Data Glacier

Your Deep Learning Partner

Machine Learning Model

Healthcare - Persistency of a drug

Name	Specialization	Country	Email
Ms. Larisa Popa	Data Science	Romania	Larisapopa4@gmail.com
Ms. Afshan Hashmi	Data Science	Kingdom of Saudi Arabia	afshanhashmi786@gmail.com
Mr. Omar Safwat	Data Science	Egypt	omarksafwat@gmail.com
Mr. Roger Burek-Bors	Data Science	Poland	roger.burek-bors@hotmail.com

Team: DG_team_project_PL-RO-KSA-EGY

Date: May 10, 2021

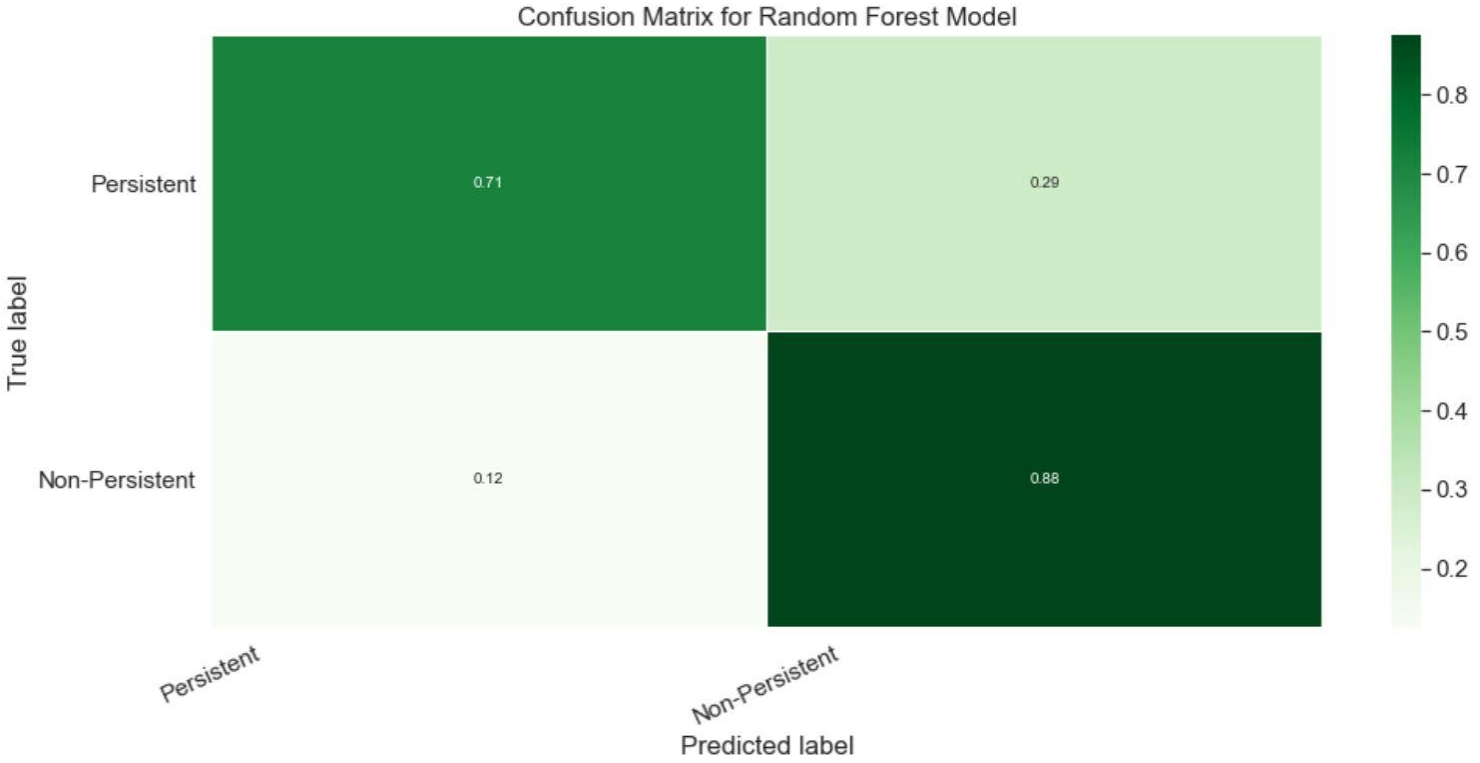
Executive Summary

- This report highlights the machine learning building phase and selection.
- ML models experimented:
  - Support Vector Machine
  - Logistic Regression
  - KNN
  - Random Forest
- **The highest model accuracy and precision were attained using the logistic regression statistical model**.

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: 79 – 81 %

Random Forest Model



	precision	recall	f1-score	support
0.0	0.76	0.71	0.74	371
1.0	0.84	0.88	0.86	657
accuracy			0.82	1028
macro avg	0.80	0.79	0.80	1028
weighted avg	0.81	0.82	0.81	1028

- 0.0 correlates to “Persistent” flag.
- 1.0 correlates to “Non-Persistent” flag.

Support Vector Machine Model

- Model Trade-offs:
  - Advantages:
    - Can successfully handle high dimensional data
    - Can successfully handle imbalanced classes
  - Disadvantages:
    - Difficult to diagnose and improve
    - Quite sensitive to outliers – training on dataset with outliers decreased model accuracy
    - Not suitable for large datasets since the training time will be higher
- Results obtained:
  - Accuracy: 84 %

    Techniques applied for improving worsened the model performance:
    - Upsampling
    - Downsampling
    - PCA dimension reduction

