Data Science Internship

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

Week 12 deliverables

Group name: Gold Standard Team
Names:
1) Harshith Sakala Santhosh
2) Mario Rodriques Peres
3) Alexis Michael-Igbokwe
Email:sakala.harshith@gmail.com
Country: United Kingdom
Specialization: Data Science
Report date: 31st of May 2023
Internship Batch: LISUM19
Project home: GitHub

Contents

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. Analyzing 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 Assigned for these tests are 0.05.

4.1.1. 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.

4.1.2. 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 the 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:

- Dexa_During_Rx
- Comorb_Long_Term_Current_Drug_Therapy
- Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
- Comorb_Encounter_For_Immunization
- Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx
- Comorb_Other_Disorders_Of_Bone_Density_And_Structure ,
- Concom_Systemic_Corticosteroids_Plain
- Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified
- Concom_Anaesthetics_General
- Concom_Viral_Vaccines

4.2. Elimination of unwanted and Similar correlated Variables

It is always better to remove features that have weak correlation with output label and removing a few features as they have high correlation with other features as this increases speed of training and most importantly it increases model's accuracy. There are features that constitute high correlation with others and slightly less correlation with target variables. 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:**
   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, 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_Structure,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.

5. Feature Selection and Modeling

This section is all about Types of feature selection and model techniques employed for data modeling.

5.1. Feature Selection :

Several Techniques were used for feature selection they are as follows:

Selecting features based on EDA and Statistical Tests (Chisquare, Z and T Tests) against SVM, Random Forest, XGBoost and Dense Neural Network:

As mentioned above, the features selected based on calculating ChiSquare analysis between categorical variable and target variable and within themselves, these variables were selected in such a way that they possess high correlation with target variable and weak correlation among themselves. In addition, Z-test and T-test were performed between Numerical variables and Categorical variables. It was found out that all numerical variables had a good correlation with other categorical variables. So, to make data homogeneous all numerical variables were excluded from final feature selection.

The EDA features are as follows:
'Ntm_Speciality_Bucket'
'Dexa_During_Rx'
'Tscore_Bucket_During_Rx'
'Risk_Segment_During_Rx'
'Gluco_Record_During_Rx'
'Frag_Frac_Prior_Ntm'
'Adherent_Flag'
'Risk_Rheumatoid_Arthritis'
'Risk_Untreated_Chronic_Hypogonadism'
'Risk_Smoking_Tobacco'
'Risk_Excessive_Thinness'

'Risk_Immobilization'
'Comorb_Long_Term_Current_Drug_Therapy'
'Concom_Systemic_Corticosteroids_Plain'

Features selected from SelectKBest using Mutual Information against SVM, Random Forest, XGBoost and Dense Neural Networks

The features in this category were selected by utilizing a method called 'SelectKBest' Method using Mutual_Info statistical tests to calculate correlation of all input variables with target variables. Top or Best 20 variables were selected based on performance of all models. They are as follows:

'Ntm_Speciality'
'Dexa_During_Rx'
'Frag_Frac_During_Rx'
'Tscore_Bucket_During_Rx'
'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms'
'Comorb_Encounter_For_Immunization'
'Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx'
'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified'
'Comorb_Long_Term_Current_Drug_Therapy'
'Comorb_Dorsalgia'
'Comorb_Personal_History_Of_Other_Diseases_And_Conditions'
'Comorb_Other_Disorders_Of_Bone_Density_And_Structure'
'Comorb_Osteoporosis_without_current_pathological_fracture'
'Comorb_Gastro_esophageal_reflux_disease'
'Concom_Systemic_Corticosteroids_Plain'
'Concom_Cephalosporins'
'Concom_Macrolides_And_Similar_Types'
'Concom_Broad_Spectrum_Penicillins'
'Concom_Anaesthetics_General'
'Concom_Viral_Vaccines'

5.2.Model Development and Selection

A few models were tested during this stage. There were tested the Linear Regression, Random Forest, SVM, ADABoost, XGBoost and Dense Neural Network. Each member of the team tried different approaches and a single Jupiter file was prepared, which included Best three models performed across different feature selection methods they are:

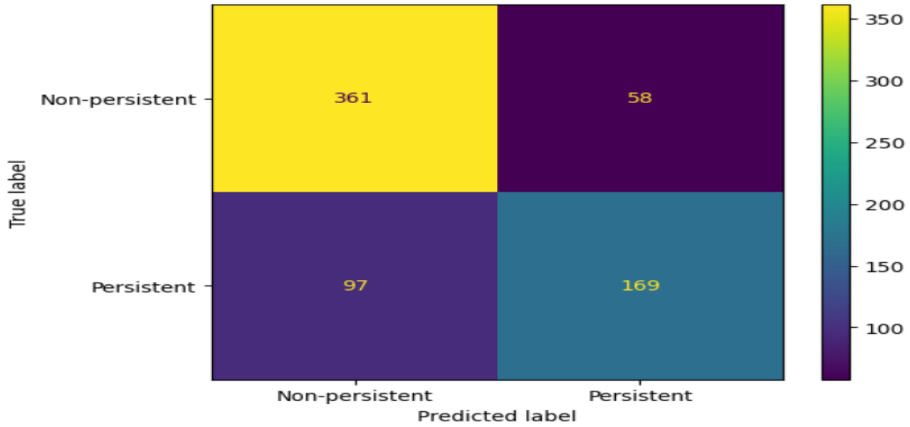
1) SVC
2) Random Forest Classifier
3) Dense Neural Network

5.2.1. **Model performance using EDA Features:**

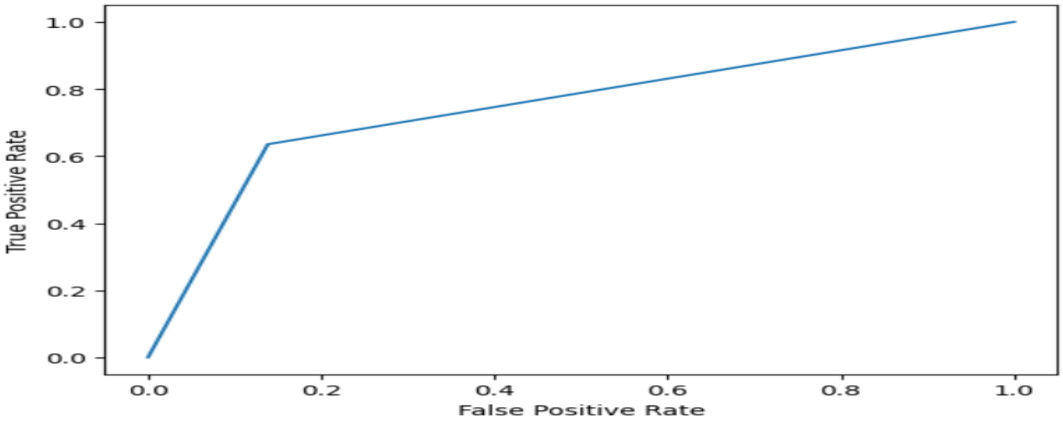
As Above Mentioned, models and in addition to these we have also used XGBoost to understand models performances. It is observed that all models produced similar accuracies, but F1-score were less compared to their accuracies and the reason for that was models have low recall compared to precision of that model. So, all models are underperforming for 'Persistent' class the possible reason is that the number of instances is very less compared to majority class instance and models are getting little biased towards major class leading to less Recall and F1 Score. The performance of models are

as follows.

