Exploratory Data Analysis

Healthcare - Persistency of a drug

Mike Wang

7/29/2023

Agenda

Problem Statement

EDA

Recommendations

Model Selection & Evaluation

Conclusion

Problem Statement

- One of the challenges for all Pharmaceutical companies is to understand the persistency of drugs as per the physician prescription. To solve this problem ABC Pharma company approached an analytics company to automate this process of identification.
- Main objective is to provide data analysis based on the dataset and build a classification model to predict persistency of drugs for patients suffering from Nontuberculous mycobacteria (NTM)

Dataset Information

```
RangeIndex: 3424 entries, 0 to 3423
Data columns (total 69 columns):
```

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Risk_F
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...	

5 rows × 69 columns

The Initial dataset Includes 3424 data entries and 69 unique columns.
There is a Ptid for each patient's identifier.
Persistency_Flag is the target variable.

Dataset Information

- Features include patient demographic information (Age, Race, etc.) and characteristics (Dexa Scan score, Risk factor, etc.)

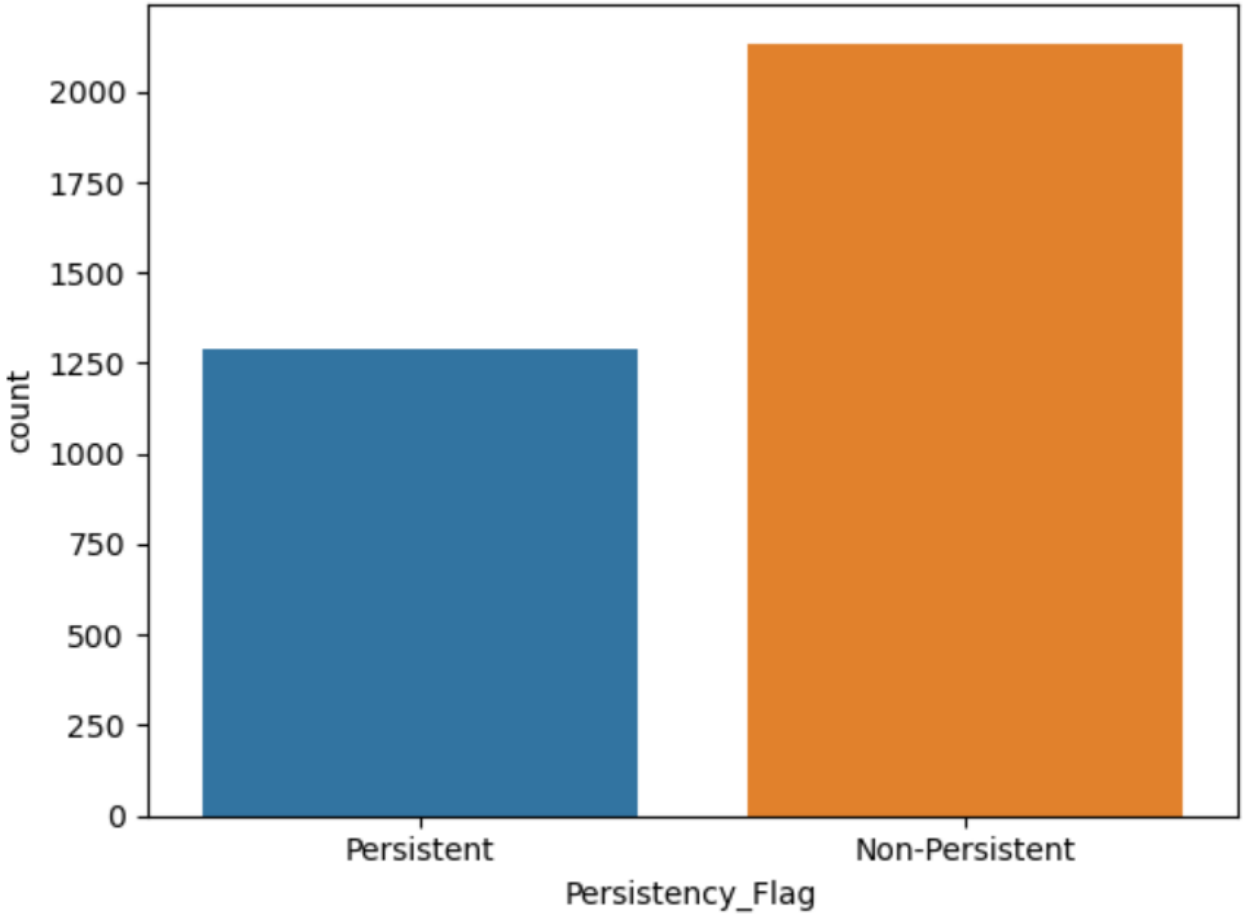
Age	Age of the patient during their therapy
Race	Race of the patient from the patient table
Region	Region of the patient from the patient table
Ethnicity	Ethnicity of the patient from the patient table
Gender	Gender of the patient from the patient table
IDN Indicator	Flag indicating patients mapped to IDN
NTM - Physician Specialty	Specialty of the HCP that prescribed the NTM Rx
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
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

