Data Glacier

Your Deep Learning Partner

HealthCare: Persistency of a Drug

Group Name: The Data Doctors

Ashish Sasanapuri, Mohammad Shehzar Khan, Tomisin Abimbola Adeniyi, Noah Gallego

30-Dec-2023

Problem Description

One challenge for all Pharmaceutical companies is to understand the persistence of a drug as per the physician's prescription. To solve this problem ABC Pharma company approached an analytics company to automate this process of identification.

Data Understanding

- The dataset provides the factors impacting the patient's persistence to New Therapy Medication (NTM) by ABC pharmaceutical company prescribed by various physicians.

- The aim is to build a machine-learning model that classifies the patient into **Persistent** (Compliant) and **Non-persistent** (Non-Compliant).

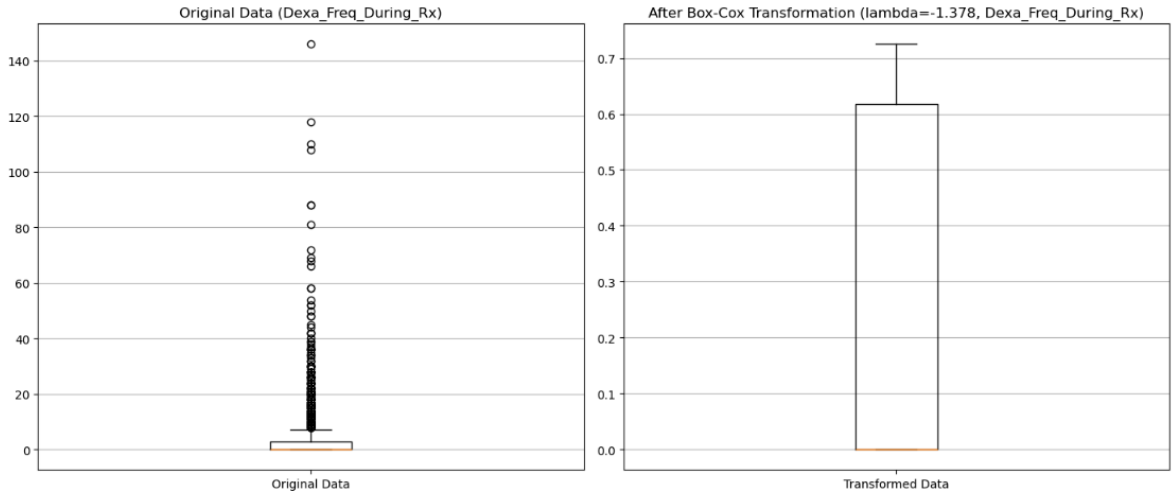
- The dataset consists of 3242 records and is a an imbalanced dataset due to low number of **Persistent** records as compared to **Non-persistent**.

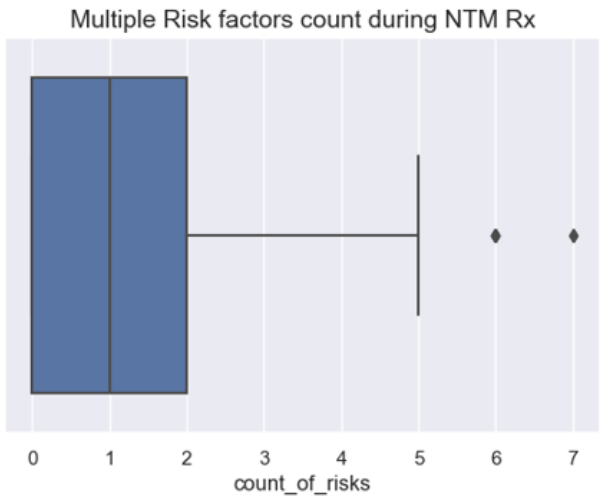
Data Understanding

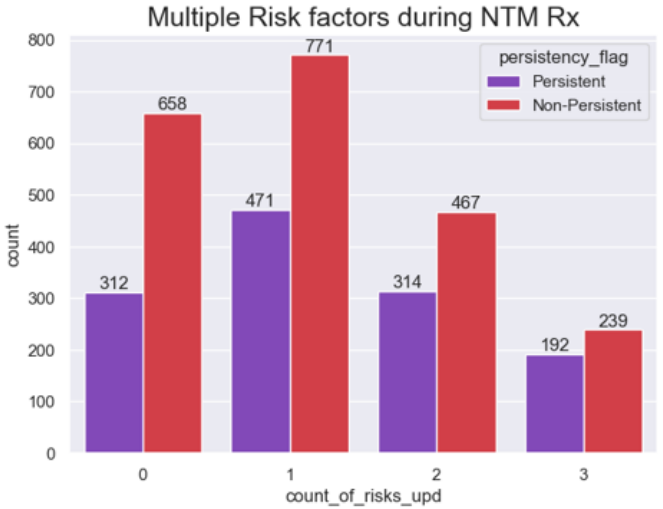
- The dataset contains a total of 69 features that are divided into multiple categories -
  - 1 Target variable: Persistency_Flag
  - 1 Unique identifier for each patient: Ptid
  - 6 Demographic variables of the each patient: Age_Bucket, Gender, Race, Ethnicity, Region, Idn_Indicator
  - 3 Physician Specialist attributes: Ntm_Speciality, Ntm_Specialist_Flag, Ntm_Specialist_Bucket
  - 13 Clinical factors: T-Score details, Risk_Segment details, Multiple risk factors count, DEXA details, Fragility fracture details, Glucocorticoid details
  - 45 Disease/Treatment factors: Injectable drugs, Risk factors, Comorbidities, Concomitancies, Adherence to therapy

Data Preprocessing

- **Outliers Detection and Handling**: 2 features contain outliers – *Dexa_Freq_During_Rx* and *Count_of_Risks*.
- Handled outliers in *Dexa_freq_During_Rx* using **Box-cox transformation**.
- Reduced category count from 0 – 7 to 0 – 3 where 3 signifies number of risks a patient suffers at the same time more than or equal to 3.