5.2.1.1. SVM Performance:



Confusion Matrix of SVC

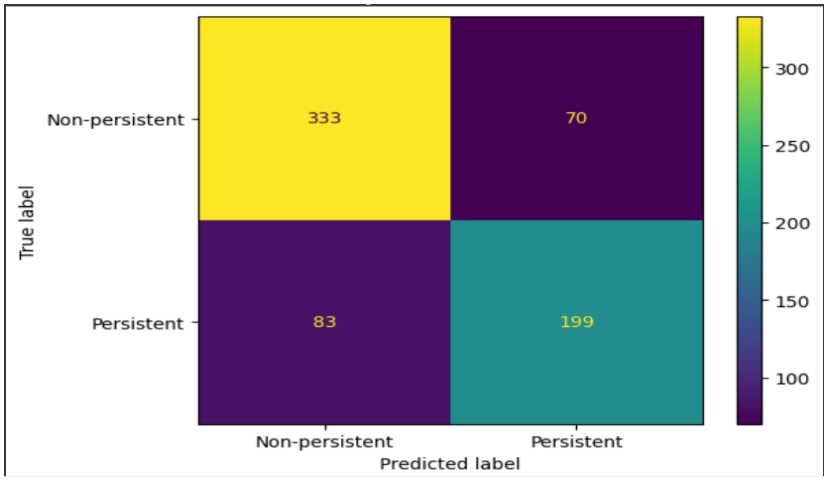


ROC-AUC Curve of SVC

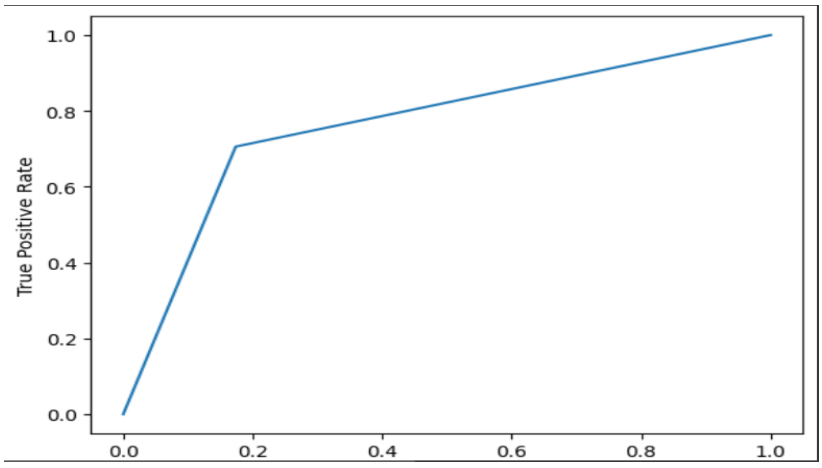
```
{'C': 100, 'gamma': 0.1, 'kernel': 'rbf'}
SVC(C=100, class_weight={0: 0.4, 1: 0.6}, gamma=0.1)
0.6855983772819473
The accuracy of the model is : 0.7737226277372263
The F1-Score of the model is:  0.6855983772819473
The Precision of the model is : 0.7444933920704846
The ROC-AUC Score for the model is 0.7484567624311376
```

F1-Score, Accuracy, Precision of SVC

5.2.1.2. Random Forest Performance:



Confusion Matrix of Random Forest



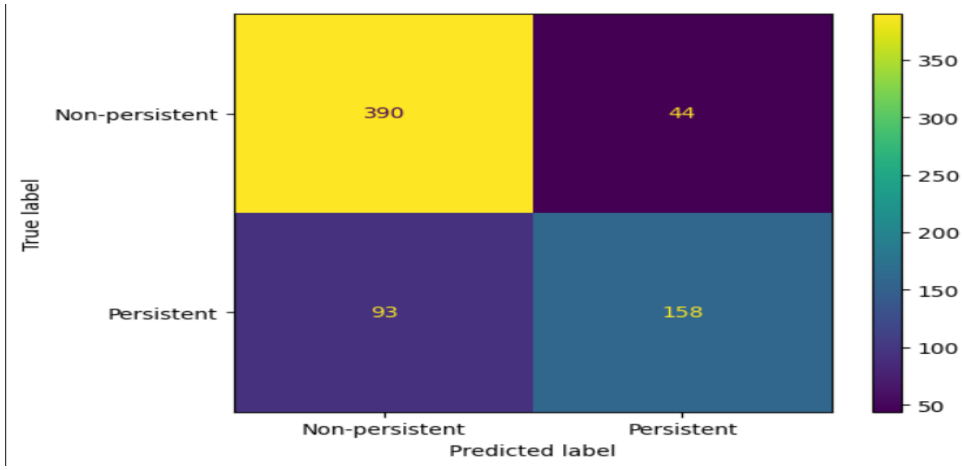
ROC-AUC of Random Forest

```
{'n_estimators': 50, 'min_samples_split': 10, 'min_samples_leaf': 6, 'max_features': 'auto', 'max_depth': 70}
RandomForestClassifier(max_depth=70, max_features='auto', min_samples_leaf=6,
                       min_samples_split=10, n_estimators=50)
0.7766423357664234
The accuracy of the model is : 0.7766423357664234
The F1-Score of the model is:  0.722323049001815
The Precision of the model is : 0.7397769516728625
The ROC-AUC Score for the model is 0.765988244196892
```

