = 7

• Ntm_Speciality = 8

• Dexa_During_Rx = 1 (Yes)

• Count_Of_Risks = 6

• Concom_Viral_Vaccines = 1 (Yes)

• Concom_Anaesthetics_General = 1 (Yes)

• Concom_Broad_Spectrum_Penicillins = 1 (Yes)

• Concom_Macrolides_And_Similar_Types = 1 (Yes)

- **Concom_Cephalosporins = 1 (Yes)**
- **Comorb_Gastro_esophageal_reflux_disease = 1 (Yes)**
- **Comorb_Other_Disorders_Of_Bone_Density_And_Structure = 1 (Yes)**
- **Comorb_Long_Term_Current_Drug_Therapy = 1 (Yes)**
- **Dexa_Freq_During_Rx = 4**
- **Dexa_Freq_During_Rx = 8**
- **Dexa_Freq_During_Rx = 9**
- **Adherent_Flag = 'Non-Adherent'**
- **Change_Risk_Segment = 'Worsened'**
- **Change_T_Score = 'Improved'**
- **Change_T_Score = 'Worsened'**

Following features are certainly (%100) has NON-PERSISTENT value so if your case has following values there is no need to focus on it anyway ;

- **Ntm_Speciality = 0**
- **Risk_Immobilization = 1 (Yes)**
- **Risk_Untreated_Chronic_Hyperthyroidism = 1 (Yes)**
- **Dexa_Freq_During_Rx = 0**

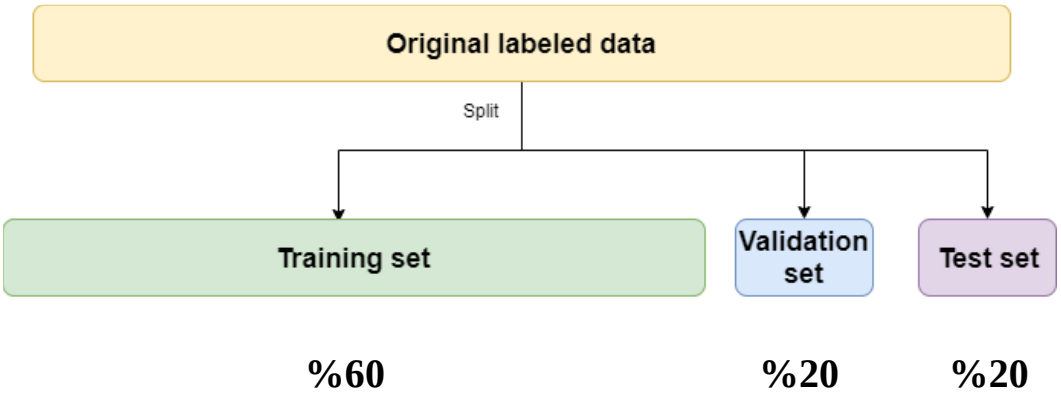
4. Data Modeling Technique

• For this dataset , **modeling** will be made with **67 features** using **OneHotEncoding** and **oversampling** methods.

• **3 features** (**Count_Of_Risks , Ntm_Speciality , Dexa_Freq_During_Rx**) have **transformed** in **feature engineering step** and **any extra column** has **not derivated** from dataset.

• I am planning to use following **machine learning algorithms** in dataset modelling step (train-validation-test);

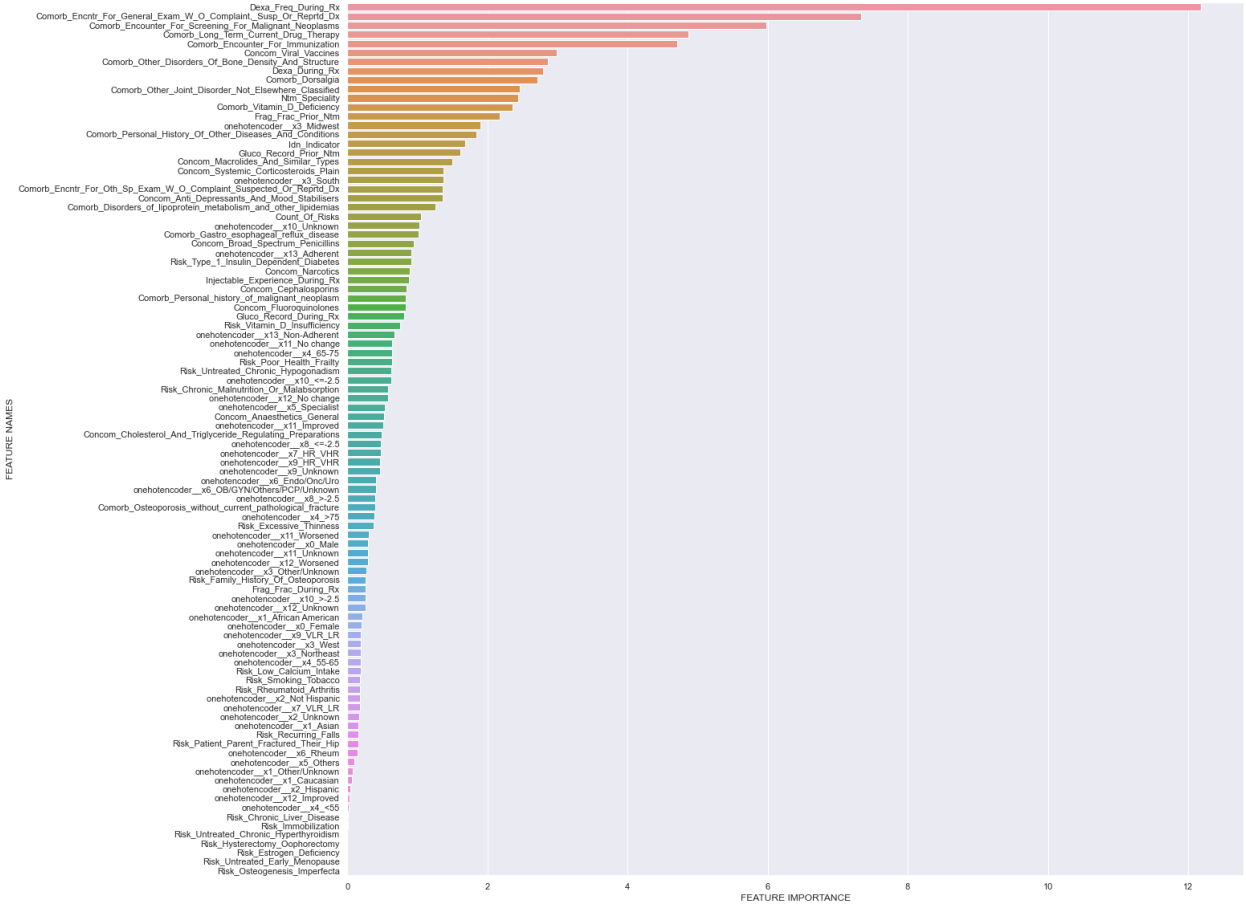
1. Decision Tree Classifier
2. Random Forest Classifier
3. Logistic Regression
4. CatBoost Classifier



