Data Glacier

Your Deep Learning Partner

Healthcare Persistency of a drug (Data Science)

Final Project

25th July 2021

Name: Reeha Khan
Email: *khanreeha22@gmail.com*
Country: Pakistan
College: National University of Sciences and Technology
Specialization: Data Science

Agenda

Executive Summary

Problem Statement

Approach

EDA

Model Deployment

Model Selection

Model Evaluation

Conclusion/Recommendations

Executive Summary

- 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.
- *ML Problem:* 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

Problem Statement

- ABC is a pharmaceutical company that wants to understand the persistency of a drug as per the physician's prescription for a patient. The company has approached an Analytics company to automate this process of identification. This Analytics company has assigned the task to Team Primero to come up with a solution to automate the persistency of a drug for the client ABC.

Business Understanding

- The pharma company ABC wants to understand about the persistency of a drug for each patient. There are a bunch of Non-Tuberculous Mycobacterial (NTM) infection data. The company wants to divide each patient as either persistent or not depending on the prescription data. Depending on the persistency count, ABC pharma company would produce medicines in that quantity so that they can run their business strategically.

Tasks

- Problem understanding
- Data Understanding
- Data Cleaning and Feature engineering
- Model Development
- Model Selection
- Model Evaluation
- Report the accuracy, precision and recall of both the class of target variable
- Report ROC-AUC as well
- Deploy the model
- Explain the challenges and model selection

Approaches taken

- Data was taken from github and analysed
- Problem understanding
- Data Understanding
- Data Cleaning and Feature engineering
- Model Development
- Model Selection
- Model Evaluation

Data Intake Report

- Name: Healthcare – Data Science
- Report date: 25th April 2021
- Internship Batch: LISUM01
- Data storage location: https://github.com/ReehaKhan/Project-Healthcare
- Total number of files: 1
- Total number of features: 26
- Base format of the file: .xlsx
- Size of the data: 898 KB

Analyzing dependency of variable (Before Transformation)

```
Non-Persistent : 62.35 %
Persistent : 37.65 %
```

- The analysis showed non-persistence of drugs was greater than persistence.

Missing Values

Missing Values

```
In [301]: df.isnull().sum()

Out[301]: ptid                                                                 0
persistency_flag                                                     0
gender                                                               0
race                                                                 0
ethnicity                                                            0
region                                                               0
age_bucket                                                           0
ntm_speciality                                                       0
ntm_specialist_flag                                                  0
ntm_speciality_bucket                                                0
gluco_record_prior_ntm                                               0
gluco_record_during_rx                                               0
dexa_freq_during_rx                                                  0
dexa_during_rx                                                       0
frag_frac_prior_ntm                                                  0
frag_frac_during_rx                                                  0
risk_segment_prior_ntm                                               0
tscore_bucket_prior_ntm                                              0
risk_segment_during_rx                                               0
tscore_bucket_during_rx                                              0
change_t_score                                                       0
change_risk_segment                                                  0
adherent_flag                                                        0
idn_indicator                                                        0
injectable_experience_during_rx                                      0
comorb_encounter_for_screening_for_malignant_neoplasms               0
comorb_encounter_for_immunization                                    0
comorb_encntr_for_general_exam_w_o_complaint,_susp_or_reprtd_dx      0
comorb_vitamin_d_deficiency                                          0
comorb_other_joint_disorder_not_elsewhere_classified                 0
comorb_encntr_for_oth_sp_exam_w_o_complaint_suspected_or_reprtd_dx   0
comorb_long_term_current_drug_therapy                                0
comorb_dorsalgia                                                     0
comorb_personal_history_of_other_diseases_and_conditions             0
comorb_other_disorders_of_bone_density_and_structure                 0
comorb_disorders_of_lipoprotein_metabolism_and_other_lipidemias      0
comorb_osteoporosis_without_current_pathological_fracture            0
comorb_personal_history_of_malignant_neoplasm                        0
comorb_gastro_esophageal_reflux_disease                              0
concom_cholesterol_and_triglyceride_regulating_preparations          0
concom_narcotics                                                     0
concom_systemic_corticosteroids_plain                                0
concom_anti_depressants_and_mood_stabilisers                         0
concom_fluoroquinolones                                              0
concom_cephalosporins                                                0
concom_macrolides_and_similar_types                                  0
concom_broad_spectrum_penicillins                                    0
concom_anaesthetics_general                                          0
concom_viral_vaccines                                                0
risk_type_1_insulin_dependent_diabetes                               0
risk_osteogenesis_imperfecta                                         0
risk_rheumatoid_arthritis                                            0
risk_untreated_chronic_hyperthyroidism                               0
risk_untreated_chronic_hypogonadism                                  0
risk_untreated_early_menopause                                       0
risk_patient_parent_fractured_their_hip                              0
risk_smoking_tobacco                                                 0
risk_chronic_malnutrition_or_malabsorption                           0
risk_chronic_liver_disease                                           0
risk_family_history_of_osteoporosis                                  0
risk_low_calcium_intake                                              0
risk_vitamin_d_insufficiency                                         0
risk_poor_health_frailty                                             0
risk_excessive_thinness                                              0
risk_hysterectomy_oophorectomy                                       0
risk_estrogen_deficiency                                             0
risk_immobilization                                                  0
risk_recurring_falls                                                 0
```

- No missing values were found.