EDA (Imbalanced data)

There is a slightly unbalanced number between the two values of target variables. (2135 non-persistent vs. 1289 Persistent)

This could lead to a naturally lower accuracy using a prediction model since there are more Nos than Yess.

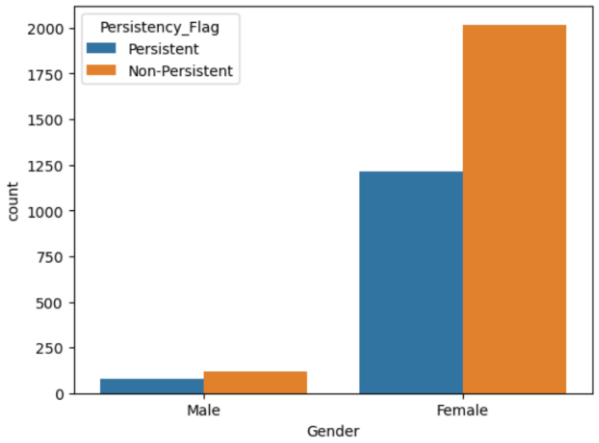
One way to solve this issue is to use a resampling technique to have a more balanced dataset.

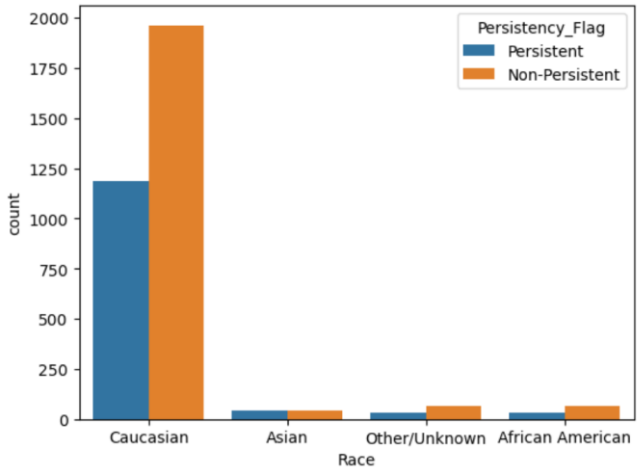


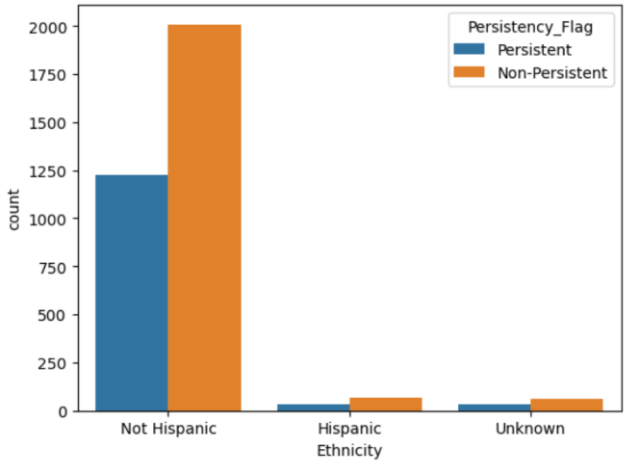
EDA (Imbalanced data)

Gender, Race, and Ethnicity all have unbalanced data between groups.