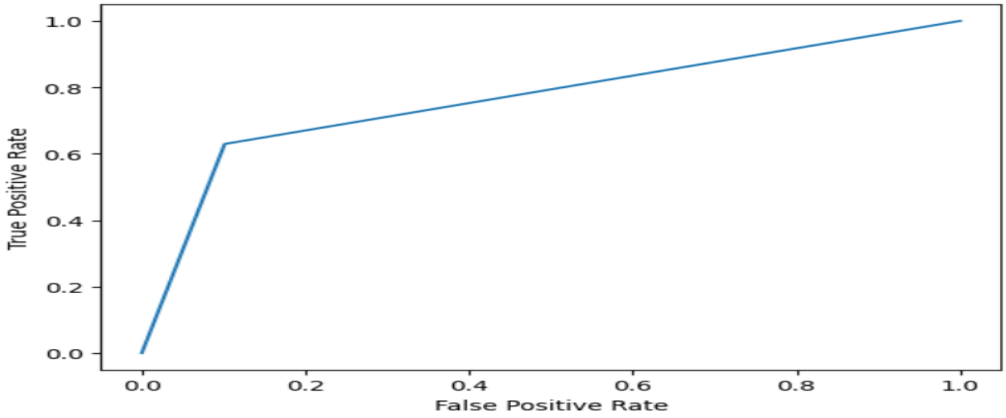
F1-Score, Accuracy, Precision of Random Forest

Note: It is observed that, Random Forest seems to be performing better than others in this category with balance across precision, recall and accuracy.

5.2.1.3. XGBoost Performance:



Confusion Matrix of XGBoost



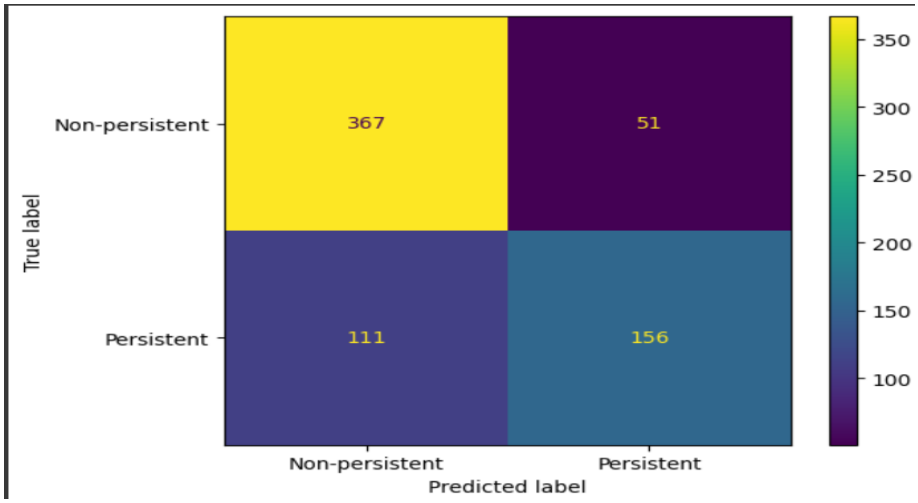
ROC-AUC Curve of XGBoost

```
The accuracy of the model is : 0.8
The F1-Score of the model is:  0.6975717439293597
The Precision of the model is : 0.7821782178217822
The ROC-AUC Score for the model is 0.7640497916169423
```