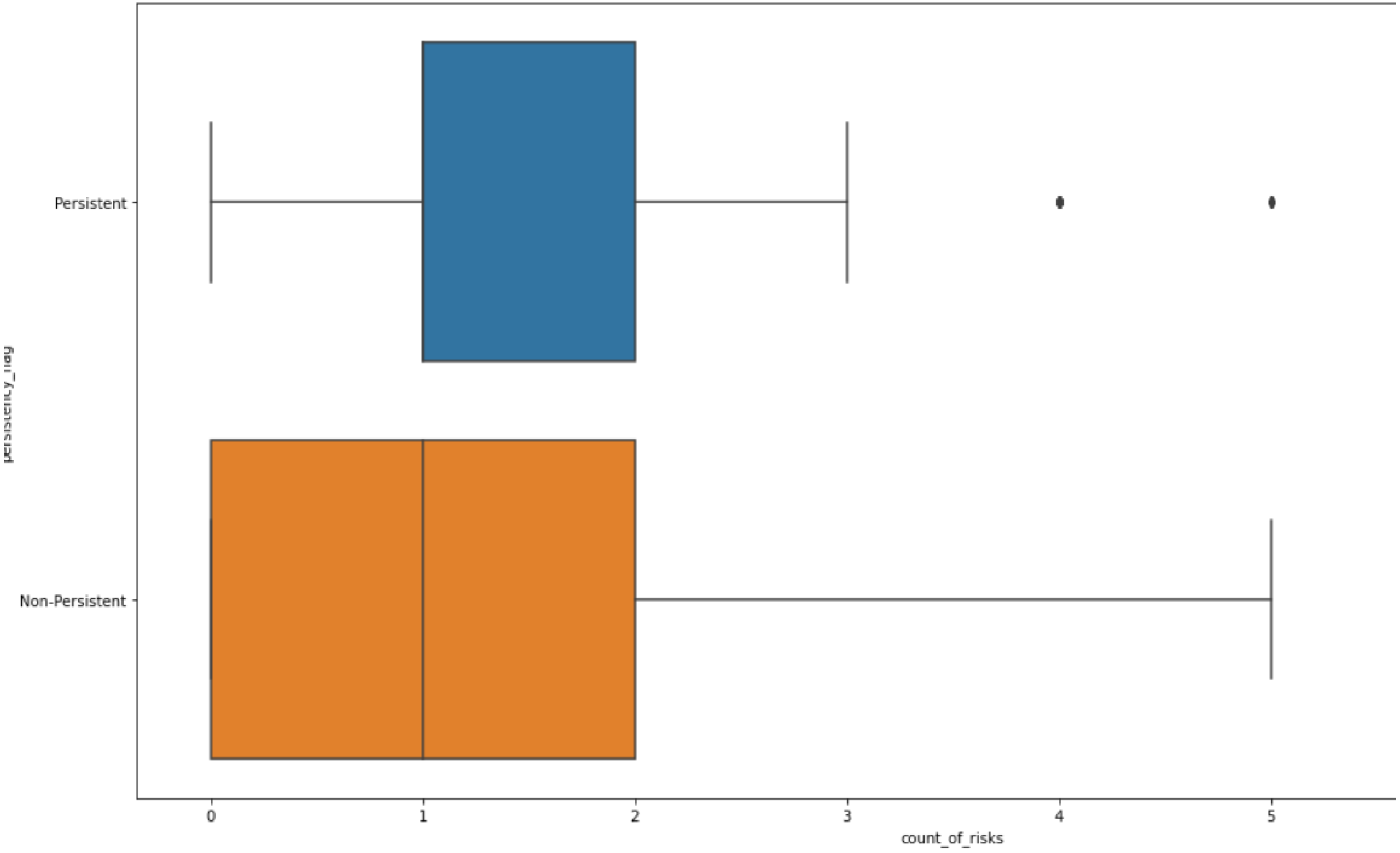
Correlation between features

Out[10]:

	persistency_flag
persistency_flag	1.000000
dexa_during_rx	0.491823
dexa_freq_during_rx	0.395247
comorb_long_term_current_drug_therapy	0.352760
comorb_encounter_for_screening_for_malignant_neoplasms	0.322320
comorb_encounter_for_immunization	0.314887
comorb_encntr_for_general_exam_w_o_complaint,_susp_or_reprtd_dx	0.289828
comorb_other_disorders_of_bone_density_and_structure	0.247283
concom_systemic_corticosteroids_plain	0.242854
comorb_other_joint_disorder_not_elsewhere_classified	0.233279
concom_anaesthetics_general	0.222293
concom_viral_vaccines	0.222241

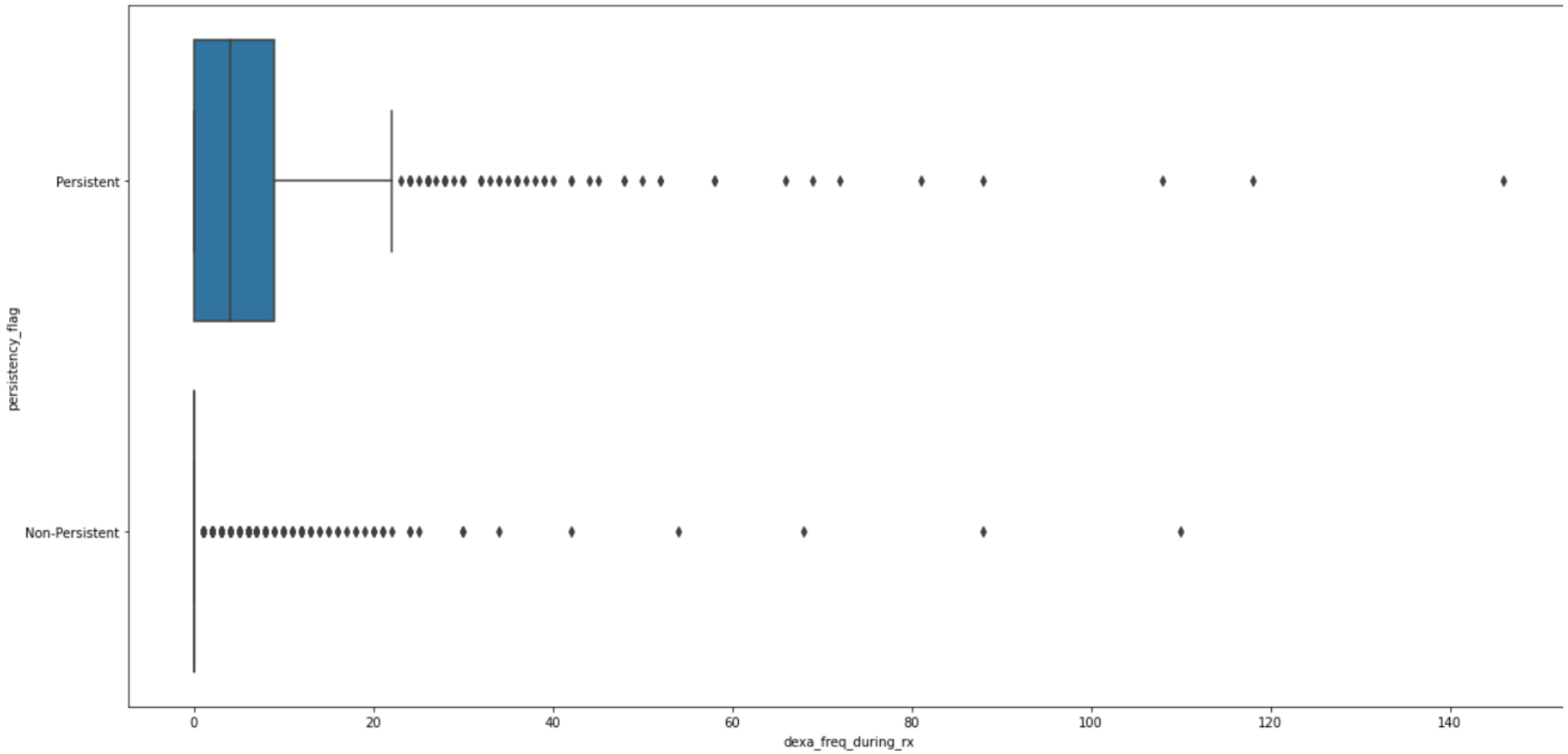
- We try to find the correlation values between the label and other features columns and it turn out that there are many columns having very less correlation value. That means it will be wise to ignore those columns to consider for model training as more number of columns that are unrelated would overfit.