To be able to use these features for further prediction, we need to combine some of the values to create new variables.



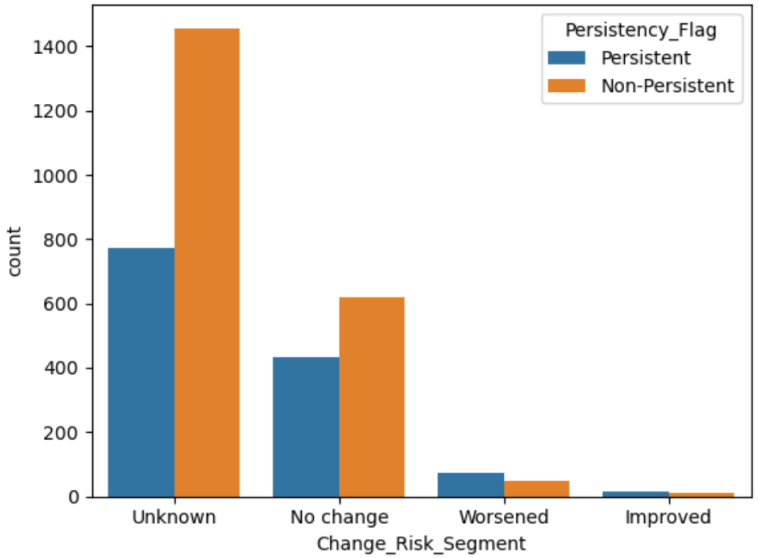


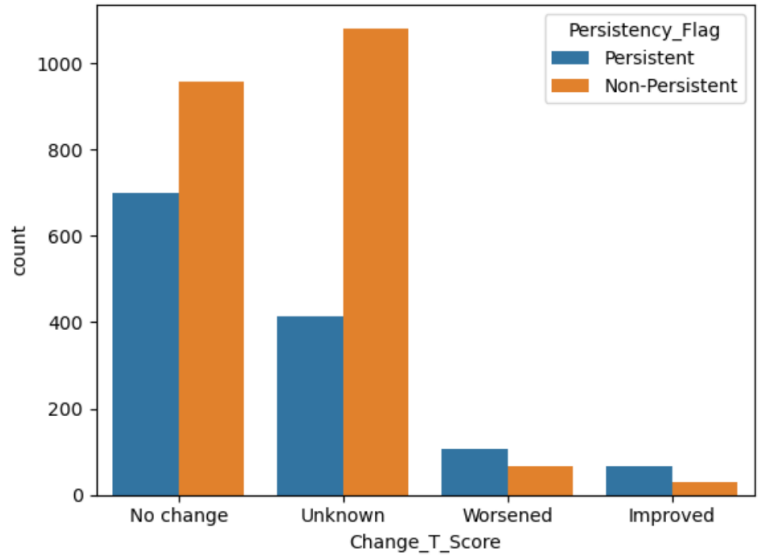


EDA (Missing Values)

There are two variables (Change of Risk Segment and Change of T Score) that have a large number of unknown values.

