رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

DEPI: ROUND 3

TRACK: DATA SCIENCE

GROUP: CAI3_AIS4_G1

FINAL TECHNICAL PROJECT

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

CUSTOMER CHURN PREDICTION & ANALYSIS

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

TEAM COGNITIX

KHALID SALIM

MARIAN MILAD

TOQA MOHAMED

YOUSSEF DIAB

SALMA MOHAMMED

MAHMOUD MOHAMED

AGENDA

A- Data Preprocessing:

a. EDA (5 min)

b. Feature Transf. (5 min)

B- Model Dev. & MLOps

- Baseline: Log Reg (3 min)
- Random Forest (3 min)
- Gradient Boosting (3 min)
- XGBoosting (3 min)
- SVM (3 min)

C- Model Deployment (5 min)

A:\Data_Prep\EDA

df . shape

- 38 Columns x 7043 Rows
- 38 Columns = 15 Num + 23 Cat
- Description → Appendix (A)

No	Column	Description
0	CustomerID	A unique ID that identifies each customer
1	Gender	The customer's gender: Male, Female
2	Age	The customer's current age, in years, at the time the fiscal quarter enc
3	Married	Indicates if the customer is married: Yes, No
4	Number of Dependents	Indicates the number of dependents that live with the customer (d parents, grandparents, etc.)
5	City	The city of the customer's primary residence in California
6	Zip Code	The zip code of the customer's primary residence
7	Latitude	The latitude of the customer's primary residence
8	Longitude	The longitude of the customer's primary residence
9	Number of Referrals	Indicates the number of times the customer has referred a frien company to date
10	Tenure in Months	Indicates the total amount of months that the customer has been wit the quarter specified above
11	Offer	Identifies the last marketing offer that the customer accepted: Non Offer D, Offer E

Info() & nunique()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 38 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   Customer ID                        7043 non-null   object 
 1   Gender                             7043 non-null   object 
 2   Age                                7043 non-null   int64  
 3   Married                            7043 non-null   object 
 4   Number of Dependents               7043 non-null   int64  
 5   City                               7043 non-null   object 
 6   Zip Code                           7043 non-null   int64  
 7   Latitude                           7043 non-null   float64
 8   Longitude                          7043 non-null   float64
 9   Number of Referrals                7043 non-null   int64  
 10  Tenure in Months                   7043 non-null   int64  
 11  Offer                              3166 non-null   object 
 12  Phone Service                      7043 non-null   object 
 13  Avg Monthly Long Distance Charges  6361 non-null   float64
 14  Multiple Lines                     6361 non-null   object 
 15  Internet Service                   7043 non-null   object 
 16  Internet Type                      5517 non-null   object 
 17  Avg Monthly GB Download            5517 non-null   float64
 18  Online Security                    5517 non-null   object 
```

```
Customer ID                          7043
Gender                                  2
Age                                    62
Married                                 2
Number of Dependents                   10
City                                 1106
Zip Code                             1626
Latitude                             1626
Longitude                            1625
Number of Referrals                    12
Tenure in Months                       72
Offer                                   5
Phone Service                           2
Avg Monthly Long Distance Charges    3583
Multiple Lines                          2
Internet Service                        2
Internet Type                           3
Avg Monthly GB Download                49
Online Security                         2
Online Backup                           2
```

describe().T

	count	mean	std	min	25%	50%	75%	max
Age	7043.0	46.509726	16.750352	19.000000	32.000000	46.000000	60.000000	80.000000
Number of Dependents	7043.0	0.468692	0.962802	0.000000	0.000000	0.000000	0.000000	9.000000
Zip Code	7043.0	93486.070567	1856.767505	90001.000000	92101.000000	93518.000000	95329.000000	96150.000000
Latitude	7043.0	36.197455	2.468929	32.555828	33.990646	36.205465	38.161321	41.962127
Longitude	7043.0	-119.756684	2.154425	-124.301372	-121.788090	-119.595293	-117.969795	-114.192901
Number of Referrals	7043.0	1.951867	3.001199	0.000000	0.000000	0.000000	3.000000	11.000000
Tenure in Months	7043.0	32.386767	24.542061	1.000000	9.000000	29.000000	55.000000	72.000000
Avg Monthly Long Distance Charges	6361.0	25.420517	14.200374	1.010000	13.050000	25.690000	37.680000	49.990000
Avg Monthly GB Download	5517.0	26.189958	19.586585	2.000000	13.000000	21.000000	30.000000	85.000000
Monthly Charge	7043.0	63.596131	31.204743	-10.000000	30.400000	70.050000	89.750000	118.750000
Total Charges	7043.0	2280.381264	2266.220462	18.800000	400.150000	1394.550000	3786.600000	8684.800000
Total Refunds	7043.0	1.962182	7.902614	0.000000	0.000000	0.000000	0.000000	49.790000
Total Extra Data Charges	7043.0	6.860713	25.104978	0.000000	0.000000	0.000000	0.000000	150.000000
Total Long Distance Charges	7043.0	749.099262	846.660055	0.000000	70.545000	401.440000	1191.100000	3564.720000
Total Revenue	7043.0	3034.379056	2865.204542	21.360000	605.610000	2108.640000	4801.145000	11979.340000

رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

df_info = info_plus(df)