Analysis of Outliners



- Visual analysis showing the outliners in one column by box plot Analysis.

Analysis of Outliners



- Box plot analysis showing the outliners.

Analysis of Skewness and kurtosis

```
Count of risks skewness: 0.8797905232898707
Count of risks Kurtosis: 0.9004859968892842
```

- Data shows a moderate positive skewed data on this column and fairly platykurtic so the data has little outliers.

```
dexa_freq_during_rx skweness: 6.8087302112992285
dexa_freq_during_rx Kurtosis: 74.75837754795428
```

- We can see a very high positive skewed and also with very high kurtosis(Platykurtic). This suggests presence of a lot of outliers.

Analysis showing the standardization of dexa_freq_during_rx df

```
outer range (low) of the distribution:
[[-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]
 [-0.3707352]]

outer range (high) of the distribution:
[[ 7.98784109]
 [ 8.11076133]
 [ 8.47952205]
 [ 9.58580421]
 [10.44624589]
 [10.44624589]
 [12.90465068]
 [13.15049116]
 [14.13385307]
 [17.57561978]]
```

- The distribution shows the low and high range of the distribution of dexa_freq_during_rx.

Analysis of Categorical data description

	ptid	persistency_flag	gender	race	ethnicity	region	age_bucket	ntm_speciality	ntm_specialist_flag	ntm_speciality_bucket	gluco
count	2942	2942	2942	2942	2942	2942	2942	2942	2942	2942	2942
unique	2942	2	2	4	3	5	4	35	2	3	2
top	P2611	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	N
freq	1	2047	2769	2701	2784	1210	1262	1345	1774	1855	2241

- The following analysis shows the distribution of categorical data.

Analysis of Means group-wise

persistency_flag	Non-Persistent	Persistent
dexa_freq_during_rx	0.085491	0.662570
count_of_risks	0.074744	0.155866

gender	Female	Male
dexa_freq_during_rx	0.263874	0.215800
count_of_risks	0.099494	0.098266

	dexa_freq_during_rx	count_of_risks
race		
African American	0.246377	0.168478
Asian	0.135266	0.021739
Caucasian	0.266445	0.098297
Other/Unknown	0.204167	0.125000

- The analysis shows means group wise analysis of persistency, gender and race during administration of dexa and risk count.

Analysis of Means group-wise

ethnicity	Hispanic	Not Hispanic	Unknown
dexa_freq_during_rx	0.279835	0.260417	0.264069
count_of_risks	0.265432	0.097342	0.000000

age_bucket	55-65	65-75	<55	>75
dexa_freq_during_rx	0.242229	0.297880	0.273973	0.242208
count_of_risks	0.118167	0.097039	0.089041	0.093106

ntm_specialist_flag	Others	Specialist
dexa_freq_during_rx	0.215145	0.330765
count_of_risks	0.056370	0.164812

- The analysis shows clearly group-wise analysis according to Ethnicity, Age and NTM specialist during administration of dexa and risk count.

Analysis of Means group-wise cont..

ntm_speciality_bucket	Endo/Onc/Uro	OB/GYN/Others/PCP/Unknown	Rheum
dexa_freq_during_rx	0.442907	0.215274	0.221349
count_of_risks	0.170415	0.053639	0.185658

risk_chronic_liver_disease	N	Y
dexa_freq_during_rx	0.260132	0.452381
count_of_risks	0.096482	0.714286

- Mean group wise analysis of NTM Speciality and Risk due to Chronic liver disease during administration of dexa and risk count.

Analysis of Means group-wise cont..

risk_family_history_of_osteoporosis	N	Y
dexa_freq_during_rx	0.258113	0.287671
count_of_risks	0.045283	0.590753

risk_low_calcium_intake	N	Y
dexa_freq_during_rx	0.261069	0.259259
count_of_risks	0.090502	0.819444

risk_vitamin_d_insufficiency	N	Y
dexa_freq_during_rx	0.223363	0.303468
count_of_risks	-0.175866	0.409321

- Mean group wise analysis of risk due to family history of osteoporosis, risk due to low calcium intake and Risk due to Vitamin D insufficiency during administration of dexa and risk count.

Analysis of Means group-wise

risk_chronic_liver_disease	N	Y
dexa_freq_during_rx	0.260132	0.452381
count_of_risks	0.096482	0.714286

risk_family_history_of_osteoporosis	N	Y
dexa_freq_during_rx	0.258113	0.287671
count_of_risks	0.045283	0.590753

risk_low_calcium_intake	N	Y
dexa_freq_during_rx	0.261069	0.259259
count_of_risks	0.090502	0.819444

- Mean group wise analysis of risk due to chronic liver disease, risk due to family history of osteoporosis and risk due to low calcium intake during administration of dexa and risk count.

