Data Glacier

Your Deep Learning Partner

Data Science Virtual Internship

Data Science :: Healthcare
Persistency of a Drug :: Final Project
Final Presentation

Ugur Selim Ozen

18-Dec-2021

Methodology

1	• Data Cleaning and Transformation
2	• Exploratory Data Analysis
3	• Feature Engineering
4	• Model Evaluation
5	• Interpreting Results

Utilized Technologies

jupyter

NumPy

pandas

seaborn

matplotlib

plotly

scikit learn

CatBoost

Background – Healthcare :: Persistency of a Drug

- 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.

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

The analysis has been divided into 3 parts:

- Data-Business Understanding
- Finding key insights from dataset about features
- Recommendations for pharmaceutical companies

Data Cleaning and Transformation

- Total features : **69**
- Total observations : **3424**
- Null or Missing values : **0**
- Dataset size : 1.8 MB

Assumptions:

- By checking null values in all features, we can see there is **no null values** but now we need to focus on much more to observe any missing or unknown value assigned for null values.

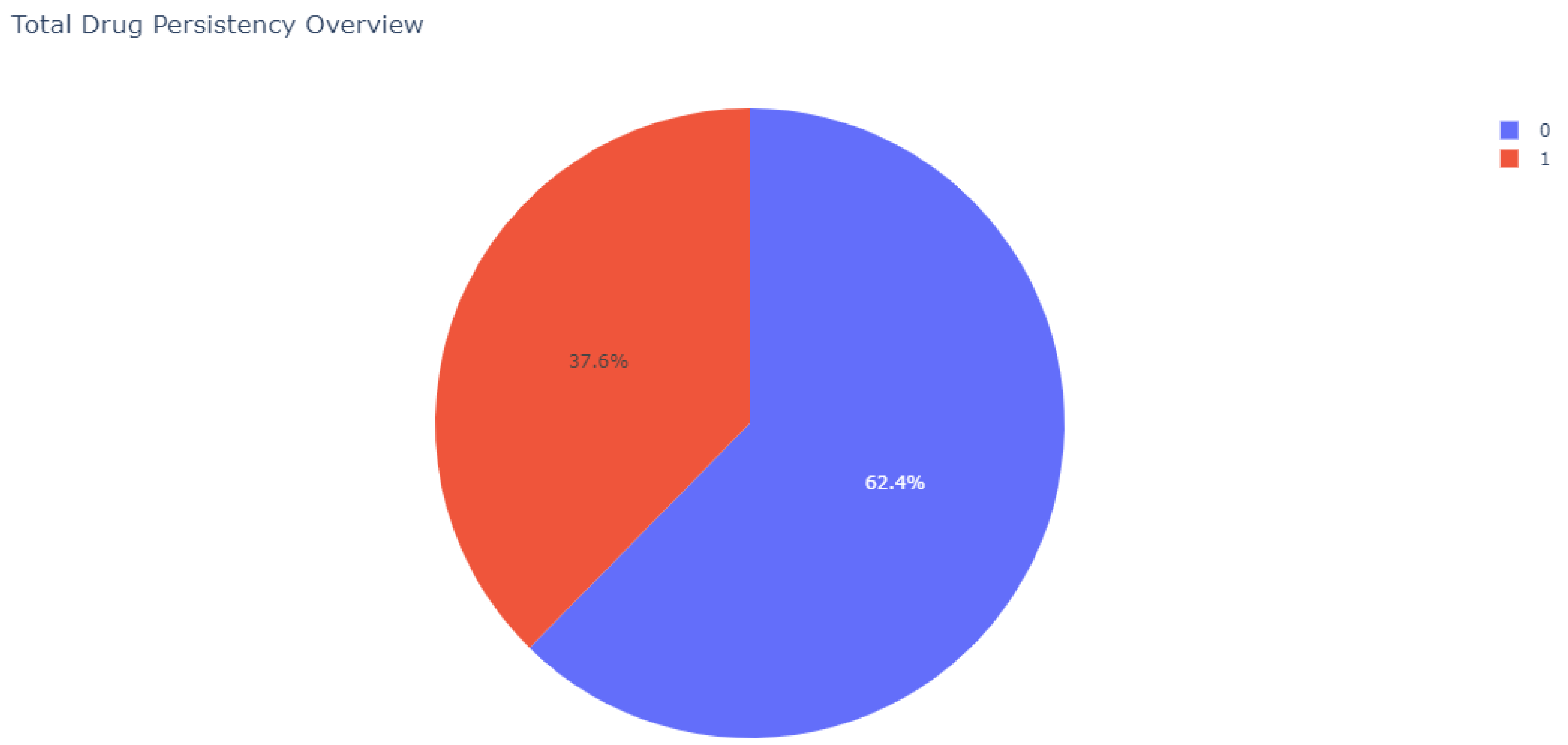
- This is a **classification problem** so we can impute null or missing values generally with **two approaches** ; first one is **filling with most recurring value (mode)** and second one is we can **categorize the missing values** with some value like **'Missing' or 'Unkown'.**

- In this dataset **null or missing values were filled ' Unkown'** value therefore we can apply first method which is filling with mode.

- Filling with mode operation is made for only **4 columns** ; **Race , Ethnicity , Region and Ntm_Speciality** because in other columns , ratio of 'Unknown' is more than %50 that means 'Unknown' itself is mode in column so it can be meaningless and not correct operation for other columns.

- Generally, for the classification problems we can have imbalanced dataset in real-life , we can say that this **dataset is imbalanced**, so we need to apply **oversampling or undersampling methods** in model building step.

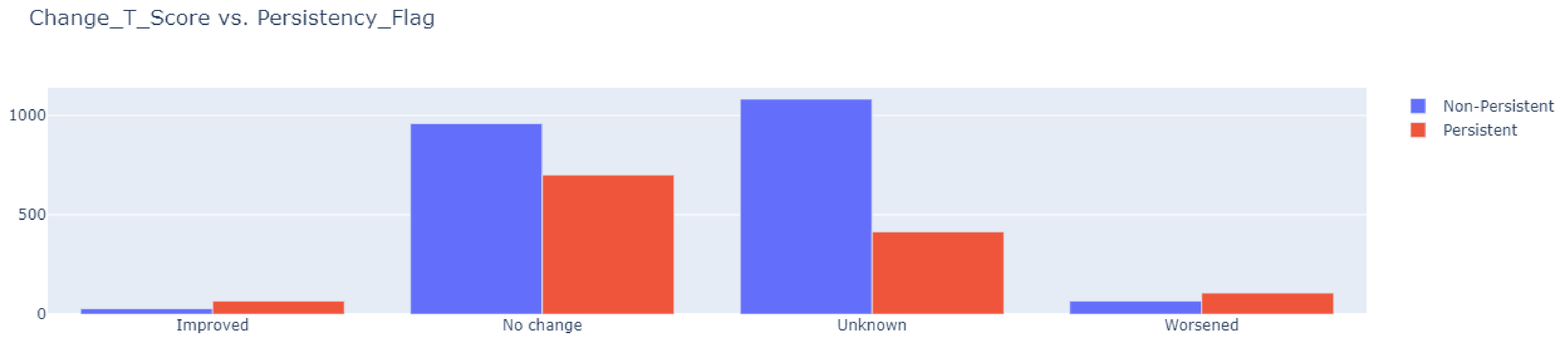
Total Drug Persistency Analysis



- As seen from this Pie Chart; The total **number of non-persistent drug** is approximately **1.7 times** that of **persistent drug**, so it means that dataset is imbalanced.

