Data Glacier

Your Deep Learning Partner

Final Project Report

Healthcare: Persistency of a drug (Data Science)

Name: Chooladeva Piyasiri

University: National Institute of Business Management (NIBM)

Email: *chooladevapiyasiri@gmail.com*

Country: Sri Lanka

Specialization: Data Science

Batch Code: LISUM18

Date: 30 April 2023

Agenda

- Problem Statement
- Datasets Exploration
- EDA
- Model Selection
- Model Evaluation
- Conclusion

Your Deep Learning Partner

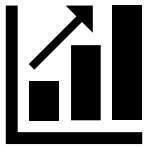
Problem Statement

- ABC is a pharmaceutical company. Medication persistency following a patient’s doctor's prescription is a problem that ABC, a pharmaceutical company, wants to comprehend. Persistence of drugs is the duration of time a patient takes medication, from initiation to discontinuation of therapy. To automate this identifying process, this organization has contacted an analytics firm.
- The analytics firm has to create a classification for the given dataset with the aim of gathering insights on the factors that are affecting the persistency.

Datasets Exploration

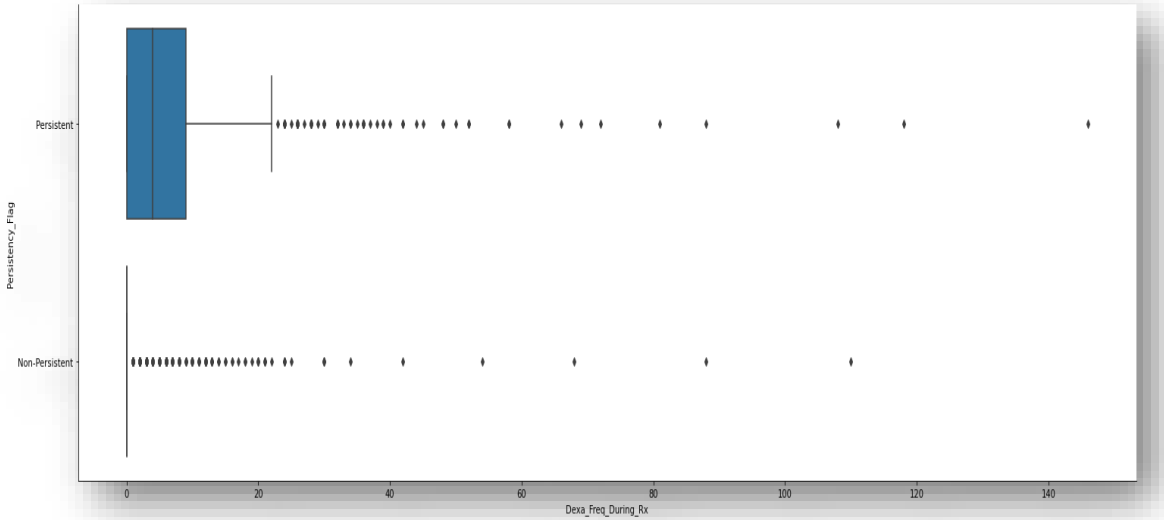
- The dataset contains 3423 observations with 26 columns and 68 features.
- It contains demographic information, clinical data, other diseases as risk factor information, and information on their physician's specialty for each patient.
- There are two numerical columns in the dataset, per the data type description. Which are:
  - **Dexa_Freq_During_Rx**
  - **Count_Of_Risks**
- The Target Variable: **Persistency_Flag**
- The dataset does not contain any NA values.

EDA

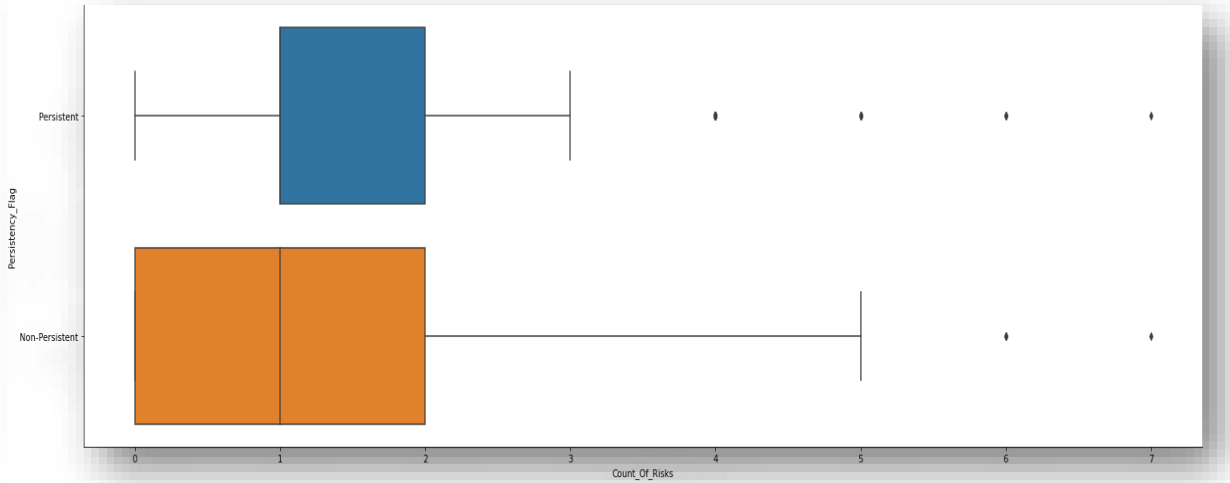


Outlier Analysis

Dexa_Freq_During_Rx



Count_Of_Risks



- Both numerical features contain some outliers.
- I performed the Capping and the Trimming/Removing techniques to deal with the outliers.
- After removing the outliers in the Count_Of_Risks column, the number of rows in the dataset was reduced from 3424 to 3401.