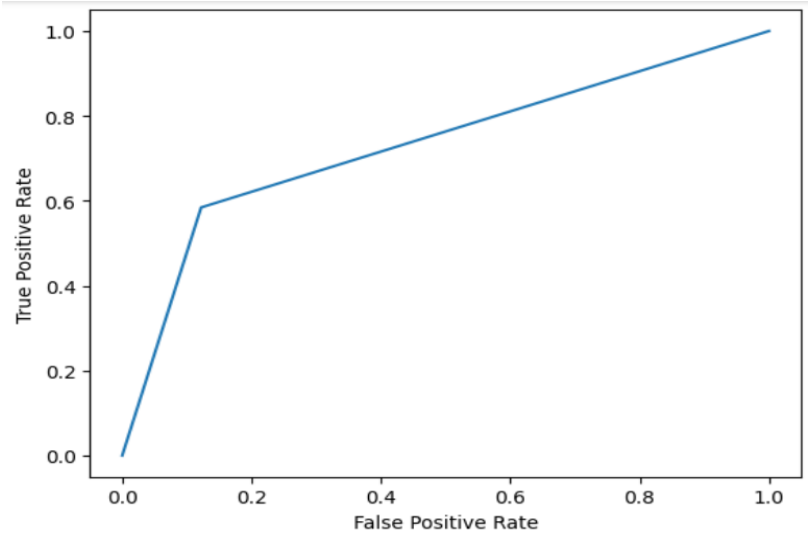
Accuracy, Precision and F1-Score of XGBoost

Note: Though the accuracy of this model is little better than other models, the recall of the model is the second lowest among others in this category. So, it is not an ideal model to go for.

5.2.1.4. Dense Neural Network Performance:



Confusion Matrix of Dense Neural Network



ROC-AUC Curve of Dense Neural Network

```
The F1-score of the Dense Neural Network is  0.6582278350218616
The accuracy of the model is : 0.7635036496350365
The F1-Score of the model is:  0.6582278481012659
The Precision of the model is : 0.7536231884057971
The ROC-AUC Score for the model is 0.731130046771679
```

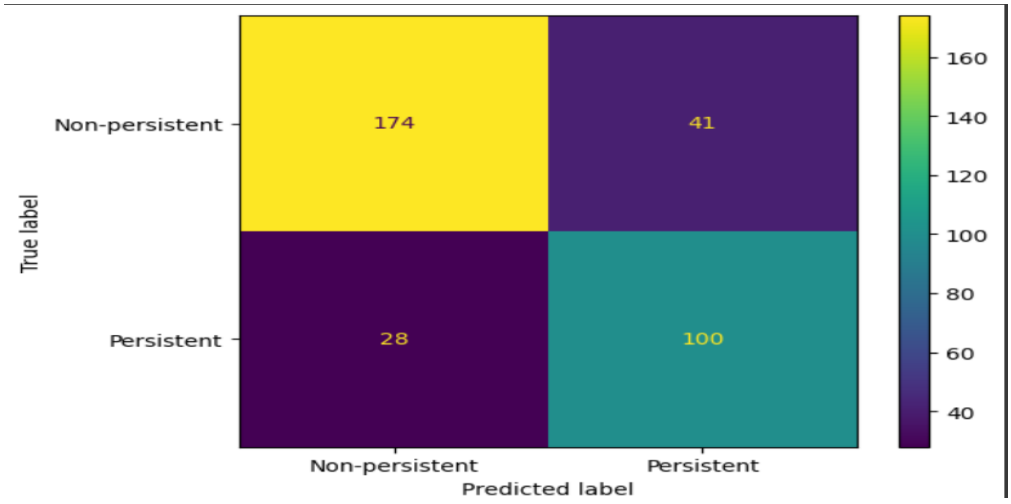
Accuracy, Precision and F1-Score of Dense Neural Network

Note: It is observed that the model though has similar accuracy to others, but it has the lowest F1-score in this category and the reason is having the lowest Recall Score. So, all models are biased towards majority class in prediction.