```
[22]: df_info[df_info['Nulls'] > 0]
```

[22]:

	Column	Non-Null	Nulls	DType	N Unique	First 5 Unique	% Missing
11	Offer	3166	3877	object	5	[nan, Offer E, Offer D, Offer A, Offer B]	55
13	Avg Monthly Long Distance Charges	6361	682	float64	3583	[42.39, 10.69, 33.65, 27.82, 7.38]	9
14	Multiple Lines	6361	682	object	2	[No, Yes, nan]	9
16	Internet Type	5517	1526	object	3	[Cable, Fiber Optic, DSL, nan]	21
17	Avg Monthly GB Download	5517	1526	float64	49	[16.0, 10.0, 30.0, 4.0, 11.0]	21
18	Online Security	5517	1526	object	2	[No, Yes, nan]	21
19	Online Backup	5517	1526	object	2	[Yes, No, nan]	21
20	Device Protection Plan	5517	1526	object	2	[No, Yes, nan]	21
21	Premium Tech Support	5517	1526	object	2	[Yes, No, nan]	21
22	Streaming TV	5517	1526	object	2	[Yes, No, nan]	21
23	Streaming Movies	5517	1526	object	2	[No, Yes, nan]	21
24	Streaming Music	5517	1526	object	2	[No, Yes, nan]	21
25	Unlimited Data	5517	1526	object	2	[Yes, No, nan]	21
36	Churn Category	1869	5174	object	5	[nan, Competitor, Dissatisfaction, Other, Price]	73
37	Churn Reason	1869	5174	object	20	[nan, Competitor had better devices, Product dissatisfaction, Network reliability, Limited range of services]	73

A:\Data_Prep\Cleaning

Drop: Rows

- Duplicates → No
- %1 - %5 Nulls → No (Min. %9)
- Customer Status:
  - Joined → Tenure in Months 1, 2, 3 → Drop or turn to Stayed?