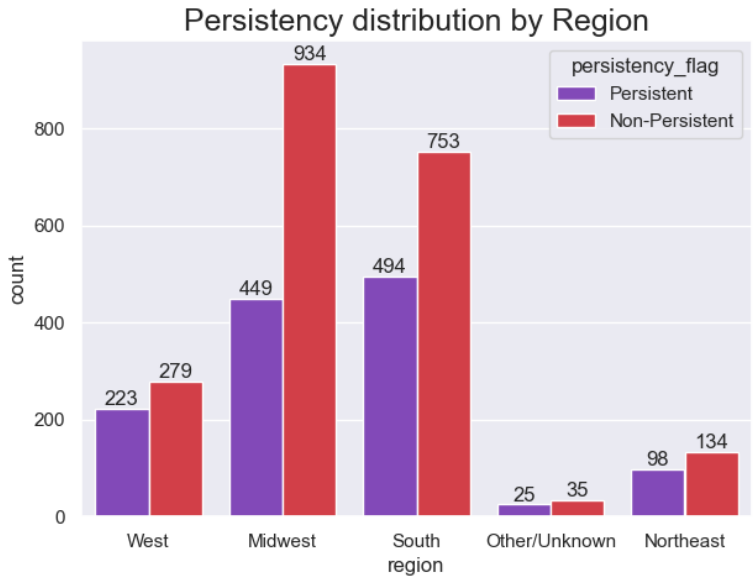


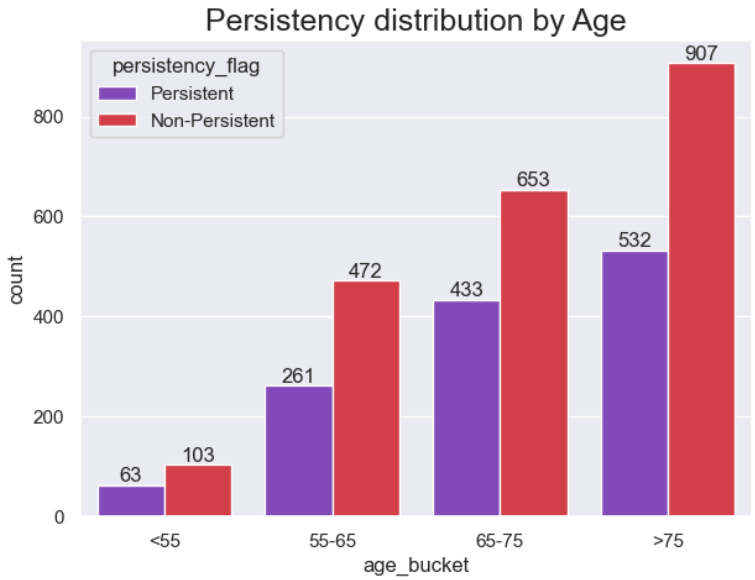
Data Analysis

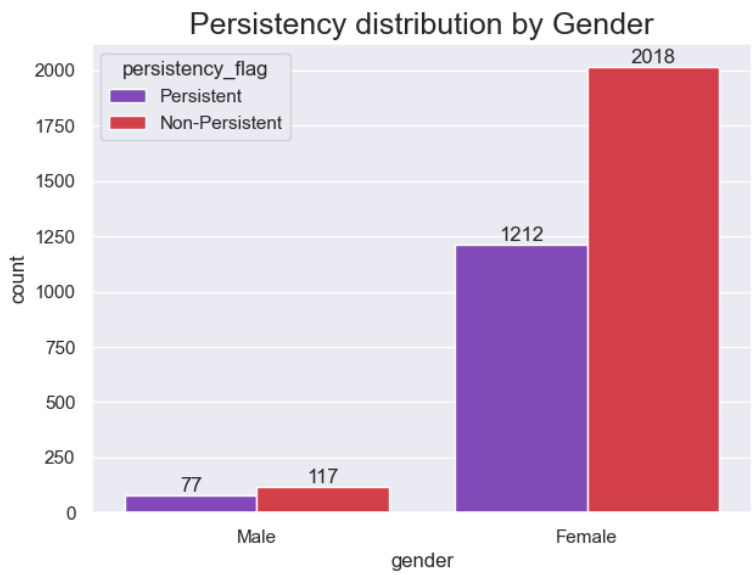
Data Glacier

Your Deep Learning Partner

Demographic Data

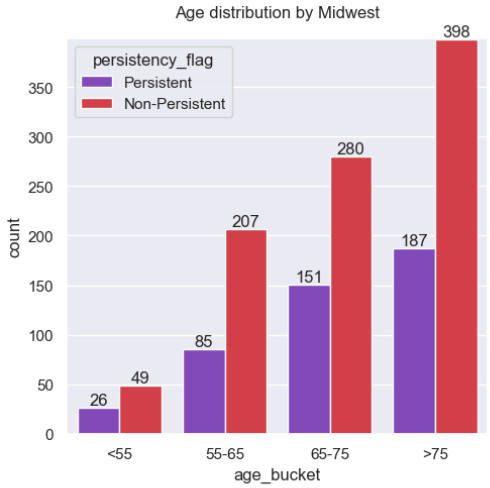


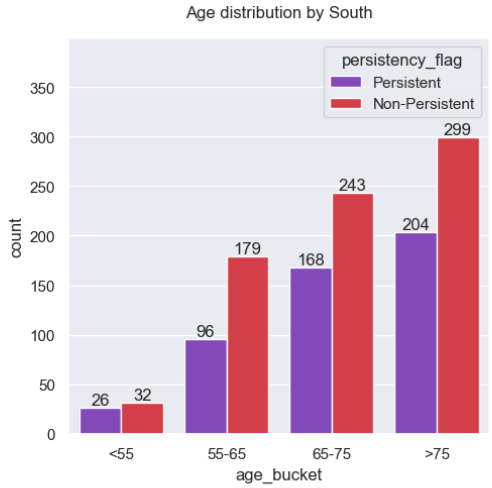


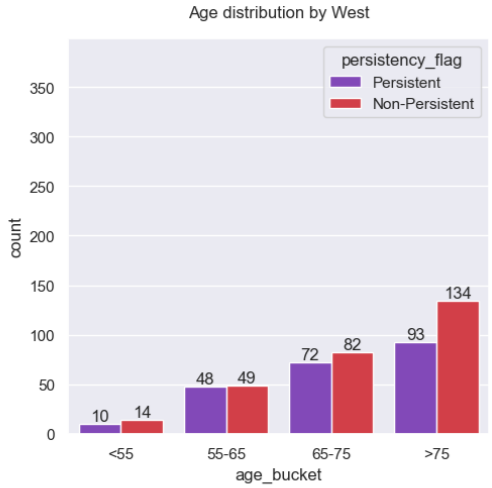


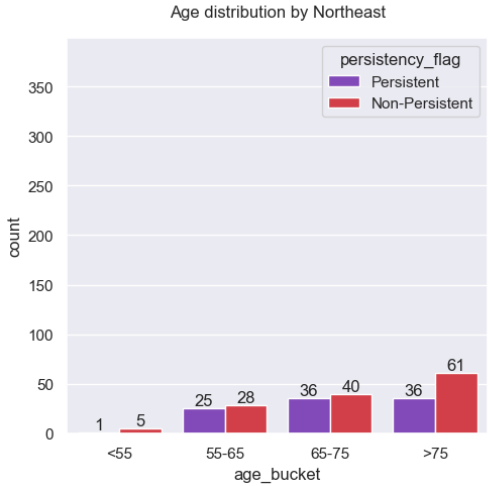
- Majority of the patients recorded are **Females** and most of them are **Non-Persistent** to NTM therapies.
- We can observe that majority of the patients are aged above *55 years* and majority **Non-Persistent** patients fall in the age group of more than *75 years* of age.
- *Midwest*, *South*, and *West* regions display majority of the patients recorded.