Skewness & Kurtosis Analysis

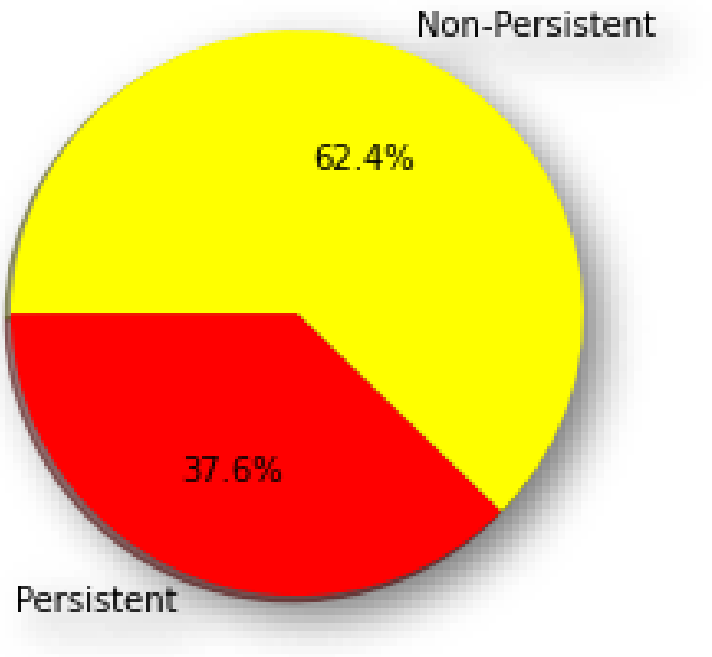
Count_Of_Risks

- ✓The Count_Of_Risks distribution is moderately skewed. (0.879)
- ✓The Count_Of_Risks distribution is *Platykurtic* (kurtosis <3).
- ✓Compared to a normal distribution, its tails are shorter and thinner, and often its central peak is lower and broader. (0.900)

Dexa_Freq_During_Rx

- ✓The Dexa_Freq_During_Rx distribution is highly skewed. (6.808)
- ✓The Dexa_Freq_During_Rx distribution is *Leptokurtic* (kurtosis >3).
- ✓Compared to a normal distribution, its tails are longer and fatter, and often its central peak is higher and sharper. (74.758)

Univariate Analysis: Exploring the target variable



- The target variable of the dataset is *Persistency_Flag*, - a flag indicating if a patient was persistent or not. The number of unique values in Persistency_Flag variable is 2. The two unique values are Persistent and Non_Persistent.
- There are 2135 non-persistent patients and 1289 persistent patients in this sample.
- Out of the total number of Persistency_Flag values, Non_Persistent appears 62.5% times and Persistent appears 37.5% times.

Bivariate Analysis: Demographical Feature Analysis according to the Persistency_Flag

Gender

Persistency_Flag	Persistency_Flag	
Gender	Non-Persistent	Persistent
Female	2010	1199
Male	115	77

Age Bucket

Persistency_Flag	Persistency_Flag	
Age_Bucket	Non-Persistent	Persistent
55-65	470	259
65-75	647	427
<55	101	63
>75	907	527

- Non-persistent patients are more common in both genders than persistent patients.
- The majority of persistent and non-persistent patients are over the age of 75.

Ethnicity

Persistency_Flag	Non-Persistent	Persistent
Ethnicity		
Hispanic	65	32
Not Hispanic	1999	1215
Unknown	61	29

Race

Persistency_Flag	Non-Persistent	Persistent
Race		
African American	65	29
Asian	43	41
Caucasian	1953	1174
Other/Unknown	64	32

Region

Persistency_Flag	Non-Persistent	Persistent
Region		
Midwest	932	447
Northeast	133	97
Other/Unknown	35	25
South	749	485
West	276	222

- The majority of persistent and non-persistent patients are Not Hispanic.
- Caucasian patients make up the vast majority of both persistent and non-persistent patients.
- The majority of persistent patients are from the Midwest, while the majority of non-persistent patients are from the South.

Risk, Comorbidity and Concomitant feature Analysis

- Comorbidity factors are present in the majority of patients, whereas risk factors are less common.
- The disorders of lipoprotein metabolism and other lipidemias (51.4%) is the most common comorbidity trait.
- Vitamin D deficiency is the leading risk factor (47.4%).
- Narcotics were found in 35.7% of the people. It is also the main concomitant feature.

