Healthcare Data Science

Name: Mustafa Fakhra
Location : Dubai, UAE
Project: Data Science
Date : 26-May-2022

Executive Summary

Problem Statement:

- 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.

The highest model accuracy and precision were attained using the Random Forest model.

Project Steps

1. Understanding the case
2. Importing Required libraries and dataset
3. Understanding our data (data exploratory)
4. Data processing and transformation
5. Model Building
6. Model evaluation
7. Model Deployment

Data Processing

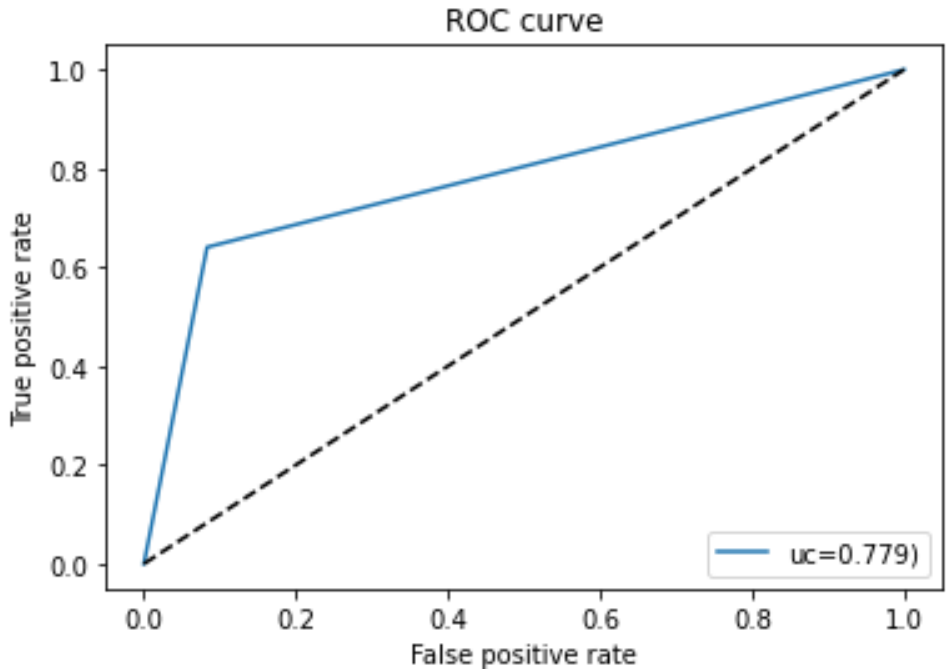
- File Used: Healthcare_dataset.xlsx
- Correlation between all variables and the predictor.
- No missing data or nulls exist.
- Data wrangling transformation included normalizing data and standardize them.
    - This has increased the correlation between the features and the predictor variable.
- Dummy variables have been created (Categorical variables to 0 and 1).

Model Building

- The case is classification so we will be using three models:
- Logistic Regression, Ridge Regression, & Random Forest classifier.

Logistic Regression Model Results

	precision	recall	f1-score	support
Non-Persistent	0.85	0.92	0.88	505
Persistent	0.78	0.64	0.70	231
accuracy			0.83	736
macro avg	0.81	0.78	0.79	736
weighted avg	0.83	0.83	0.83	736