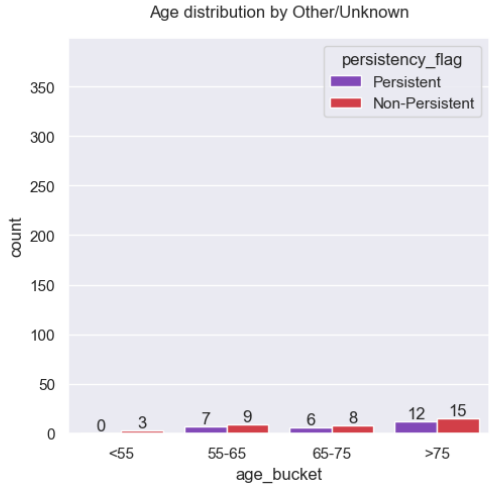
Demographic Data







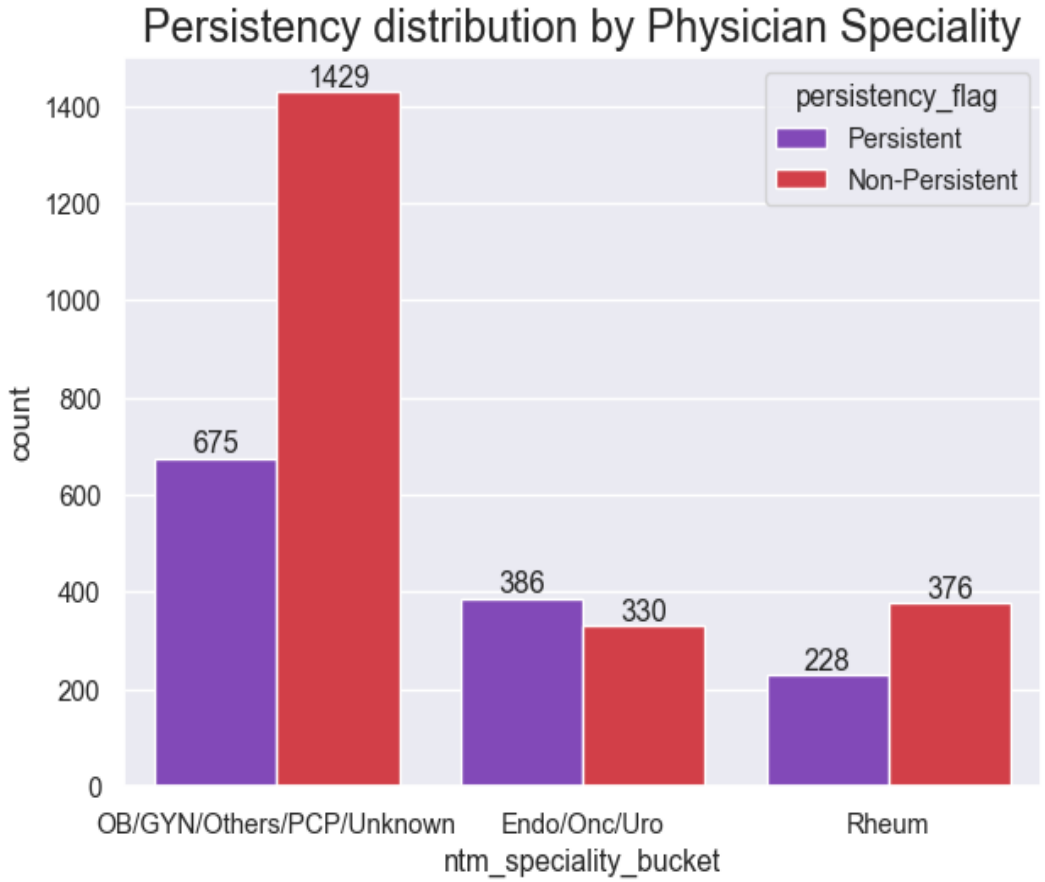


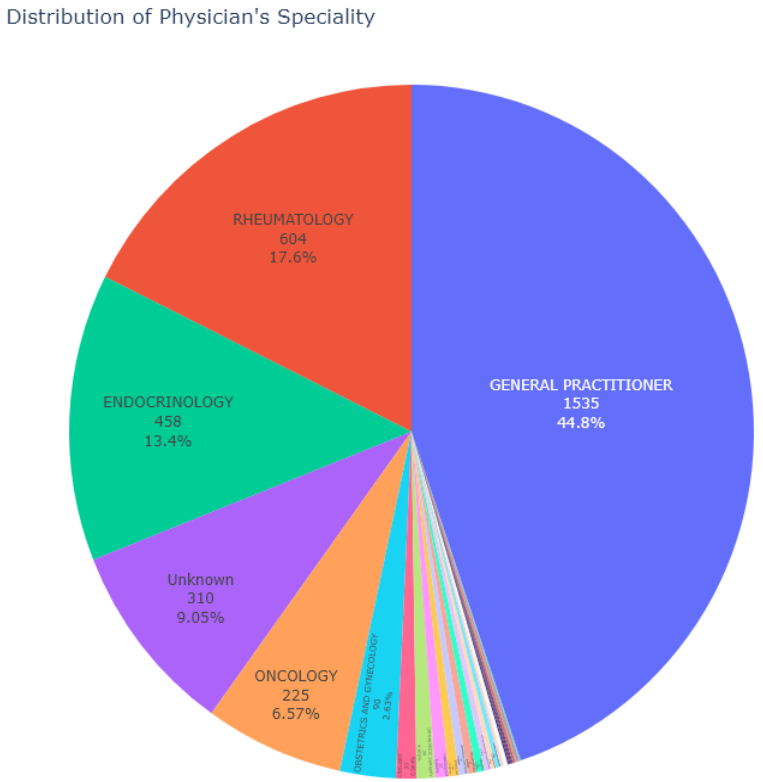


- Majority of **Non-Persistent** patients belong to the age group above *75 years* in the **Midwest** region.