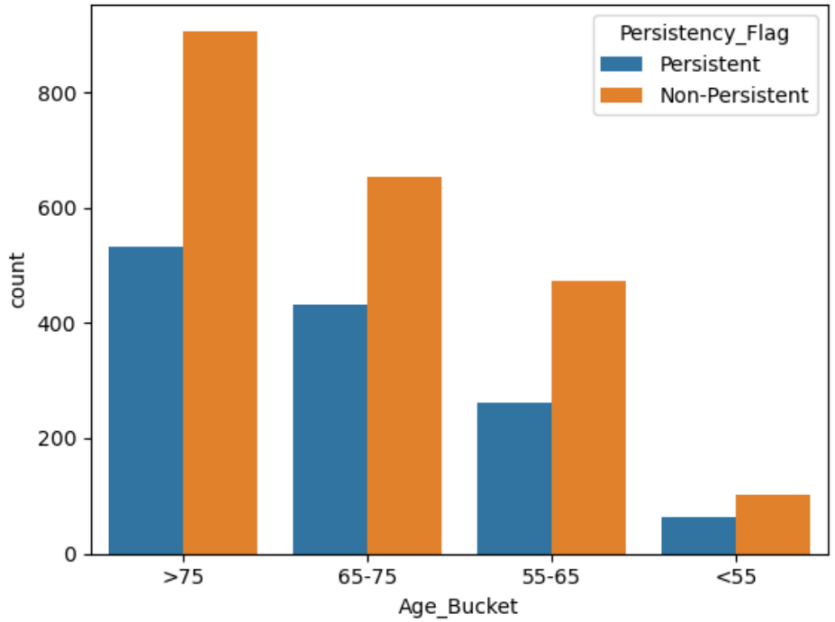
One way to treat these values is to group them into No Change since most of the values are in the No Change group.



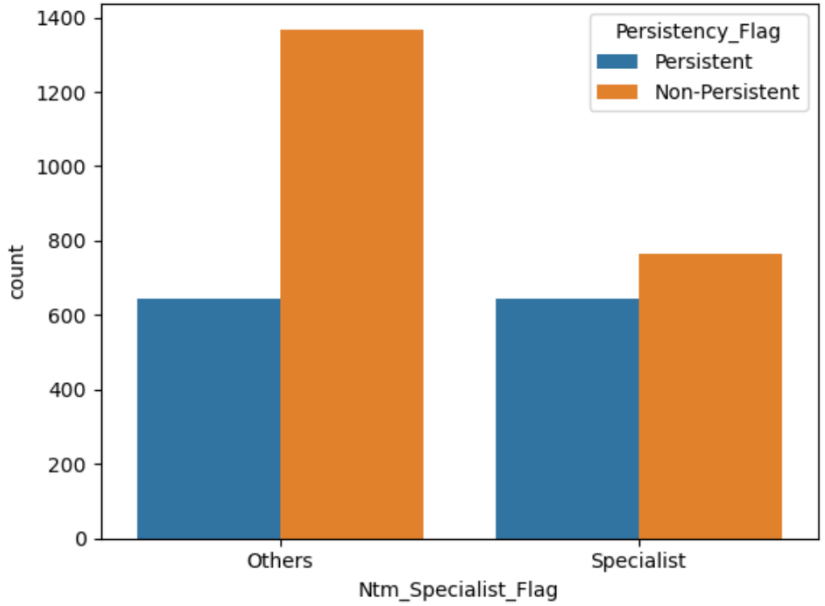


EDA

There are more older patients in the dataset as expected since older people are more likely to be sick. There was no significant difference between the proportion of drug persistency between different age groups.



Patients who received prescriptions from who are specialists have a more persistent proportion (~44%) than Other physicians (~30%). It could be due to that specialists provide more detailed and easier-followed instructions for the prescriptions.



EDA Summary

1. The dataset contains several imbalanced groups which will potentially have impacts on model prediction. It will be recommended to ignore these variables or try to merge and create more balanced groups
2. The imbalanced target variable might cause more predictions towards Non-persistent values, resampling could be implemented to balance the number of observations from both groups.
3. Two variables have too many unknown values, recommended approach is to change them to unchanged values.

Model Selection

All models performed better compared to the baseline model (dummy).
Logistic Regression, Random Forest, and LinearSVC Models have the best accuracy scores.

```
DummyClassifier: 0.6160583941605839
SVC: 0.8145985401459854
NuSVC: 0.8
LinearSVC: 0.8364963503649635
SGDClassifier: 0.8291970802919708
KNeighborsClassifier: 0.8043795620437956
LogisticRegression: 0.8408759124087591
LogisticRegressionCV: 0.8394160583941606
DecisionTreeClassifier: 0.7372262773722628
BaggingClassifier: 0.7941605839416058
ExtraTreesClassifier: 0.8321167883211679
RandomForestClassifier: 0.8394160583941606
```