Change_T_Score

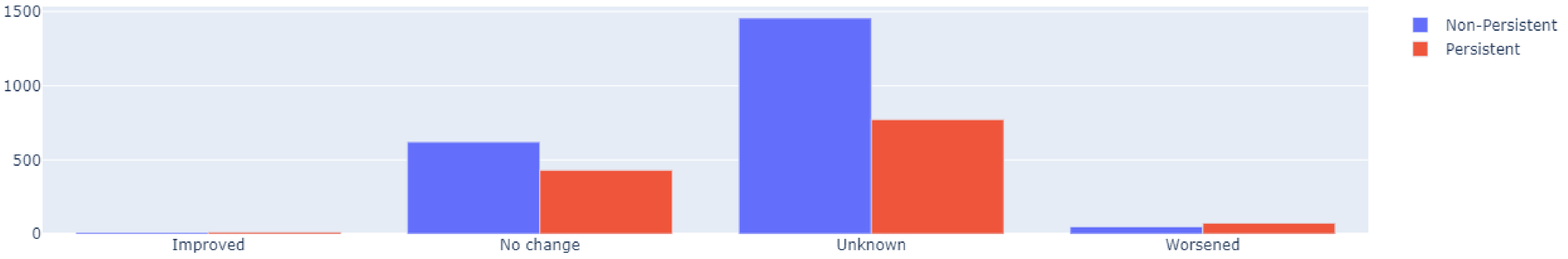
	Change_T_Score	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
3	Improved	66.0	94.0	70.212766
2	Worsened	107.0	173.0	61.849711
0	No change	701.0	1660.0	42.228916
1	Unknown	415.0	1497.0	27.722111



Change_Risk_Segment

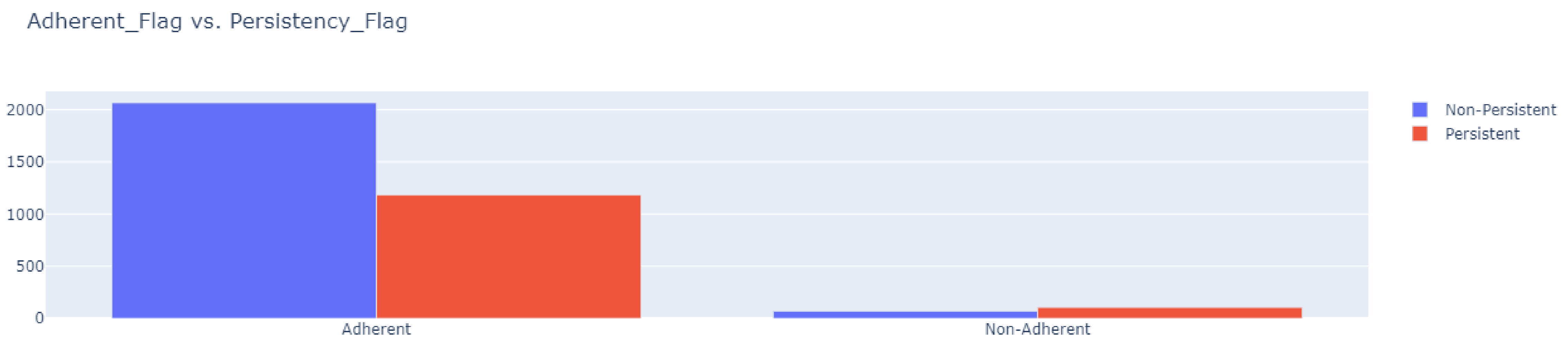
	Change_Risk_Segment	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
2	Worsened	73.0	121.0	60.330579
3	Improved	13.0	22.0	59.090909
1	No change	431.0	1052.0	40.969582
0	Unknown	772.0	2229.0	34.634365

Change_Risk_Segment vs. Persistency_Flag



Adherent_Flag

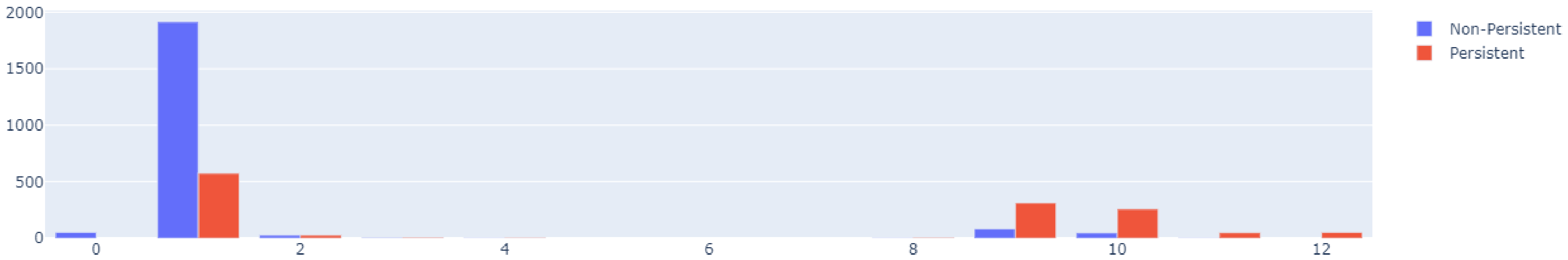
	Adherent_Flag	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	Non-Adherent	106.0	173.0	61.271676
0	Adherent	1183.0	3251.0	36.388803