Patients' Health improvement Analysis according to Persistency_Flag

Change_T_Score

Persistency_Flag Change_T_Score	Non-Persistent	Persistent
Improved	28	66
No change	951	692
Unknown	1080	413
Worsened	66	105

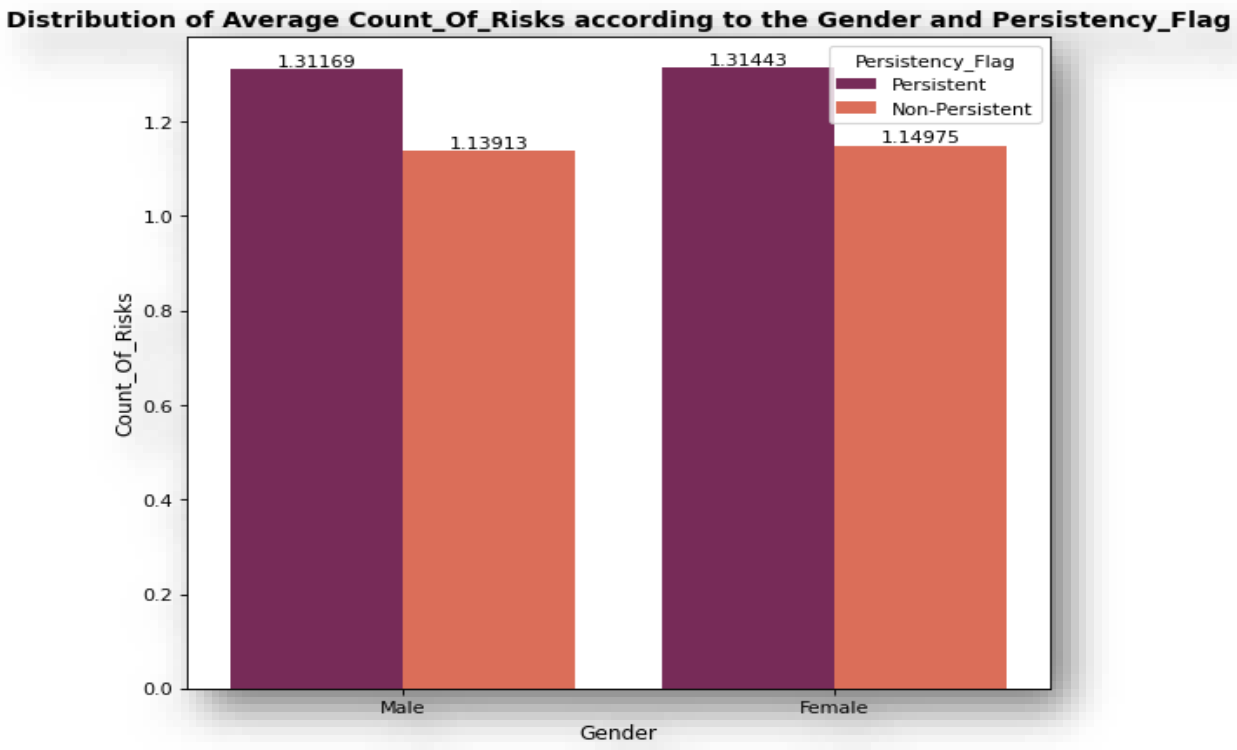
Change_Risk_Segment

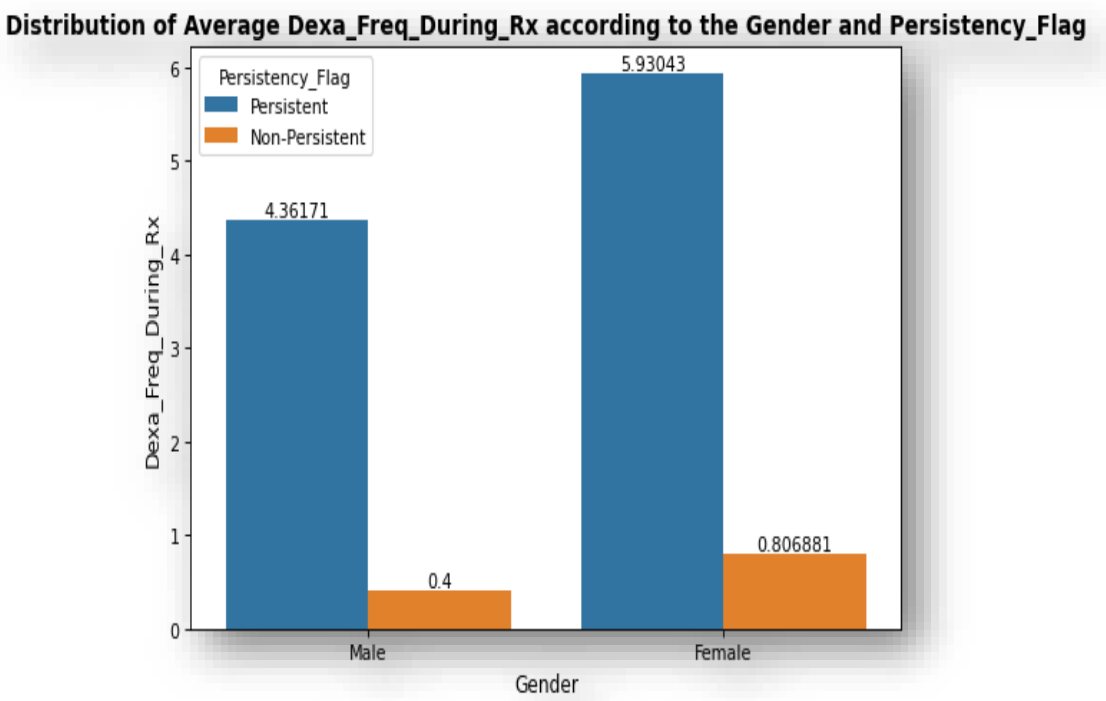
Persistency_Flag Change_Risk_Segment	Non-Persistent	Persistent
Improved	9	13
No change	615	425
Unknown	1453	765
Worsened	48	73

- The majority of persistent patients have T scores that stay unchanged after starting treatment.
- A least amount of persistent and non persistent patients have improved change in Risk Segment.

