Persistency of a drug

Project Report

Name: Amima Shifa

University: Osmania University

Country: India

Specialization: Data Science

Batch: LISUM12

Submitted to: Data Glacier

Table of Contents:

1. Project Plan..3

2. Problem Statement...3

3. Data Understanding..3

4. Data Preprocessing………………………………….........................6

5. Building the Models.………………….……………………………...7

6. Evaluation of Models…..……………………………………….......7

7. Model Selection ………………………..………………………...9

8. Conclusion………………………………………………….…11

1. Project Plan:

TASKS	17th Sept Week 0	24th Sept Week 1	1st Oct Week 2	8th Oct Week 3	15th Oct Week 4	21st Oct Week 5	30th Oct Week 6
Week 7	🟩						
Week 8		🟩					
Week 9			🟩				
Week 10				🟩			
Week 11					🟩		
Week 12						🟩	
Week 13							🟩

2. Problem Description:

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

3. Data Understanding:

The dataset contains 70 columns and 3424 rows in total. The target variable is Persistency_Flag which is a binary variable having either True or False as its value. One of the features of the given dataset is 'Ptid' that is patient identity number and it plays no role in prediction of persistence of a drug on a patient and hence is removed. Overall the datatypes of most of the features is either binary or string.

3.1. Feature Description:

Bucket	Variable	Variable Description
Unique Row Id	Patient ID	Unique ID of each patient
Target Variable	Persistency_Flag	Flag indicating if a patient was persistent or not
Demographics	Age	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	Ethnicity	Ethnicity of the patient from the patient table
	Gender	Gender of the patient from the patient table
	IDN Indicator	Flag indicating patients mapped to IDN
Provider Attributes	NTM - Physician Specialty	Specialty of the HCP that prescribed the NTM Rx
Clinical Factors	NTM - T-Score	T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate)
	Change in T Score	Change in Tscore before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Risk Segment	Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate)

	Change in Risk Segment	Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown)
	NTM - Multiple Risk Factors	Flag indicating if patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate)
	NTM - Dexa Scan Frequency	Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)
	NTM - Dexa Scan Recency	Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable)
	Dexa During Therapy	Flag indicating if the patient had a Dexa Scan during their first continuous therapy
	NTM - Fragility Fracture Recency	Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate)
	Fragility Fracture During Therapy	Flag indicating if the patient had fragility fracture during their first continuous therapy
	NTM - Glucocorticoid Recency	Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx
	Glucocorticoid Usage During Therapy	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy

Disease/Treatment Factor	NTM - Injectable Experience	Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx
	NTM - Risk Factors	Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx
	NTM - Comorbidity	Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied
	NTM - Concomitancy	Concomitant drugs recorded prior to starting with a therapy(within 365 days prior from first rxdate)
	Adherence	Adherence for the therapies

4. Data Preprocessing:

The dataset contains 70 columns and 3424 rows in total. One of the features of the given dataset is ‘Ptid’ which is a patient identification number and it plays no role in the prediction of the persistence of a drug on a patient and hence is removed. There are no missing values present in any column of the dataset. 'risk_segment_during_rx', 'tscore_bucket_during_rx', 'change_t_score' and 'change_risk_segment' contained unknown values, so these columns were dropped from the dataset. Binary values present in the dataset were mapped from Y, N to 1,0 respectively. A lot of outliers were detected in Dexa Frequency during RX which were then transformed using log transformation. Up sampling was done to increase the records of the minority class, to have same count of records of each class.

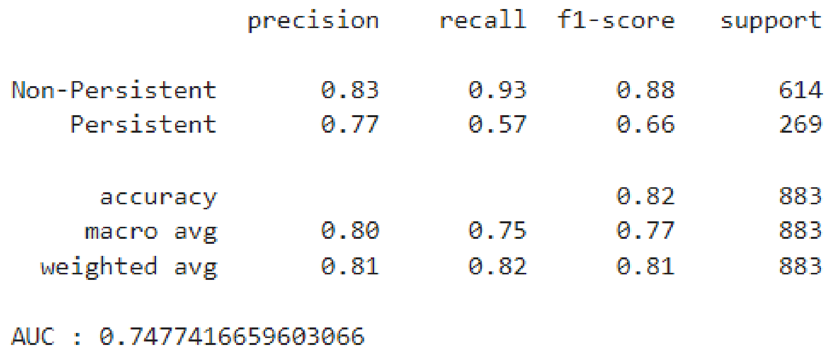
5. Building the Models:

For building models, the dataset was split into training and testing sets. Various regression techniques were used to perform classification to determine the persistance of the drug. The models trained and tested include Logistic Regression, Random Forest Classifier, Gradient Boosting Model, XGBoost and AdaBoost Classifer.