Physician Attributes

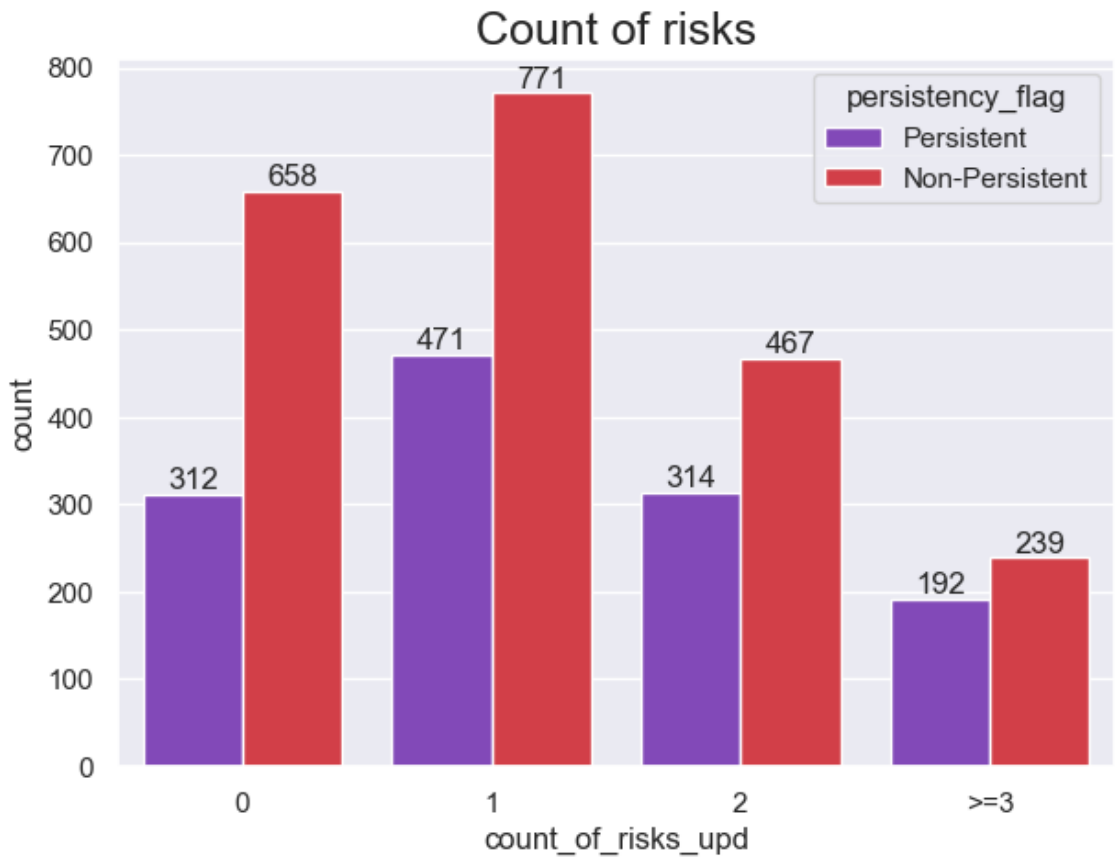
- Around **45%** of Physicians who have prescribed new medication to the patients are *'General Practitioners'*.





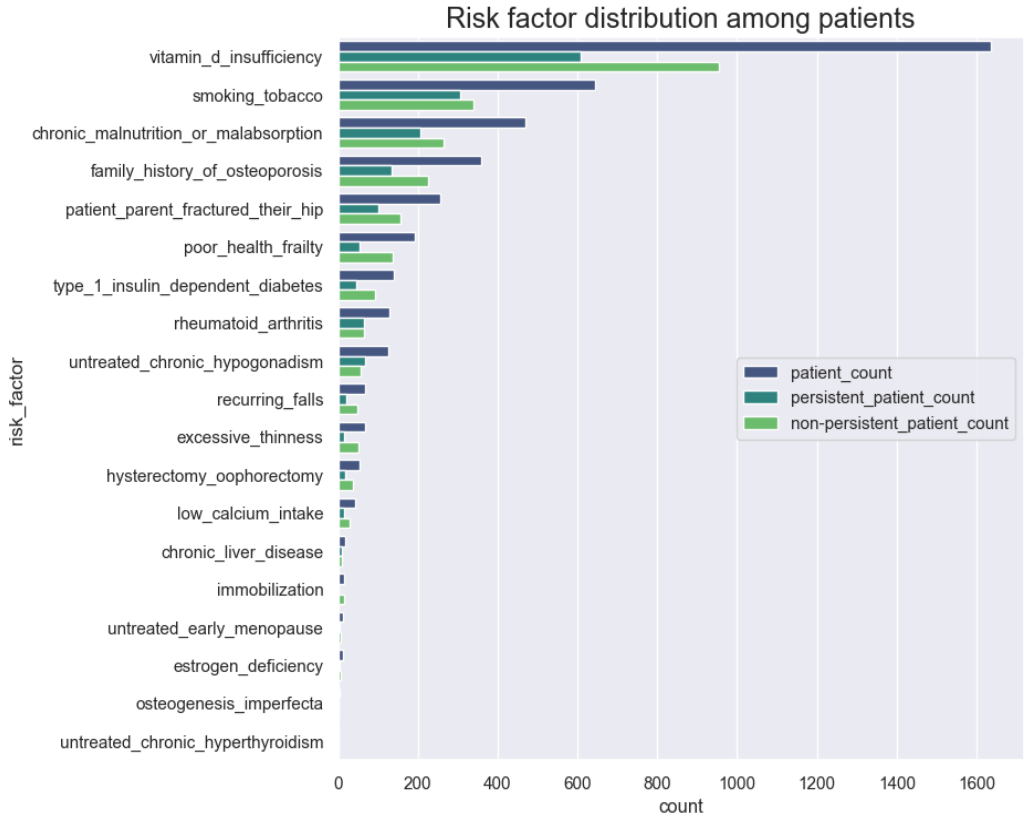
Risk Factors

- As the number of risks per patient increases, the number of **Non-Persistent** patients decreases.