Analysis of Means group-wise

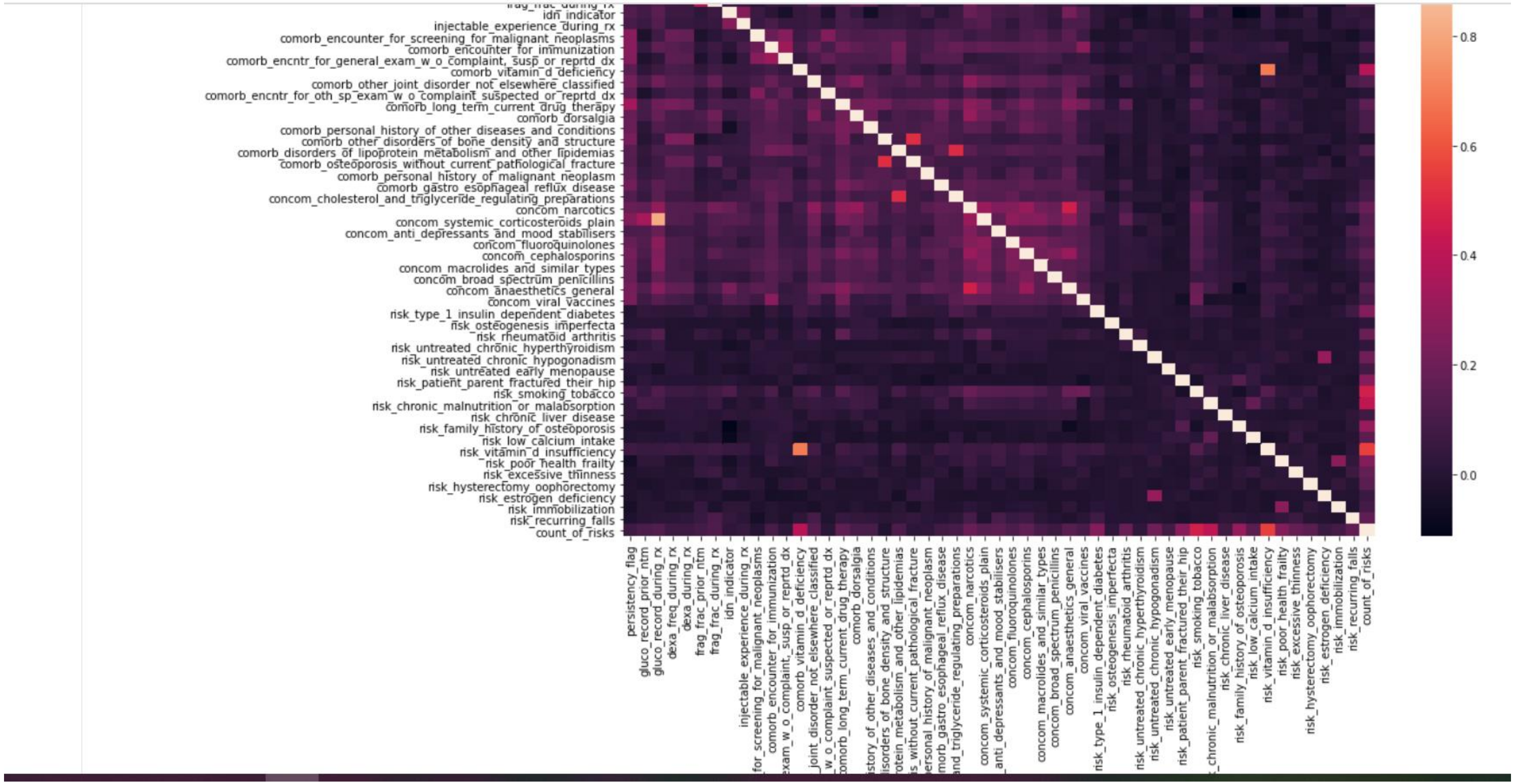
risk_excessive_thinness	N	Y
dexa_freq_during_rx	0.261946	0.218579
count_of_risks	0.085908	0.737705

risk_hysterectomy_oophorectomy	N	Y
dexa_freq_during_rx	0.261650	0.222222
count_of_risks	0.089748	0.722222

risk_immobilization	N	Y
dexa_freq_during_rx	0.262002	0.027778
count_of_risks	0.096416	0.833333

- Mean group wise analysis of risk due to excessive thinness, risk due to hysterectomy oophorectomy and risk due to immobilization during administration of dexa and risk count

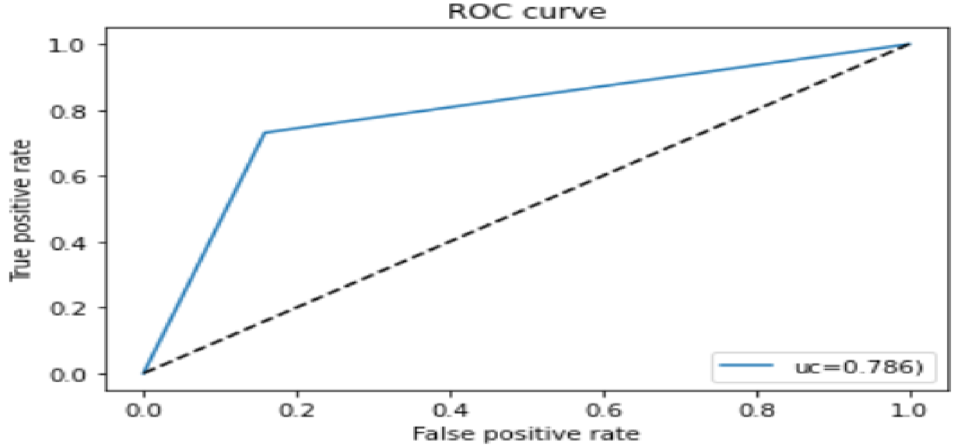
Analyzing dependency of variable (After Transformation)