  - Churned → Tenure in Months 1, 2, 3
  - Decision: Joined → Stayed

Drop: Columns 1

- Irrelevant → Cust. ID, Online Sec., Online Backup, Pap. Billing, Payment Method → Drop
- Categorical + Many Categories: Lat., Long., Zip Code → Drop
  City → Keep (demographics)
- High % Nulls: Offers, Churn Category, Chrun Reason → Drop

Drop: Columns 2

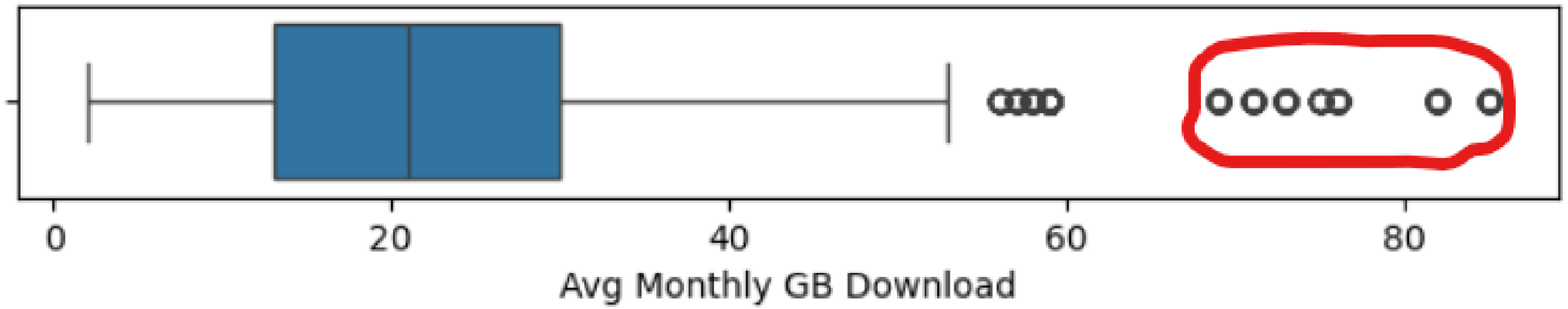
- Related: Total Ch. ≈ Monthly Ch. x Tenure → Drop Monthly Ch.
- Related: T. Rev = T. Ref. + T. Ex. Data Ch. + T. Long. Dist. Ch.
  → Drop Total Revenue
- Related: T. Long Dist. Ch. = Avg Mon. L. Dist. Ch. x Tenure
  → Drop Avg Mon. L. Dist. Ch.

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Nulls: Num Cols → Median

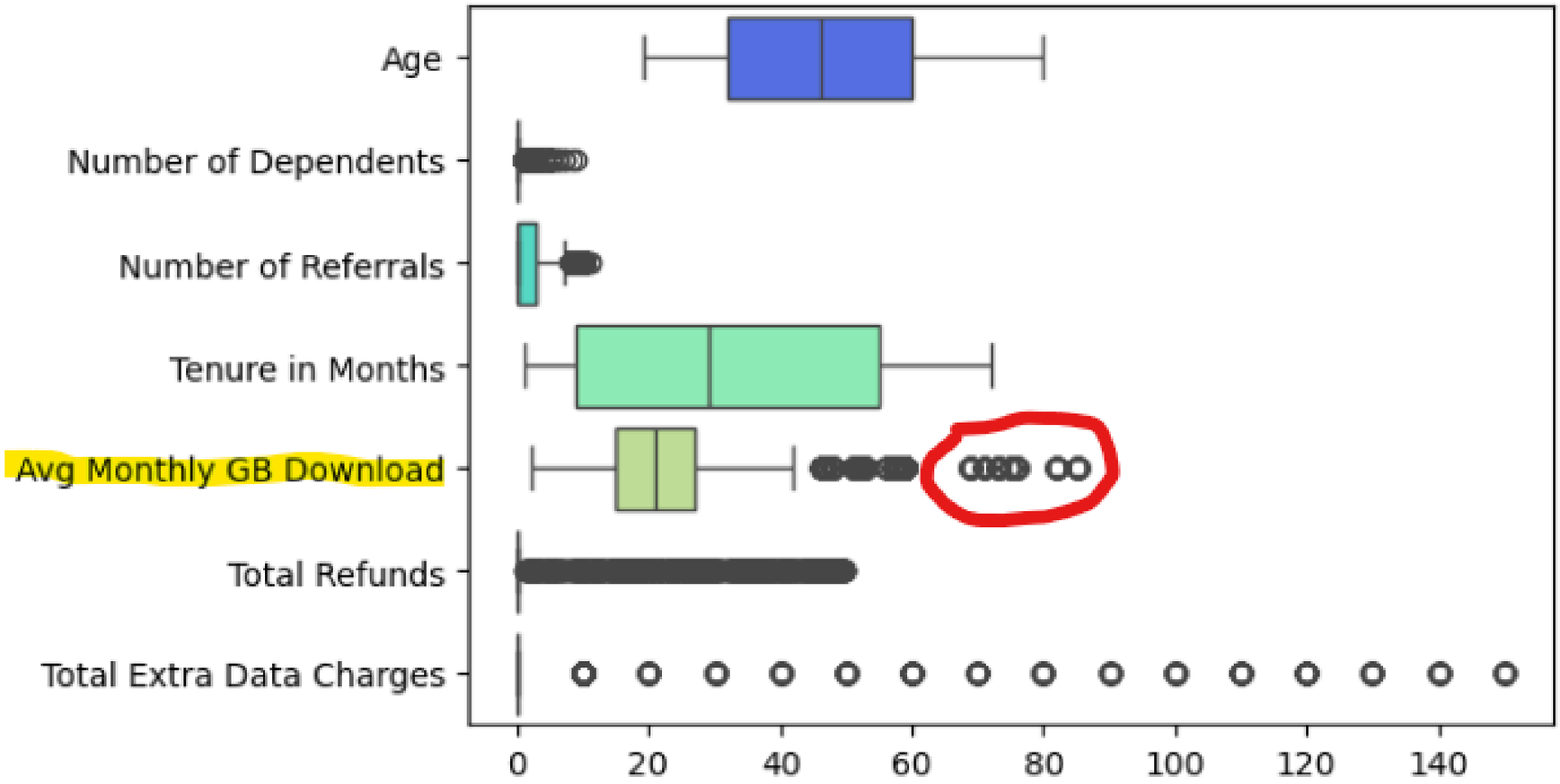
	Column	Non-Null	Nulls	DType	N Unique	First 5 Unique	% Missing
11	Avg Monthly GB Download	5517	1526	float64	49	[16.0, 10.0, 30.0, 4.0, 11.0]	21



Nulls: Cat Cols → Mode

	Column	Non-Null	Nulls	DType	N Unique	First 5 Unique	% Missing
8	Multiple Lines	6361	682	object	2	[No, Yes, nan]	9
10	Internet Type	5517	1526	object	3	[Cable, Fiber Optic, DSL, nan]	21
12	Device Protection Plan	5517	1526	object	2	[No, Yes, nan]	21
13	Premium Tech Support	5517	1526	object	2	[Yes, No, nan]	21
14	Streaming TV	5517	1526	object	2	[Yes, No, nan]	21
15	Streaming Movies	5517	1526	object	2	[No, Yes, nan]	21
