Exploratory Data Analysis

HealthCare-DataScience Project

Halit Ayberk DEMIR
LISUM19

30.05.2023

Problem Statement

Data Glacier
Your Deep Learning Partner

ABC is a pharmaceutical company and desires to understand the persistency of the drug

per physician description. To solve this issue, ABC company reached an analytics company to

automate this process of identification.

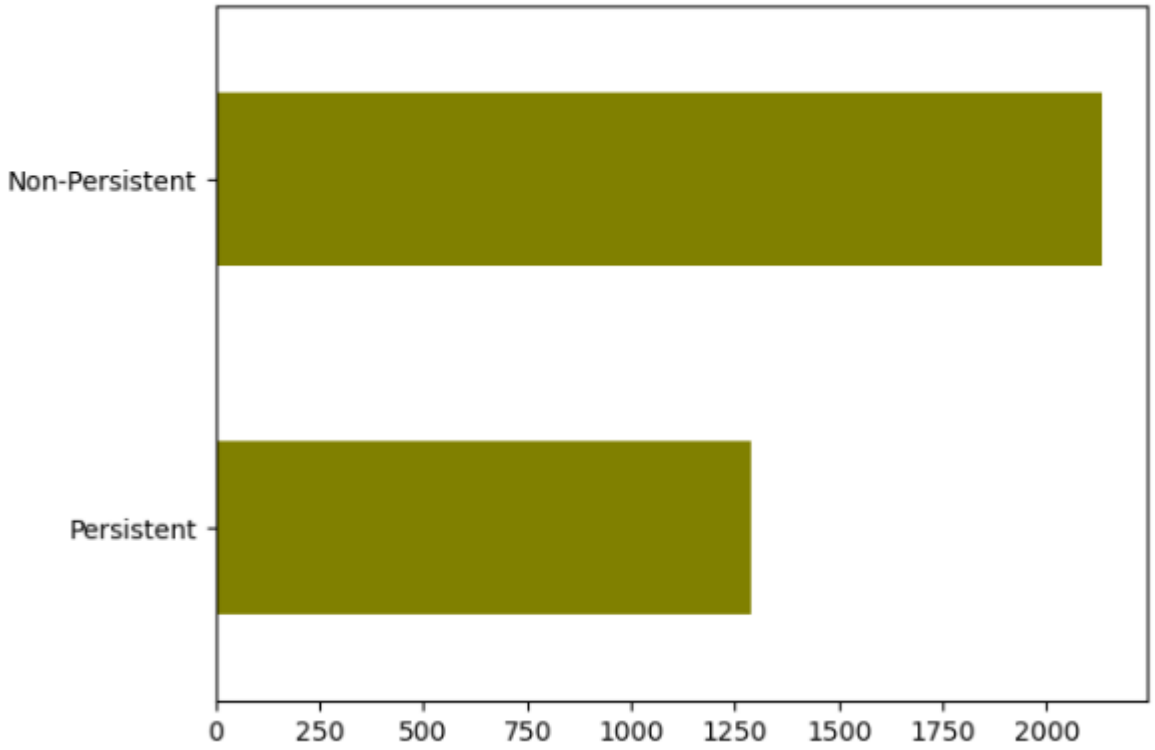
Data Understanding

Data Glacier
Your Deep Learning Partner

#	Column	Non-Null Count	Dtype
0	Ptid	3424 non-null	object
1	Persistency_Flag	3424 non-null	object
2	Gender	3424 non-null	object
3	Race	3424 non-null	object
4	Ethnicity	3424 non-null	object
5	Region	3424 non-null	object
6	Age_Bucket	3424 non-null	object
7	Ntm_Speciality	3424 non-null	object
8	Ntm_Specialist_Flag	3424 non-null	object
9	Ntm_Speciality_Bucket	3424 non-null	object
10	Gluco_Record_Prior_Ntm	3424 non-null	object
11	Gluco_Record_During_Rx	3424 non-null	object
12	Dexa_Freq_During_Rx	3424 non-null	int64
13	Dexa_During_Rx	3424 non-null	object
14	Frag_Frac_Prior_Ntm	3424 non-null	object
15	Frag_Frac_During_Rx	3424 non-null	object
16	Risk_Segment_Prior_Ntm	3424 non-null	object
17	Tscore_Bucket_Prior_Ntm	3424 non-null	object
18	Risk_Segment_During_Rx	3424 non-null	object
19	Tscore_Bucket_During_Rx	3424 non-null	object
20	Change_T_Score	3424 non-null	object
21	Change_Risk_Segment	3424 non-null	object
22	Adherent_Flag	3424 non-null	object
23	Idn_Indicator	3424 non-null	object
24	Injectable_Experience_During_Rx	3424 non-null	object
25	Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	3424 non-null	object
26	Comorb_Encounter_For_Immunization	3424 non-null	object
27	Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	3424 non-null	object
28	Comorb_Vitamin_D_Deficiency	3424 non-null	object
29	Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	3424 non-null	object
30	Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	3424 non-null	object
31	Comorb_Long_Term_Current_Drug_Therapy	3424 non-null	object
32	Comorb_Dorsalgia	3424 non-null	object
33	Comorb_Personal_History_Of_Other_Diseases_And_Conditions	3424 non-null	object
34	Comorb_Other_Disorders_Of_Bone_Density_And_Structure	3424 non-null	object
35	Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	3424 non-null	object
36	Comorb_Osteoporosis_without_current_pathological_fracture	3424 non-null	object
37	Comorb_Personal_history_of_malignant_neoplasm	3424 non-null	object
38	Comorb_Gastro_esophageal_reflux_disease	3424 non-null	object
39	Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	3424 non-null	object
40	Concom_Narcotics	3424 non-null	object
41	Concom_Systemic_Corticosteroids_Plain	3424 non-null	object
42	Concom_Anti_Depressants_And_Mood_Stabilisers	3424 non-null	object
43	Concom_Fluoroquinolones	3424 non-null	object
44	Concom_Cephalosporins	3424 non-null	object
45	Concom_Macrolides_And_Similar_Types	3424 non-null	object
46	Concom_Broad_Spectrum_Penicillins	3424 non-null	object
47	Concom_Anaesthetics_General	3424 non-null	object
48	Concom_Viral_Vaccines	3424 non-null	object
49	Risk_Type_1_Insulin_Dependent_Diabetes	3424 non-null	object
50	Risk_Osteogenesis_Imperfecta	3424 non-null	object
51	Risk_Rheumatoid_Arthritis	3424 non-null	object
52	Risk_Untreated_Chronic_Hyperthyroidism	3424 non-null	object
53	Risk_Untreated_Chronic_Hypogonadism	3424 non-null	object
54	Risk_Untreated_Early_Menopause	3424 non-null	object
55	Risk_Patient_Parent_Fractured_Their_Hip	3424 non-null	object
56	Risk_Smoking_Tobacco	3424 non-null	object
57	Risk_Chronic_Malnutrition_Or_Malabsorption	3424 non-null	object
58	Risk_Chronic_Liver_Disease	3424 non-null	object
59	Risk_Family_History_Of_Osteoporosis	3424 non-null	object
60	Risk_Low_Calcium_Intake	3424 non-null	object
61	Risk_Vitamin_D_Insufficiency	3424 non-null	object
62	Risk_Poor_Health_Frailty	3424 non-null	object
63	Risk_Excessive_Thinness	3424 non-null	object
64	Risk_Hysterectomy_Oophorectomy	3424 non-null	object
65	Risk_Estrogen_Deficiency	3424 non-null	object
66	Risk_Immobilization	3424 non-null	object
67	Risk_Recurring_Falls	3424 non-null	object

- Dataset has 68 columns.
- Our dataset has only one numerical column which is dexa frequency during rx.
-66 Columns are categorical values most of them being simply «Y» or «N»

Understanding Important Columns.



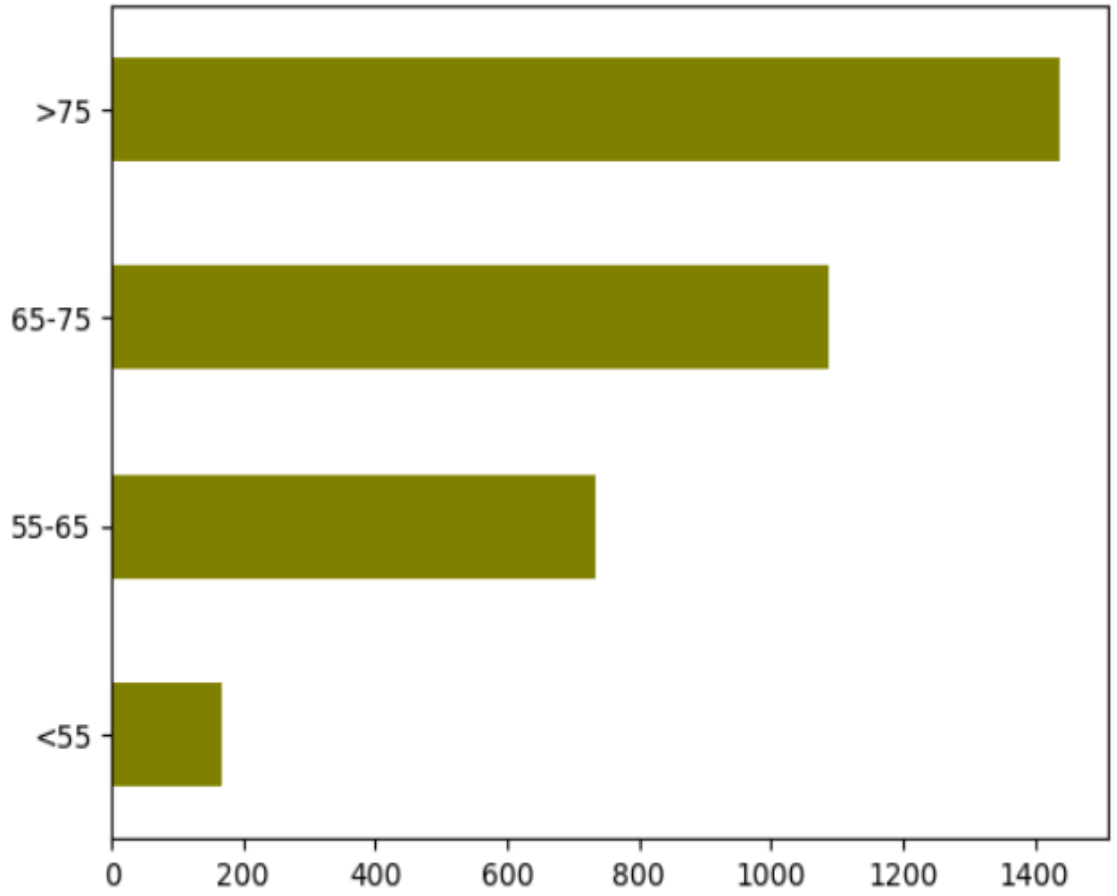
Target Feature, Persistent Flag

Uneven Data;