6. Evaluation of Models:

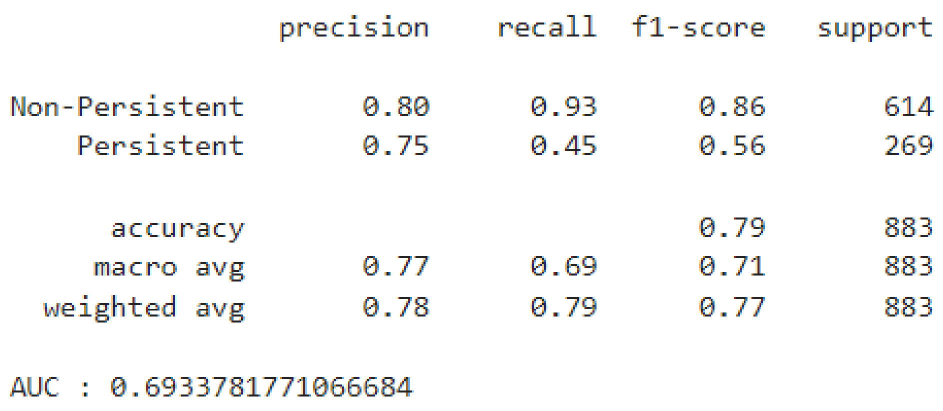
The models were evaluated based on Accuracy, Precision, Recall, f1 Score, Support and AUC scores.

6.1. Logistic Regression



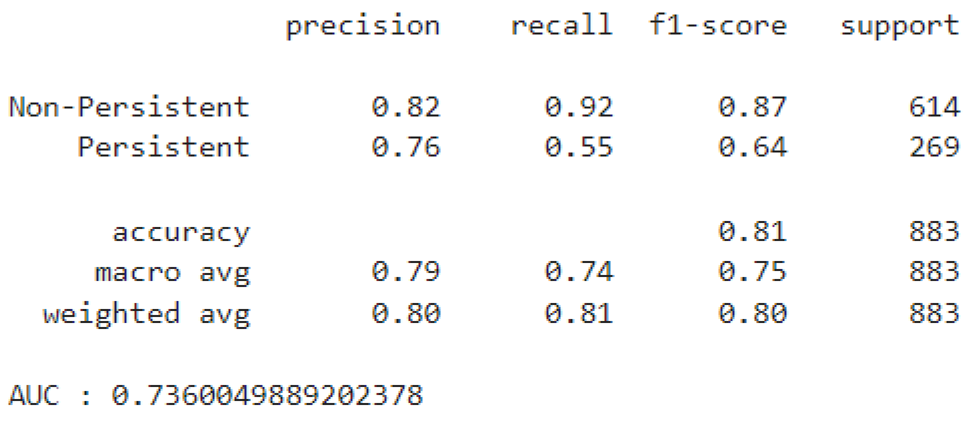
```
                precision    recall  f1-score   support

Non-Persistent       0.83      0.93      0.88       614
    Persistent       0.77      0.57      0.66       269

      accuracy                           0.82       883
     macro avg       0.80      0.75      0.77       883
  weighted avg       0.81      0.82      0.81       883

AUC : 0.7477416659603066
```

6.2. Random Forest Classifier:



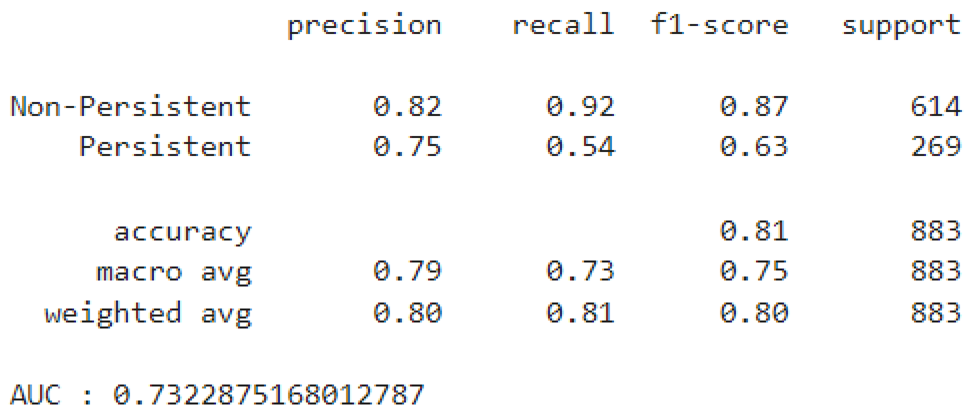
```
                precision    recall  f1-score   support

Non-Persistent       0.80      0.93      0.86       614
    Persistent       0.75      0.45      0.56       269

      accuracy                           0.79       883
     macro avg       0.77      0.69      0.71       883
  weighted avg       0.78      0.79      0.77       883

AUC : 0.6933781771066684
```