16	Streaming Music	5517	1526	object	2	[No, Yes, nan]	21
17	Unlimited Data	5517	1526	object	2	[Yes, No, nan]	21

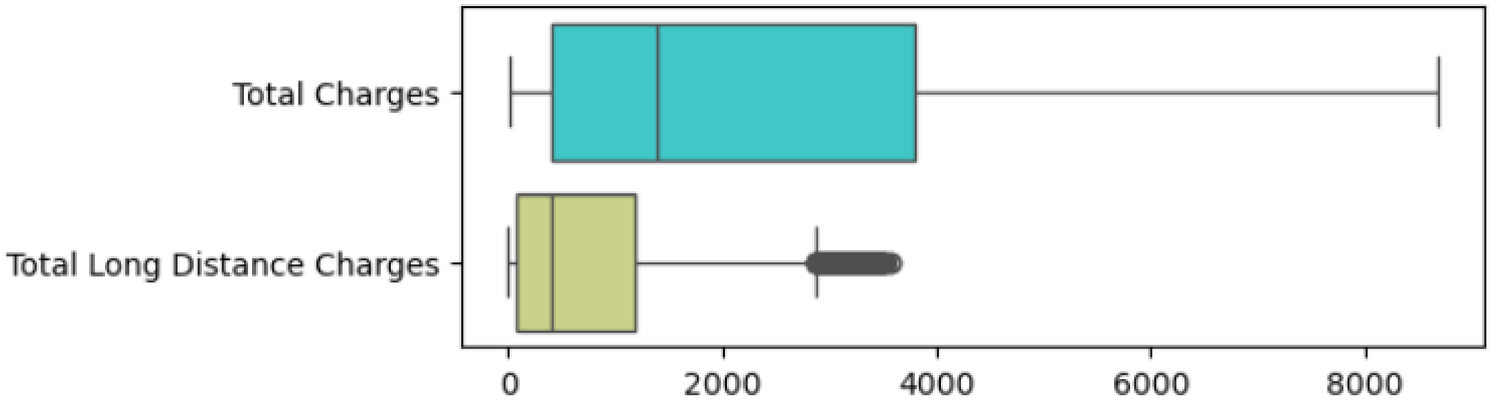
Outliers: Small Values

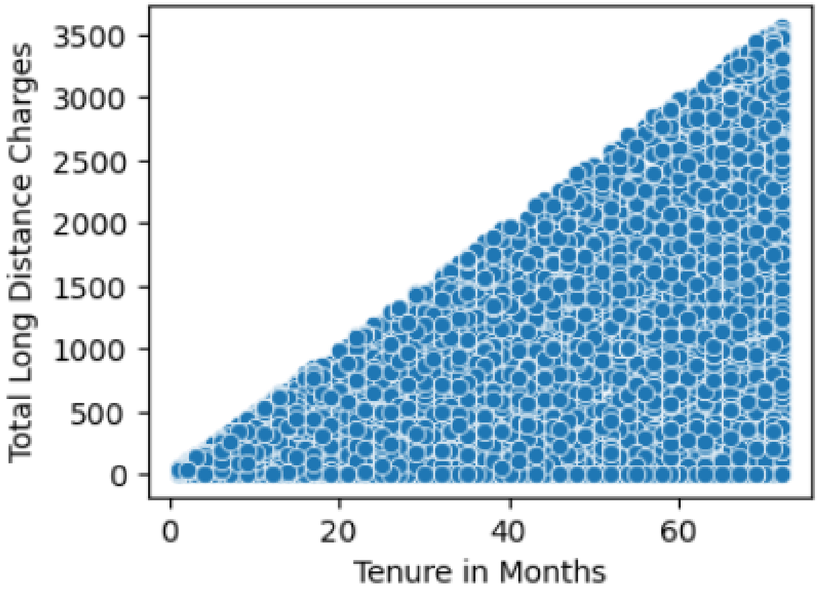


رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Outliers: Large Values





Encoding Cat. Cols

	Column	Non-Null	Nulls	DType	N Unique	First 5 Unique	% Missing
0	Gender	7043	0	object	2	[Female, Male]	0
1	Married	7043	0	object	2	[Yes, No]	0
2	City	7043	0	object	1106	[Frazier Park, Glendale, Costa Mesa, Martinez, Camarillo]	0
3	Phone Service	7043	0	object	2	[Yes, No]	0
4	Multiple Lines	7043	0	object	2	[No, Yes]	0
5	Internet Service	7043	0	object	2	[Yes, No]	0
6	Internet Type	7043	0	object	3	[Cable, Fiber Optic, DSL]	0
7	Device Protection Plan	7043	0	object	2	[No, Yes]	0
8	Premium Tech Support	7043	0	object	2	[Yes, No]	0
9	Streaming TV	7043	0	object	2	[Yes, No]	0
10	Streaming Movies	7043	0	object	2	[No, Yes]	0
11	Streaming Music	7043	0	object	2	[No, Yes]	0
12	Unlimited Data	7043	0	object	2	[Yes, No]	0
13	Contract	7043	0	object	3	[One Year, Month-to-Month, Two Year]	0
14	Customer Status	7043	0	object	2	[Stayed, Churned]	0

Enc.: Ordinal → Mapping

Before	Encoded
Month-to-Month	1
One Year	2
Two Year	3

Enc.: Nominal 2-State → Mapping

Before			Encoded
Male	Churned	Yes	1
Female	Stayed	No	0

Enc.: Nominal M-State → Binary