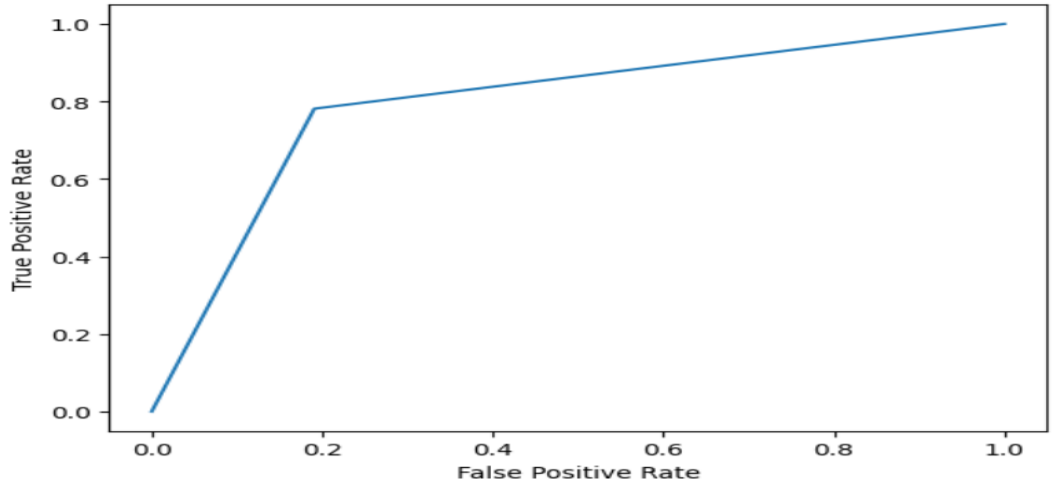
5.2.2. **Model Performance using SelectKBest Features:**

Top 20 features have been selected for data modelling. All three models mentioned above utilized these features for data training and it was found that all models seem to perform better than previous category models. Though the accuracy didn't increase much but the balance across Recall and Precision is better, leading to a good F1-Score. Random Forest is the best performer in both categories and has better accuracy and better F1 Score. So, considering all these Random Forest Classifier using SelectKBest features will be the pick for model training and prediction and this model is deployed in AWS EC2 as a Fast-API.

5.2.2.1 SVC Performance:



Confusion Matrix of SVC



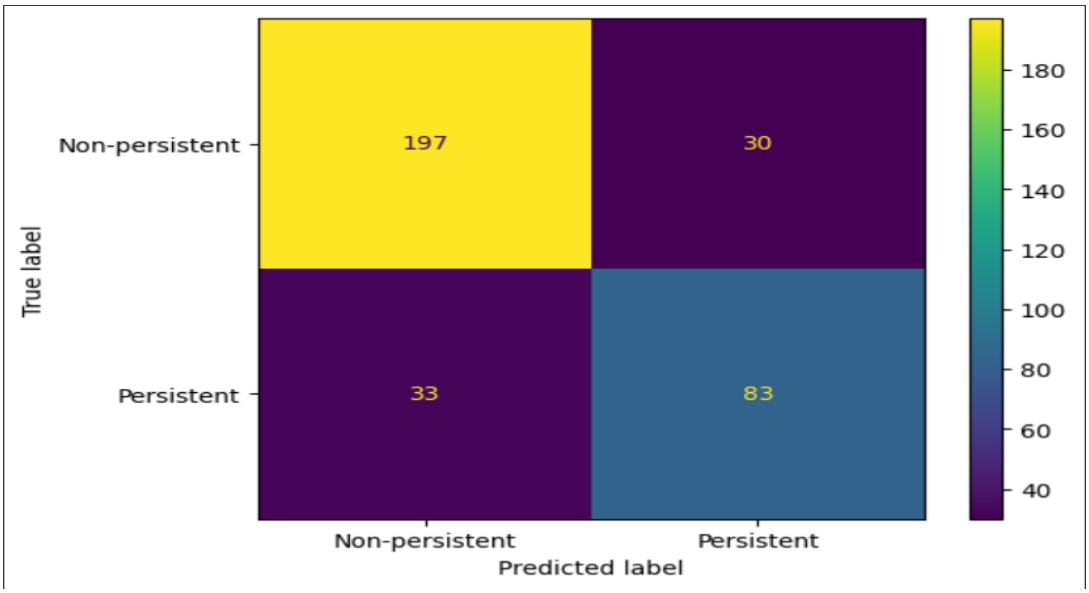
ROC-AUC Curve of SVC

```
The accuracy of the model is : 0.7988338192419825
The F1-Score of the model is:  0.7434944237918215
The Precision of the model is : 0.7092198581560284
The ROC-AUC Score for the model is 0.7952761627906978
```