Dexa_Freq_During_Rx

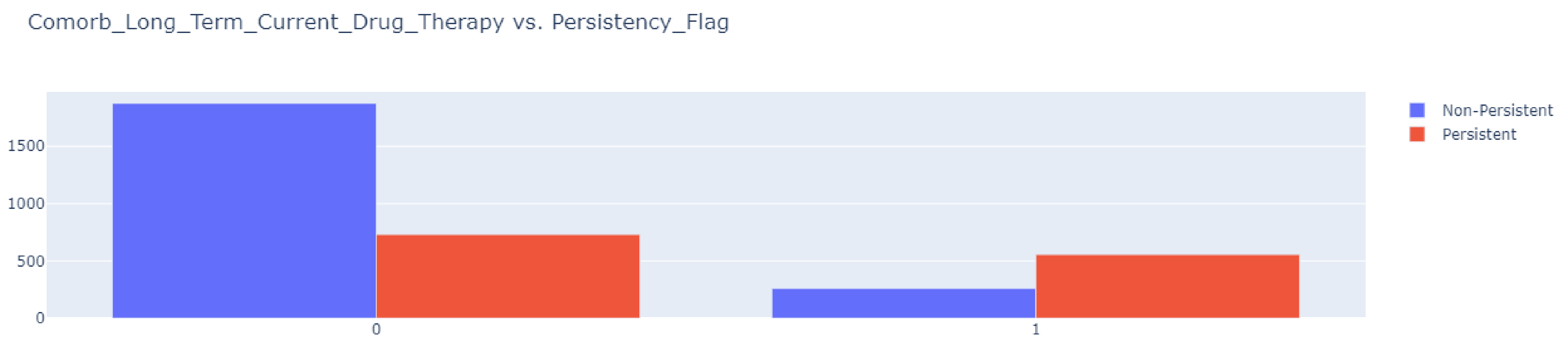
	Dexa_Freq_During_Rx	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
6	12	51.0	51.0	100.000000
7	11	48.0	51.0	94.117647
2	10	258.0	304.0	84.868421
4	9	313.0	396.0	79.040404
9	8	7.0	10.0	70.000000
8	4	6.0	9.0	66.666667
5	3	8.0	14.0	57.142857
3	2	25.0	50.0	50.000000
0	1	573.0	2488.0	23.030547
1	0	0.0	51.0	0.000000

Dexa_Freq_During_Rx vs. Persistency_Flag



Comorb_Long_Term_Current_Drug_Theraphy

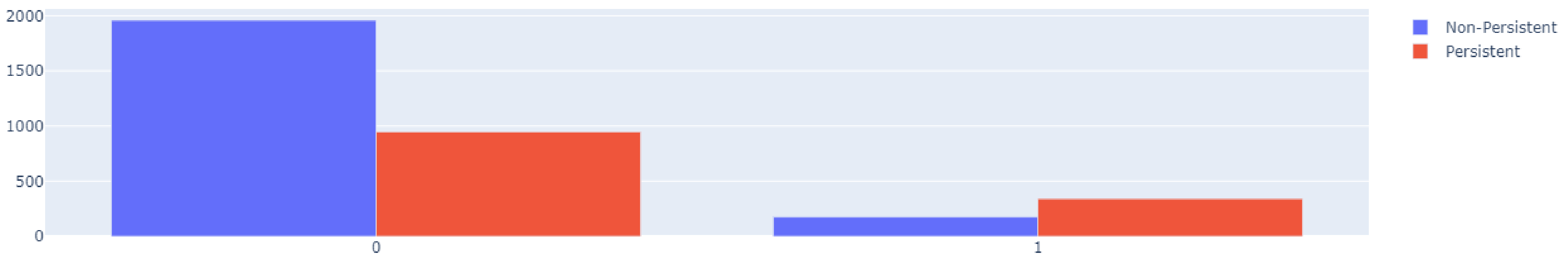
	Comorb_Long_Term_Current_Drug_Therapy	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	557.0	817.0	68.176255
0	0	732.0	2607.0	28.078251



Comorb_Other_Disorders_of_Bone_Density_and_Structure

	Comorb_Other_Disorders_Of_Bone_Density_And_Structure	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	342.0	518.0	66.023166
0	0	947.0	2906.0	32.587749

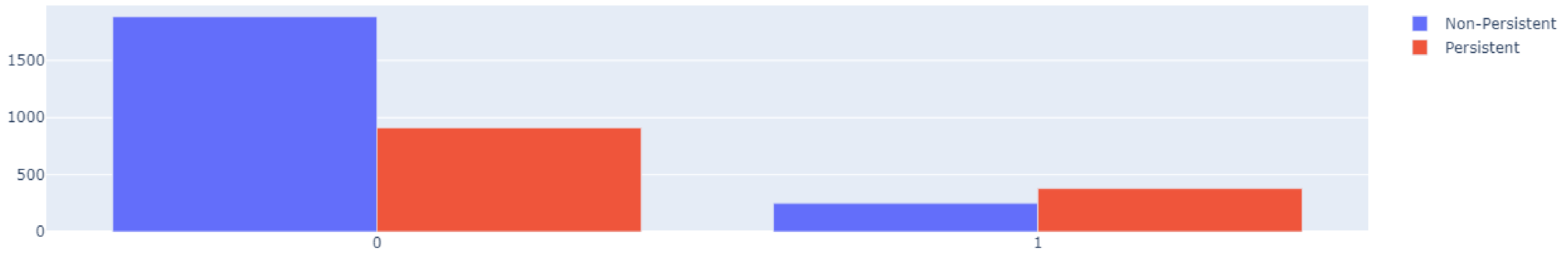
Comorb_Other_Disorders_Of_Bone_Density_And_Structure vs. Persistency_Flag



Comorb_Gastro_Esophageal_Reflux_Disease

	Comorb_Gastro_esophageal_reflux_disease	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	379.0	630.0	60.158730
0	0	910.0	2794.0	32.569792

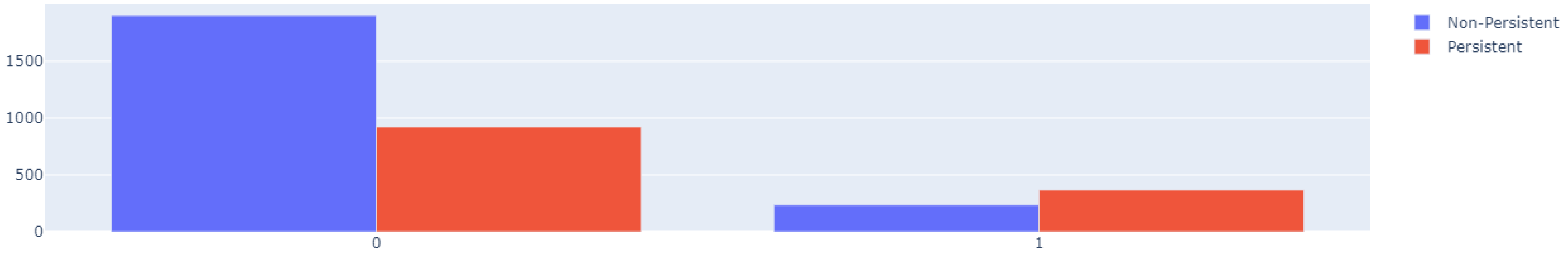
Comorb_Gastro_esophageal_reflux_disease vs. Persistency_Flag