Non Persistent: 2135

Persistent: 1289

Understanding Important Columns.



Age Bucket;

>75 : 1439

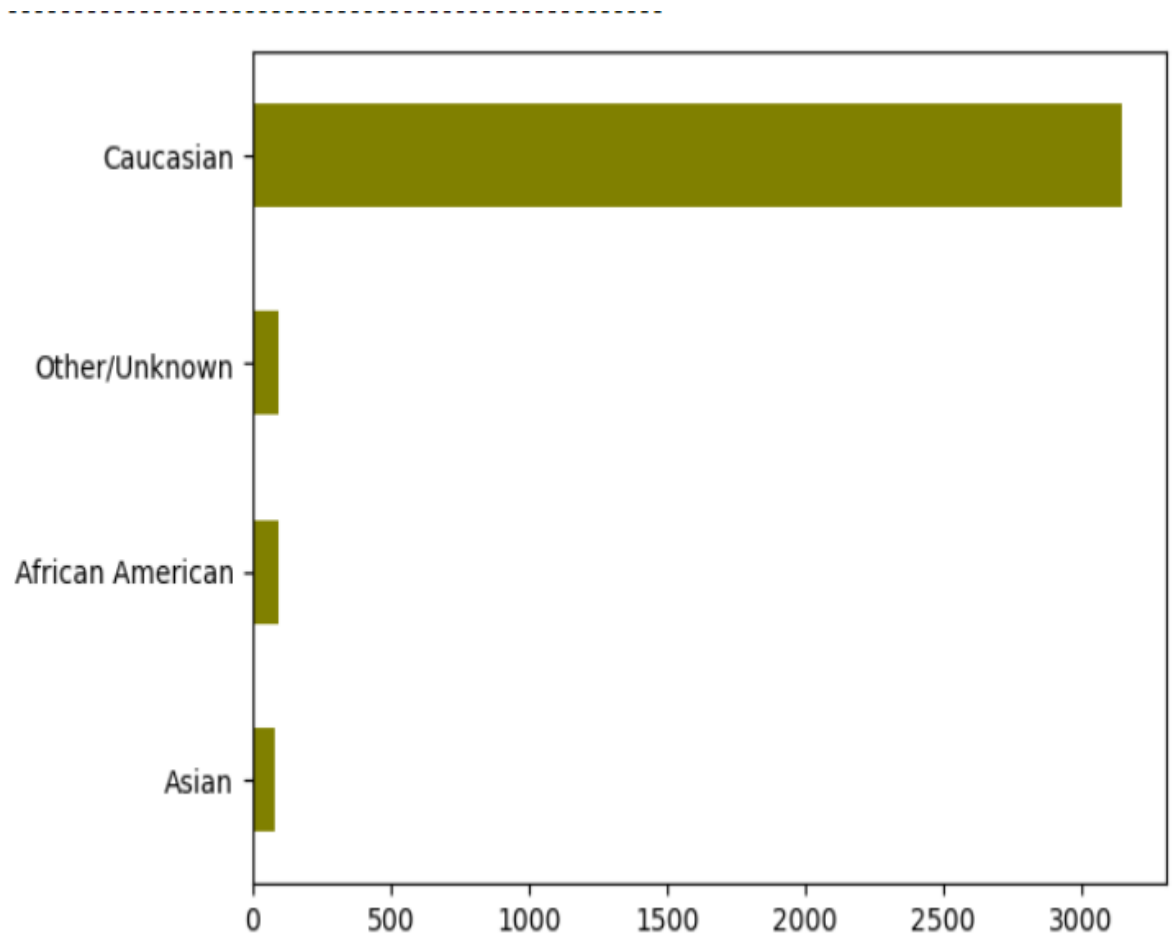
65-75: 1086

55-65: 733

<55: 166

Understanding Important Columns.