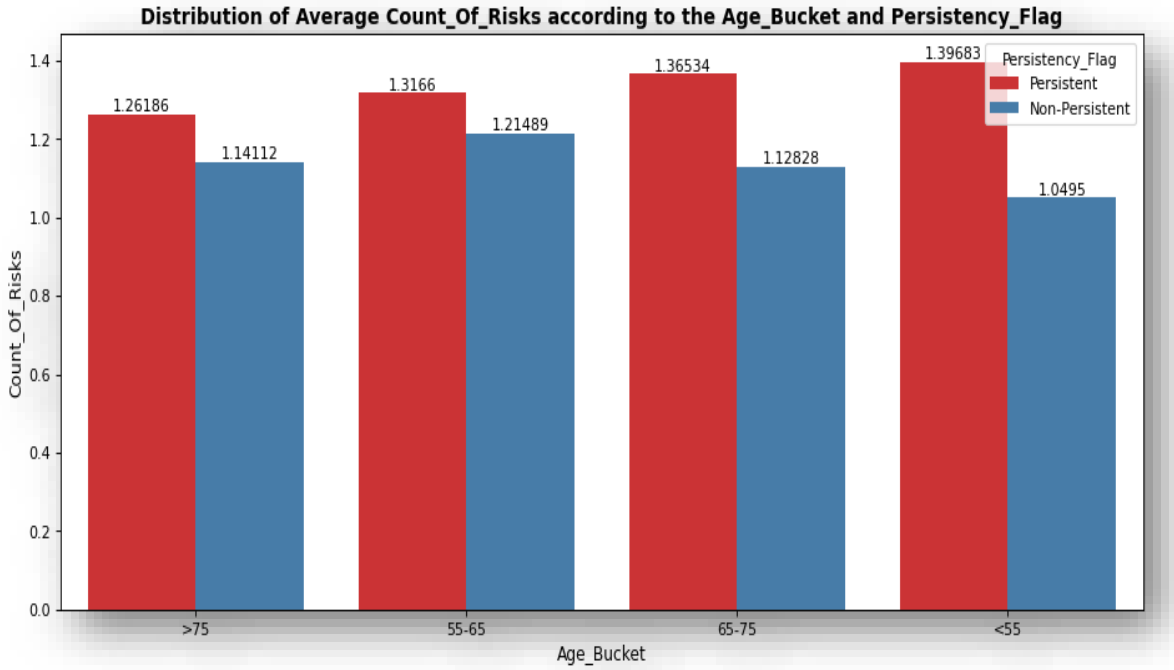
Analysis of average Count_Of_Risks, Dexa_Freq_During_Rx according to the Gender, Age Bucket and Persistency_Flag

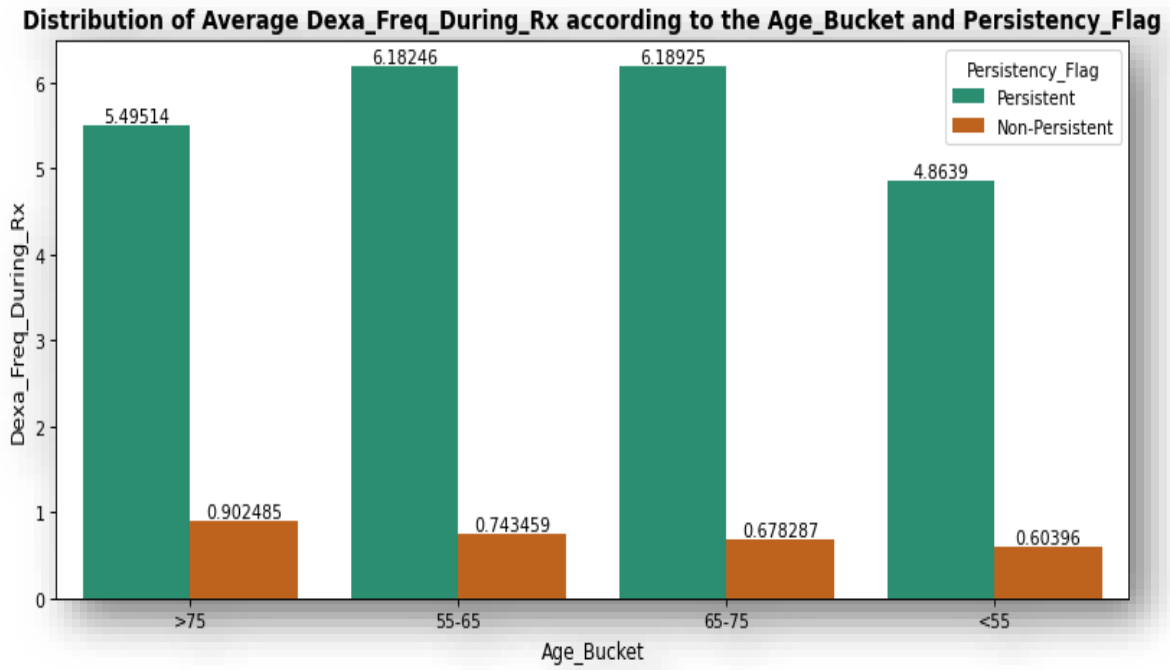
Gender





Age Bucket





- The analysis represents the average distribution analysis of persistency with the gend er and age bucket during Dexa_Freq_During_Rx and Count_Of_Risks administration.