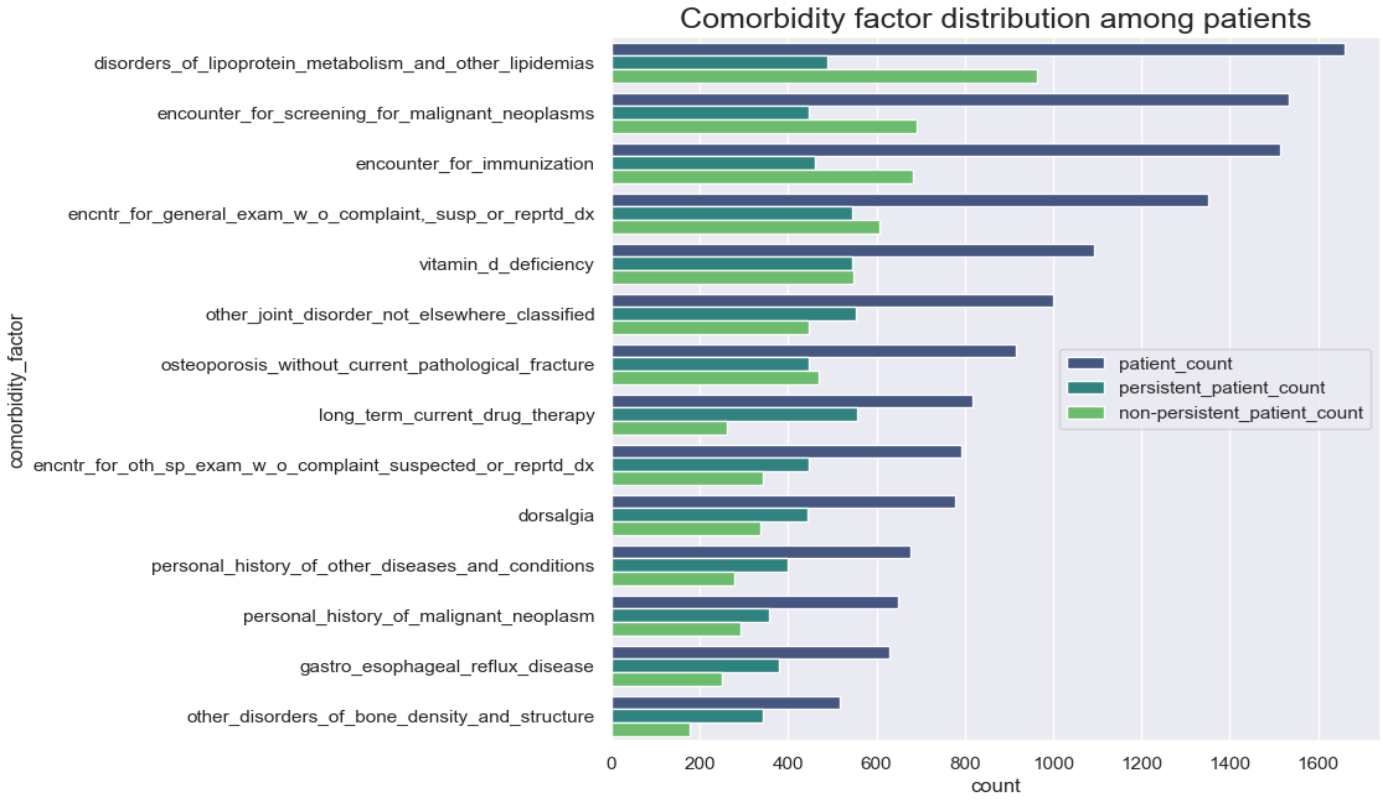
Risk Factors

- Majority of the patients have been susceptible to **Risk Factors** such as *'Vitamin D insufficiency'*, '*smoking tobacco*', *'chronic malnutrition or malabsorption'* and have a *'family history of osteoporosis'*.
- Due to heavy imbalance of data in **Risk Factor** categories, we can reduce dimensionality by reducing the categories capturing less data into a single category.



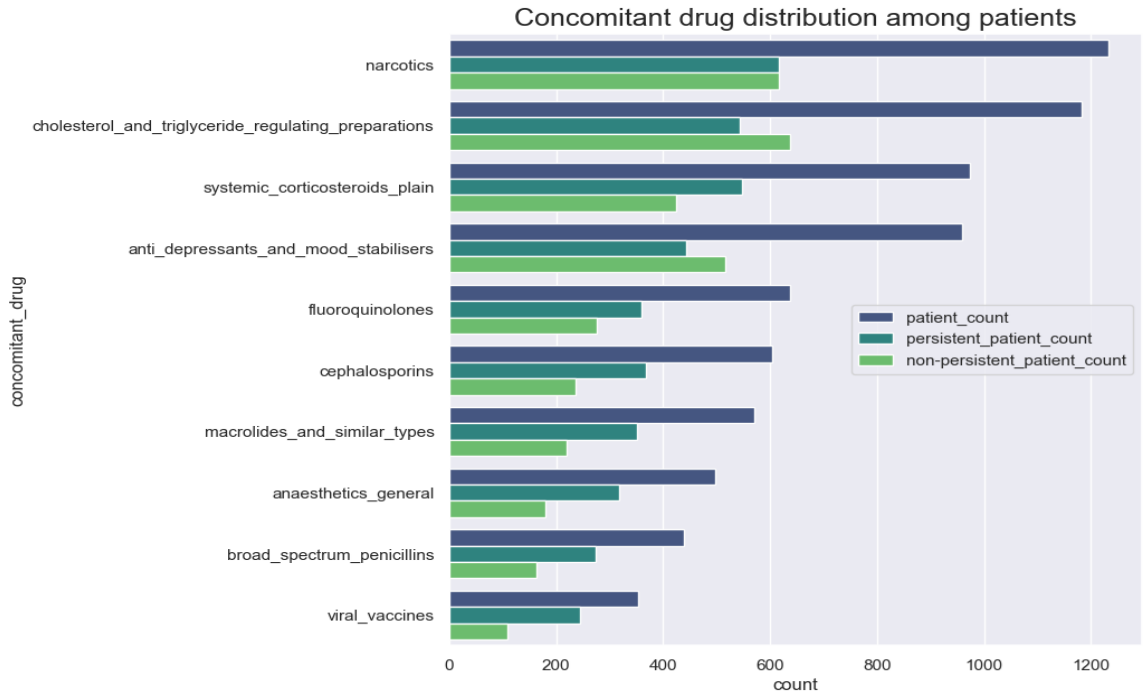
Comorbidity Factors

- There are total 14 **Comorbidity Factors** recorded for each patient.
- The top **Comorbidity Factors** include *disorders_of_lipoprotein_metabolism_and_other_lipidemias*, *encounter_for_screening_for_malignant_neoplasms*, *encounter_for_immunization,* and *encntr_for_general_exam_w_o_complaint,_susp_or_reprtd_dx.*



Concomitant Drugs

- We can see that the graph shows the distribution of patients who have received **Concomitant Drugs** 1 year prior to start therapy.
- The count for **Non-Persistent** patients who have been given **Concomitant Drugs** such as *Narcotics, cholesterol_and_triglyceride_regulating_preparations,* and *anti_depressants_and_mood_stabilisers* is greater compared to the other categories.