Data Glacier
Your Deep Learning Partner



Race

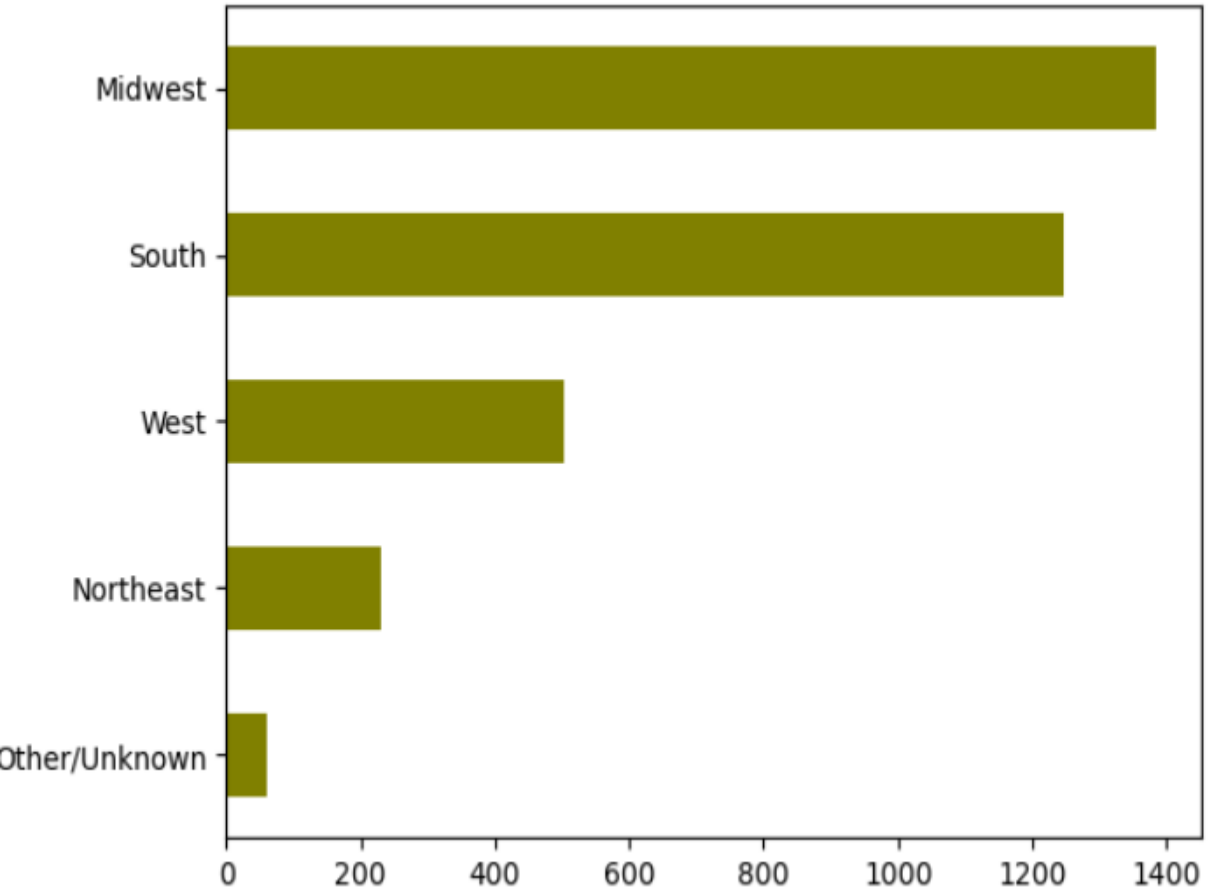
Caucasian: 3148,

Other/Unknown: 97,

African American: 95,

Asian: 84,

Understanding Important Columns



Region

Midwest: 1383

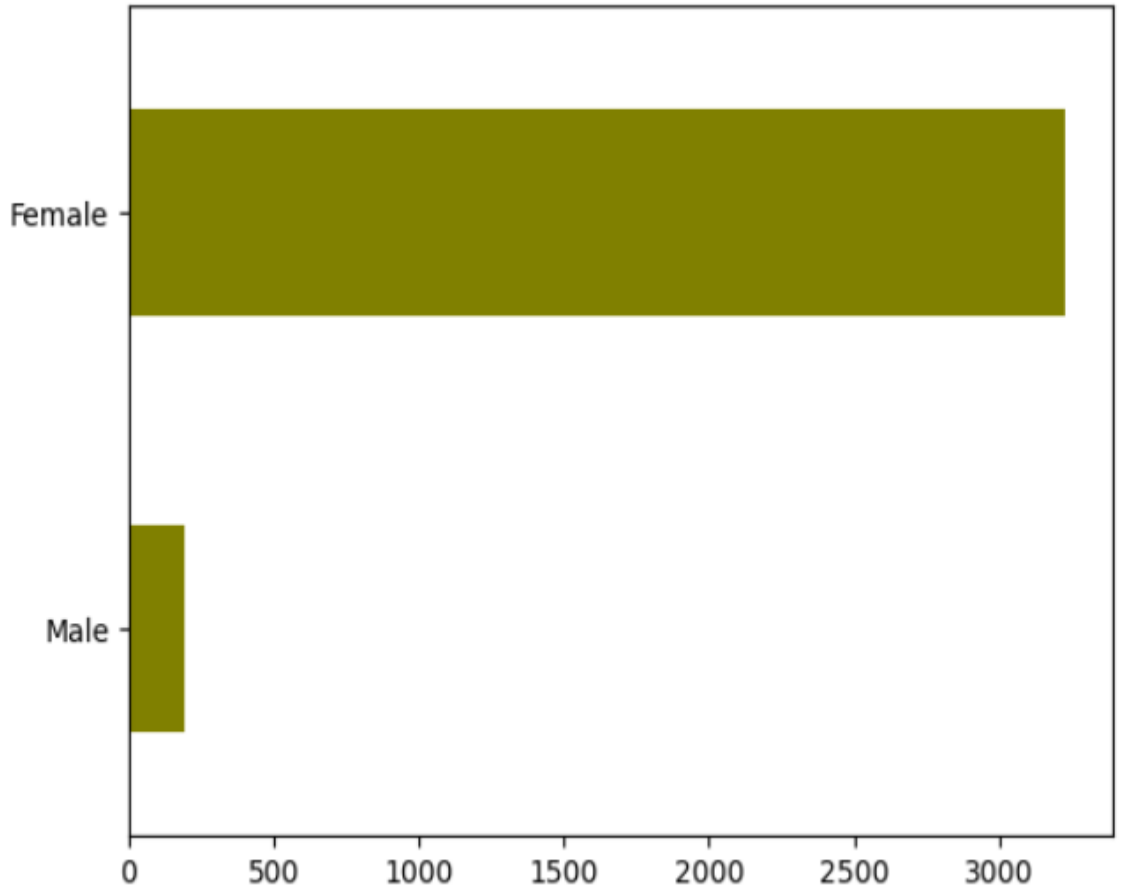
South: 1247

West: 502

Other/Unkown:60

Understanding Important Columns.

Data Glacier
Your Deep Learning Partner



Gender

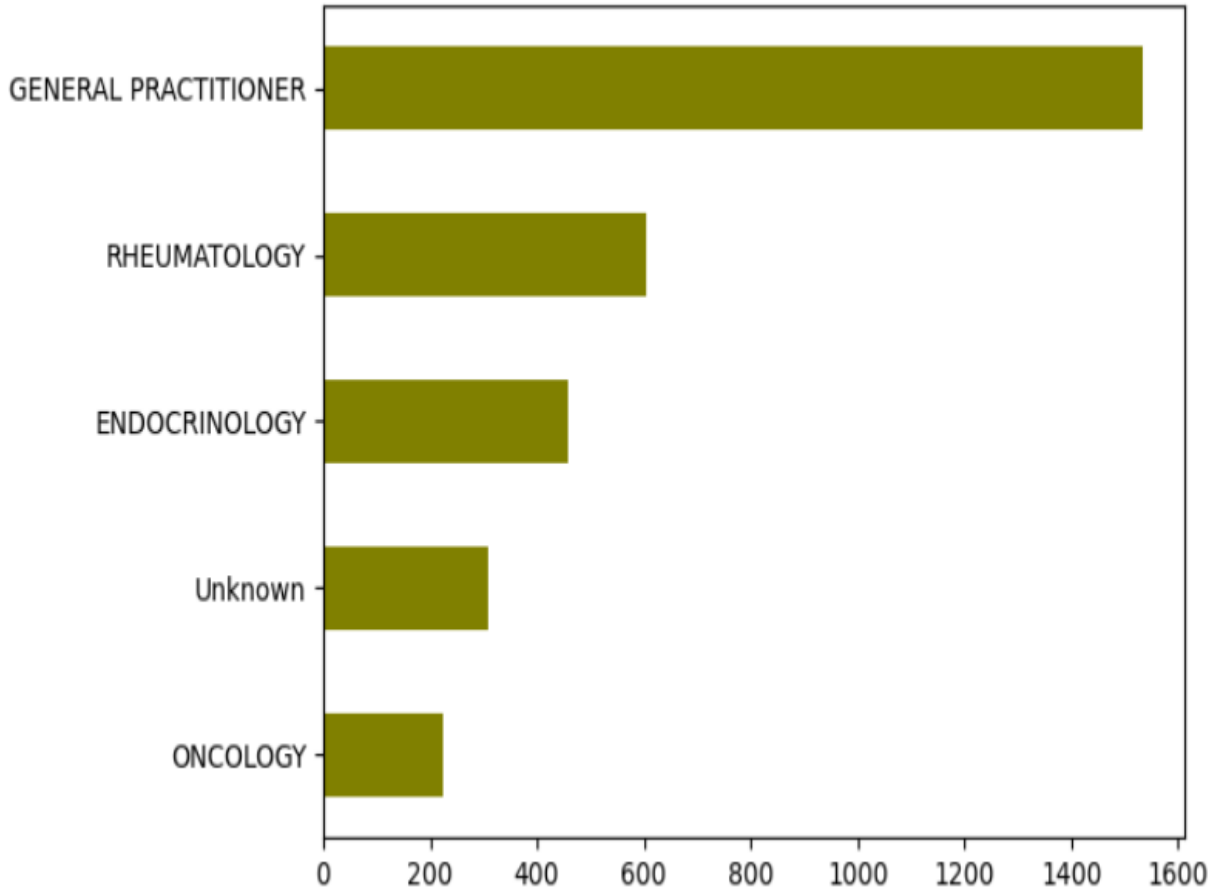
Uneven Dataset

Female Dominating Dataset

Female: 3424

Male: 194

Understanding Important Columns.



Ntm Speciality

36 Unique Values

With Leading;

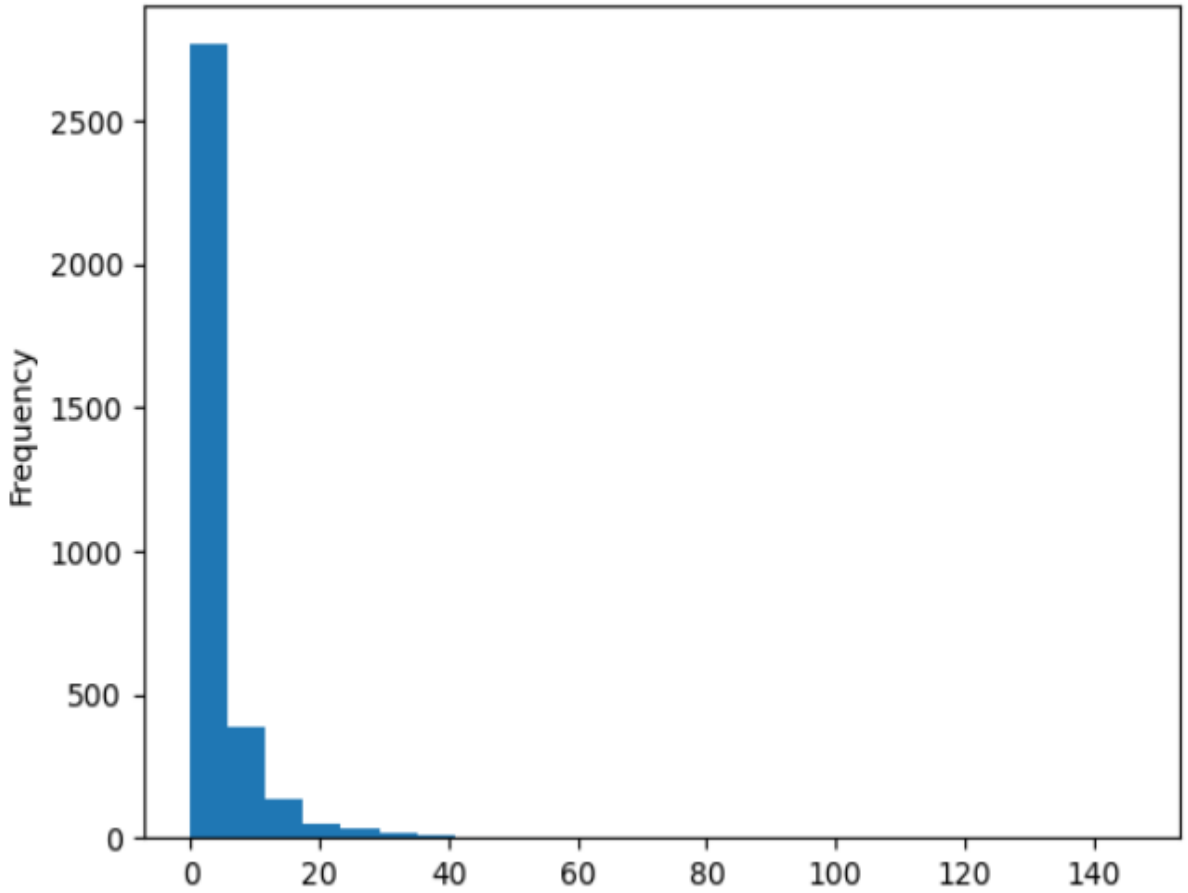
General Practitioner: 1535

Rheumatology: 604

Endocrinology: 458

Understanding Important Columns.