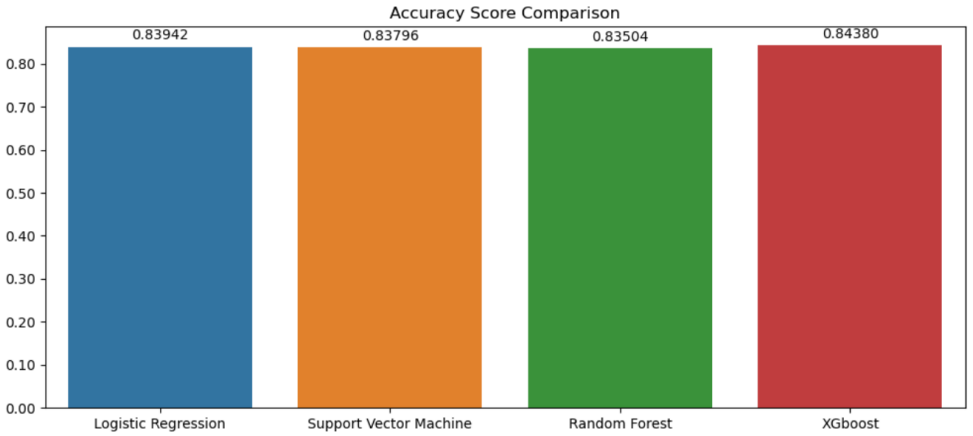
Logistic Regression, Linear SVC, Random Forest also have the highest F1 Score

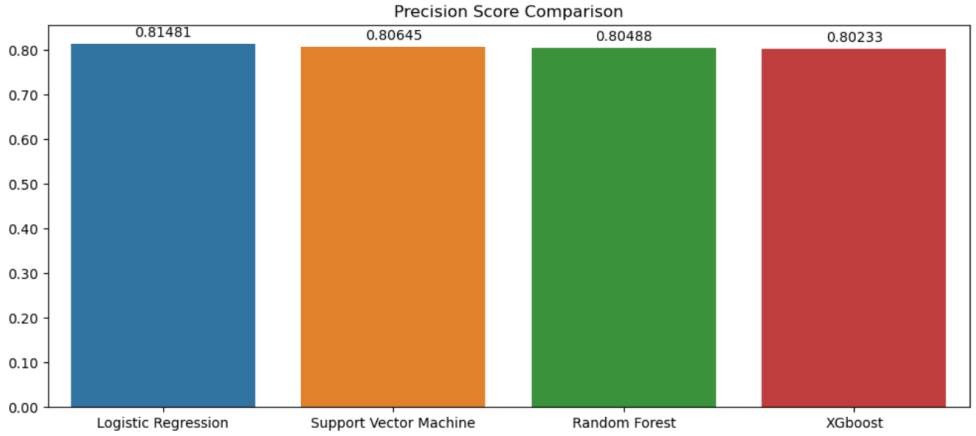
```
DummyClassifier: 0.0
SVC: 0.744466800804829
NuSVC: 0.702819956616052
LinearSVC: 0.7858546168958742
SGDClassifier: 0.5804878048780487
KNeighborsClassifier: 0.7309236947791166
LogisticRegression: 0.787524366471735
LogisticRegressionCV: 0.7826086956521738
DecisionTreeClassifier: 0.68796992481203
BaggingClassifier: 0.7544910179640719
ExtraTreesClassifier: 0.7775628626692457
RandomForestClassifier: 0.7836257309941521
```