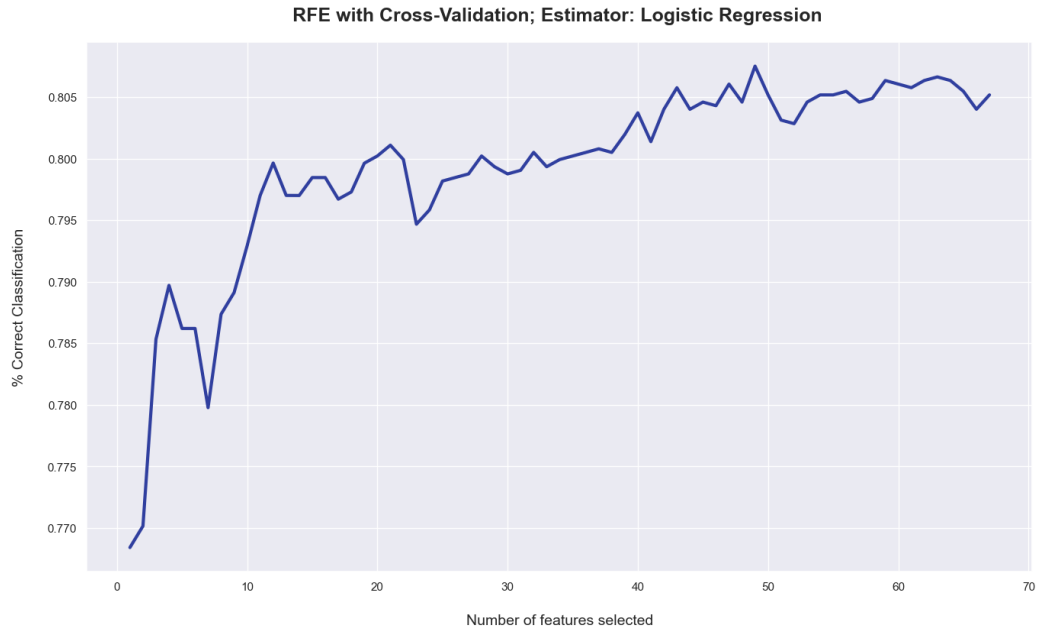
Model Building

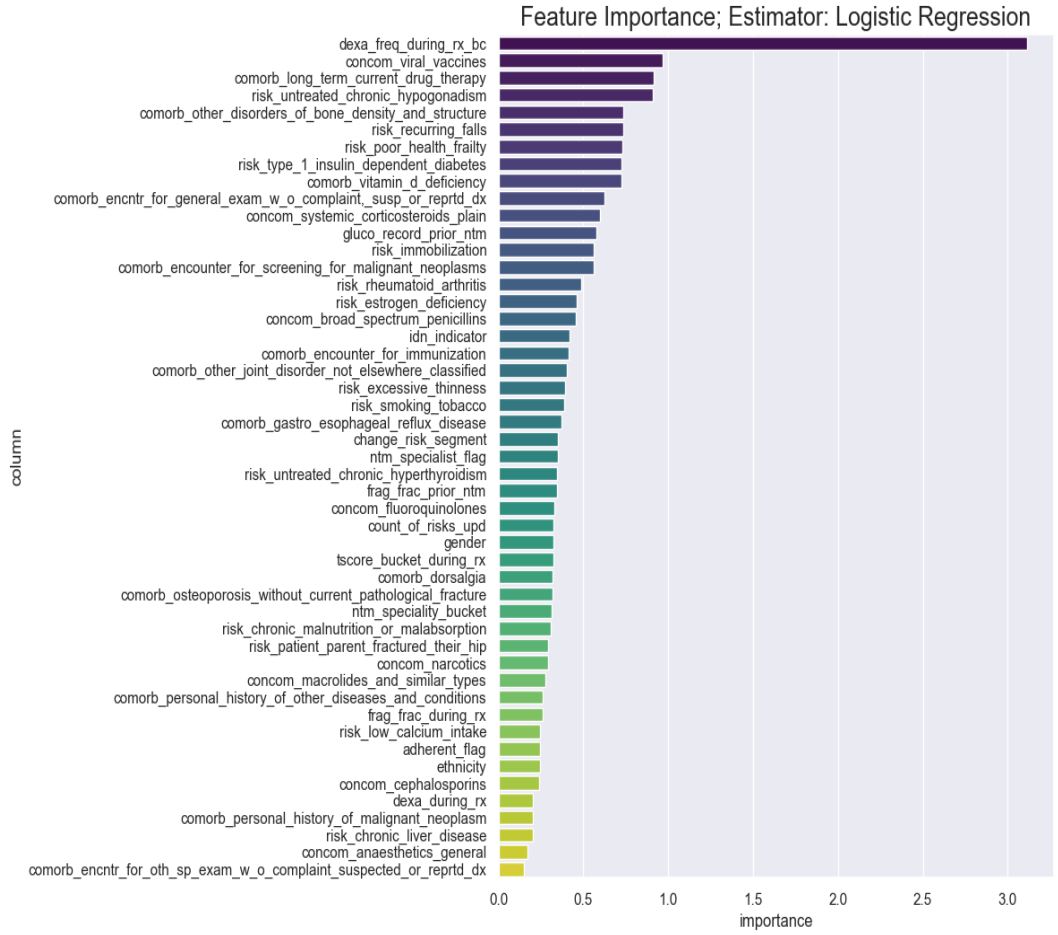
Your Deep Learning Partner

Feature Selection

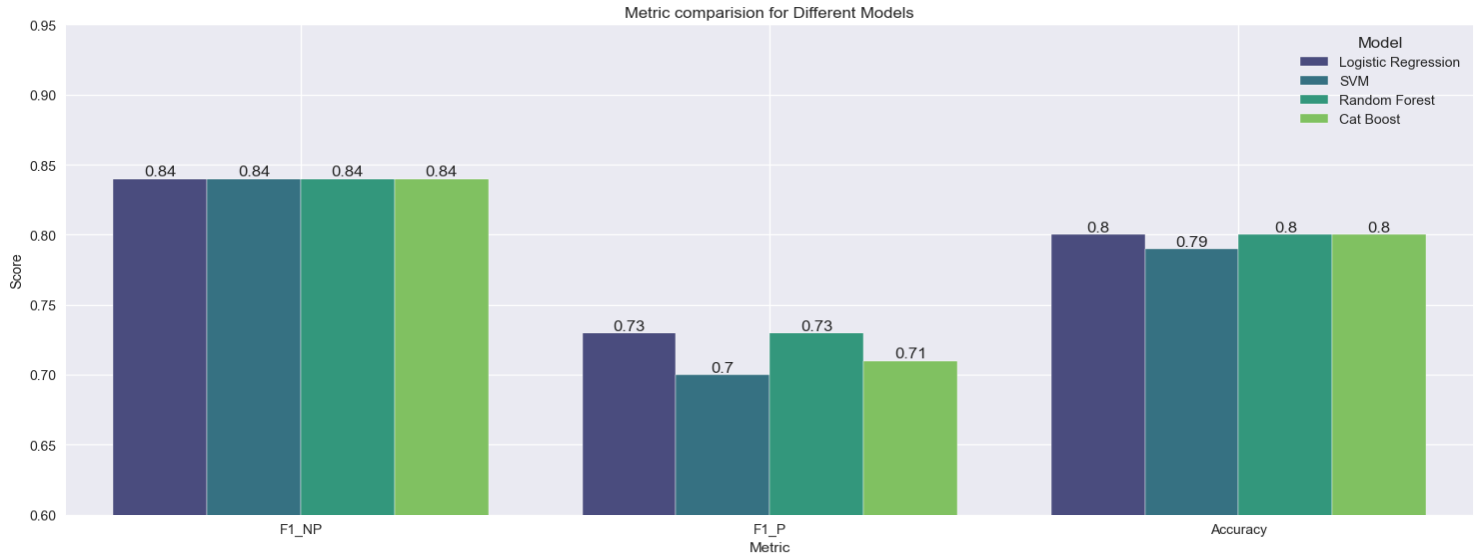
- Applied **Recursive Feature Elimination with Cross Validation(RFECV)** methods for feature selection.