Data Glacier
Your Deep Learning Partner



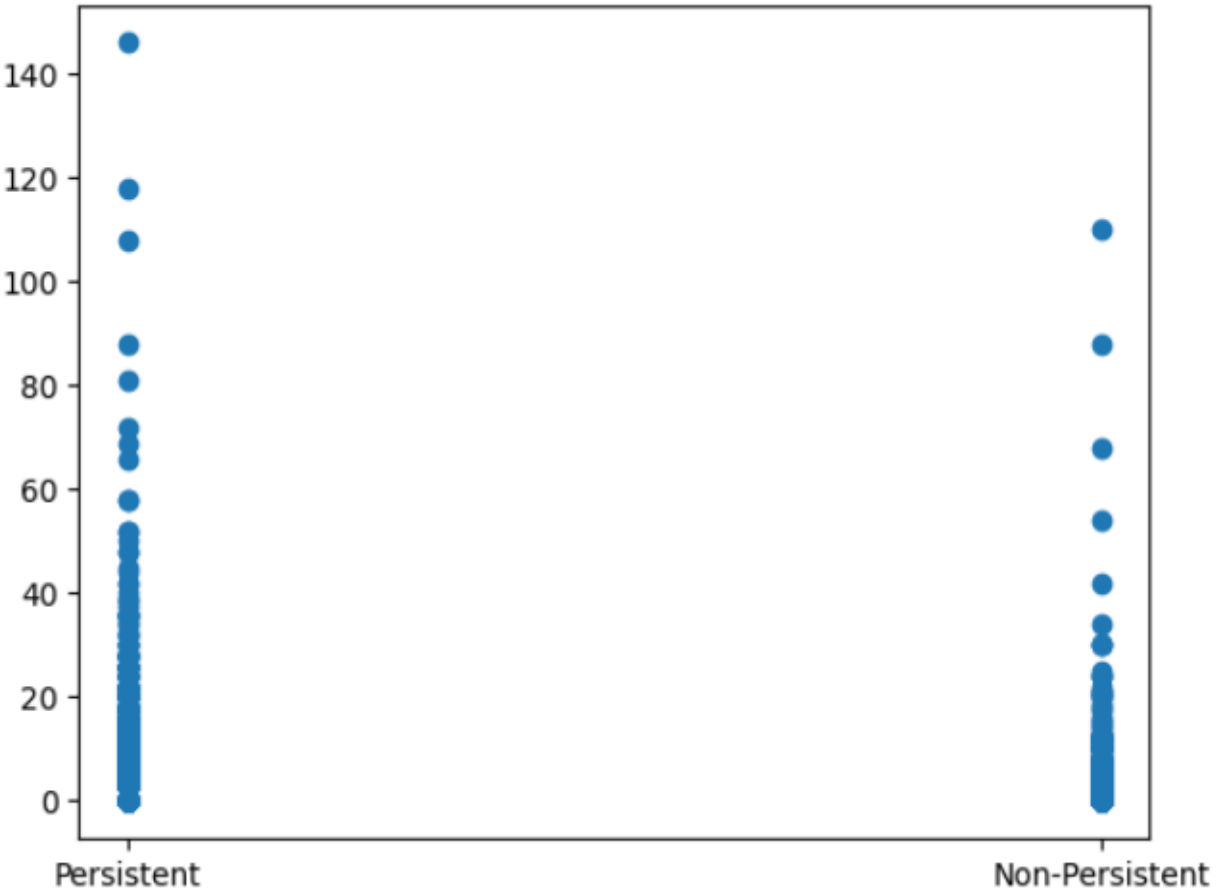
Dexa Frequancy During Rx

Only Numerical Column In The Dataset

	Dexa_Freq_During_Rx
count	3424.000000
mean	3.016063
std	8.136545
min	0.000000
25%	0.000000
50%	0.000000
75%	3.000000
max	146.000000

Finding Outliers.

Data Glacier
Your Deep Learning Partner



Distribution of Dexa Frequancy During Rx with target column persistancy flag.

Clearly outliers are visible.

This outliers are dropped.

Unknown Values In the Dataset.

Data Glacier
Your Deep Learning Partner

Race	2.832944
Ethnicity	2.657710
Region	1.752336
Ntm_Speciality	9.053738
Risk_Segment_During_Rx	43.720794
Tscore_Bucket_During_Rx	43.720794
Change_T_Score	43.720794
Change_Risk_Segment	65.099299

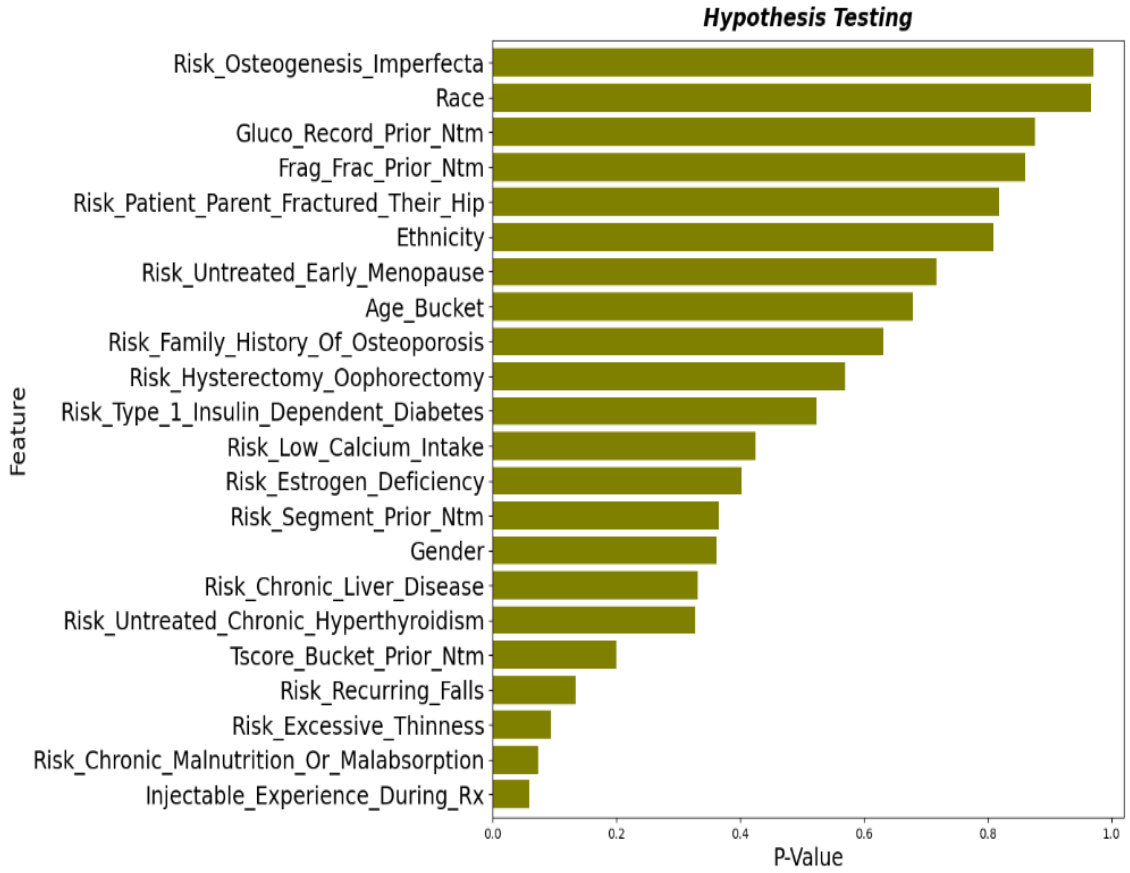
Percentage of Unknown Values In the columns.

If the percentage is higher than 40, column is dropped.