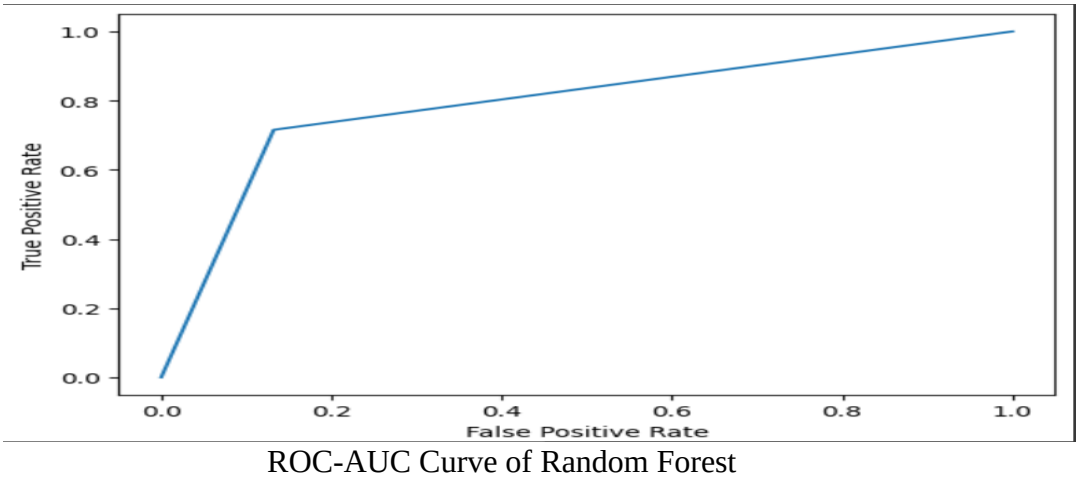
Accuracy, Precision and F1-score of SVC

Note: It is observed that accuracy is almost 80% which is better than other models of previous category and it has a good recall as well, Precision seems to be little less than recall but the balance between precision and recall is better than those previous models. So, it has led to a good F1-Score.

5.2.2.2 Random Forest Performance:



Confusion Matrix of Random Forest



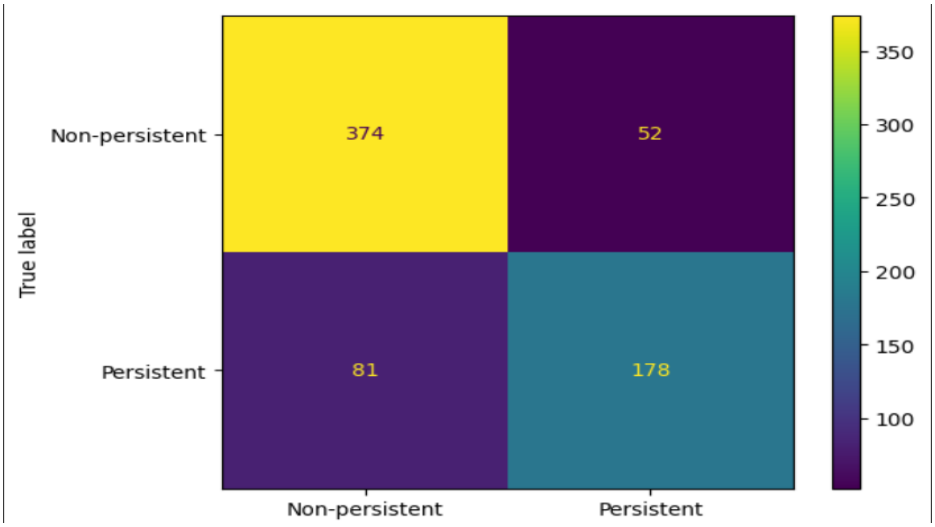
ROC-AUC Curve of Random Forest

```
The accuracy of the model is : 0.8163265306122449
The F1-Score of the model is:  0.7248908296943231
The Precision of the model is : 0.7345132743362832
The ROC-AUC Score for the model is 0.7916793255354703
```