Support Vector Machine Model Results

```
1 data_training.shape
```

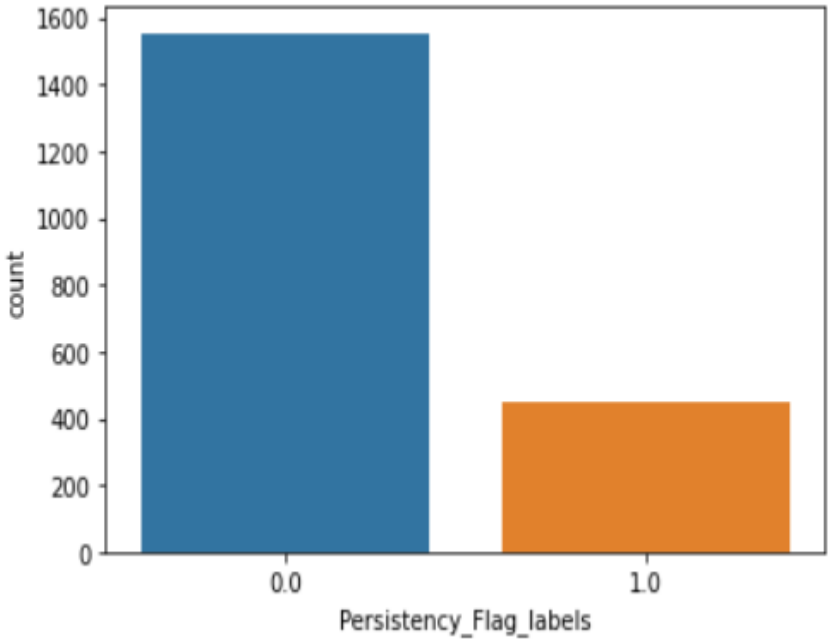
```
(2004, 51)
```

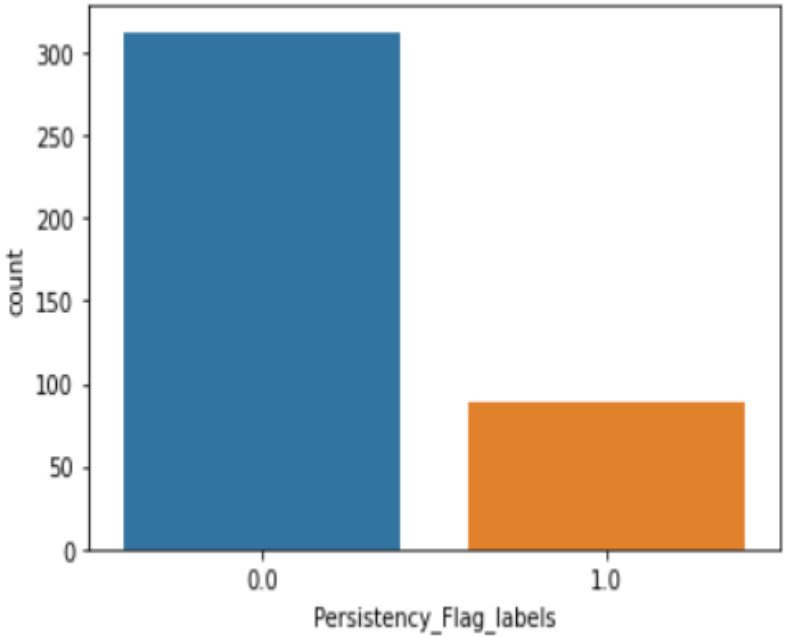
```
1 X_train.shape
```

```
(1603, 50)
```

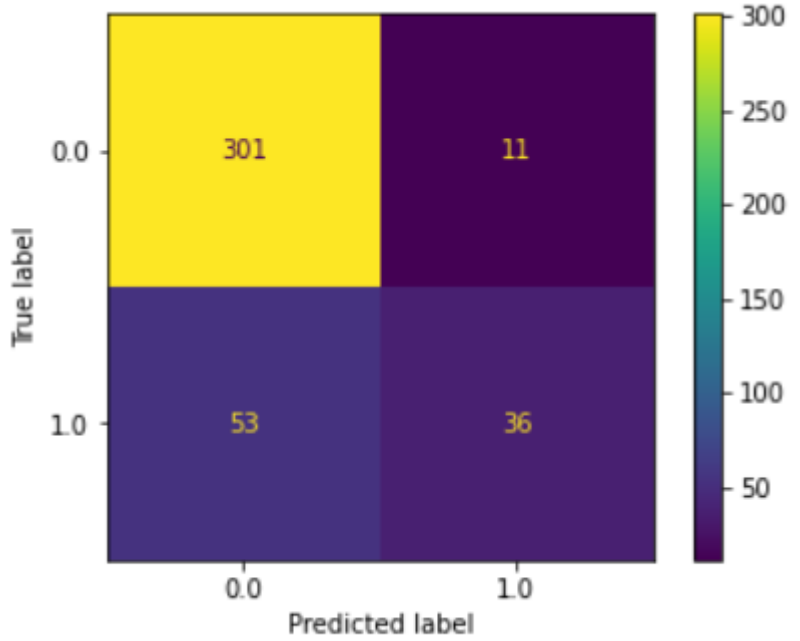
```
1 X_test.shape
```

```
(401, 50)
```