6.3. Gradient Boosting Model:



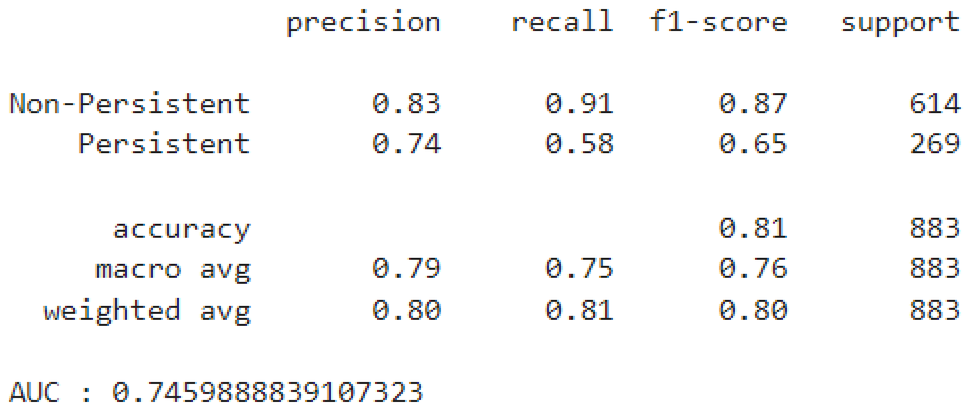
```
                precision    recall  f1-score   support

Non-Persistent       0.82      0.92      0.87       614
    Persistent       0.76      0.55      0.64       269

      accuracy                           0.81       883
     macro avg       0.79      0.74      0.75       883
  weighted avg       0.80      0.81      0.80       883

AUC : 0.7360049889202378
```

6.4. XGBoost Model:



```
                precision    recall  f1-score   support

Non-Persistent       0.82      0.92      0.87       614
    Persistent       0.75      0.54      0.63       269

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

AUC : 0.732287516801278
```

6.5. AdaBoost Model:

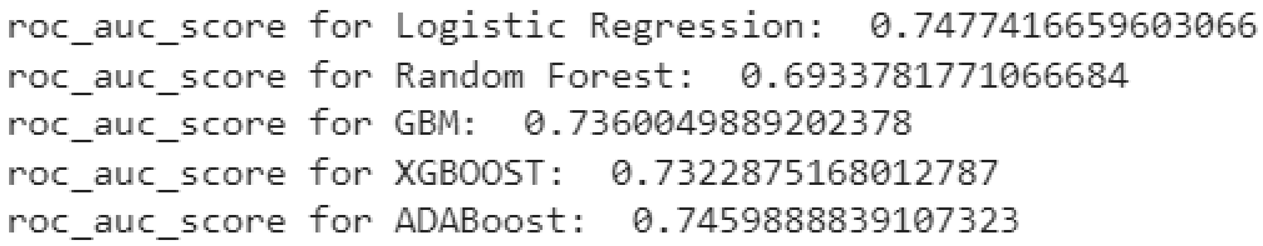


```
                precision    recall  f1-score   support

Non-Persistent       0.83      0.91      0.87       614
    Persistent       0.74      0.58      0.65       269

      accuracy                           0.81       883
     macro avg       0.79      0.75      0.76       883
  weighted avg       0.80      0.81      0.80       883

AUC : 0.7459888839107323
```