Else, the unknown data is filled with the most frequent value

Chi2 Analysis

Data Glacier
Your Deep Learning Partner

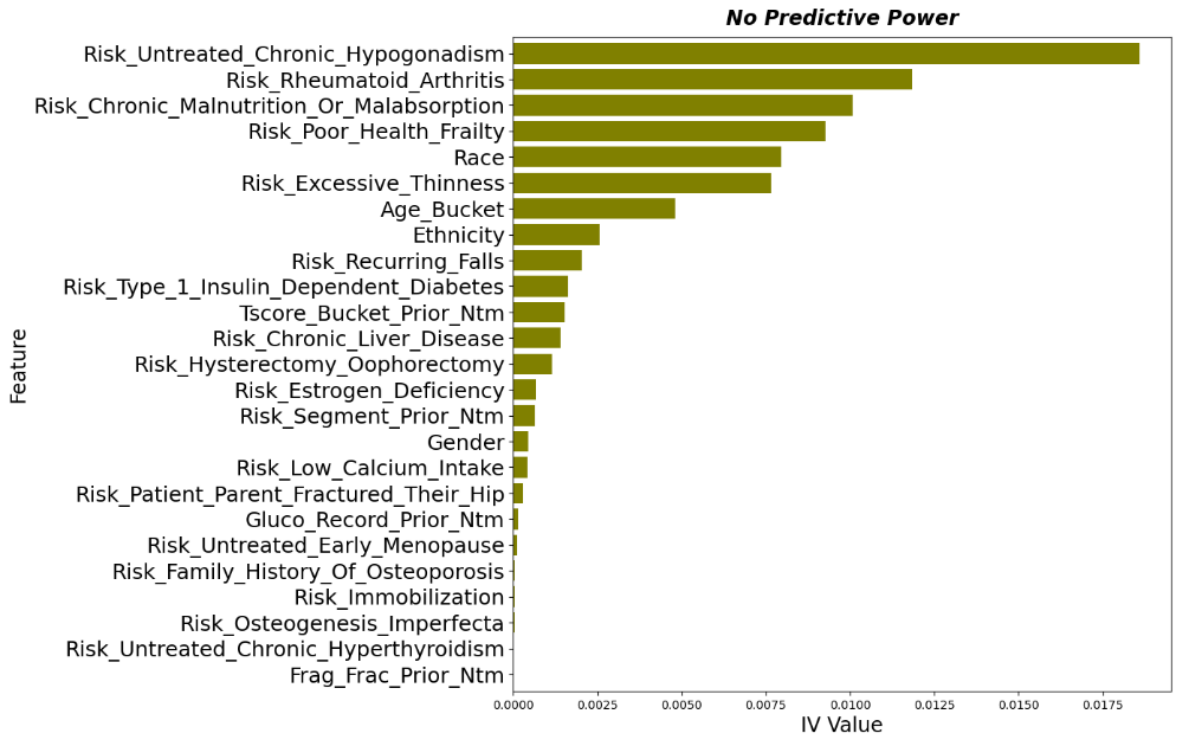


Reduction of Columns

At the left columns with low assosiation are shown

If P value is higher it has a higher probability of being not assosiated with the target variable

Information Value



IV technique for feature selection

Columns with no predictive power.

Suggested ML Models:

Data Glacier
Your Deep Learning Partner

Logistic Regression: WOE/IV

DecisionTree/RandomForest

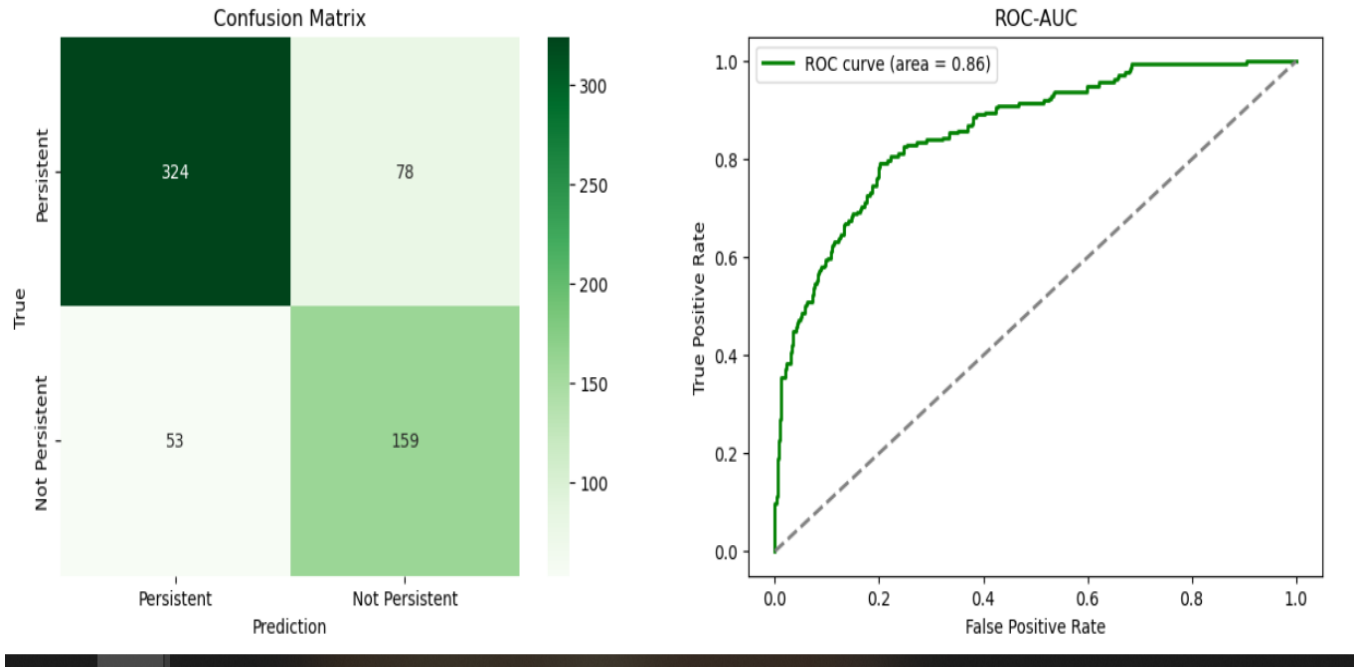
Gradient Boosting Machines (GBM) / XGBoost / LightGBM

Support Vector Machines (SVM):

Neural Networks

Model 1 Logistic Regression (Base Model)

Data Glacier
Your Deep Learning Partner



Accuracy: 0.7866449511400652

Precision: 0.6708860759493671

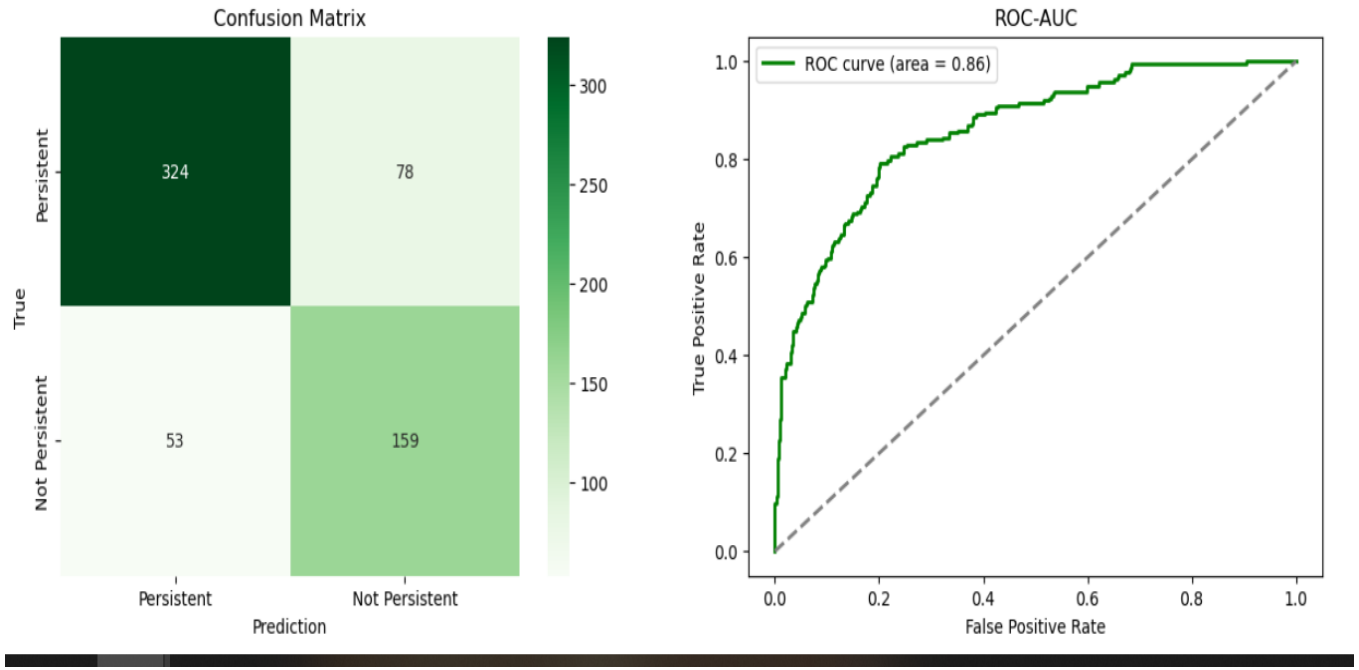
Recall: 0.75

F1 Score: 0.7082405345211581

AUC-ROC: 0.7779850746268656

Model 1 Logistic Regression (Base Model)

Data Glacier
Your Deep Learning Partner



Accuracy: 0.7866449511400652

Precision: 0.6708860759493671

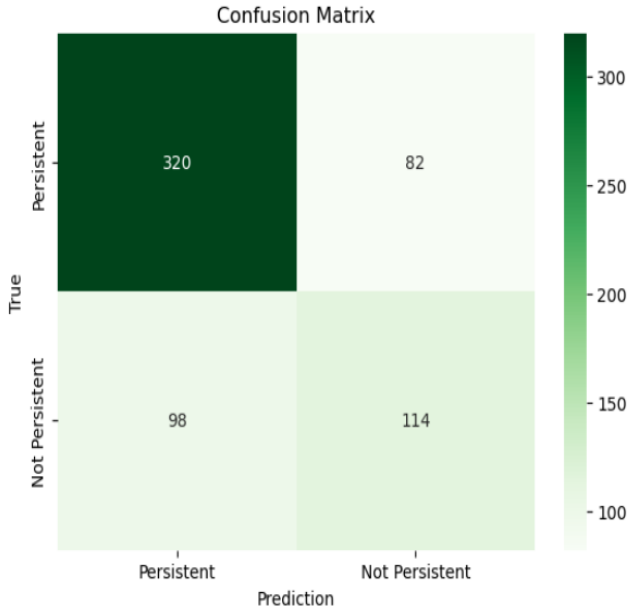
Recall: 0.75

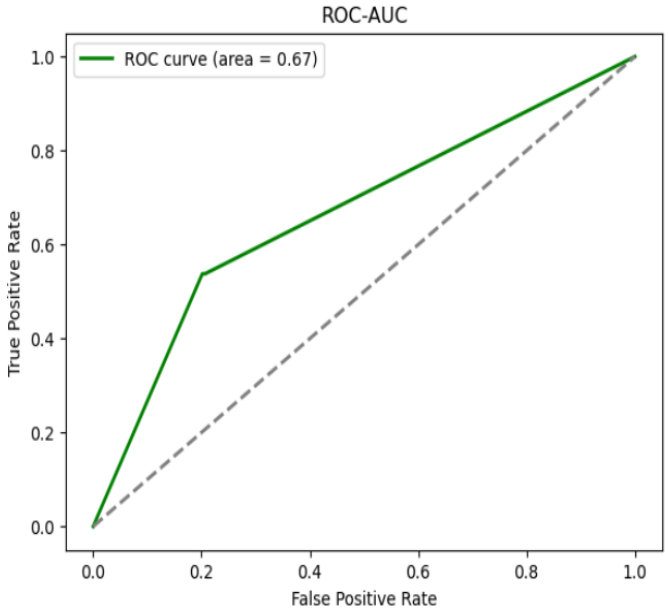
F1 Score: 0.7082405345211581

AUC-ROC: 0.7779850746268656

Model 2 Decision Tree.