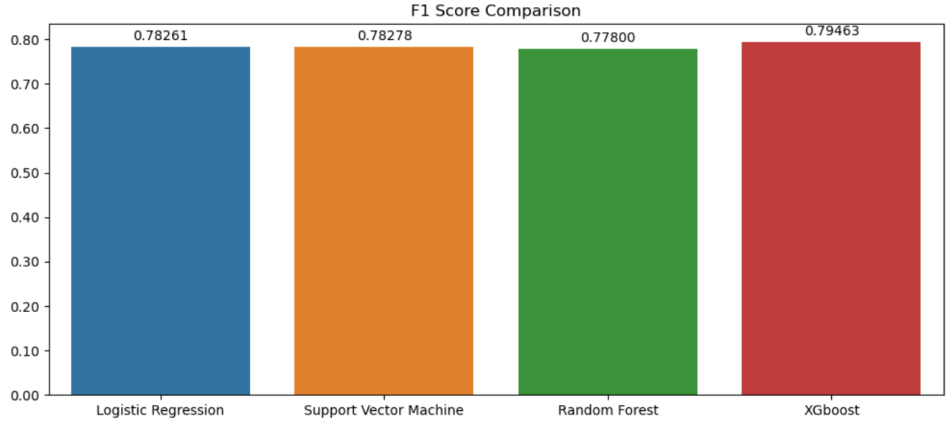
In general, Logistic Regression, Random Forest, and Linear SVC models have the best performance in predicting the result. It is recommended to use them as base models to evaluate and tune hyper-parameters to further increase the model performance.

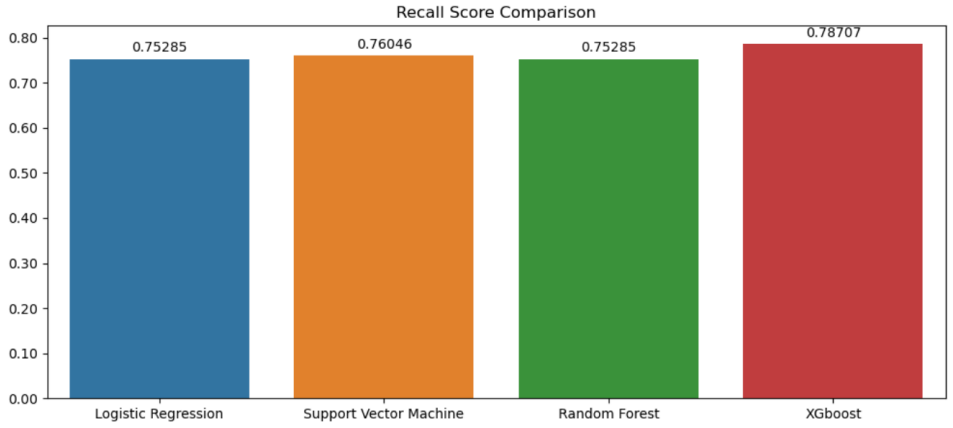
Model Evaluation

- Four different models were selected and their performances were tested as below, including two Linear Models (Logistic Regression & Support Vector Machine), one Ensemble model using Bagging (Random Forest), and one Ensemble model using Boosting (Extreme Gradient Boosting):









Bagging vs. Boosting

- Bagging Algorithm contains numerous decision trees and takes the algorithms' average to get the final score.
- Boosting Algorithm relies on training decision trees consequently and finally arrives to a powerful predictive model.