7. Model Selection:

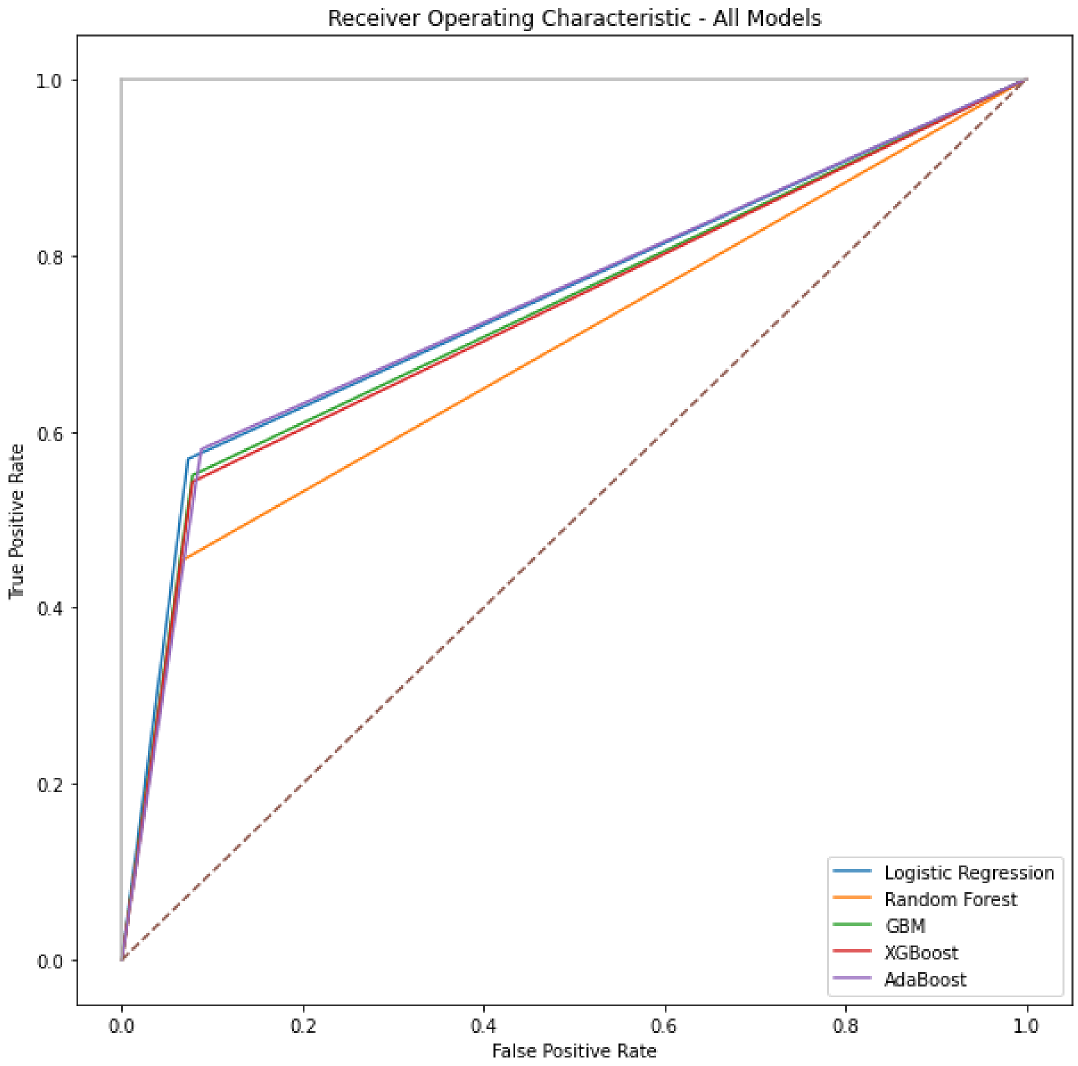
ROC (Receiver Operating Characteristics) curve and AUC (Area Under The Curve) are the most important evaluation metrics for evaluating any classification model's performance at various threshold settings. Hence, for different models area under the curve was calculated and ROC curves were plotted.

7.1. Area under the curve values:



```
roc_auc_score for Logistic Regression:  0.7477416659603066
roc_auc_score for Random Forest:  0.693378177106684
roc_auc_score for GBM:  0.7360049889202378
roc_auc_score for XGBOOST:  0.7322875168012787
roc_auc_score for ADABoost:  0.7459888839107323
```

7.2. ROC Curves:



7.3. Model Finalisation:

Based on previous evaluations, Logistic Regression Model was selected to build the final model for the given problem statement due to its high accuracy and ROC value.

8. Conclusion:

The given dataset was cleaned and transformed for the classification problem.Then it was split into two sets as train set and test set. Next, different models were trained and tested like Logistic Regression, Random Forest Model, Gradient Boosting Model, XGBoost and AdaBoost Models. From the previous comparisons, it can be concluded that: Logistic Regression Model is the best fit model to the dataset with accuracy score 0.817. Closely followed by AdaBoost Classification Model with accuracy score 0.81 and Gradient Boosting Model with accuracy score 0.808.