- Internet Type [Cable, Fiber Optic, DSL]: 3 States
  → 2 bits (Binary Cols)

- City [1106 Cities]: 1106 States > 1024 (2^10)
  → 11 bits → 11 columns


رواد مصر الرقمية


وزارة الاتصـــــالات
وتكنولوجيا المعلومات

B:\MLM+MLOps\

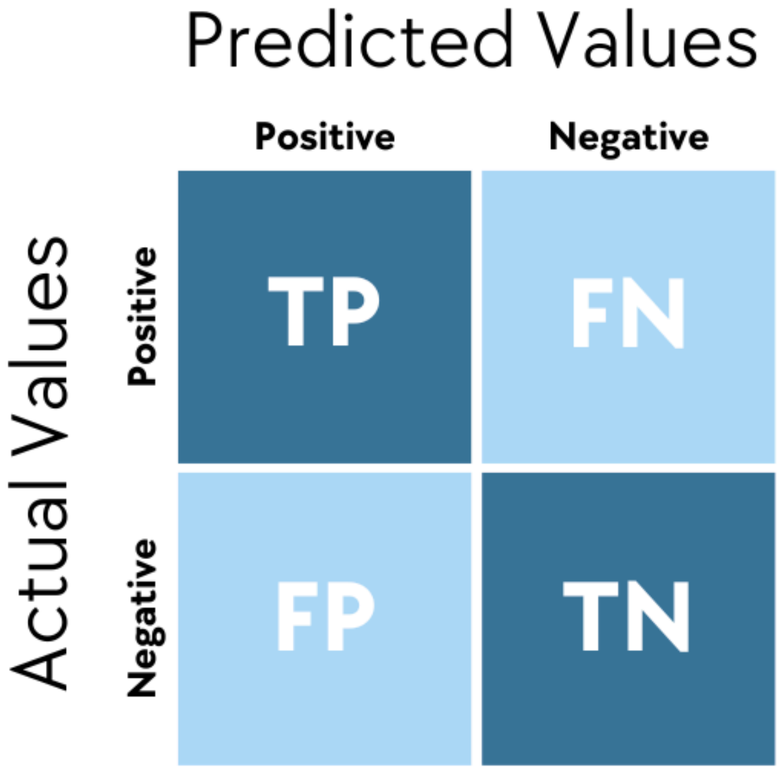
Data & MLMs

- Clean Encoded Data: 35 x 7043
  - 35 Cols [34 Feat. + 1 Target]
  - Train/Test: 80/20 or 70/30

- MLMs:
  - LR (BL) + RF + GB + XGB + SVM
  - MLM: Basic & Grid Search CV
  - LR & SVM: Scaled Features
  - RF, GB & XGB → Tree-based (No Scaling)

MLM Performance Eval.

- Metrics:
  - Accuracy
  - Precision
  - **Recall**
  - F1-Score



1. Accuracy = $\frac{T_P + T_N}{T_P + T_N + F_P + F_N}$
2. Precision = $\frac{T_P}{T_P + F_P}$
3. Recall (sensitivity) = $\frac{T_P}{T_P + F_N}$
4. F1-Score = 2 $\times$ $\frac{Precision \times Recall}{Precision + Recall}$

Where:

T_P: True Positive

T_N is the True Negative

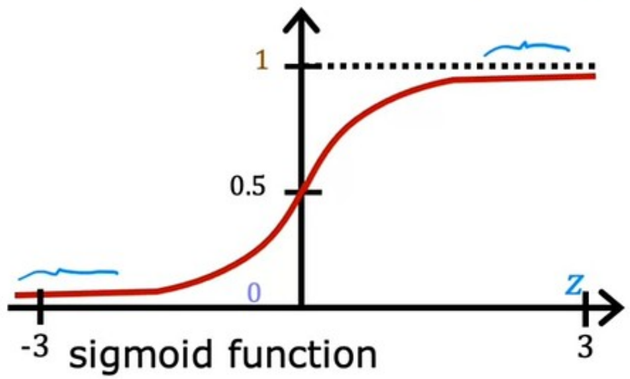
F_P is the False Positive (Type I Error)

F_N is the False Negative (Type II Error)

B:\MLM+MLOps\LogReg

Logistic Regression (Basic)

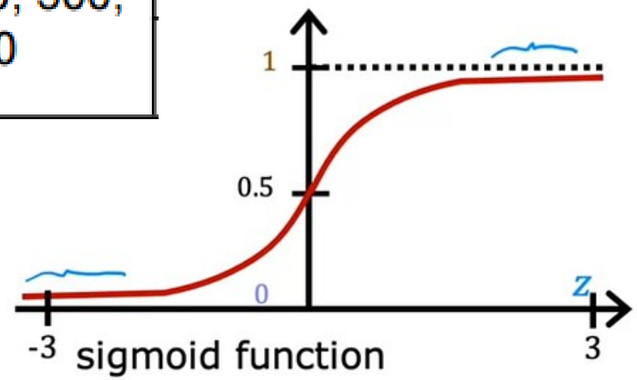
Hyper Parameter	Value
Random State	30
Max Iterations	300



Results	Value
Accuracy	0.85
Precision	0.74
Recall	0.73
F1 Score	0.73

Logistic Regression (GS-CV)

Hyper Parameter	Value
C (Tunable)	0.1, 1, 10, 100
Penalty	L2 Reg.