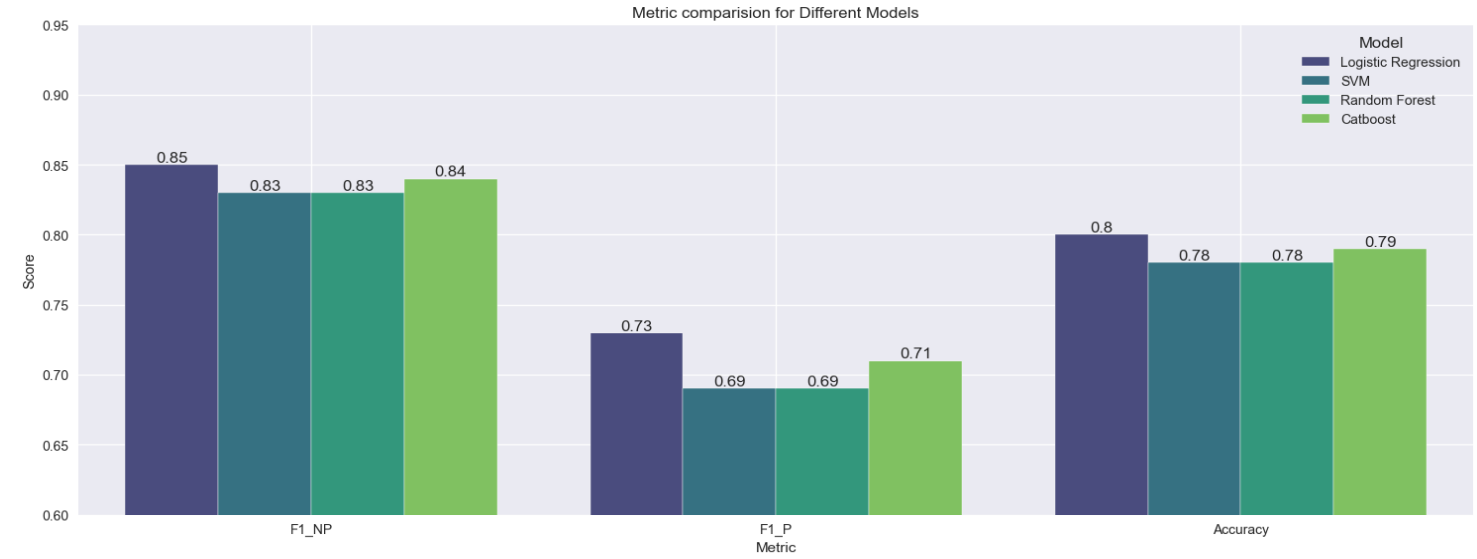
- Obtained 49 optimal features among which 14 features were picked based on feature importance threshold of 0.5 for training the model.