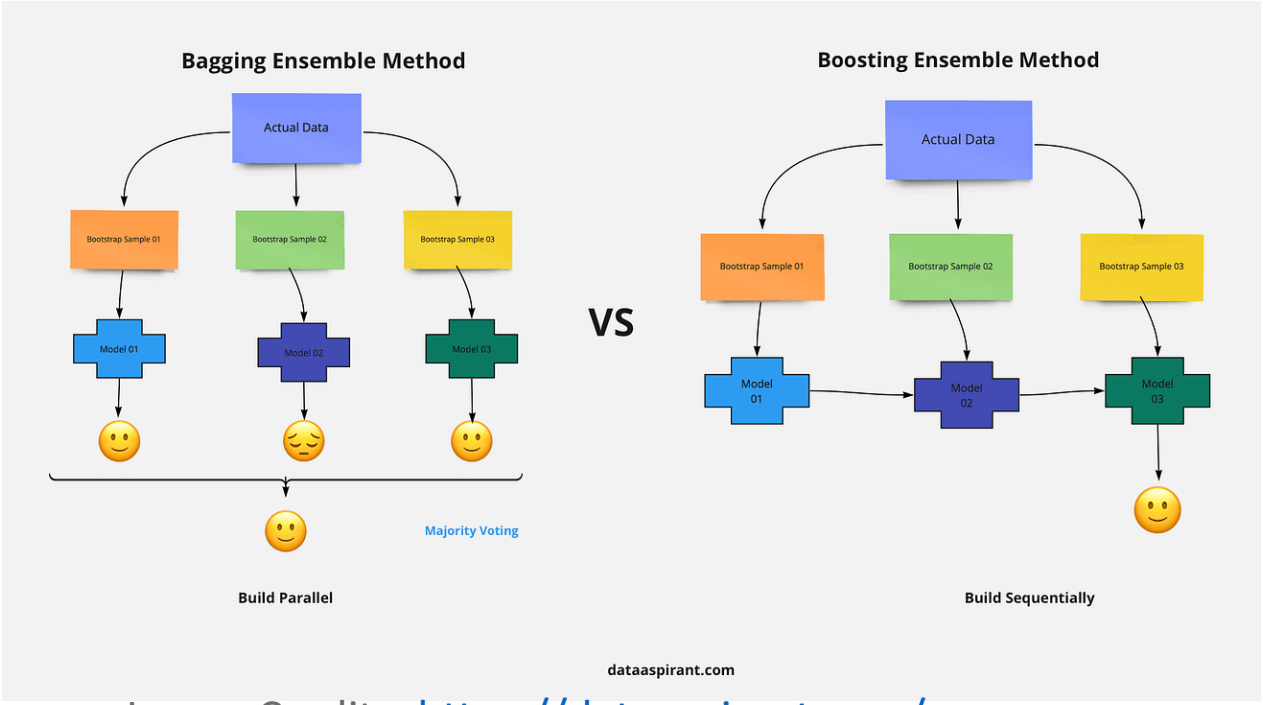
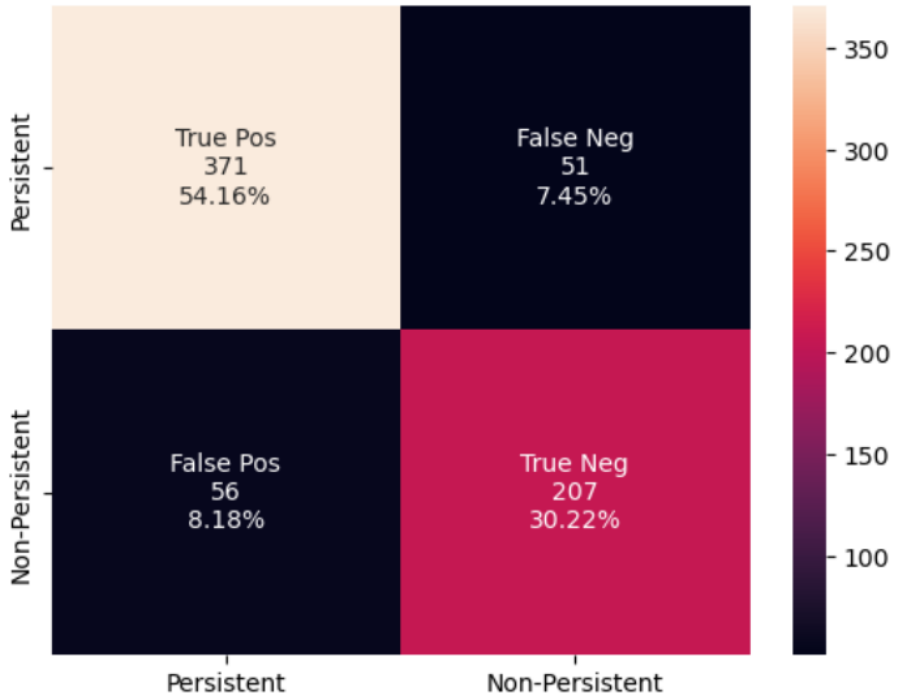


Image Credits: https://dataaspirant.com/

Model Evaluation

- Overall, the XG Boost model has the best performance scores among all the models.
- Out of 685 samples, the model predicted 578 correctly(371 Persistent, 207 Non-Persistent).



Conclusion

Using the trained predictive model, we were able to predict the persistence flag with about 84.3% accuracy.
The future direction could aims to fill more missing values for some of the features to increase the performance.

Thank You