Data Glacier
Your Deep Learning Partner





Accuracy: 0.706840390879478 8

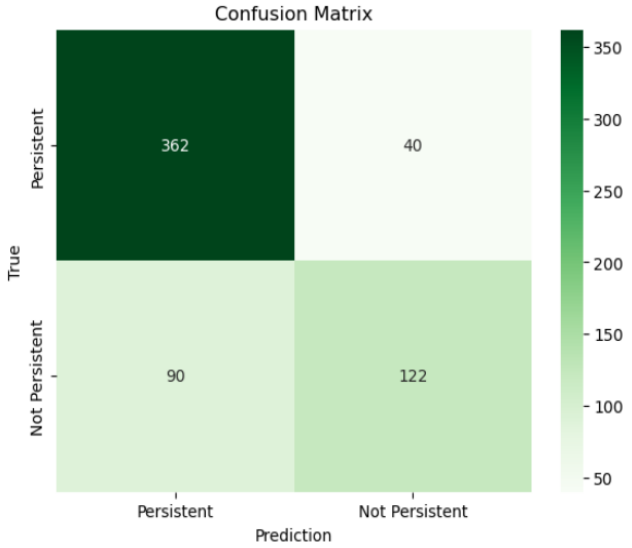
Precision: 0.581632653061224 5

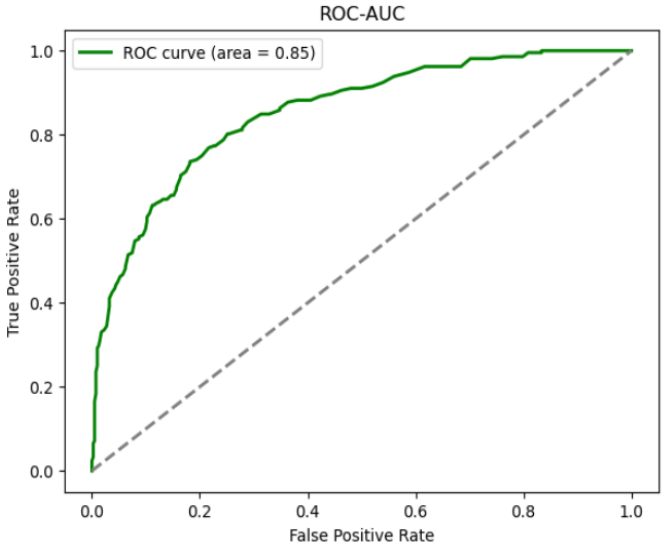
Recall: 0.537735849056603 8

F1 Score: 0.558823529411764 6

AUC-ROC: 0.666877874777058 2

Model 3 Random Forest Classifier





Accuracy: 0.7882736156351792

Precision: 0.7530864197530864

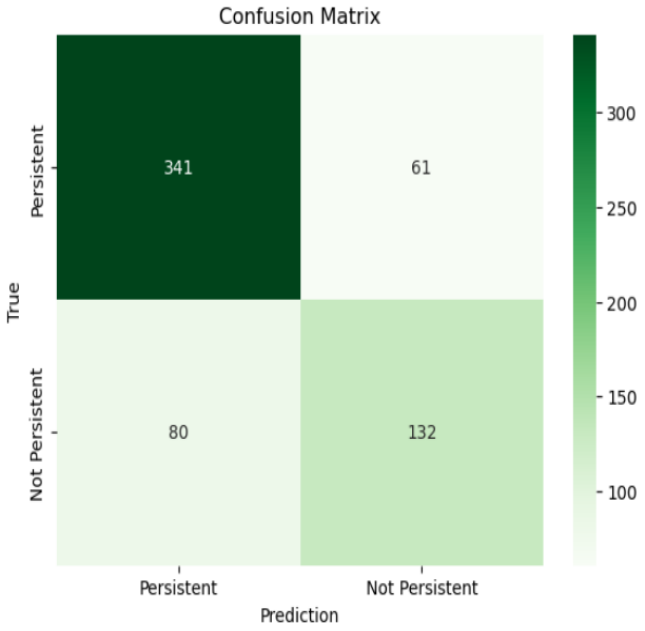
Recall: 0.7554716981132075

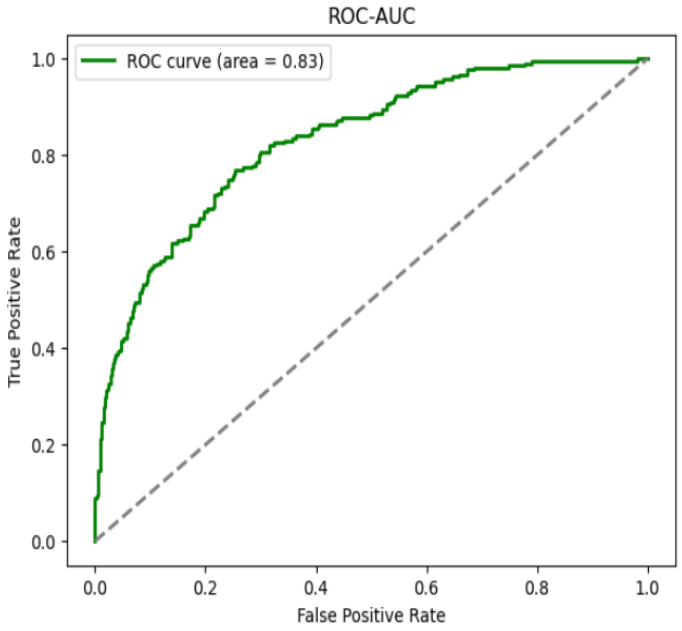
F1 Score: 0.7524064171122994

AUC-ROC: 0.7379846052755092

Model 4 Gradient Boosting

Data Glacier
Your Deep Learning Partner





Accuracy: 0.7703583061889251

Precision: 0.6839378238341969

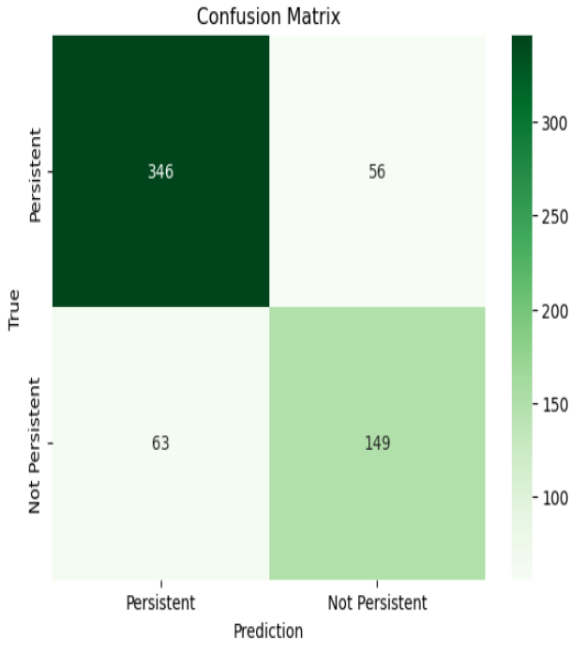