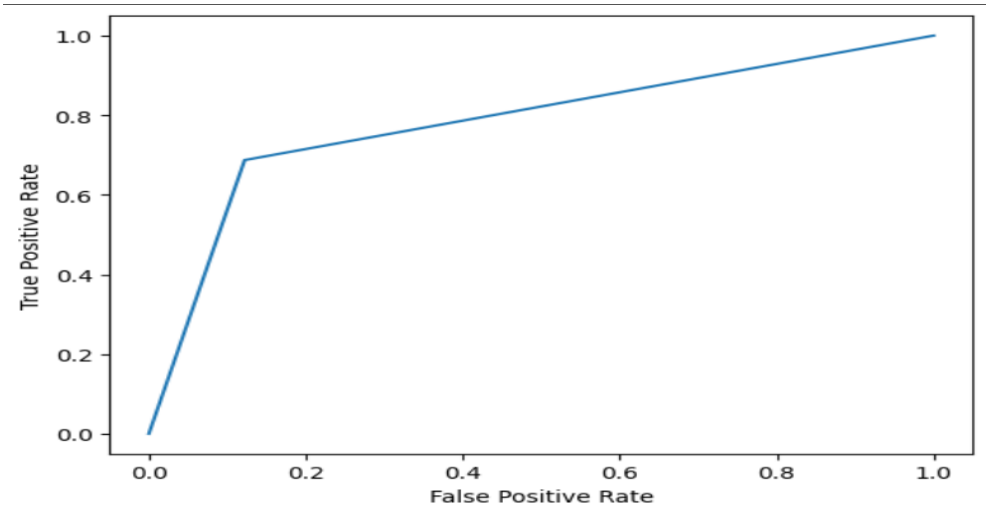
Accuracy, Precision and F1-Score of Random Forest

Note: As mentioned before, the model is the best performer amongst all and the reason is obvious it has best accuracy, one of the best Precision and Recall. As the model is not biased towards majority class and seems like giving more equal or more priority to the minority class.

5.2.2.3. Dense Neural Network Performance:



Confusion Matrix of Dense Neural Network



ROC-AUC Curve of Dense Neural Network

```
The F1-score of the Dense Neural Network is  0.7280163372724187
The accuracy of the model is : 0.8058394160583942
The F1-Score of the model is:  0.7280163599182004
The Precision of the model is : 0.7739130434782608
The ROC-AUC Score for the model is 0.7825964797795784
```

Accuracy, Precision and F1-Score of Dense Neural Network

Note: The model seems to have a good accuracy and Precision but recall seems to be little than precision but it is better than previous category models.

6. Recommendations

- It is observed that 62% of data is Non-persistent and rest is persistent and in addition to that input features are not properly distributed among themselves. Due to this model got biased a bit towards majority class variable. Though best model has accuracy of 82% that may have low bias but for this data but the chances that it may have more variance for other training and dataset. So it is better to have more data and a balance of class variables as much as possible.

- It is clearly visible that Comorb Features are very much important to decide the persistency of drug but given collinearity among themselves leads to having more redundant or un-wanted information and it is better to include those features that better include diversity in dataset and lead to better prediction.

- The number of records compared to columns are less which is not ideal for model training which led to accuracy less than 90%.

- There was a data imbalance observed and even after trying different oversampling, undersampling methods the model training didn't seem to improve much and also there are features like race, Region etc... where the most of the data is concentrated on one type of category which may not be ideal for output predictions as model would be more suitlable for one particular type of data or gets biased during prediction. It is always recommended to include features that have even distribution if possible among all variables.

Note: The best performer model i.e., Random Forest Classifier has been deployed on AWS EC2 as an API using Fast-API. It is available to access through this link :**http://35.179.88.8/docs**