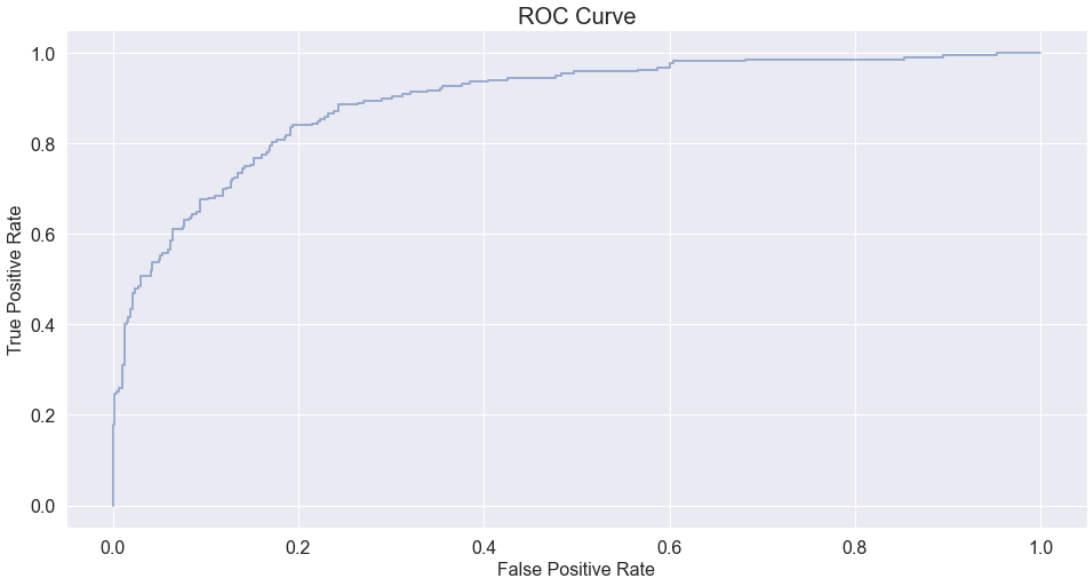
5. Model Prediction Results

Model Algorithm	Recall - 0	Recall - 1	F1 Score - 0	F1 Score - 1	5-Fold Cross Validation Recall	5-Fold Cross Validation F1 Score
Decision Tree Classifier	0.79	0.57	0.79	0.57	0.624	0.622
Random Forest Classifier	0.89	0.67	0.87	0.70	0.615	0.674
Logistic Regression	**0.86**	**0.75**	**0.87**	**0.73**	**0.651**	**0.686**
CatBoost Classifier	0.91	0.67	0.88	0.72	0.635	0.695

6. Feature Importance

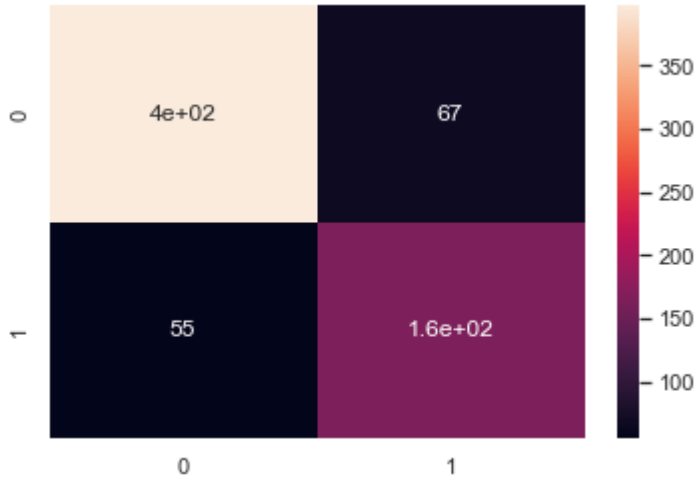


7. ROC Curve



8. Interpreting Results

I have decided to use **Logistic Regression** model.It is very **fast and stable** according to CatBoost Classifier model also it's **Recall- 1 score** is greater than CatBoost one. That means we can have **more gain** by catching the **true positive cases**.



Selected Model Confusion Matrix