Some other Clinical and Provider Attributes Features Analysis according to the Persistency_Flag

Ntm_Specialist_Flag

Persistency_Flag Ntm_Specialist_Flag	Non-Persistent	Persistent
Others	0.681138	0.318862
Specialist	0.544023	0.455977

Ntm_Speciality_Bucket

Persistency_Flag Ntm_Speciality_Bucket	Non-Persistent	Persistent
Endo/Onc/Uro	0.463932	0.536068
OB/GYN/Others/PCP/Unknown	0.680191	0.319809
Rheum	0.621035	0.378965

- The normalized percentage distribution of the target variable with the NTM_Specialist_Flag and the NTM_Speciality_Bucket variables are analyzed above.

Persistency_Flag	Non-Persistent	Persistent
Adherent_Flag		
Adherent	0.637349	0.362651
Non-Adherent	0.389535	0.610465

Persistency_Flag	Non-Persistent	Persistent
Risk_Chronic_Liver_Disease		
0	0.625702	0.374298
1	0.437500	0.562500

Persistency_Flag	Non-Persistent	Persistent
Risk_Excessive_Thinness		
0	0.622342	0.377658
1	0.758065	0.241935

Persistency_Flag	Non-Persistent	Persistent
Risk_Estrogen_Deficiency		
0	0.6243	0.3757
1	0.8000	0.2000

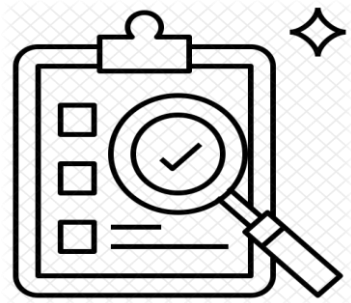
Persistency_Flag	Non-Persistent	Persistent
Risk_Type_1_Insulin_Dependent_Diabetes		
0	0.622861	0.377139
1	0.674419	0.325581

Persistency_Flag	Non-Persistent	Persistent
Risk_Smoking_Tobacco		
0	0.646698	0.353302
1	0.528571	0.471429

Model Selection

- **Logistic Regression**: A type of linear model used for binary classification. So, whether it belongs to one of the classes or is either a 0 or a 1. It attempts to predict the output value when given several input variables, placing the example into the correct category.
- **Decision Tree**: It builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while, at the same time, an associated decision tree is incrementally developed.
- **Random Forest**: It builds decision trees on different samples and takes their majority vote for classification and average in the case of regression.
- **XGBoost Classifier**: It is an implementation of gradient boosted decision trees designed for speed and performance in competitive machine learning. XGboost is an ensemble learning algorithm, meaning that it combines the results of many models.
- **AdaBoost Classifier**: It is used as an ensemble method. The most common estimator used with AdaBoost is decision trees with one level, which means decision trees with only one split.