Concom_Cephalosporins

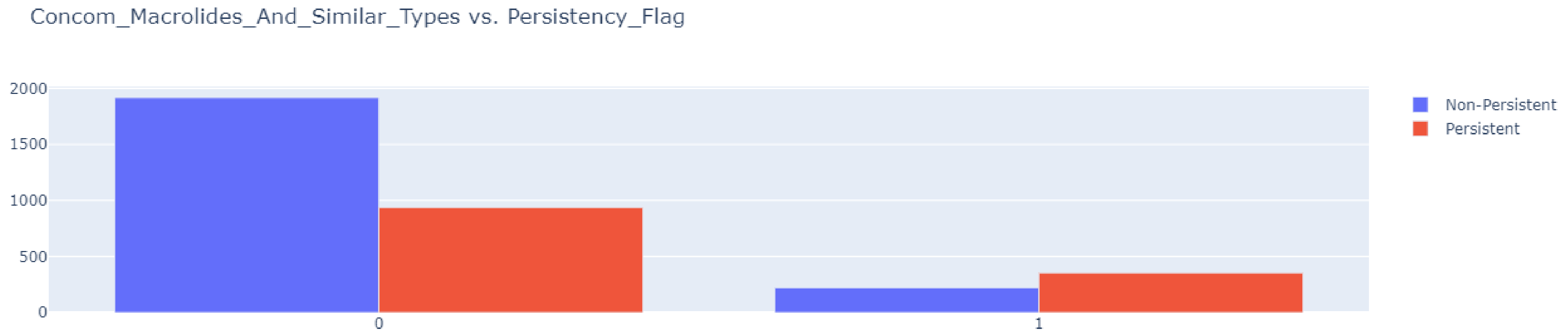
	Concom_Cephalosporins	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	367.0	603.0	60.862355
0	0	922.0	2821.0	32.683446

Concom_Cephalosporins vs. Persistency_Flag



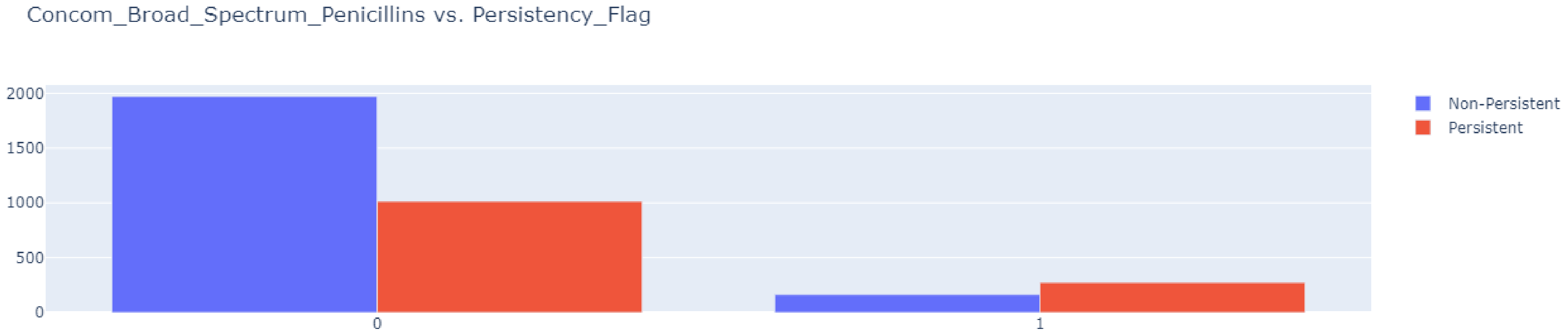
Concom_Macrolides_and_Similar_Types

	Concom_Macrolides_And_Similar_Types	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	352.0	571.0	61.646235
0	0	937.0	2853.0	32.842622



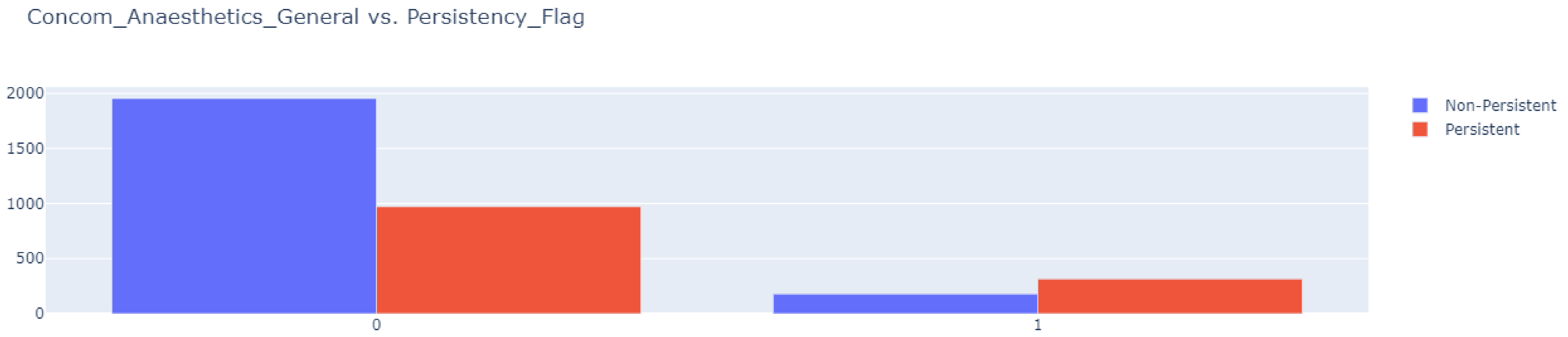
Concom_Broad_Spectrum_Penicillins

	Concom_Broad_Spectrum_Penicillins	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	275.0	439.0	62.642369
0	0	1014.0	2985.0	33.969849