Max Iterations	200, 300, 500, 700

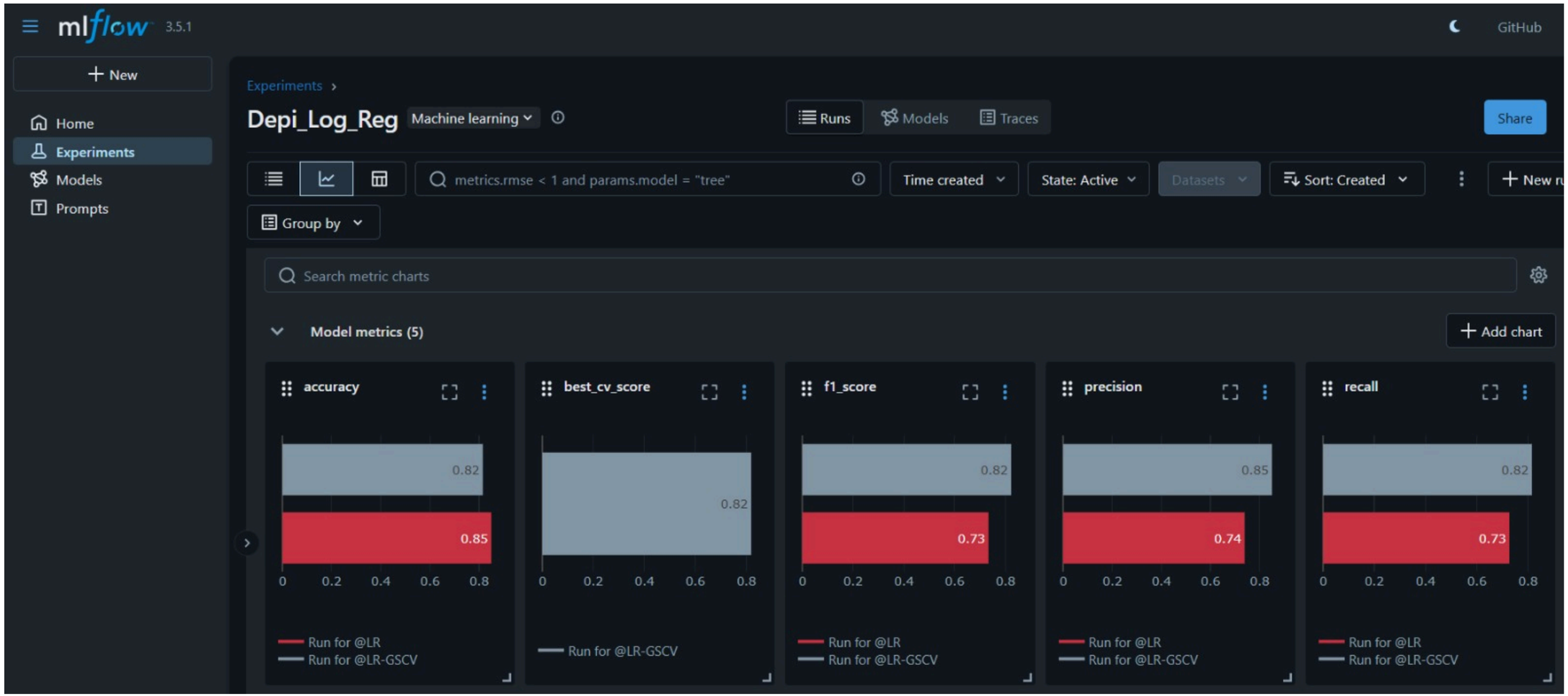


Results	Value
Accuracy	0.82
Precision	0.85
Recall	0.82
F1 Score	0.82

رواد مصر الرقمية

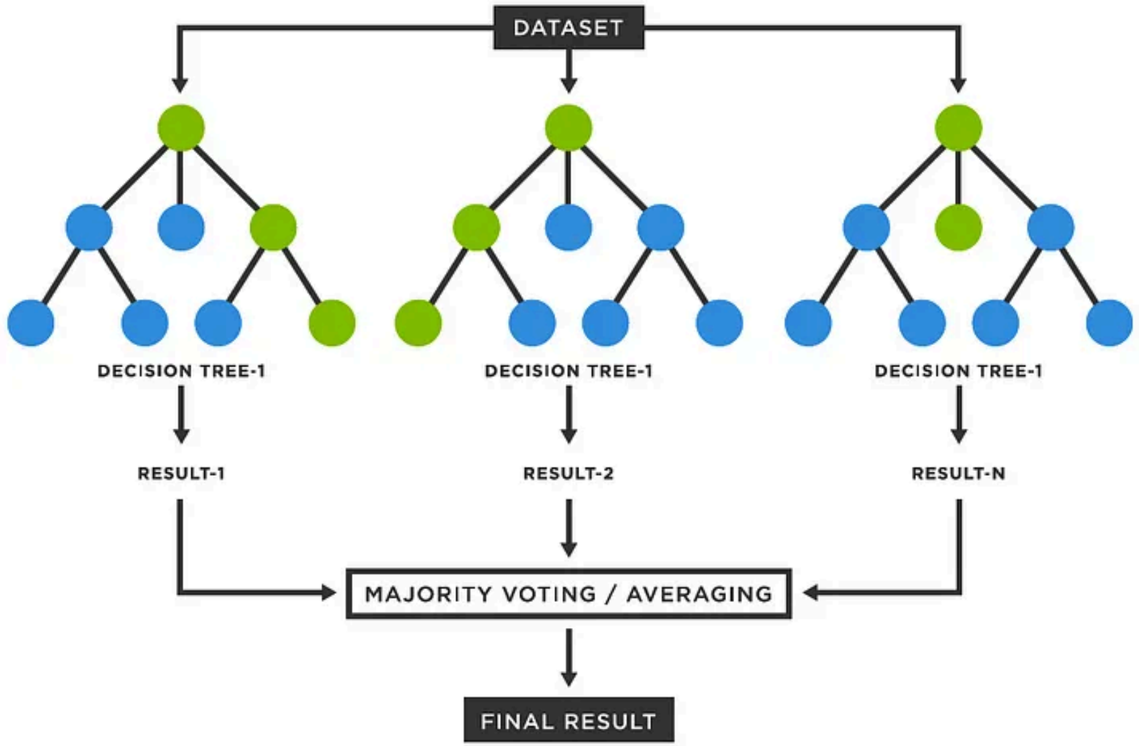
وزارة الاتصالات
وتكنولوجيا المعلومات

Logistic Regression MLFlow



B:\MLM+MLOps\RF

Random Forest (Basic)



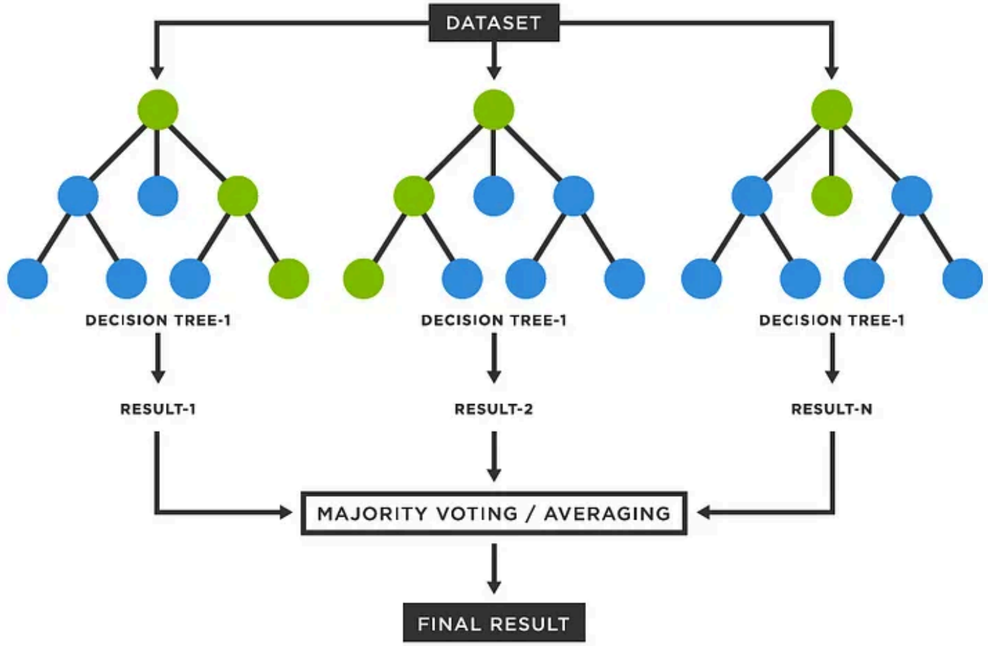
Hyper Parameter	Value
Random State	30
Class Weight	Balanced

Results	Value
Accuracy	0.87
Precision	0.86
Recall	0.87
F1 Score	0.86

Random Forest (GS-CV)

Hyper Parameter	Value
n_estimators / # Trees (Tune)	200, 300, 400, 500
Max Depth (Tune)	10, 20, 30, None
Min Samples Split (Tune)	2, 5, 10
Min Samples Leaf (Tune)	1, 2, 4
Max Features (Tune)	Sqrt, Log2

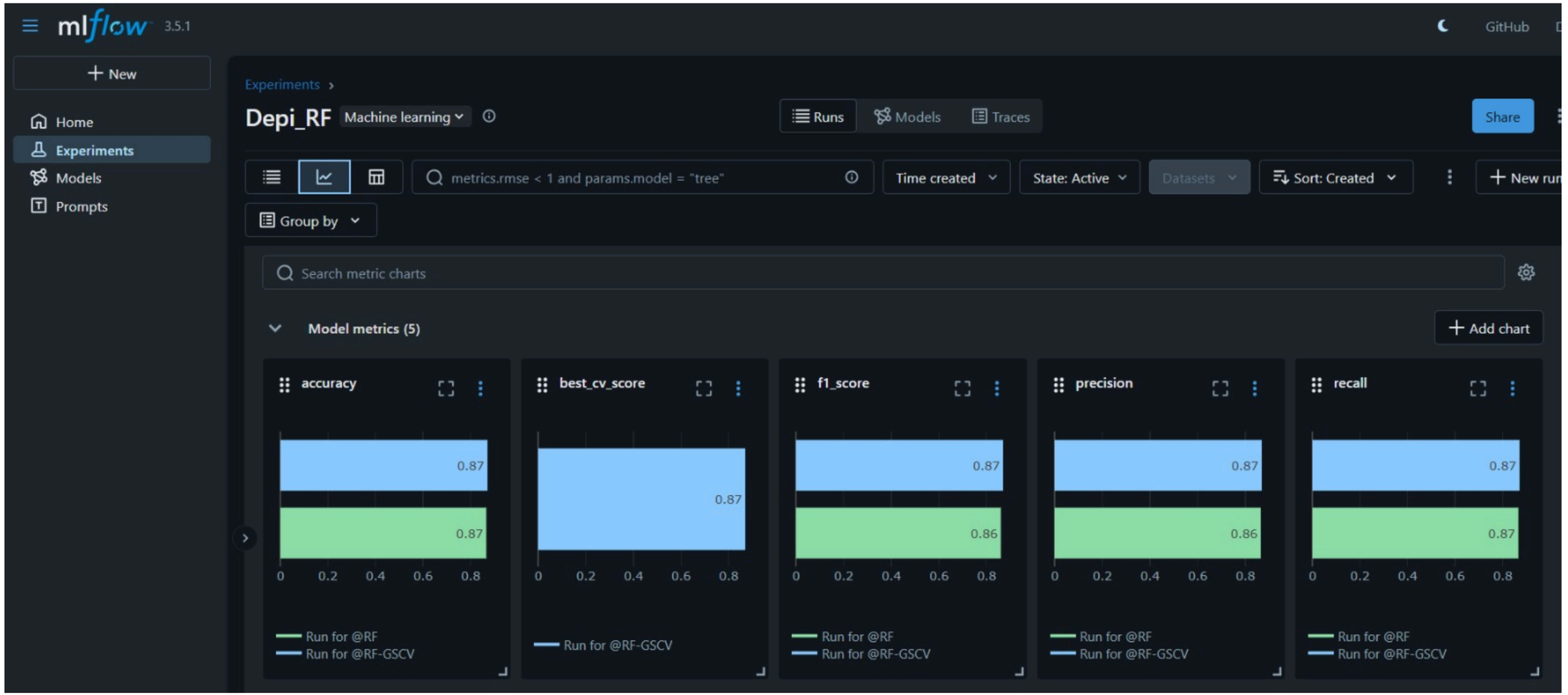
Results	Value
Accuracy	0.87
Precision	0.87
Recall	0.87
F1 Score	0.87



رواد مصر الرقمية

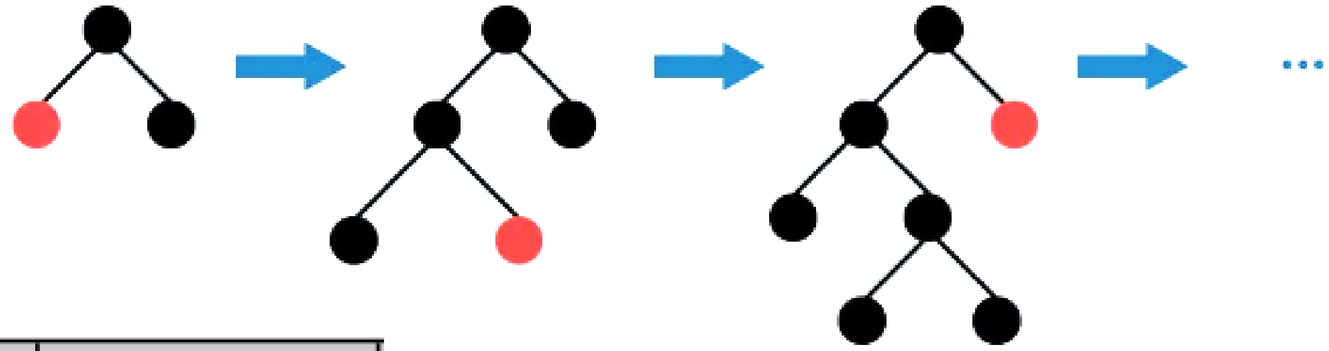
وزارة الاتصالات وتكنولوجيا المعلومات

Random Forest MLFlow



B:\MLM+MLOps\GBoost

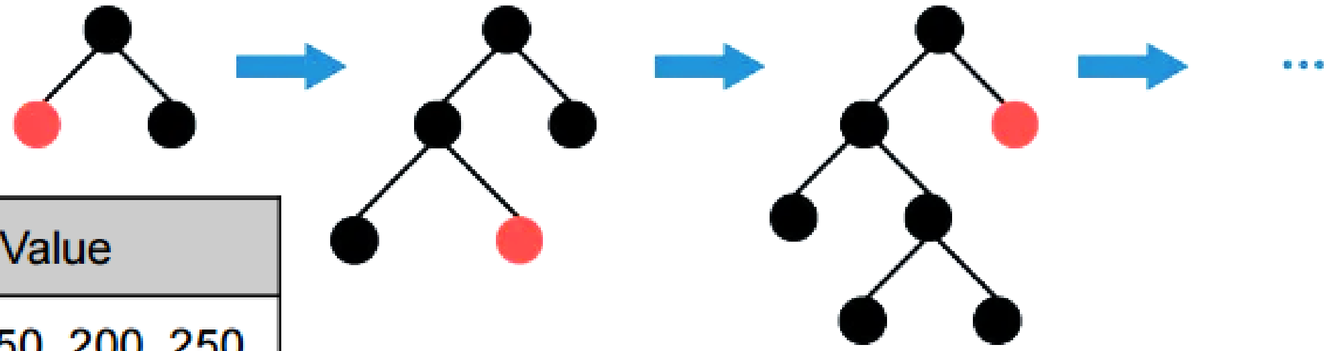
Gradient Boosting (Basic)



Hyper Parameter	Value
n_estimators / # Trees	100
Learning Rate	0.1
max_depth	3
random_state	42

Results	Value
Accuracy	0.87
Precision	0.86
Recall	0.67
F1 Score	0.86

Gradient Boosting (GS-CV)



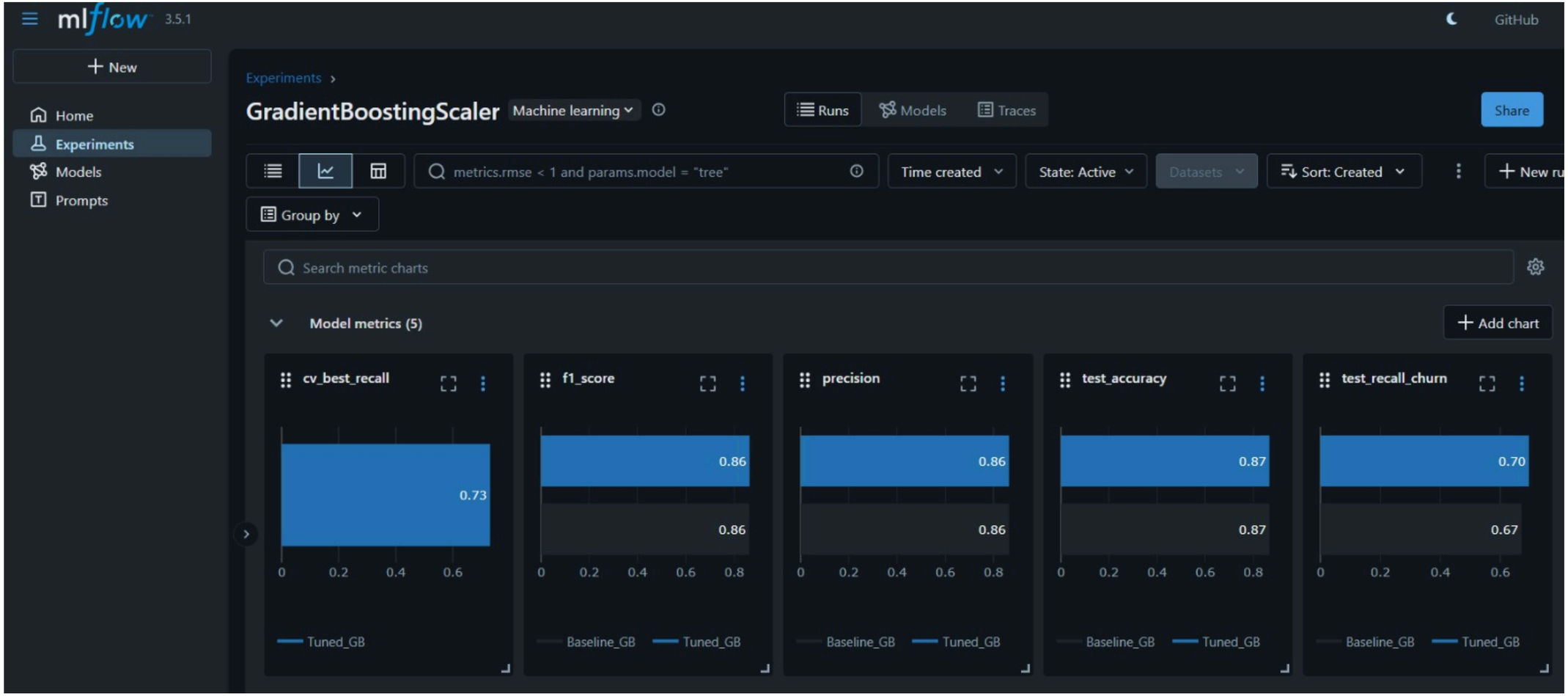
Hyper Parameter	Value
n_estimators / # Trees (Tune)	100, 150, 200, 250
Max Depth (Tune)	3, 4, 5, 6