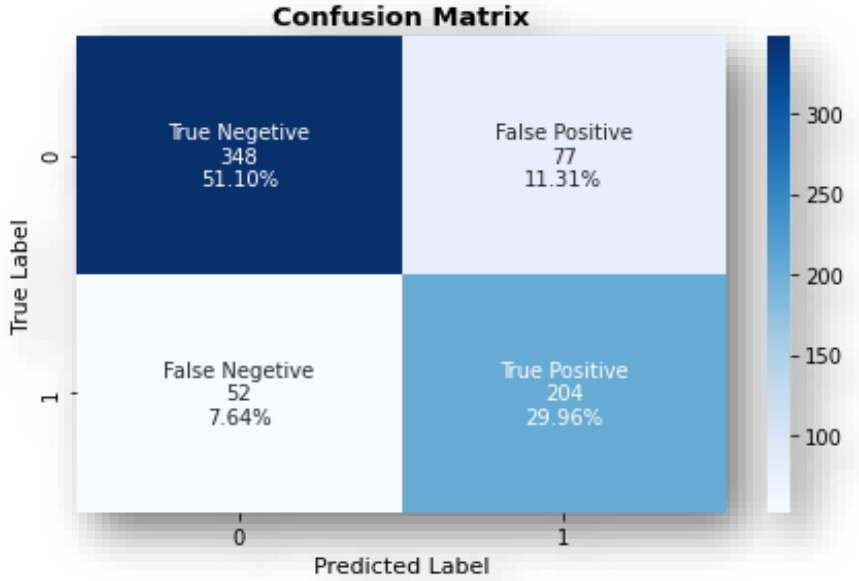
Model Evaluation



Model Evaluation: Metrics

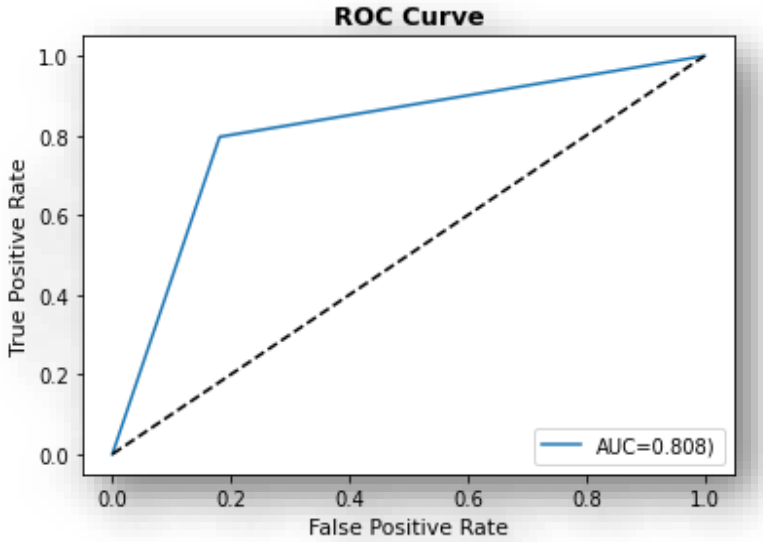
- **Accuracy** is the ratio of the number of correctly classified instances to the total number of instances. It gives an overall idea of how well the model is predicting the correct class.
- **Precision** is the ratio of the number of true positives (correctly predicted positive instances) to the number of instances predicted as positive (both true positives and false positives). It gives an idea of how well the model is predicting the positive class and avoiding false positives.
- **Recall** is the ratio of the number of true positives to the number of actual positive instances (both true positives and false negatives). It gives an idea of how well the model is capturing the positive class and avoiding false negatives.
- The **F1 score** is a harmonic mean of precision and recall, calculated as 2 x (precision x recall) / (precision + recall). It provides a single score that balances both precision and recall.
- **Cohen's Kappa** is a statistical measure that is used to assess the level of agreement between two raters or evaluators who are rating the same set of items.
- The **AUC score** represents the area under the receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various classification thresholds.

Model Evaluation: Logistic Regression

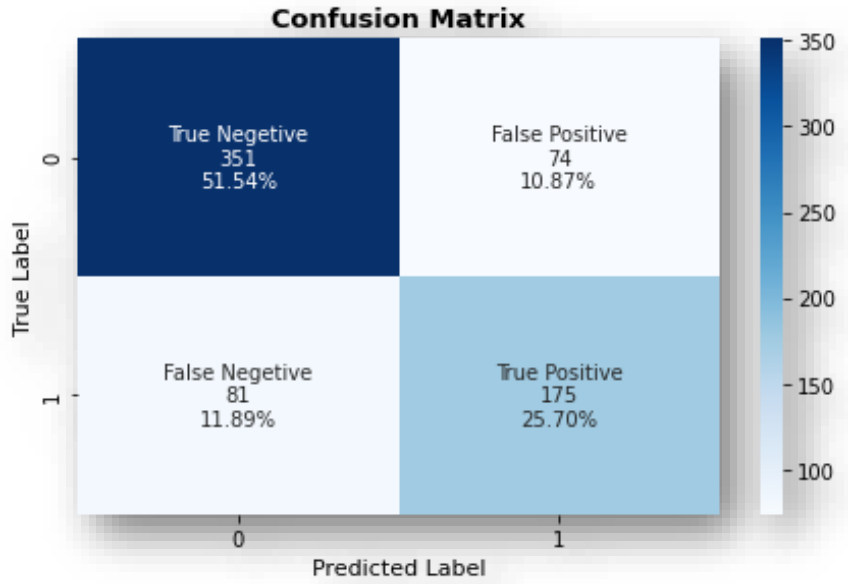


	precision	recall	f1-score	support
Non-Persistent	0.87	0.82	0.84	425
Persistent	0.73	0.80	0.76	256
accuracy			0.81	681
macro avg	0.80	0.81	0.80	681
weighted avg	0.82	0.81	0.81	681

```
Accuracy : 0.8105726872246696
Precision : 0.7259786476868327
Recall : 0.796875
F1 Score : 0.7597765363128492
Cohens Kappa: 0.603972
AUC : 0.8078492647058824
```