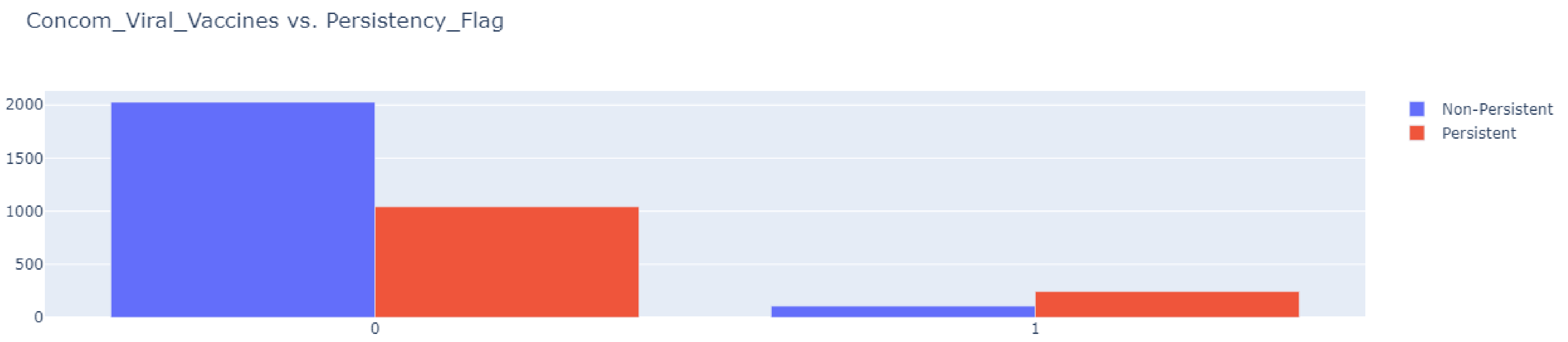
Concom_Anaesthetics_General

	Concom_Anaesthetics_General	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	317.0	497.0	63.782696
0	0	972.0	2927.0	33.208063



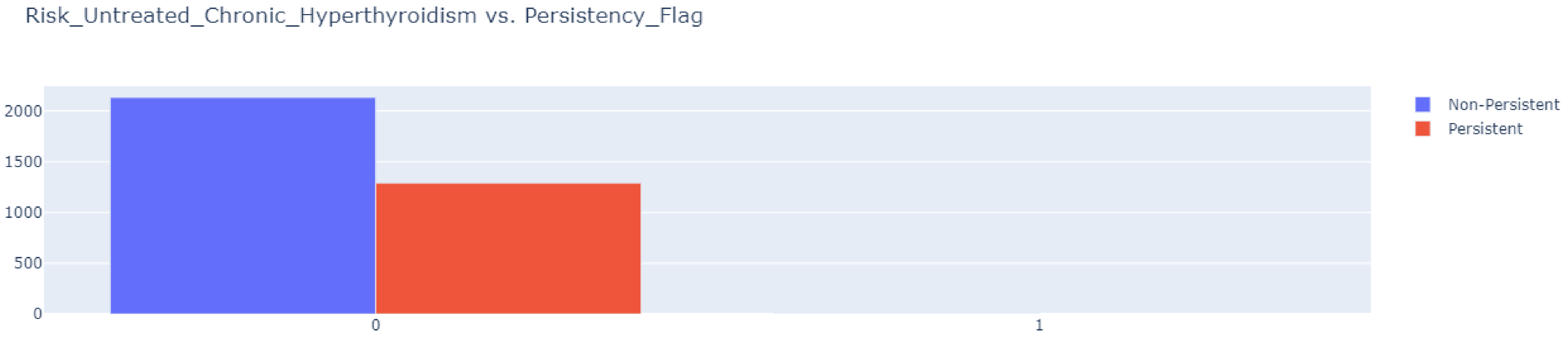
Concom_Viral_Vaccines

	Concom_Viral_Vaccines	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	245.0	353.0	69.405099
0	0	1044.0	3071.0	33.995441



Risk_Untreated_Chronic_Hyperthyroidism

	Risk_Untreated_Chronic_Hyperthyroidism	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
0	0	1289.0	3422.0	37.66803
1	1	0.0	2.0	0.00000



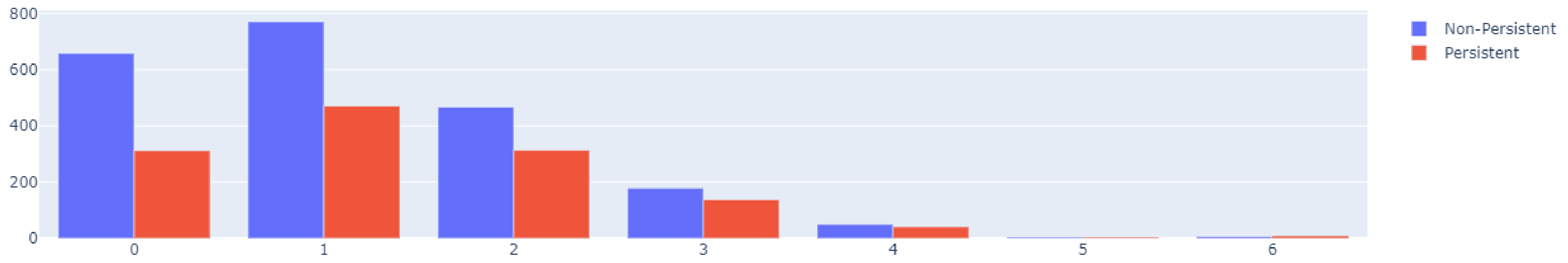
Risk_Immobilization

	Risk_Immobilization	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
0	0	1289.0	3410.0	37.800587
1	1	0.0	14.0	0.000000

Count_of_Risks

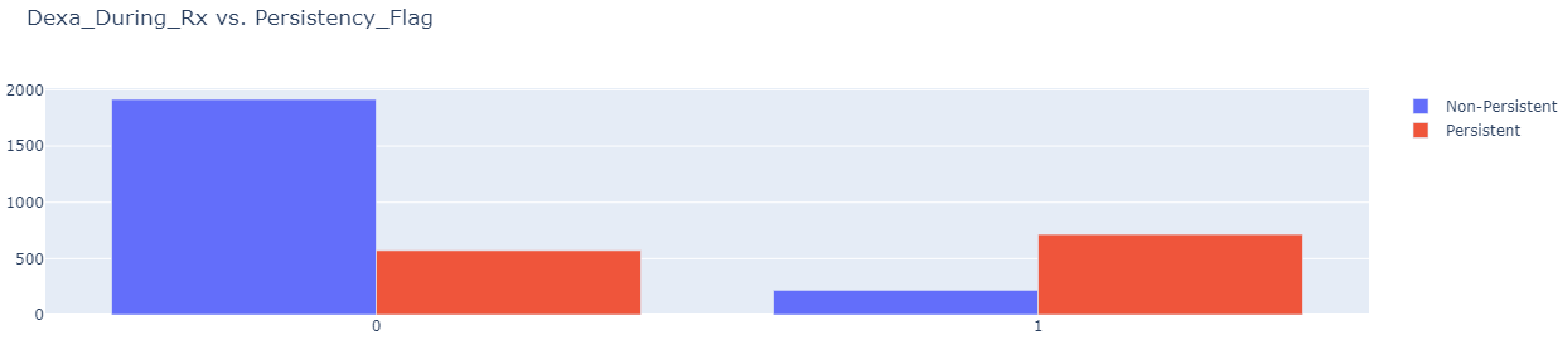
	Count_Of_Risks	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
5	6	9.0	15.0	60.000000
6	5	4.0	8.0	50.000000
4	4	41.0	91.0	45.054945
3	3	138.0	317.0	43.533123
1	2	314.0	781.0	40.204866
2	1	471.0	1242.0	37.922705
0	0	312.0	970.0	32.164948