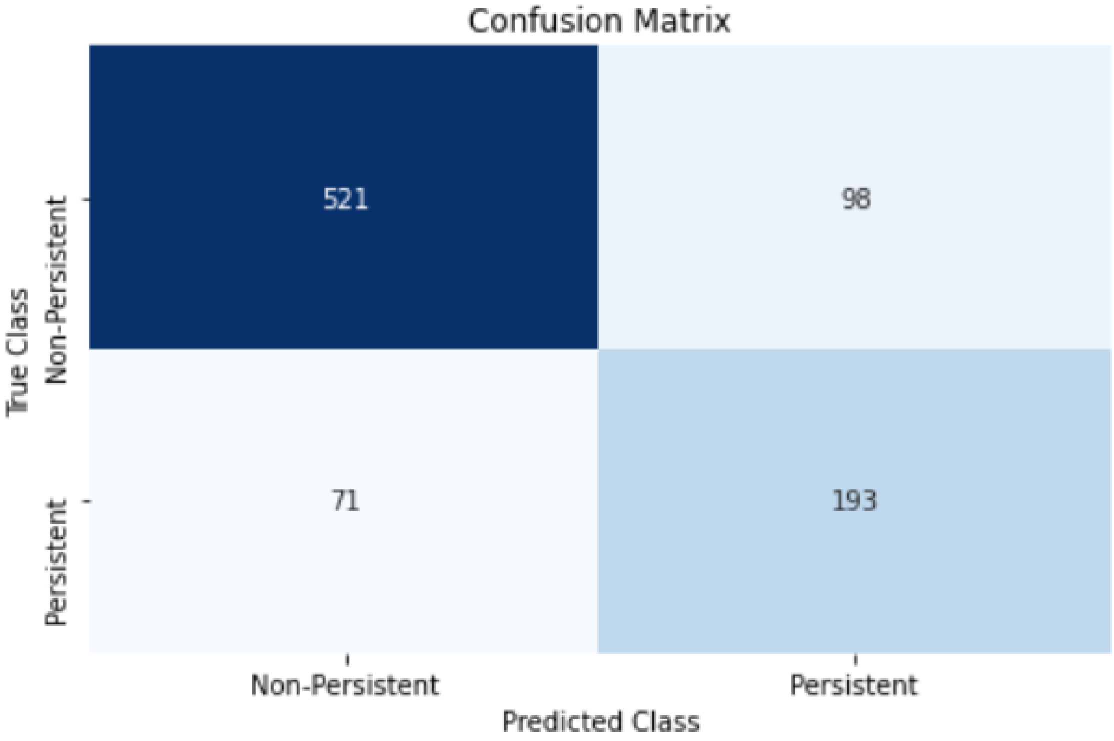
Model Creation

Logistic Regression

```
Accuracy : 0.8086070215175538
Precision : 0.6632302405498282
Recall : 0.7310606060606061
F1 Score : 0.6954954954954955
                precision    recall  f1-score   support

Non-Persistent       0.88      0.84      0.86       619
    Persistent       0.66      0.73      0.70       264

      accuracy                           0.81       883
     macro avg       0.77      0.79      0.78       883
  weighted avg       0.82      0.81      0.81       883

AUC : 0.7863703676506584
```



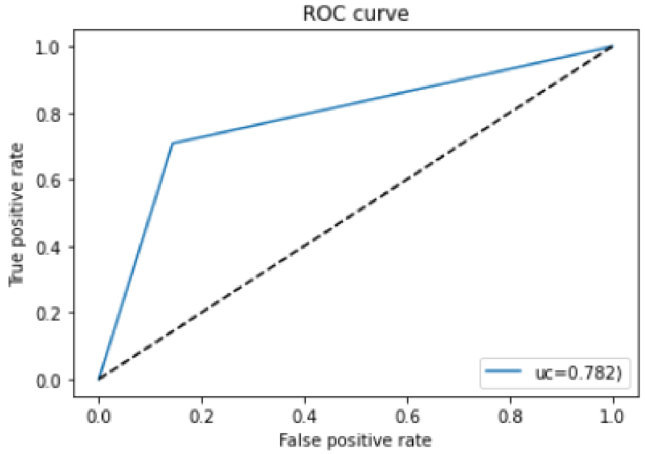
Confusion Matrix

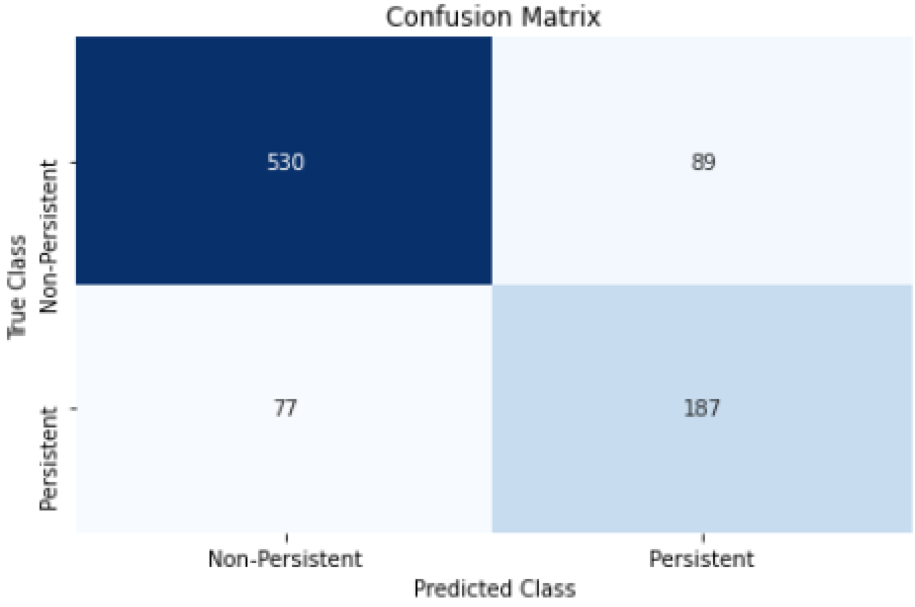


- Logistic Regression Model shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

Ridge Classifier

```
Accuracy : 0.812004530011325
Precision : 0.677536231884058
Recall : 0.7083333333333334
F1 Score : 0.6925925925925926
              precision    recall  f1-score   support

Non-Persistent       0.87      0.86      0.86       619
    Persistent       0.68      0.71      0.69       264

      accuracy                           0.81       883
     macro avg       0.78      0.78      0.78       883
  weighted avg       0.81      0.81      0.81       883

AUC : 0.782276521270867
```