Min Samples Split (Tune)	2, 5, 10
Min Samples Leaf (Tune)	1, 2, 4
Learning Rate	0.01, 0.02, 0.05, 0.1
Subsample	0.8, 0.85, 0.9, 0.095

Results	Value
Accuracy	0.87
Precision	0.86
Recall	0.70
F1 Score	0.86

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

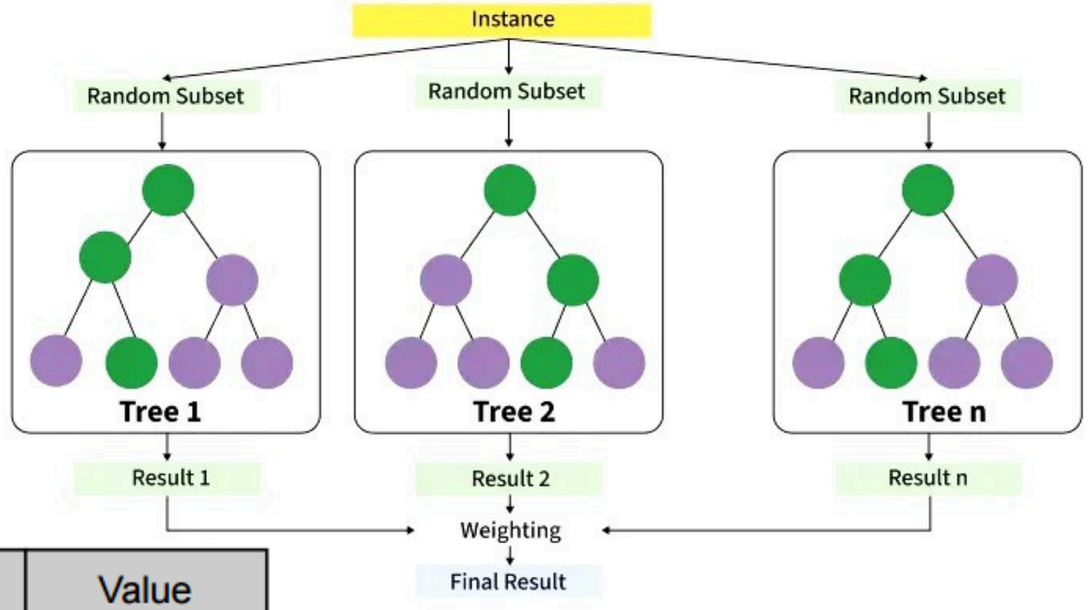
Gradient Boosting MLFlow



B:\MLM+MLOps\XGBoost

XGBoosting (Basic)

Hyper Parameter	Value
n_estimators / # Trees	Default = 100
Max Depth	Default = 6
colsample_bytree	Default = 1.0
scale_pos_weight	2.53
Learning Rate	Default = 0.3
Subsample	Default = 1.0

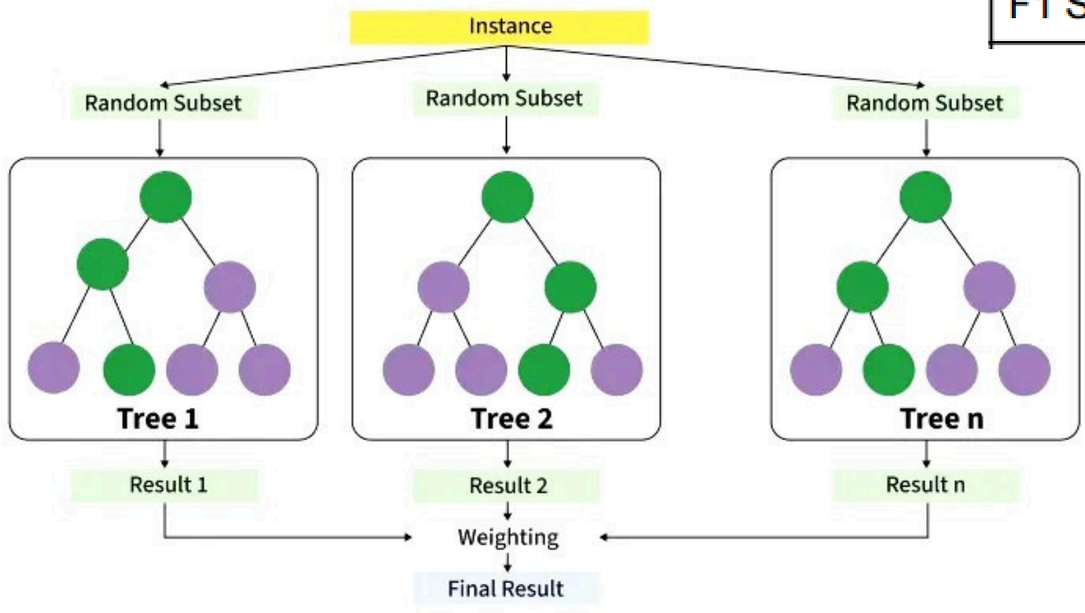


Results	Value
Accuracy	0.84
Precision	0.84
Recall	0.72
F1 Score	0.84

XGBoosting(GS-CV)

Hyper Parameter	Value
n_estimators / # Trees	100, 150, 200, 250
Max Depth	3, 5, 7
colsample_bytree	0.8, 0.9
scale_pos_weight	2.53
Learning Rate	0.01, 0.02, 0.05, 0.1
Subsample	0.8, 0.85, 0.9, 0.095

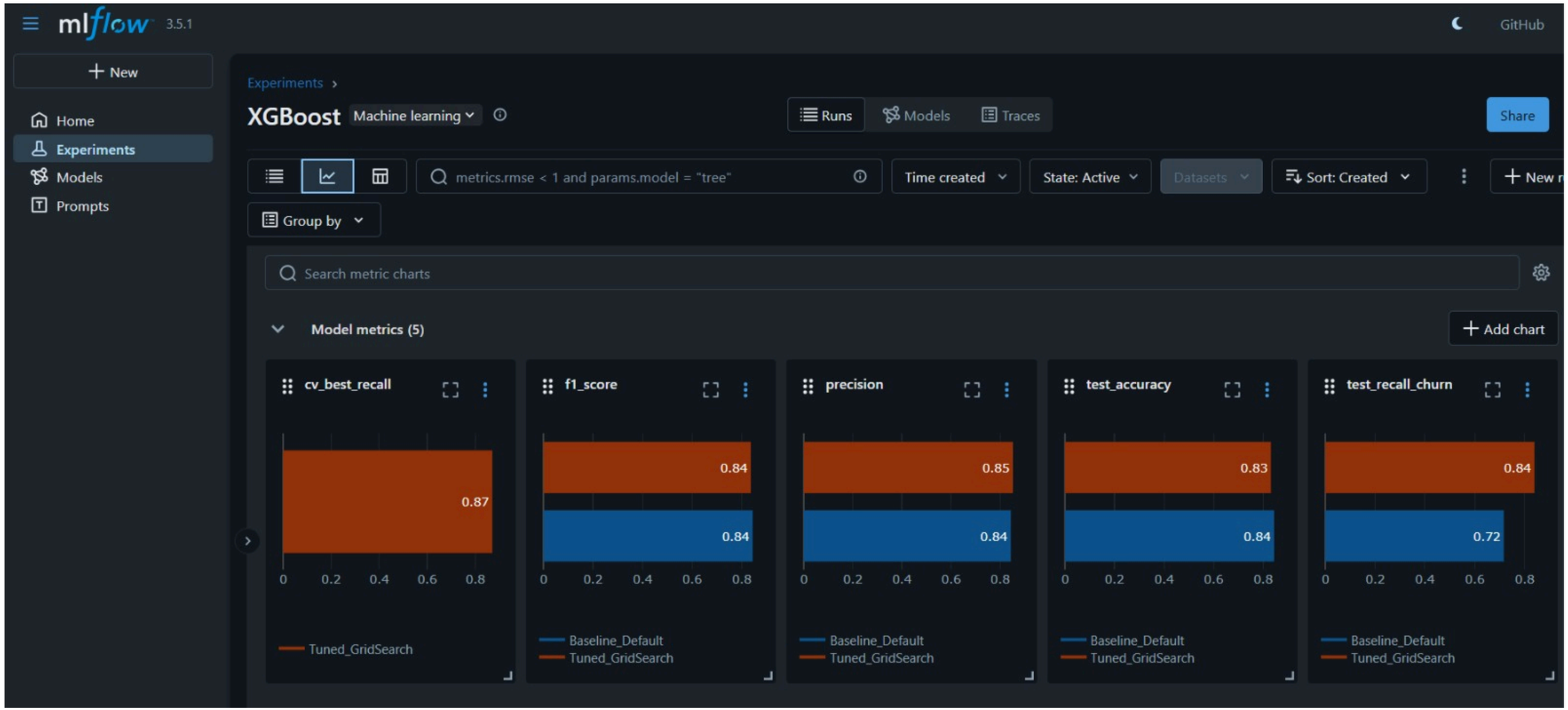
Results	Value
Accuracy	0.83
Precision	0.85
Recall	0.84
F1 Score	0.84



رواد مصر الرقمية

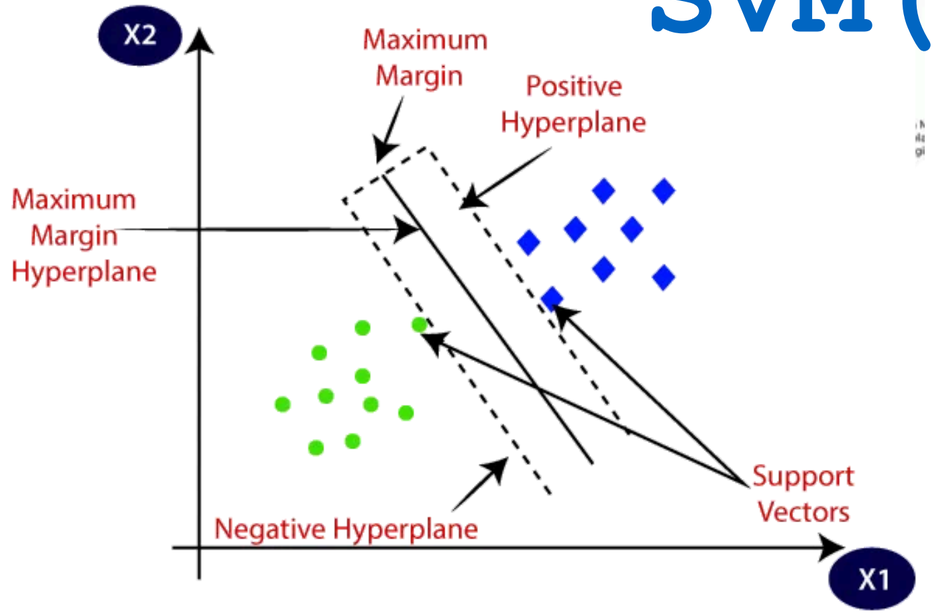
وزارة الاتصالات وتكنولوجيا المعلومات

XGBoosting MLFlow



B:\MLM+MLOps\SVM

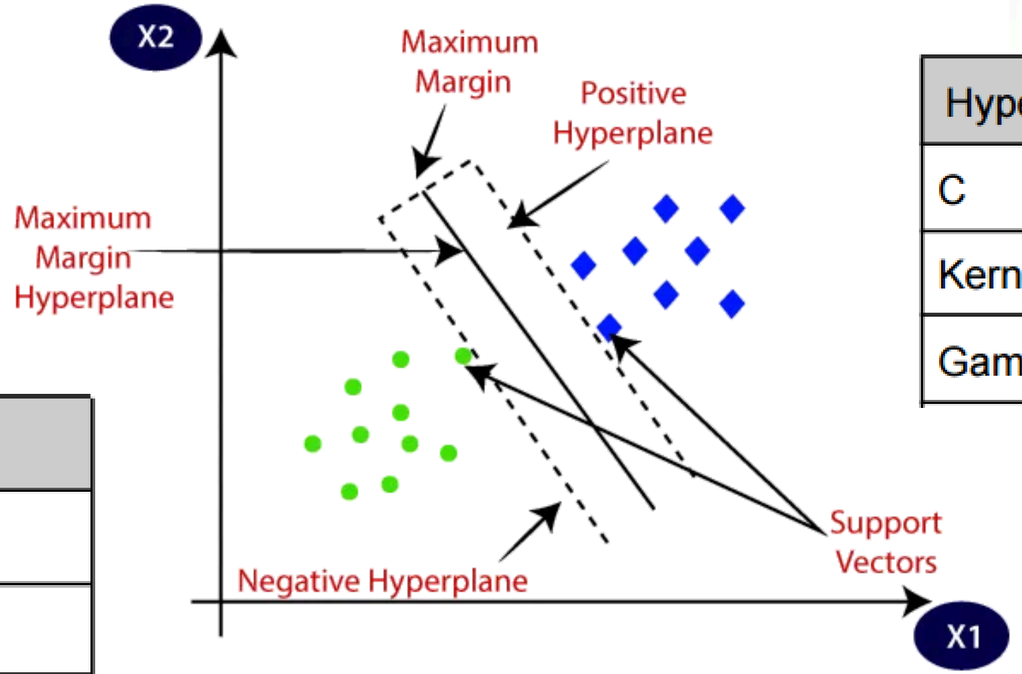
SVM(Basic)



Hyper Parameter	Value
C	0.01
Kernel	Linear

Results	Value
Accuracy	0.84
Precision	0.84
Recall	0.84
F1 Score	0.84

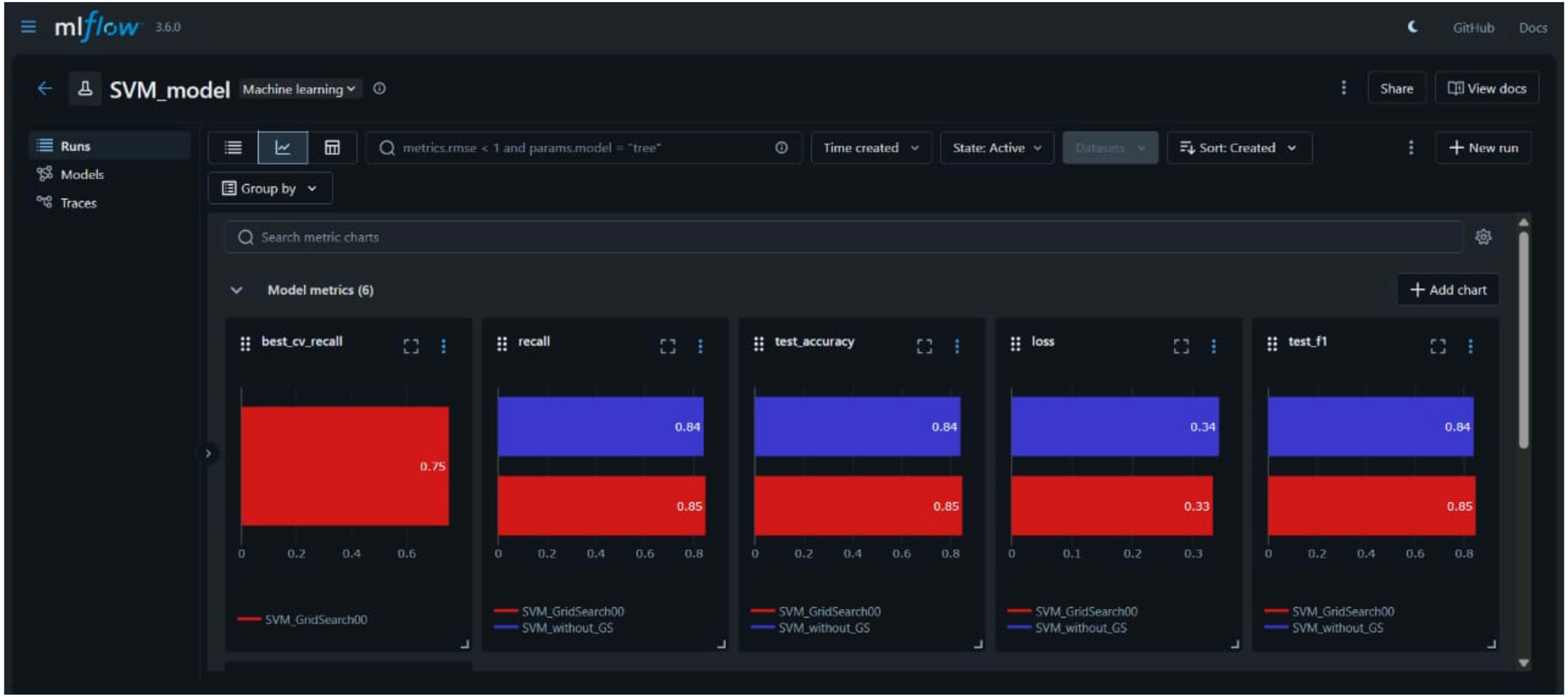