Count_Of_Risks vs. Persistency_Flag



Dexa_During_Rx

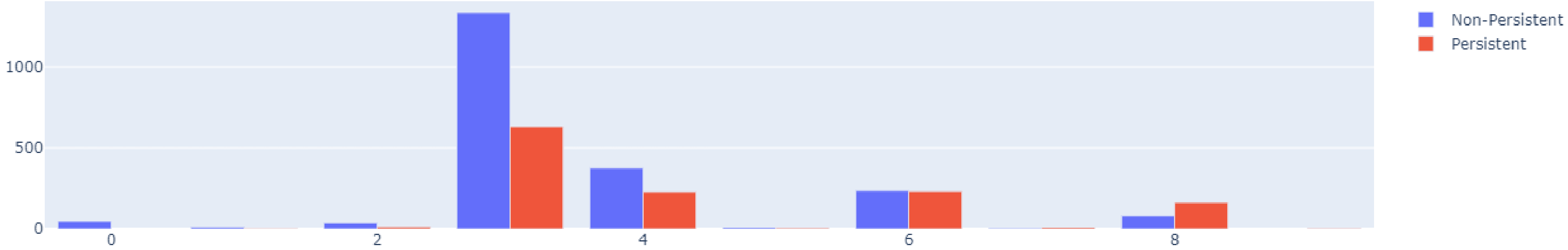
	Dexa_During_Rx	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
1	1	716.0	936.0	76.495726
0	0	573.0	2488.0	23.030547



Ntm_Speciality

	Ntm_Speciality	PERSISTENCY_NUMBER	TOTAL_CASE	PERSISTENCY_RATIO
8	9	4.0	4.0	100.000000
3	8	163.0	244.0	66.803279
6	7	8.0	13.0	61.538462
1	6	232.0	468.0	49.572650
7	5	6.0	14.0	42.857143
2	4	228.0	604.0	37.748344
0	3	632.0	1968.0	32.113821
5	2	13.0	49.0	26.530612
9	1	3.0	14.0	21.428571
4	0	0.0	46.0	0.000000

Ntm_Speciality vs. Persistency_Flag



Final Recommendations-I

1. 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

2. 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

3. 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

Final Recommendations-II

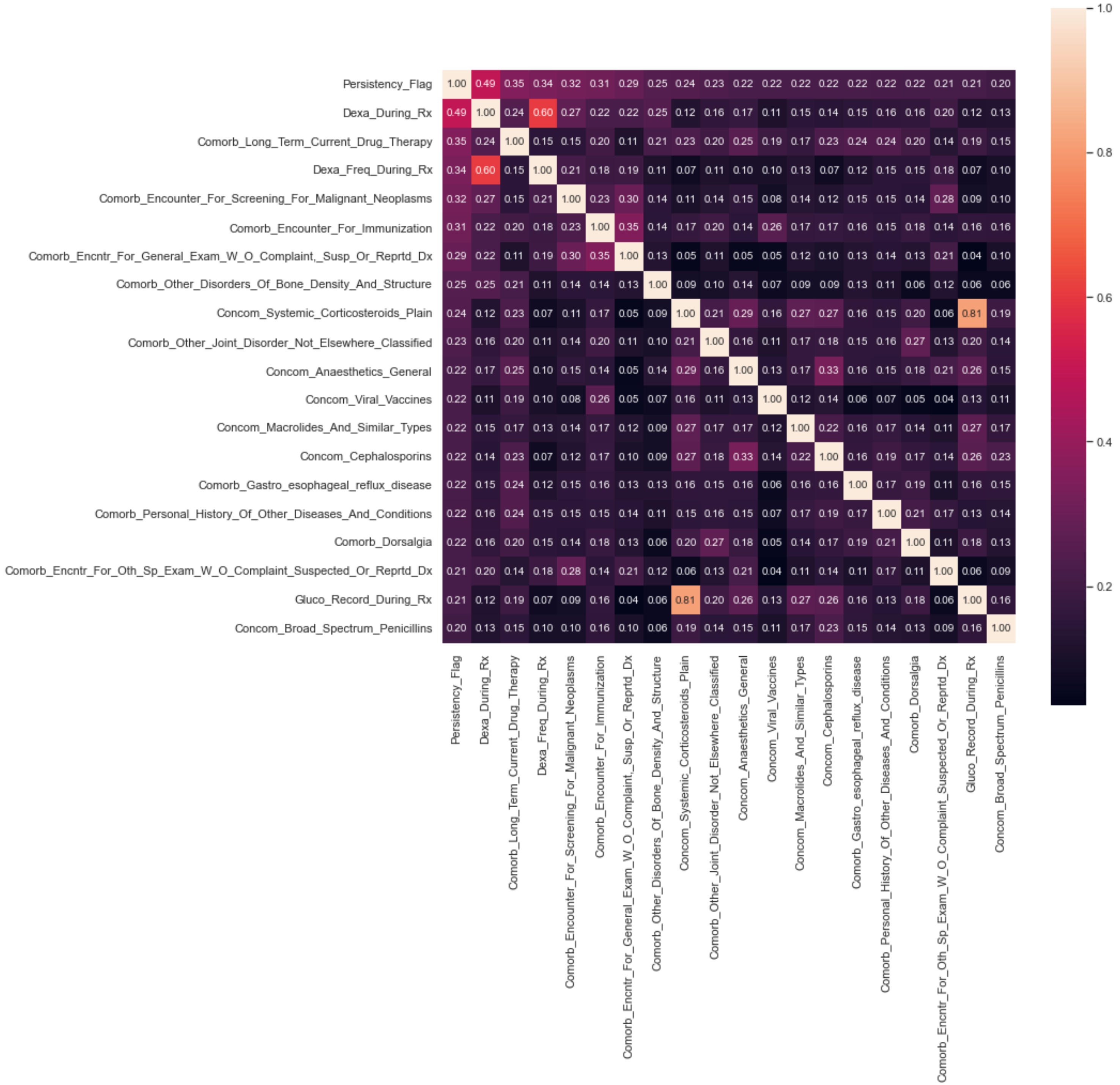
- Dexa_During_Rx = 1 (Yes)
- Count_Of_Risks = 6
- Concom_Viral_Vaccines = 1 (Yes)
- Concom_Anaesthetics_General = 1 (Yes)
- Concom_Broad_Spectrum_Penicillins = 1 (Yes)
- Concom_Macrolides_And_Similar_Types = 1 (Yes)
- Concom_Cephalosporins = 1 (Yes)
- Comorb_Gastro_esophageal_reflux_disease = 1 (Yes)
- 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'