Recall: 0.6226415094339622

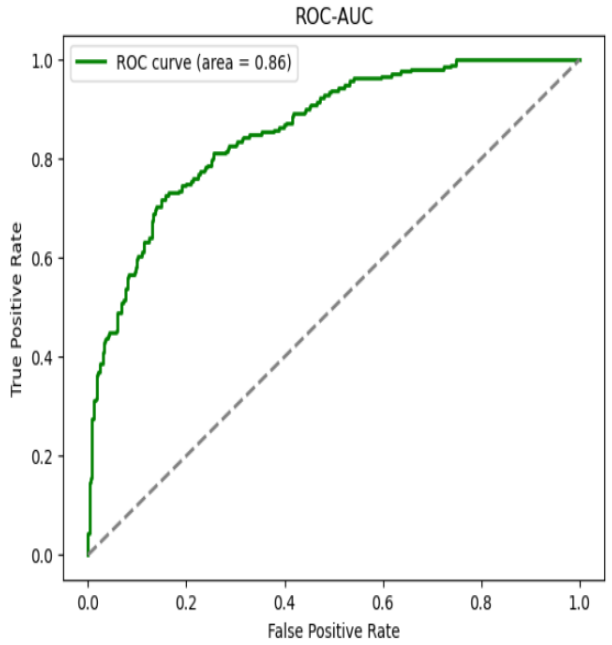
F1 Score: 0.6518518518518519

AUC-ROC: 0.735450107950812

Model 5 Support Vector Machines

Data Glacier
Your Deep Learning Partner





Accuracy: 0.8061889250814332

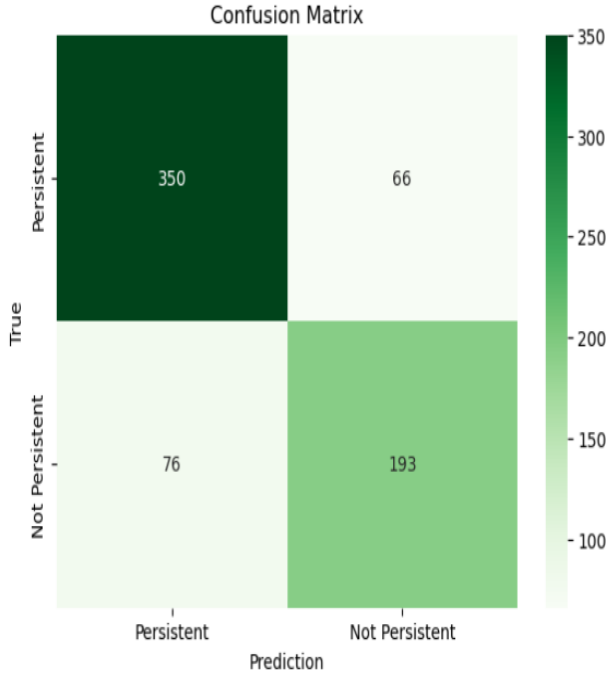
Precision: 0.7268292682926829

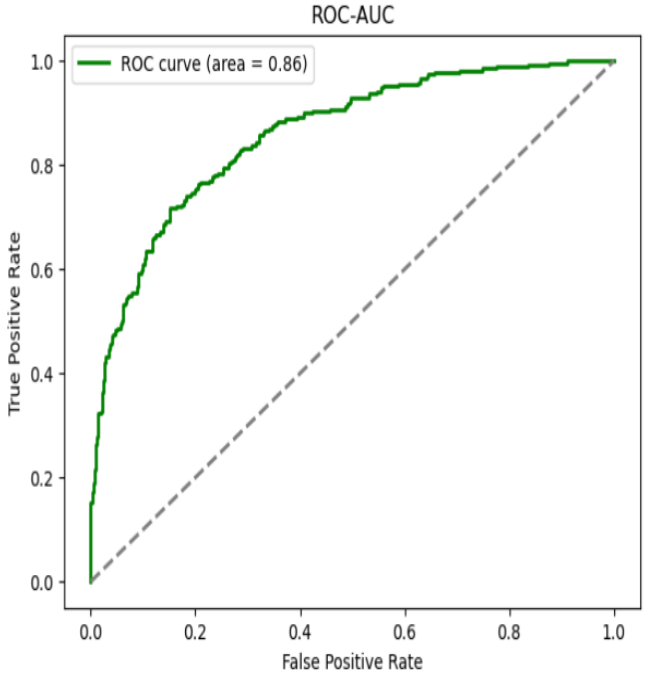
Recall: 0.7028301886792453

F1 Score: 0.7146282973621103

AUC-ROC: 0.7817633530460902

Model 6 Logistic Regression WoE





Accuracy: 0.792700729927074

Precision: 0.7451737451737451

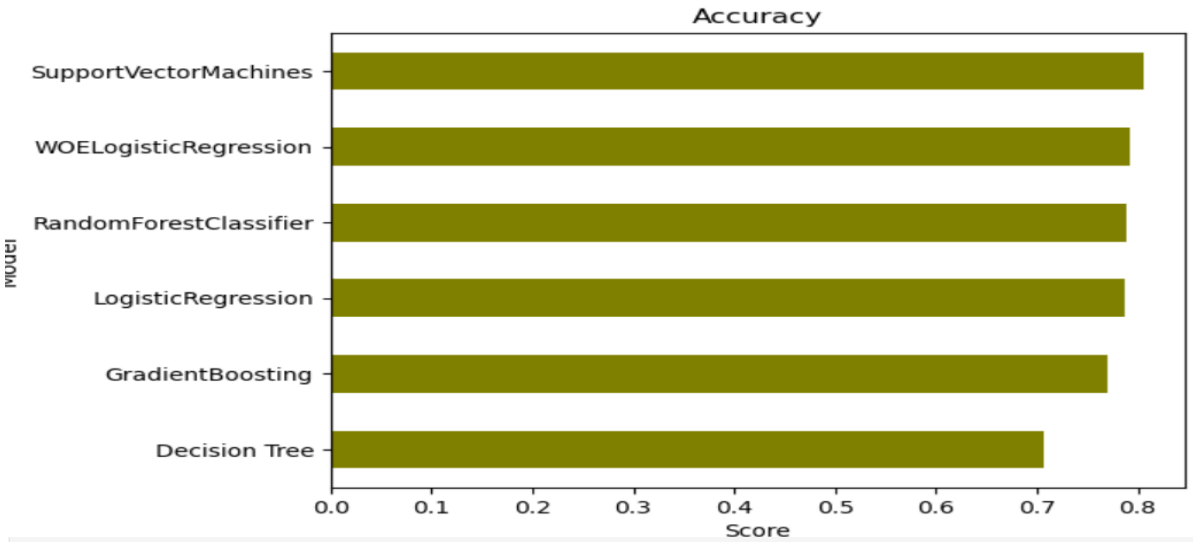
Recall: 0.7174721189591078

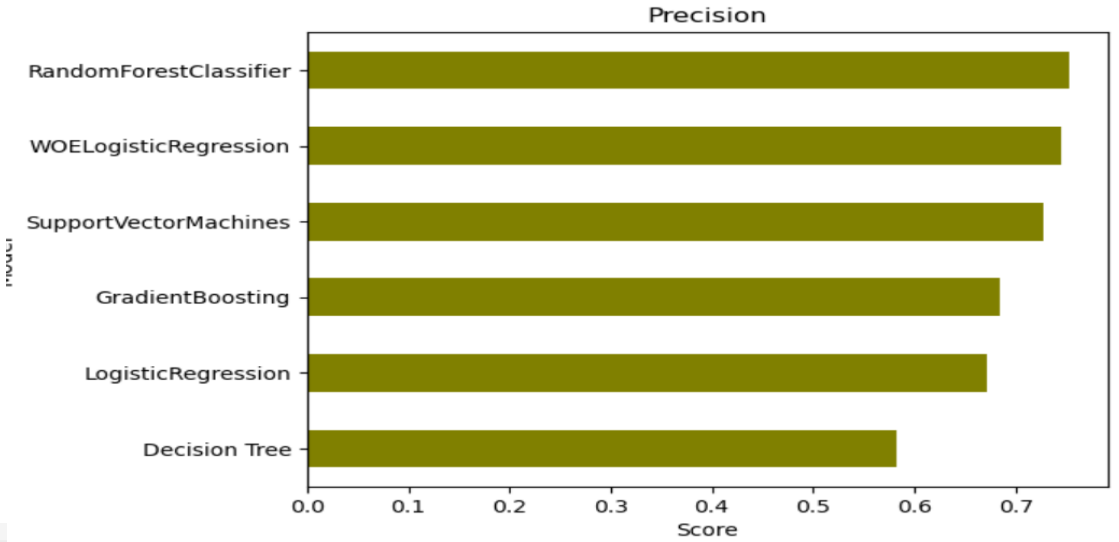
F1 Score: 0.731060606060606

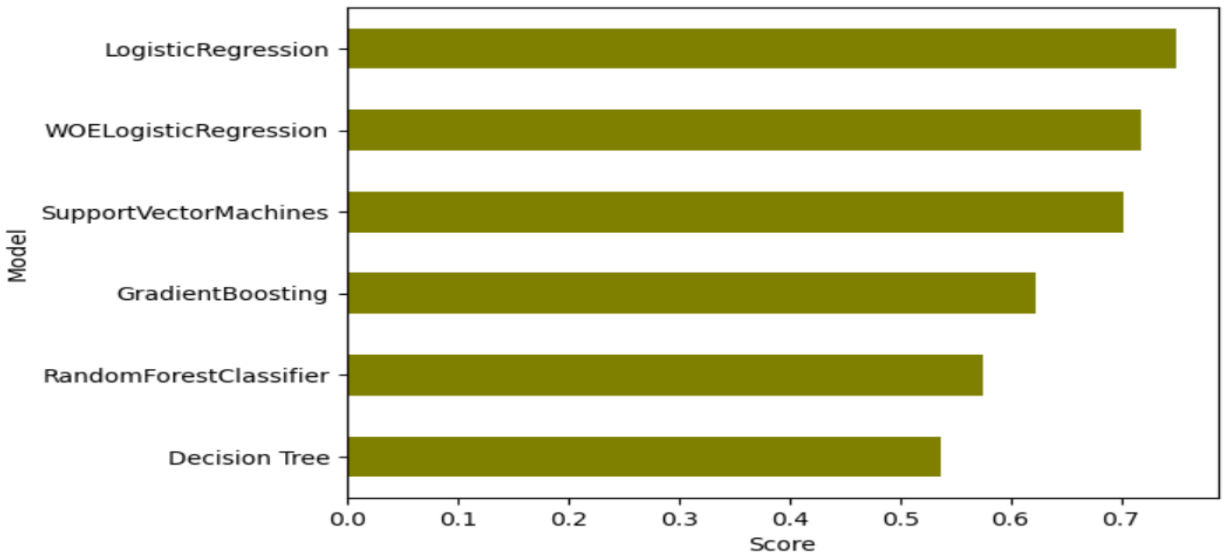