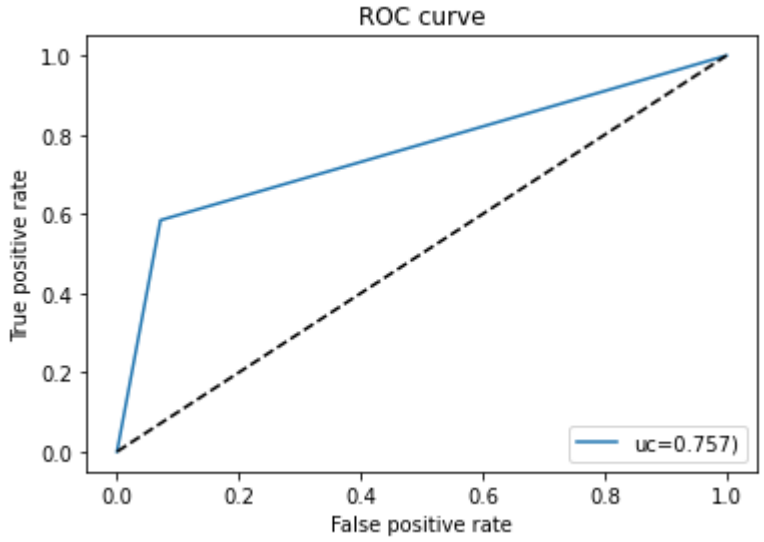


Accuracy : 0.8301630434782609
Precision : 0.7789473684210526
Recall : 0.6406926406926406
F1 Score : 0.7030878859857482

Ridge Regression Model Results

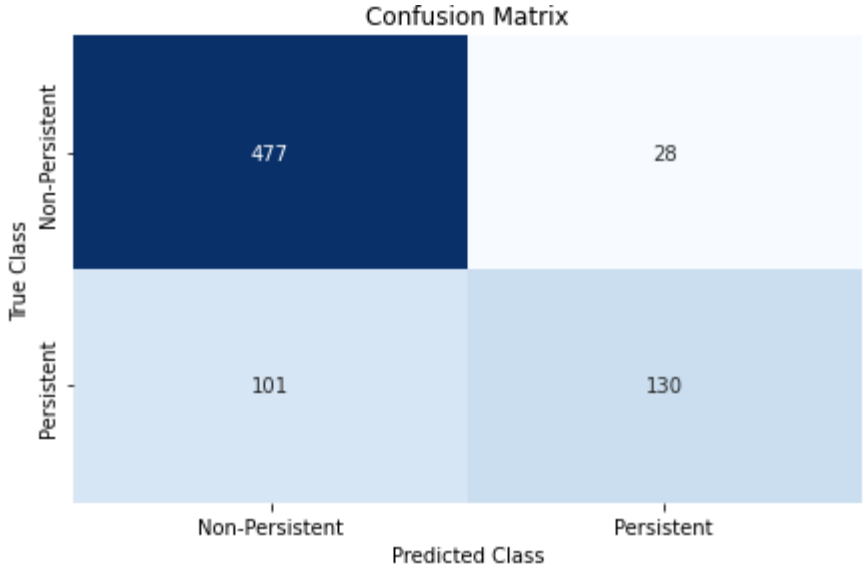
	precision	recall	f1-score	support
Non-Persistent	0.83	0.93	0.88	505
Persistent	0.79	0.58	0.67	231
accuracy			0.82	736
macro avg	0.81	0.76	0.77	736
weighted avg	0.82	0.82	0.81	736

AUC : 0.7565642278513565



Accuracy : 0.8206521739130435
Precision : 0.7894736842105263
Recall : 0.5844155844155844
F1 Score : 0.6716417910447761

Random Forest Model Results



	precision	recall	f1-score	support
Non-Persistent	0.83	0.94	0.88	505
Persistent	0.82	0.56	0.67	231
accuracy			0.82	736
macro avg	0.82	0.75	0.77	736
weighted avg	0.82	0.82	0.81	736

AUC : 0.7536625091080535

Random Forest Model

- Model Trade-offs:
  - Advantages:
    - Insensitive to Outliers.
    - Insensitive to Null values.
    - Less Prone to overfitting.
  - Disadvantages:
    - Losing Interpretability.
    - Difficult to diagnose and improve.
- Results obtained:
    - Accuracy: 82 – 83 %

Conclusion

- Approximately all the classifiers have same result, but the Ridge CLassifier and the Random Forest were the best one.
- These two models have around 82% Accuracy.
- Ridge Classifier has 78% Precision, 58% Recall, & 67% F1 Score.
- Random Forest has 82% Precision, 56% Recall, & 66% F1 Score.

Thank You