Final Recommendations-III

4. 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

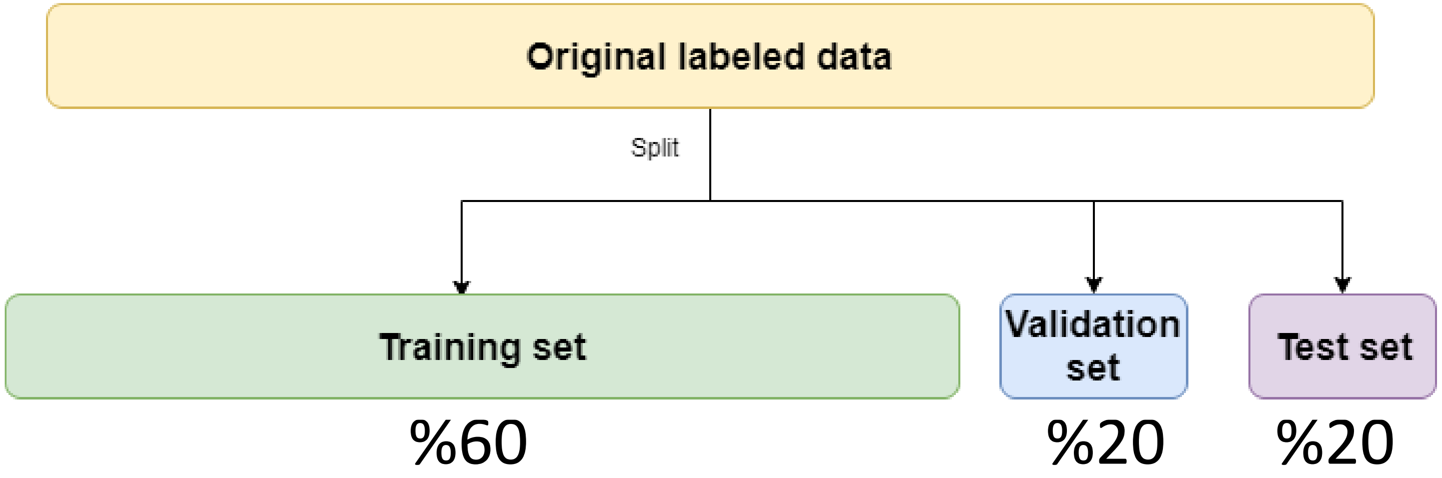
Correlation Heatmap



Recommended 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) ;

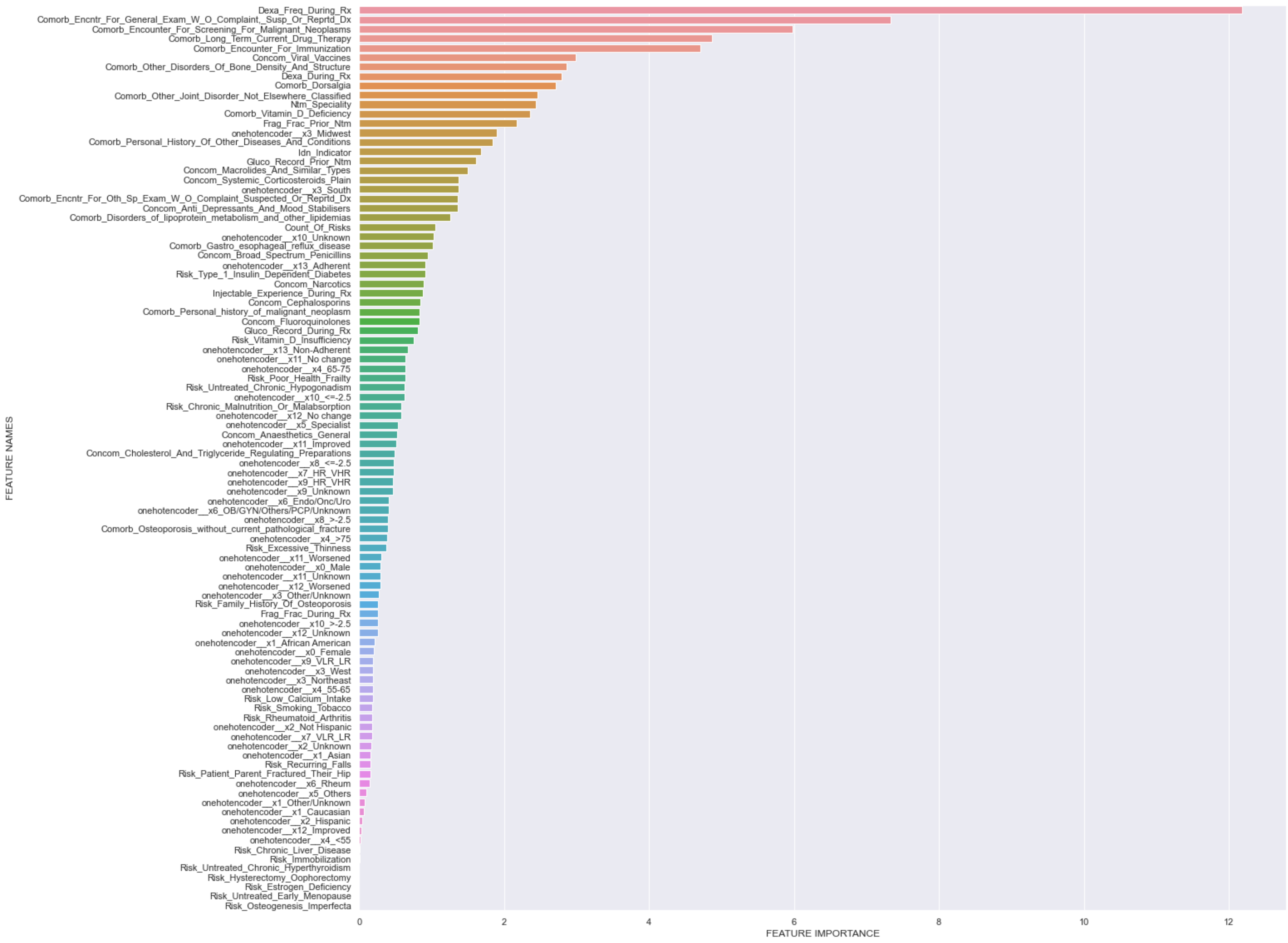
1. **Decision Tree Classifier**
2. **Random Forest Classifier**
3. **Logistic Regression**
4. **CatBoost Classifier**