AUC-ROC: 0.7794091364026308

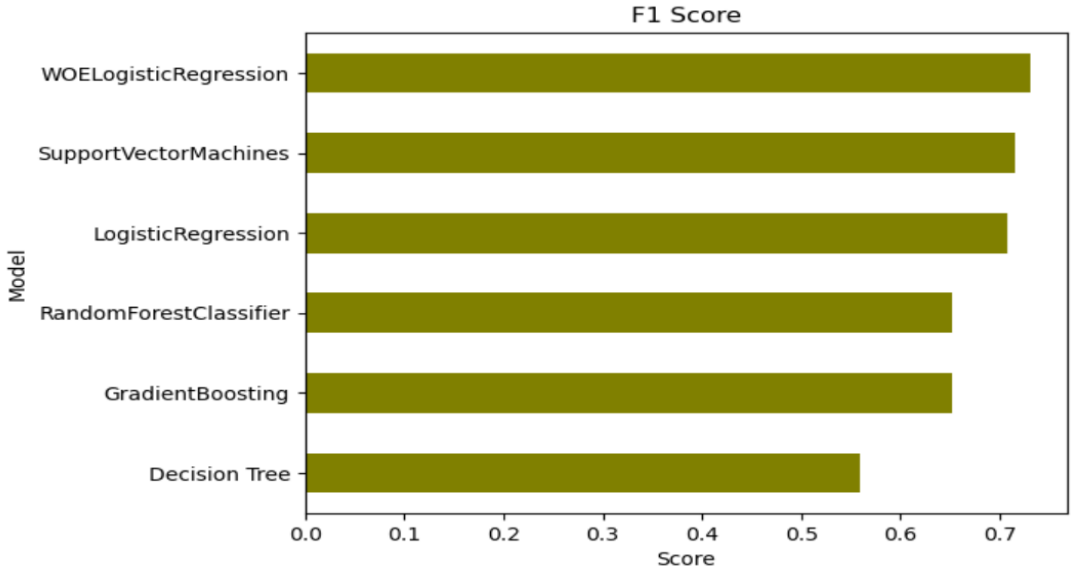
Final Metrics

Data Glacier
Your Deep Learning Partner



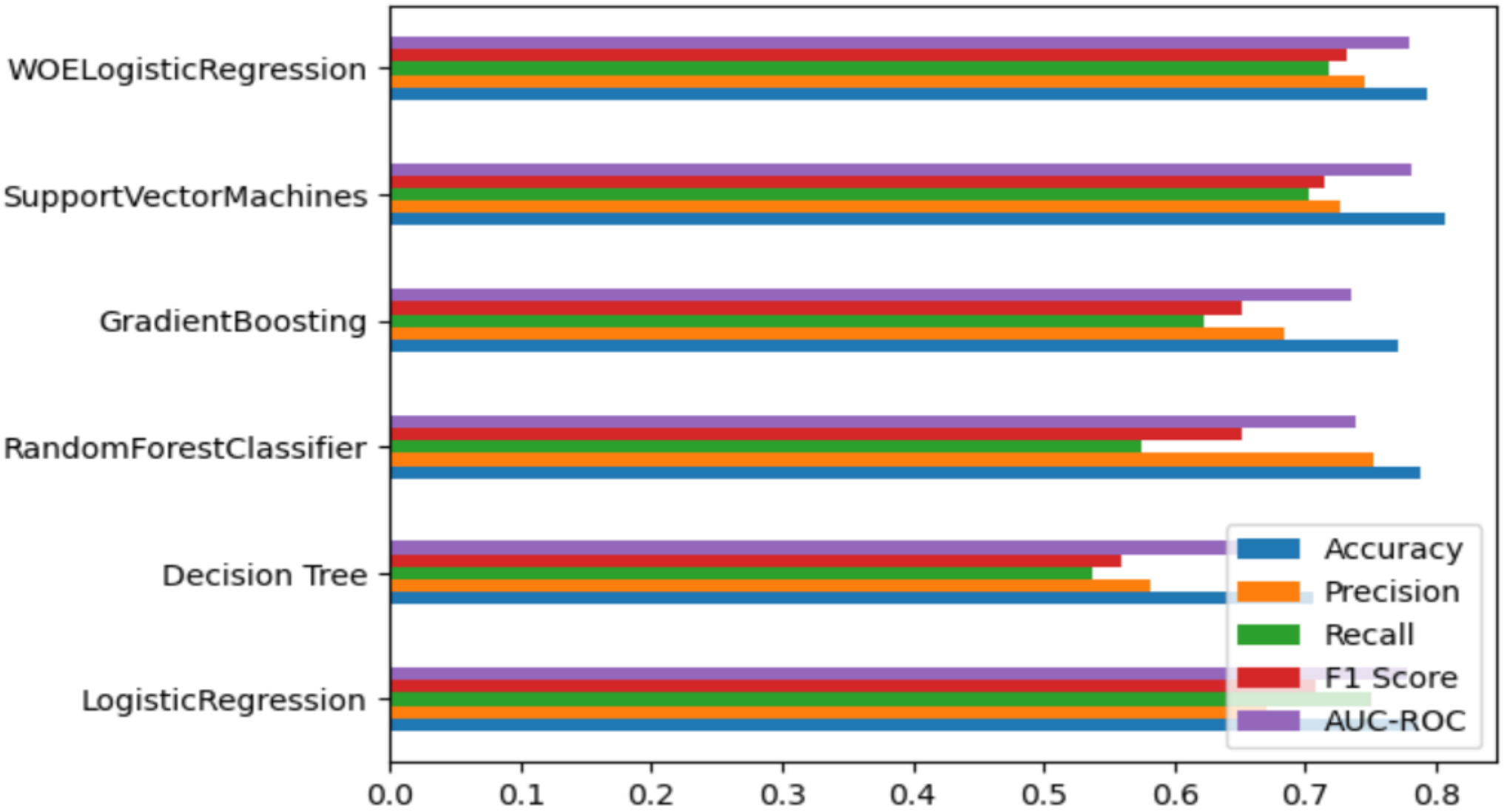






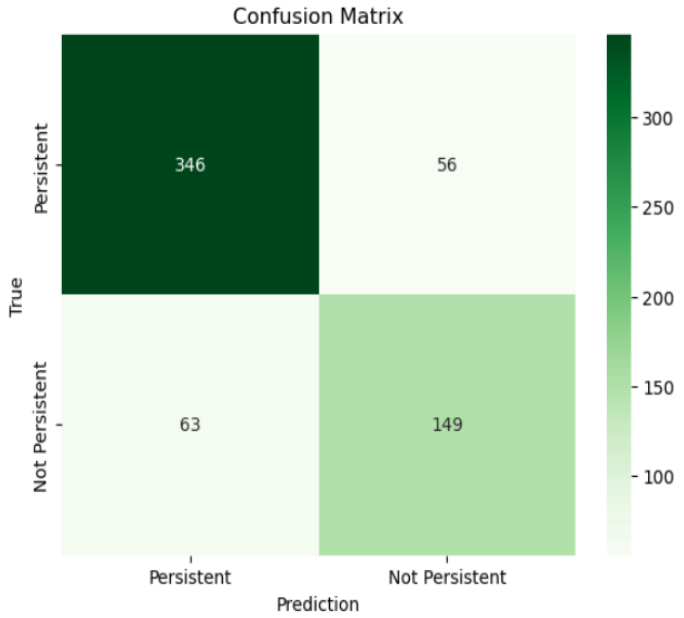
Final Metrics

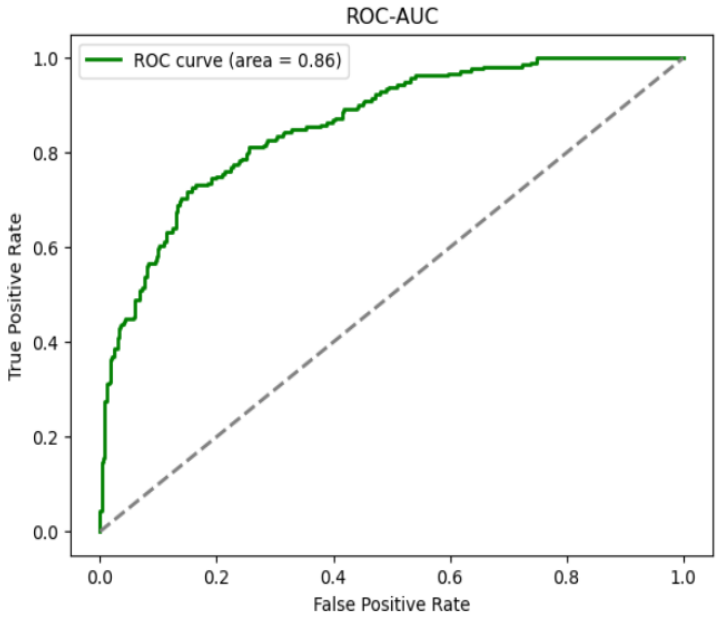
Data Glacier
Your Deep Learning Partner



Selected Model: SVM

Data Glacier
Your Deep Learning Partner





Accuracy: 0.8061889250814332
Precision: 0.7268292682926829
Recall: 0.7028301886792453
F1 Score: 0.7146282973621103
AUC-ROC: 0.7817633530460902

Since It's high accuracy, high F1 score, and ease of implementation SVM is chosen as model to deploy.

Halit Ayberk DEMIR

Data Glacier

Your Deep Learning Partner

Thank You