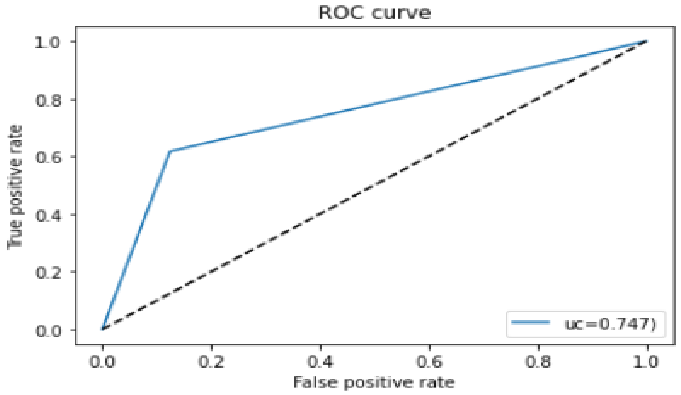


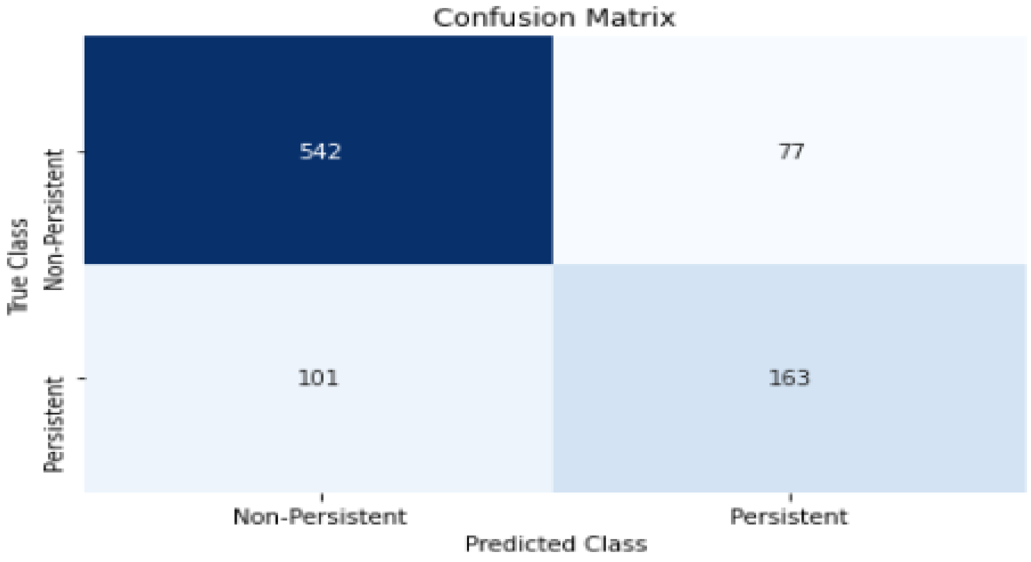


- Ridge Classifier Model shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

SGD Classifier

```
Accuracy : 0.79841449603624
Precision : 0.6791666666666667
Recall : 0.6174242424242424
F1 Score : 0.6468253968253969
                precision    recall  f1-score   support

Non-Persistent       0.84      0.88      0.86       619
    Persistent       0.68      0.62      0.65       264

      accuracy                           0.80       883
     macro avg       0.76      0.75      0.75       883
  weighted avg       0.79      0.80      0.80       883

AUC : 0.7465150291281147
```