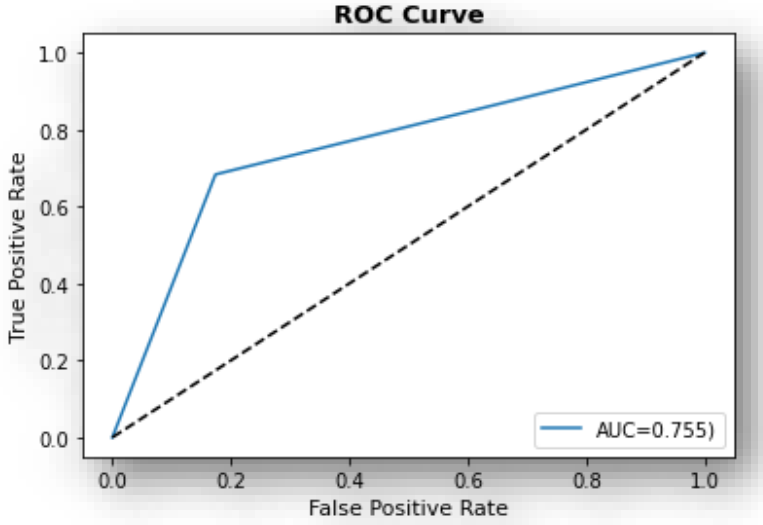


Model Evaluation: Decision Tree

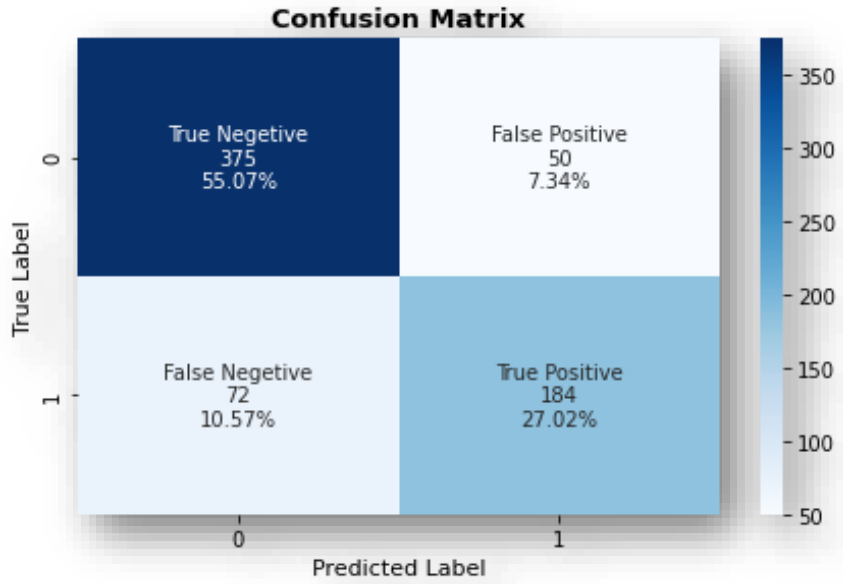


	precision	recall	f1-score	support
Non-Persistent	0.81	0.83	0.82	425
Persistent	0.70	0.68	0.69	256
accuracy			0.77	681
macro avg	0.76	0.75	0.76	681
weighted avg	0.77	0.77	0.77	681

```
Accuracy : 0.7723935389133627
Precision : 0.7028112449799196
Recall : 0.68359375
F1 Score : 0.6930693069306931
Cohens Kappa: 0.512261
AUC : 0.7547380514705884
```

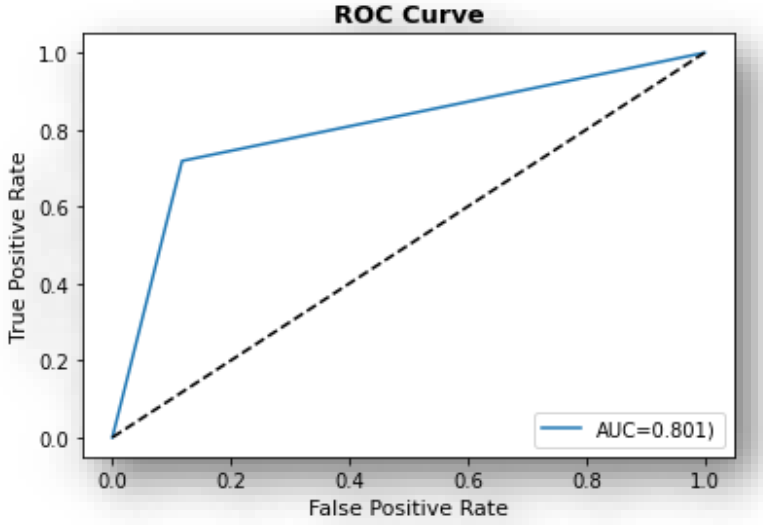


Model Evaluation: Random Forest

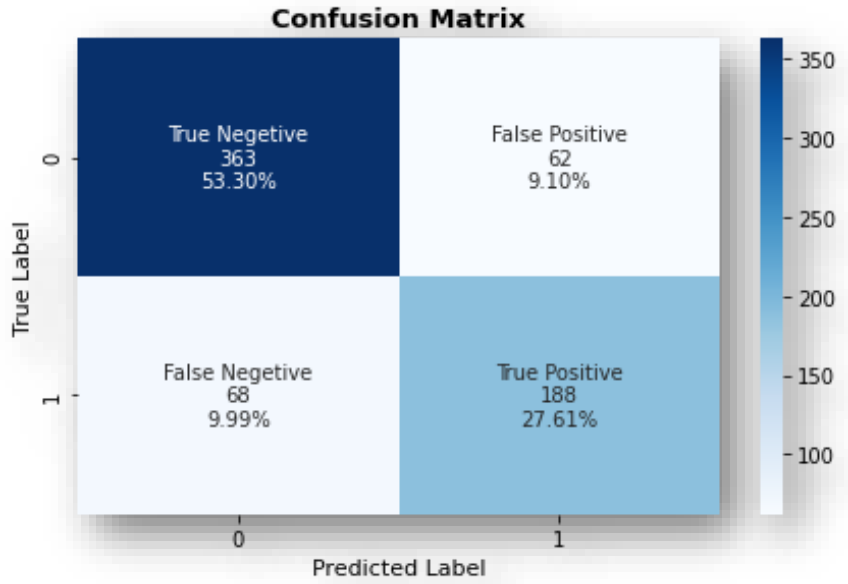


Accuracy : 0.8208516886930984
Precision : 0.7863247863247863
Recall : 0.71875
F1 Score : 0.7510204081632653
Cohens Kappa: 0.611552
AUC : 0.8005514705882353

	precision	recall	f1-score	support
Non-Persistent	0.84	0.88	0.86	425
Persistent	0.79	0.72	0.75	256
accuracy			0.82	681
macro avg	0.81	0.80	0.81	681
weighted avg	0.82	0.82	0.82	681