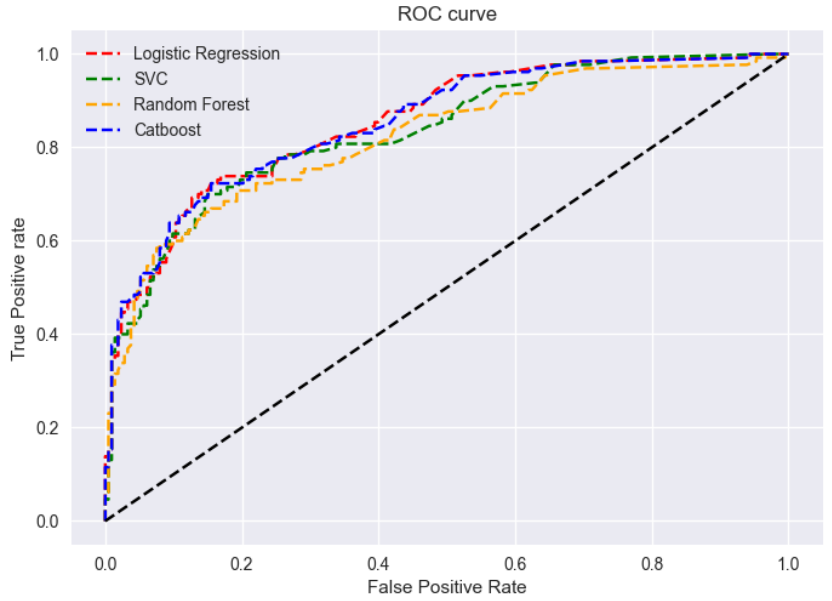


Model Evaluation and Selection



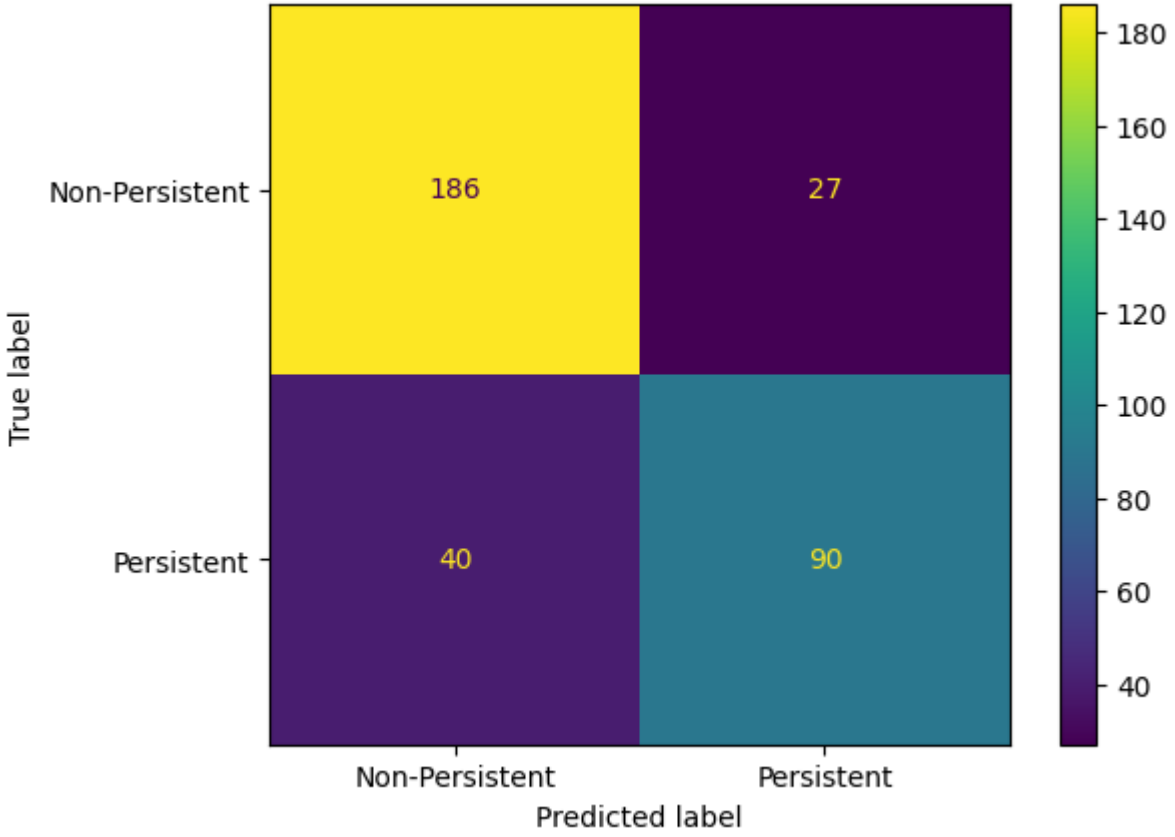


- Trained 4 machine learning models – **Logistic Regression**, **Random Forest Classifier**, **Support Vector Classifier** and **Catboost**.
- **Logistic Regression** performs better and generalises well on unseen data.



Model Evaluation and Selection

- Confusion matrix along with Accuracy, Precision, Recall and F1-scores for **Logistic Regression** on test data.



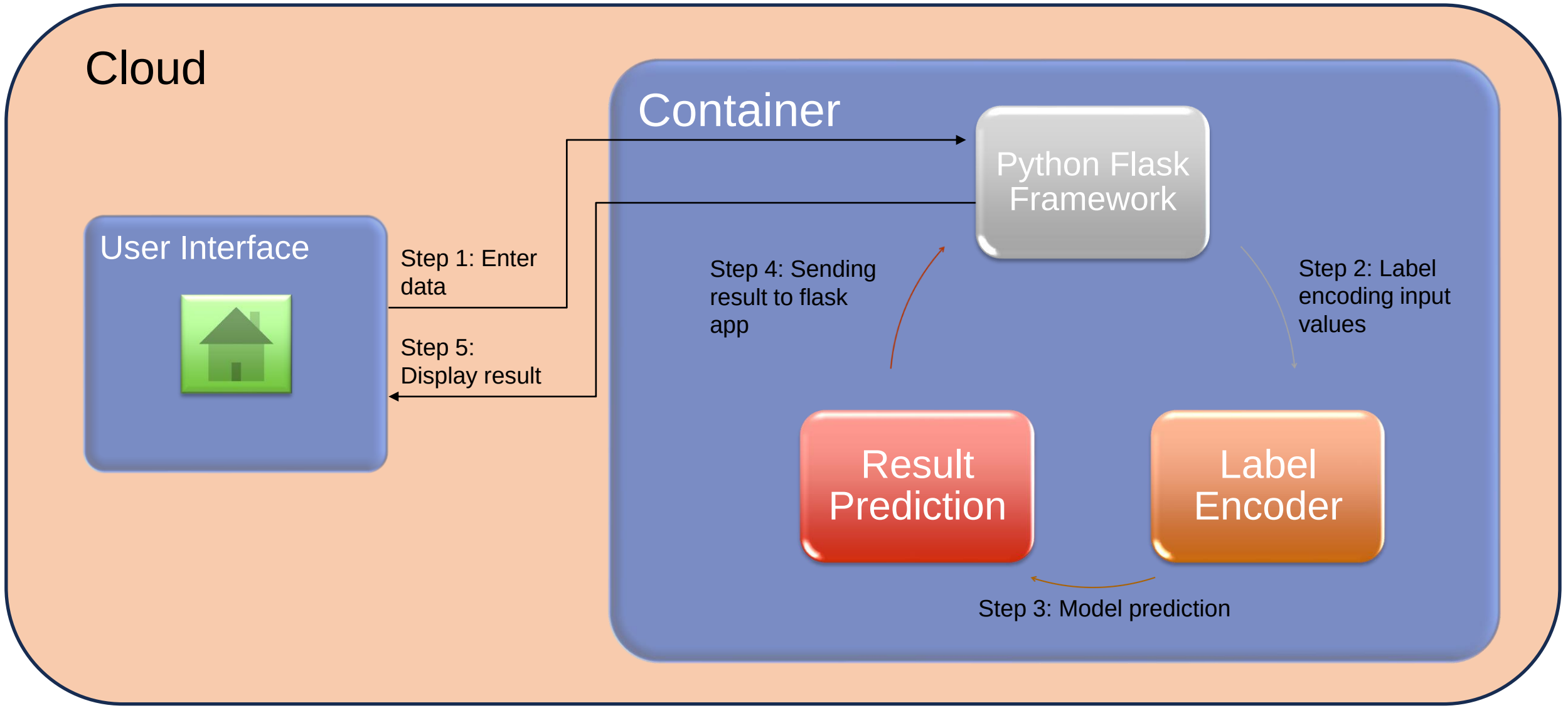
F1 score: 0.73
Accuracy: 0.8046647230320699
Classification Report:

	precision	recall	f1-score	support
0	0.82	0.87	0.85	213
1	0.77	0.69	0.73	130
accuracy			0.80	343
macro avg	0.80	0.78	0.79	343
weighted avg	0.80	0.80	0.80	343

Deployment

Your Deep Learning Partner

Architecture



Thank You

Data Glacier

Your Deep Learning Partner