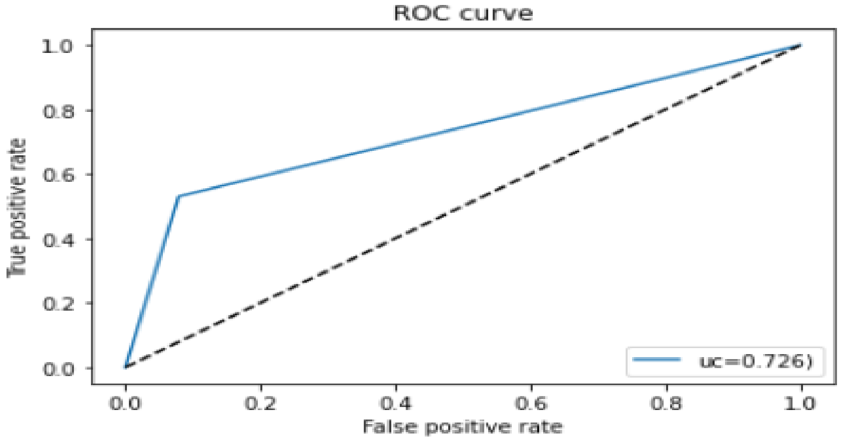


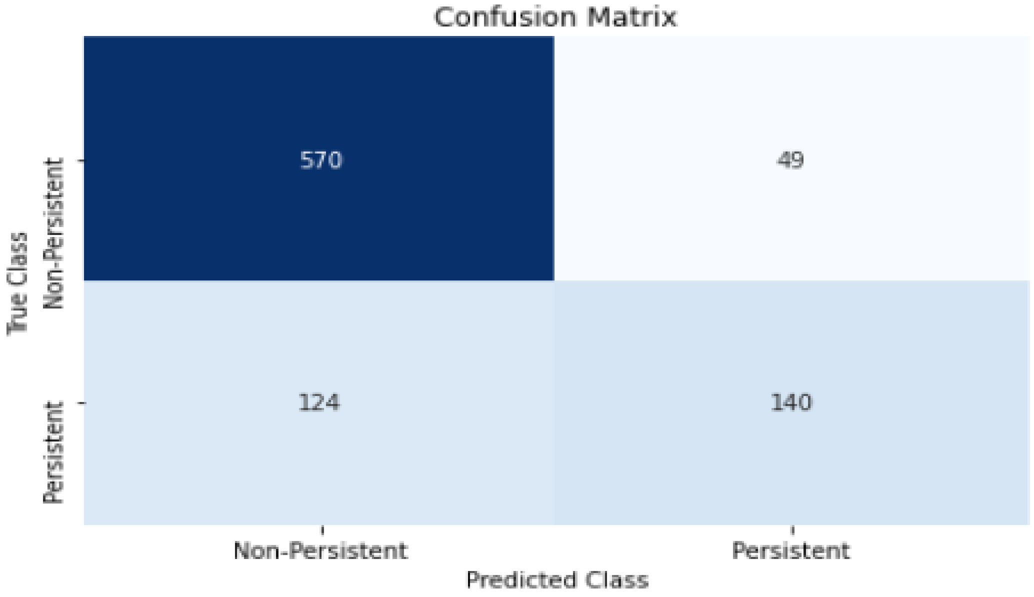


- SGD Classifier Model shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

Random Forest Classifier

```
Accuracy : 0.8040770101925255
Precision : 0.7407407407407407
Recall : 0.5303030303030303
F1 Score : 0.6181015452538631
                precision    recall  f1-score   support

Non-Persistent       0.82      0.92      0.87       619
    Persistent       0.74      0.53      0.62       264

      accuracy                           0.80       883
     macro avg       0.78      0.73      0.74       883
  weighted avg       0.80      0.80      0.79       883

AUC : 0.7255715474616928
```

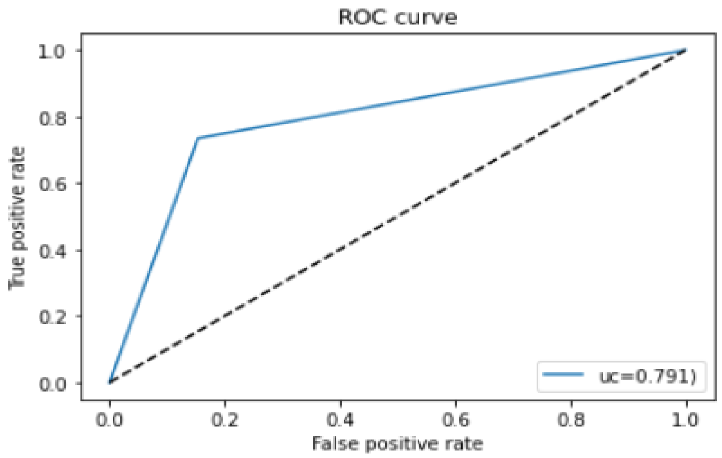


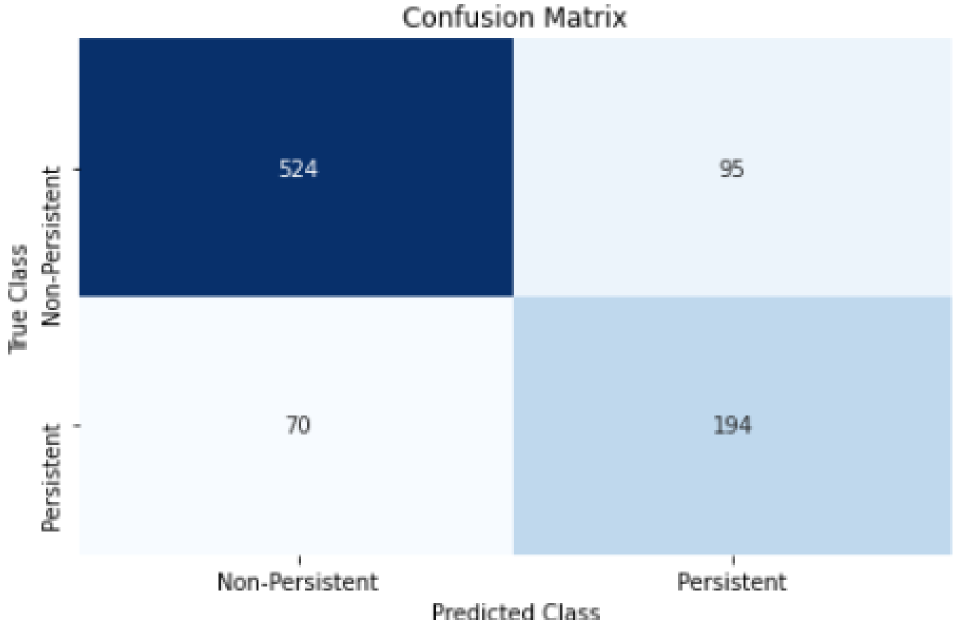


- Random Forest Classifier Model shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

Ada Boost Classifier

```
Accuracy : 0.8131370328425821
Precision : 0.671280276816609
Recall : 0.7348484848484849
F1 Score : 0.701627486437613
               precision    recall  f1-score   support

Non-Persistent       0.88      0.85      0.86       619
    Persistent       0.67      0.73      0.70       264

      accuracy                           0.81       883
     macro avg       0.78      0.79      0.78       883
  weighted avg       0.82      0.81      0.82       883

AUC : 0.7906875703725462
```