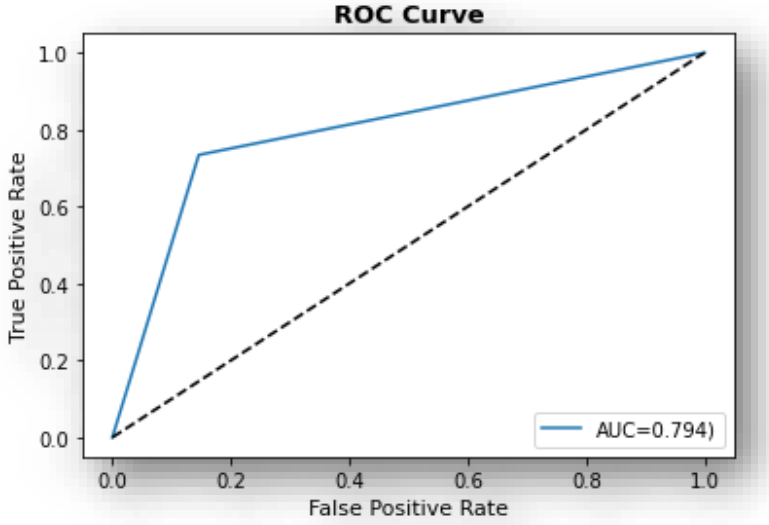


Model Evaluation: XGBoost

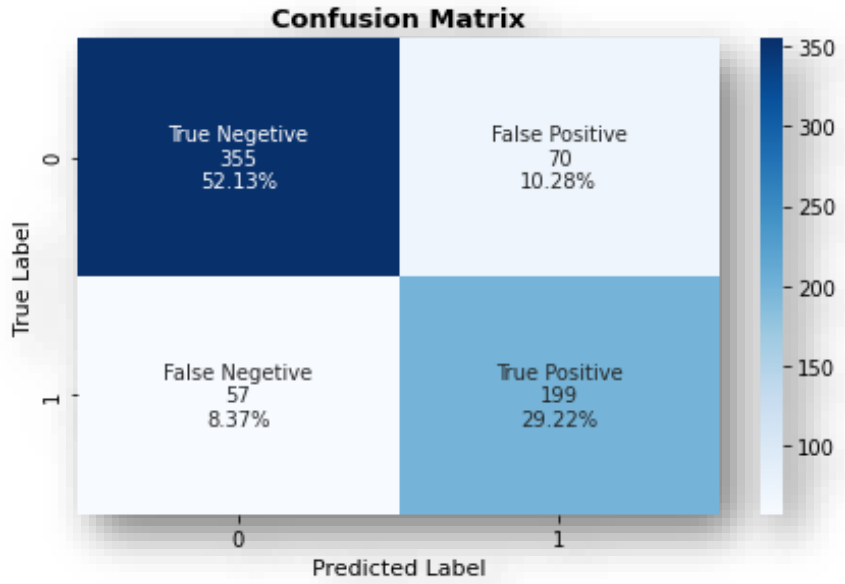


	precision	recall	f1-score	support
Non-Persistent	0.84	0.85	0.85	425
Persistent	0.75	0.73	0.74	256
accuracy			0.81	681
macro avg	0.80	0.79	0.80	681
weighted avg	0.81	0.81	0.81	681

```
Accuracy : 0.8091042584434655
Precision : 0.752
Recall : 0.734375
F1 Score : 0.7430830039525692
Cohens Kappa: 0.591248
AUC : 0.7942463235294118
```

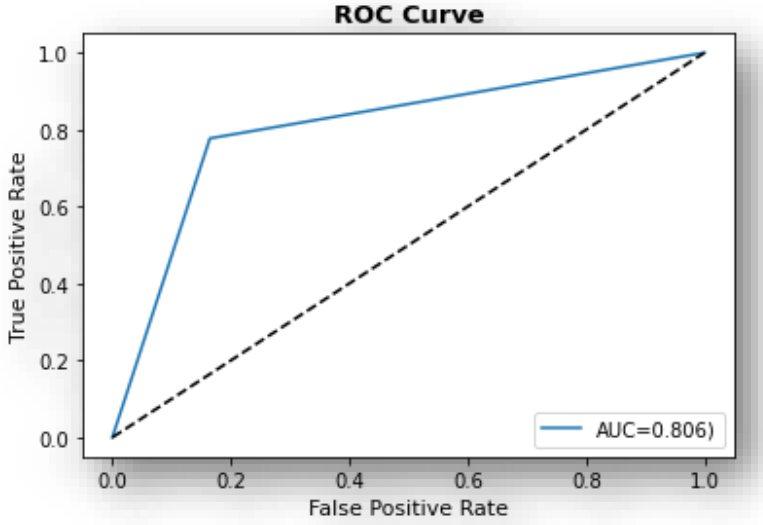


Model Evaluation: ADABoost



Accuracy : 0.8135095447870778
Precision : 0.7397769516728625
Recall : 0.77734375
F1 Score : 0.7580952380952382
Cohens Kappa: 0.606514
AUC : 0.8063189338235294

	precision	recall	f1-score	support
Non-Persistent	0.86	0.84	0.85	425
Persistent	0.74	0.78	0.76	256
accuracy			0.81	681
macro avg	0.80	0.81	0.80	681
weighted avg	0.82	0.81	0.81	681



Conclusion

Model Performance Summary

Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	AUC Score
Logistic Regression	81%	72%	79%	75%	60%	80%
Decision Tree	77%	70%	68%	69%	51%	75%
Random Forest	82%	78%	71%	75%	61%	80%
XGBoost	80%	75%	73%	74%	59%	79%
ADABoost	81%	73%	77%	75%	60%	80%

- The study found that Random Forest, Logistic Regression, and ADABoost are the top-performing models.

Thank You

Your Deep Learning Partner