SVM (GS-CV)



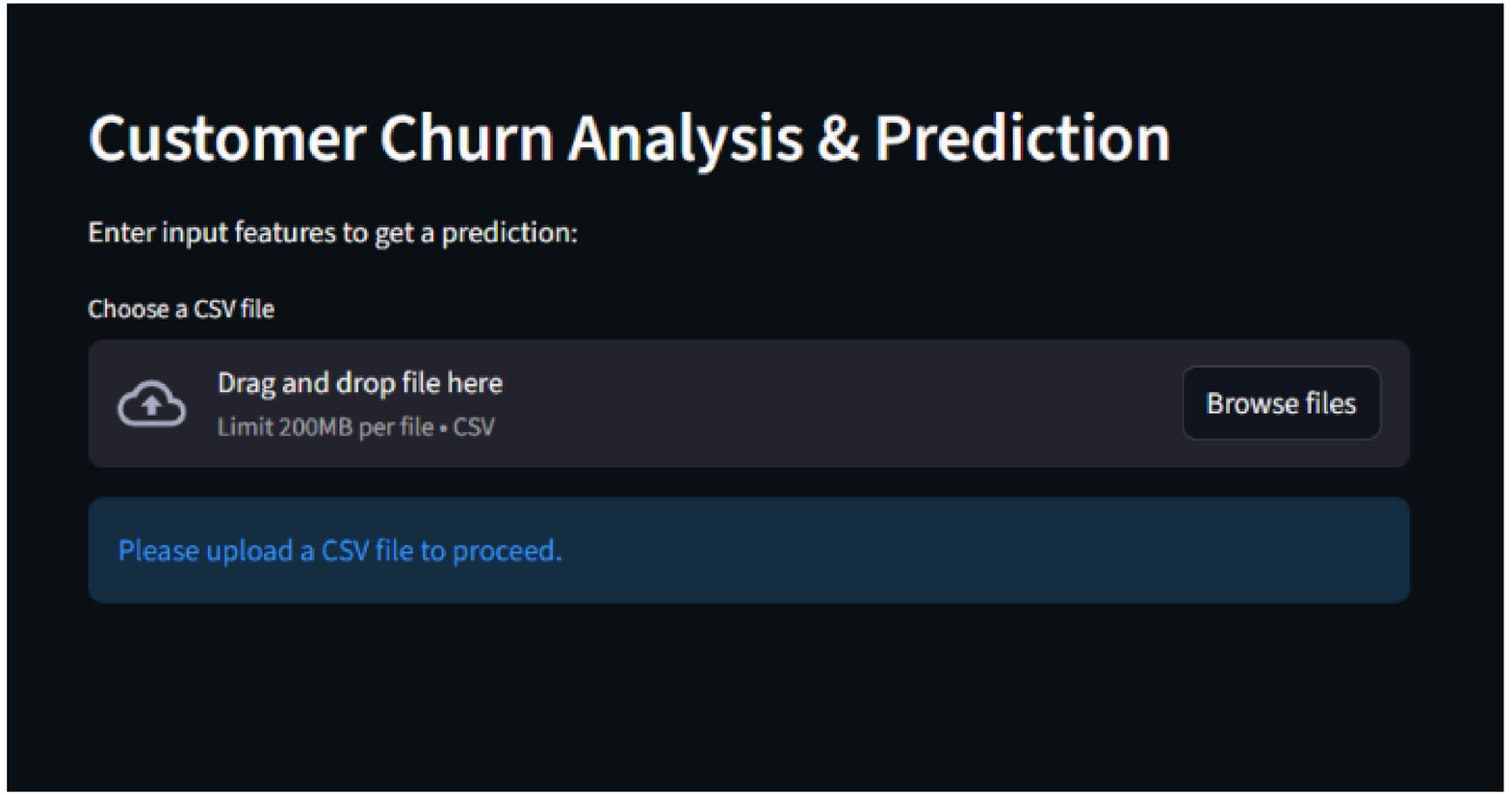
Hyper Parameter	Value
C	0.001, 0.01, 0.1, 1, 10
Kernel	Linear, RBF, Polynomial
Gamma	Scale, Auto

Results	Value
Accuracy	0.85
Precision	0.85
Recall	0.85
F1 Score	0.85

SVM MLFlow


mlflow 3.6.0
GitHub
Docs
SVM_model
Machine learning
Share
View docs
Runs
Models
Traces
metrics.rmse < 1 and params.model = "tree"
Time created
State: Active
Datasets
Sort: Created
New run
Group by
Search metric charts
Model metrics (6)
Add chart
best_cv_recall
0.75
0 0.2 0.4 0.6
SVM_GridSearch00
recall
0.84
0.85
0 0.2 0.4 0.6 0.8
SVM_GridSearch00
SVM_without_GS
test_accuracy
0.84
0.85
0 0.2 0.4 0.6 0.8
SVM_GridSearch00
SVM_without_GS
loss
0.34
0.33
0 0.1 0.2 0.3
SVM_GridSearch00
SVM_without_GS
test_f1
0.84
0.85
0 0.2 0.4 0.6 0.8
SVM_GridSearch00
SVM_without_GS

C:\Deploy


Customer Churn Analysis & Prediction
Enter input features to get a prediction:
Choose a CSV file
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Please upload a CSV file to proceed.

Model Deployment

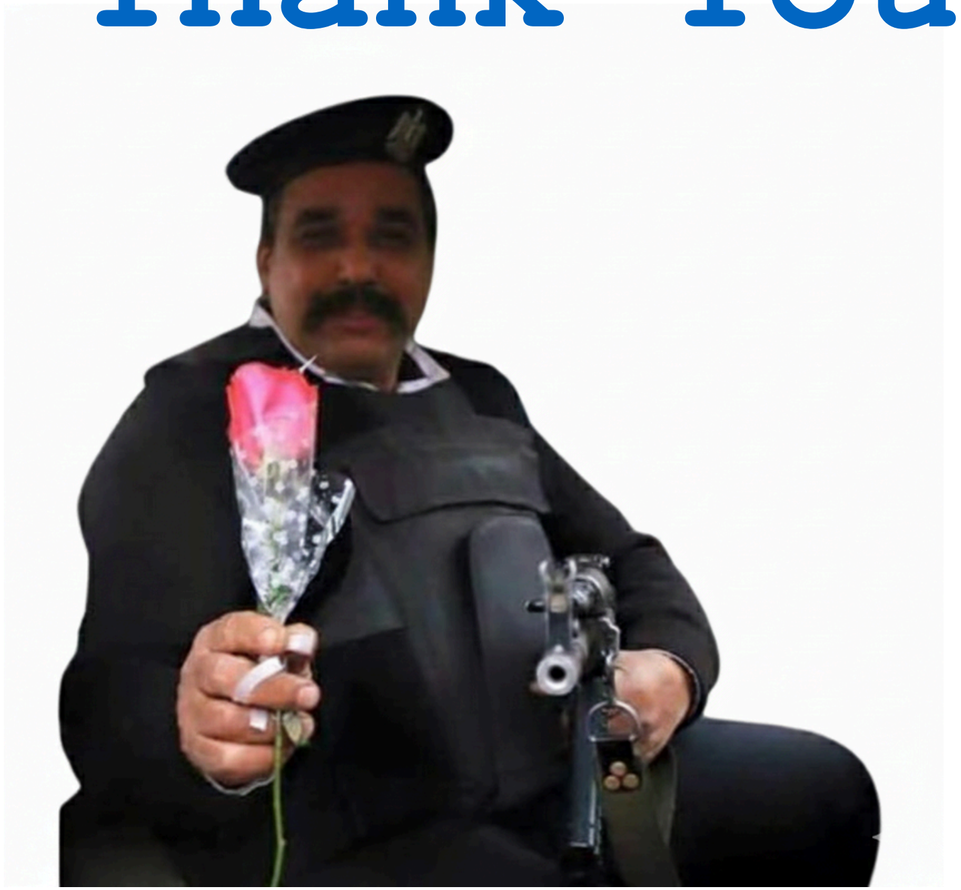
- Winner Model: Random Forest 

- Deployment Platform: Streamlit.io
- Webapp URL: cognitix.streamlit.app
- Project Repo: github.com/kbahajjaj/cognitix

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

> Thank You



رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

> Q&A