Test predictions



Support Vector Machine Model Results

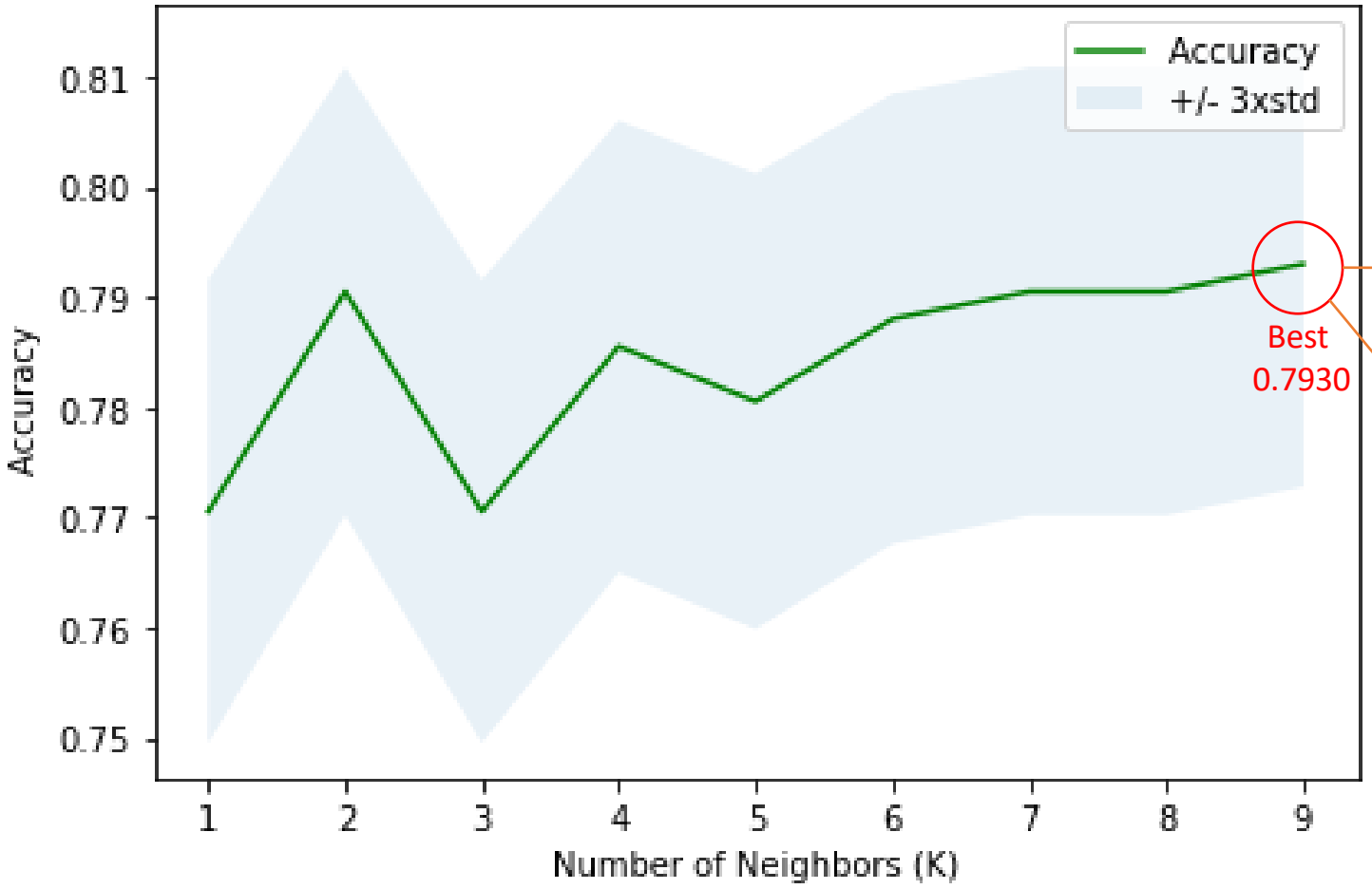
Classification report

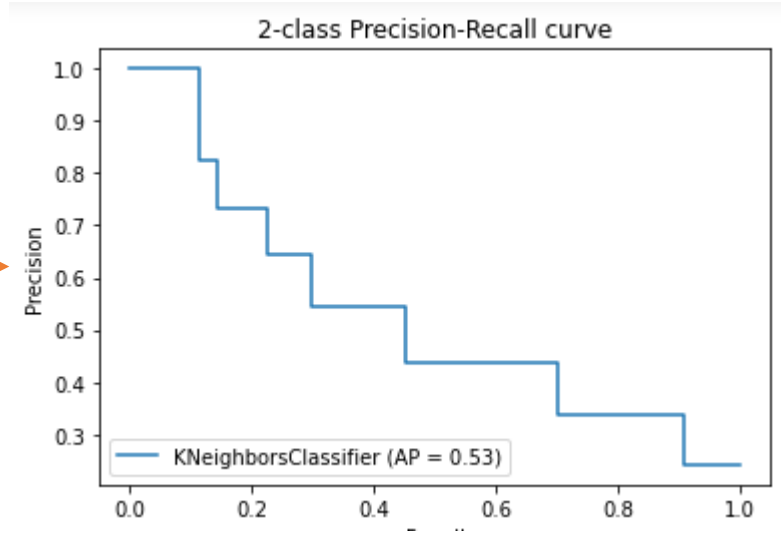
	precision	recall	f1-score	support
0.0	0.85	0.96	0.90	312
1.0	0.77	0.40	0.53	89
accuracy			0.84	401
macro avg	0.81	0.68	0.72	401
weighted avg	0.83	0.84	0.82	401

- Precision gives the percentage of the correct prediction from all values predicted positive. P= TP/(TP+FP)
- Recall measure the percentage of the correct prediction from all values that were actually positive. R= TP/(TP+FN)
- F1 score weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 score= 2(RP)/(R+P)