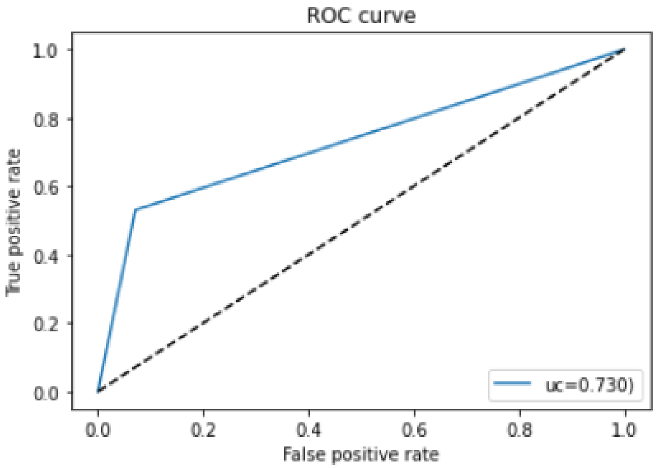


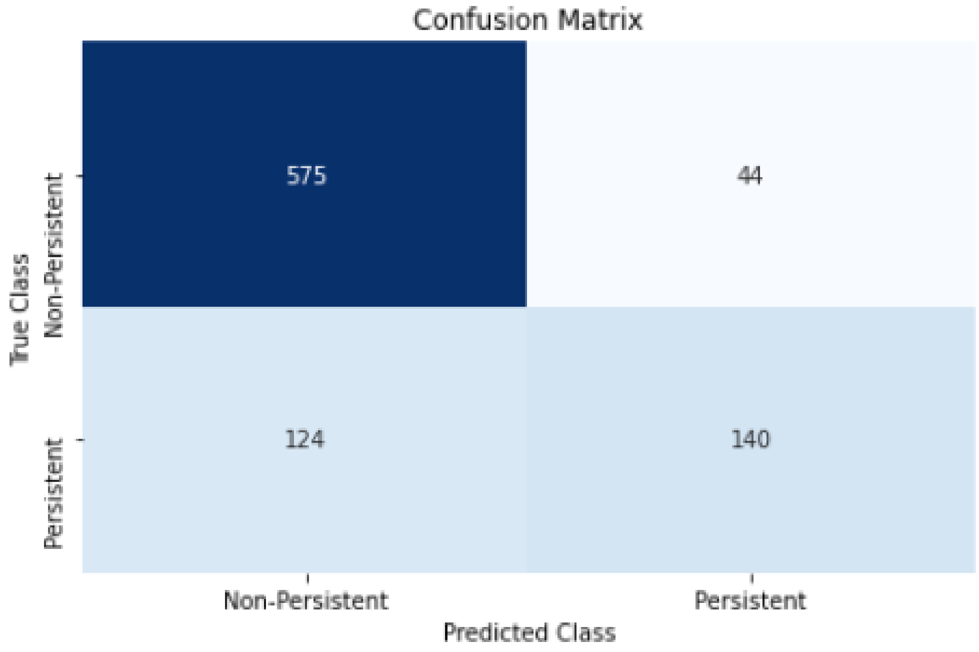


- Ada boost classifier shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

Stacking Classifier

```
Accuracy : 0.8097395243488109
Precision : 0.7608695652173914
Recall : 0.5303030303030303
F1 Score : 0.625
                precision    recall  f1-score   support

Non-Persistent       0.82      0.93      0.87       619
    Persistent       0.76      0.53      0.62       264

      accuracy                           0.81       883
     macro avg       0.79      0.73      0.75       883
  weighted avg       0.80      0.81      0.80       883

AUC : 0.7296103196749399
```

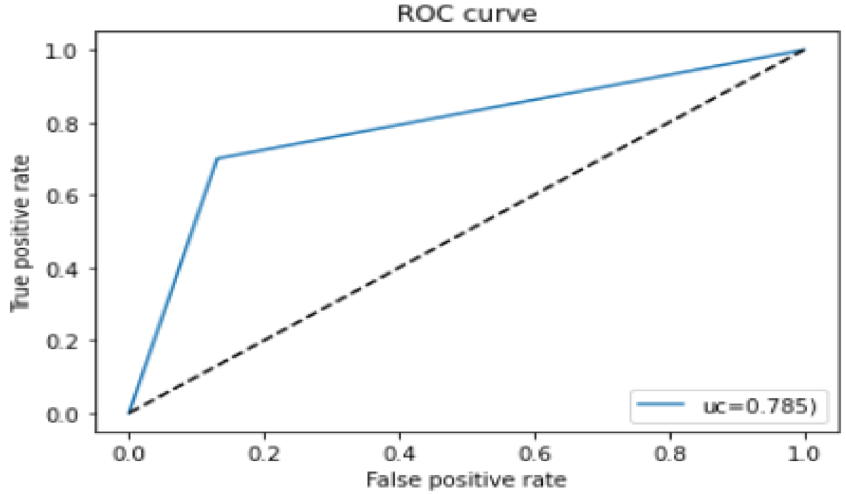


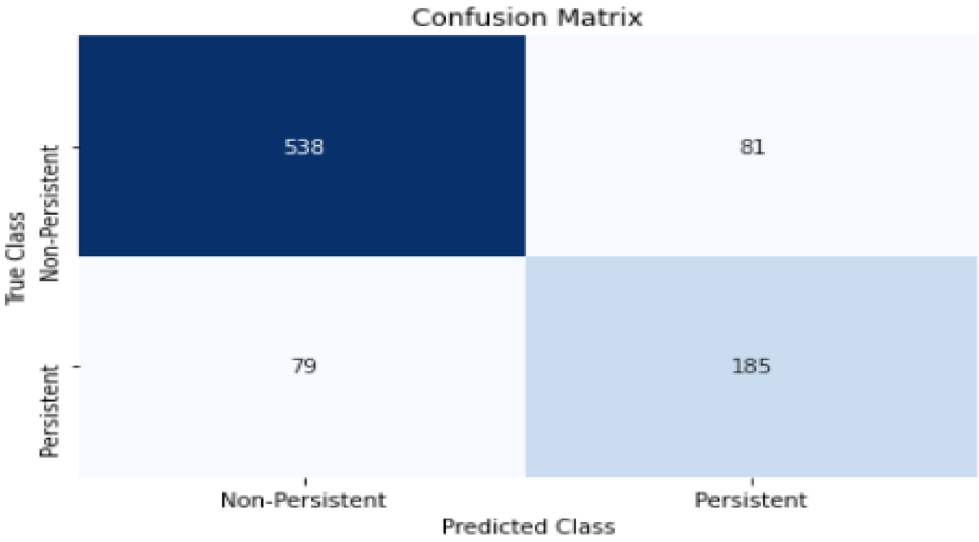


- Stacking Classifier Model shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

XG Boost Classifier

```
Accuracy : 0.8187995469988675
Precision : 0.6954887218045113
Recall : 0.7007575757575758
F1 Score : 0.6981132075471698
                precision    recall  f1-score   support

Non-Persistent       0.87      0.87      0.87       619
    Persistent       0.70      0.70      0.70       264

      accuracy                           0.82       883
     macro avg       0.78      0.78      0.78       883
  weighted avg       0.82      0.82      0.82       883

AUC : 0.7849506780241836
```





- XG Boost Classifier Model shows the Accuracy, Recall, Precision, f1 score and support of Non-Persistent and Persistence of drugs.

Conclusion

- There is not much of a significant difference between all the classifiers, however, the best can be considered to be:

1. RidgeClassifier (Linear)
2. AdaBoostClassifier (Ensemble/Boosting)
3. XGBoostClassifier (Ensemble/Boosting)

- All of these classifiers have almost

81% Accuracy, 68% Precision, 71% Recall, 70% F1 Score and 78% AUC

Thank You