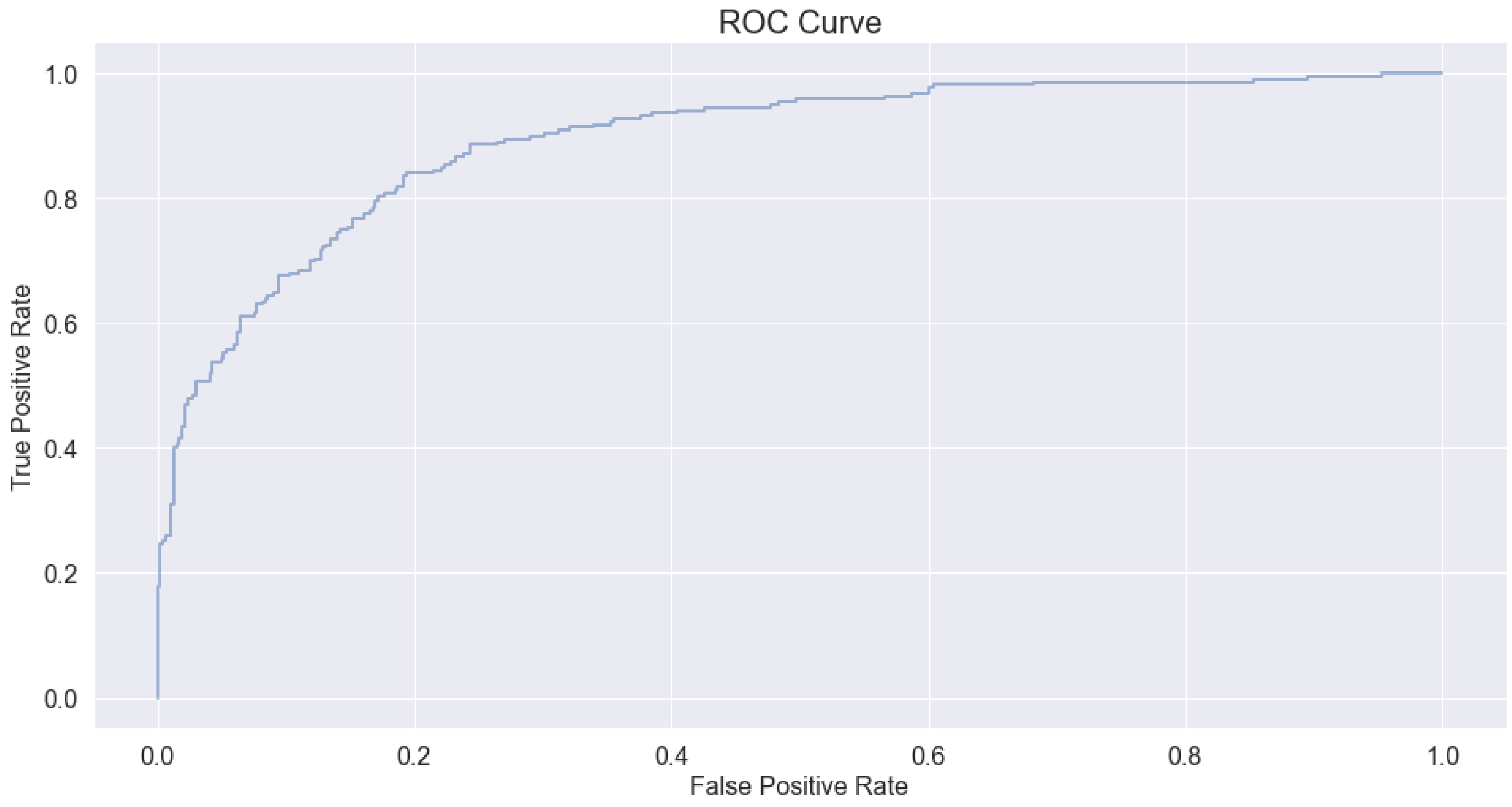
Model Prediction Test Results

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

Feature Importance

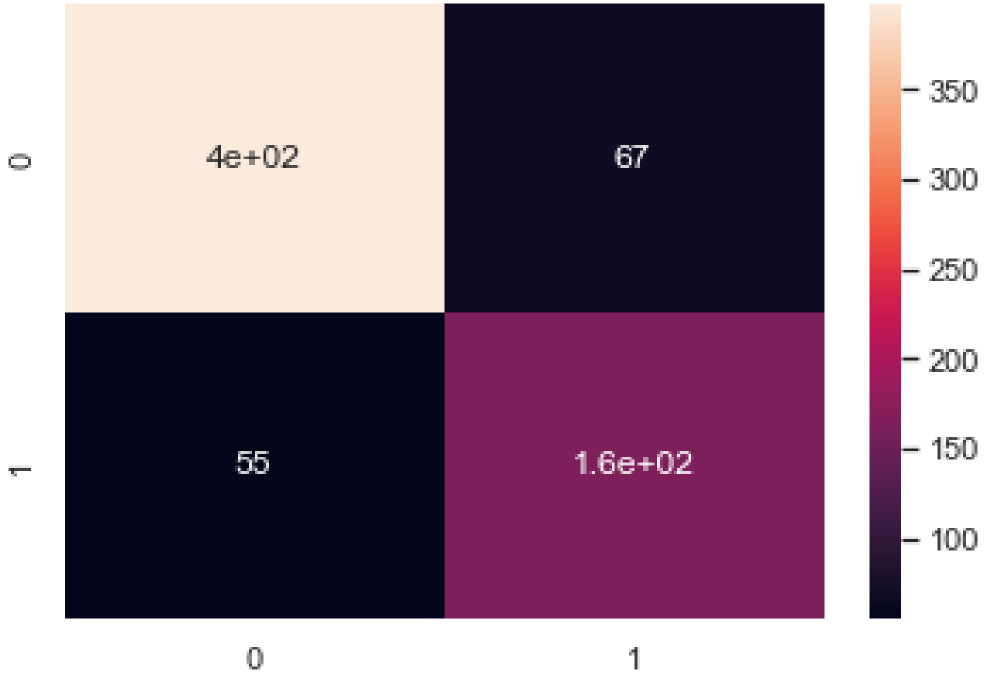


ROC Curve


ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

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

Thank You

Data Glacier

Your Deep Learning Partner