- Support is the number of actual occurrences of the class in the specified dataset.

```
ROC AUC score on test data 0.8340535868625757
ROC AUC score on the entire dataset 0.838195338195338
```

KNN Model Results






```
F1 score for macro mode 0.6117468649752115
F1 score for micro mode 0.7930174563591024
F1 score for weighted mode 0.748692165765231
```

KNN Model Results

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

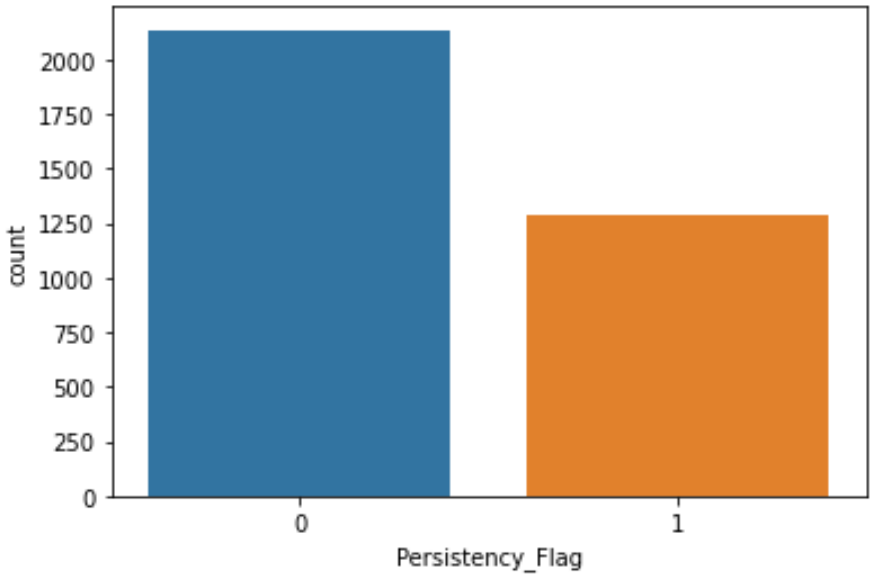
Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Imbalanced class were balanced using SMOTE

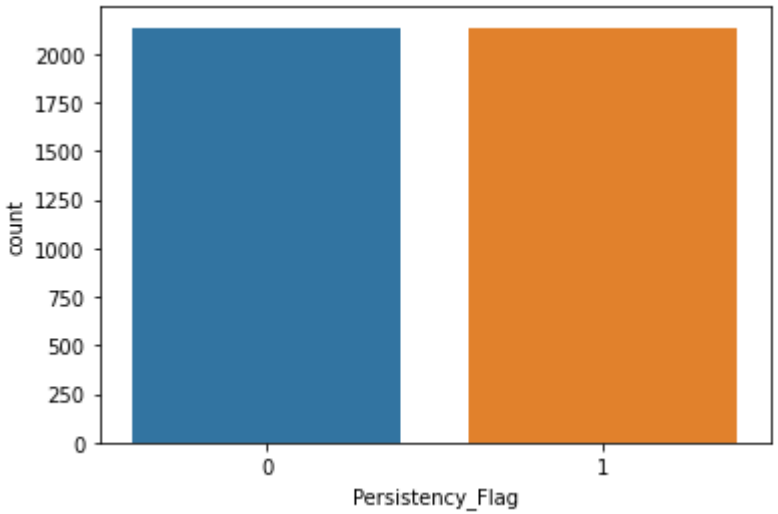
```
y['Persistency_Flag'].value_counts()
```

```
0    2135
1    1289
Name: Persistency_Flag, dtype: int64
```




```
yf['Persistency_Flag'].value_counts()
```

```
0    2135
1    2135
Name: Persistency_Flag, dtype: int64
```



SMOTE-Synthetic minority oversampling Technique

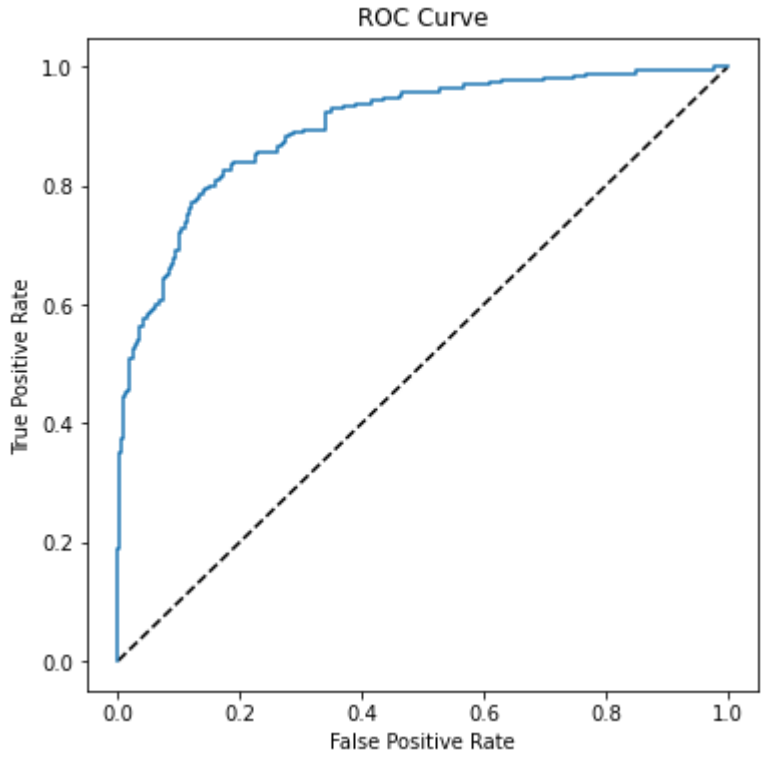
- *Increases the number of low incidence examples in a dataset using synthetic minority oversampling.*

- SMOTE takes the entire dataset as an input, but it increases the percentage of only the minority cases.

- **SMOTE** reduces the bias towards the classification ..

Logistic Regression Model Results

	precision	recall	f1-score	support
0	0.86	0.88	0.87	422
1	0.80	0.76	0.78	263
accuracy			0.84	685
macro avg	0.83	0.82	0.82	685
weighted avg	0.83	0.84	0.83	685



Overall Precision: 0.8
Overall Recall: 0.7604562737642585

AUC: 0.8993116248896257

Thank You