

Assignment: Using MLflow for Experiment Tracking and Model Management

Objective: The objective of this task is to introduce you to MLflow for experiment tracking, model management, and reproducibility in machine learning projects for the Sentiment Analysis Project.

Task

1. Integrate MLflow into your existing machine learning projects.
 2. Train machine learning models while logging relevant information with MLflow.
 3. Demonstrate how to log parameters, metrics, and artifacts using MLflow tracking APIs.
 4. Customizing MLflow UI with run names.
 5. Demonstrate metric plots.
 6. Demonstrate hyperparameter plots.
 7. Demonstrate how to register models and manage by tagging them.
 8. **(BONUS)** Build a Prefect Workflow and Auto Schedule it. Show the Prefect Dashboard with relevant outputs.
-

Task 1: Integrate MLflow into your existing machine learning projects.

Auto Logging Reviews Experiment Run Using MLFlow

Step 1 - Import MLFlow and set the experiment name

Step 1 - Import MLFlow and set the experiment name

```
import mlflow
mlflow.set_experiment("reviews_classification_ml_algos")
```

Step 2 - Start the auto logger

```
mlflow.sklearn.autolog()
```

Initialize the auto logger

max_tuning_runs=None will make sure that all the runs are recorded.

By default, top 5 runs will be recorded for each experiment

Step 3 - Start the experiment run

```
with mlflow.start_run() as run:
```

```
    clf.fit(X_train, y_train)
```

```
mlflow.sklearn.autolog(max_tuning_runs=None)
```

```
with mlflow.start_run() as run:
```

```
    %time grid_search.fit(X_train['clean_text_lemma'], y_train_le)
```

Step 4 - # Stop the auto logger

```
mlflow.sklearn.autolog(disable=True)
```

```
# Stop the auto logger
```

```
mlflow.sklearn.autolog(disable=True)
```

Task 2: Train machine learning models while logging relevant information with MLflow

Random Forest

```
***** random_forest *****
```

```
2024/03/28 09:33:28 WARNING mlflow.sklearn: Unrecognized dataset type <class 'pandas.core.series.Series'>. Dataset logging skipped.
```

```
Fitting 5 folds for each of 48 candidates, totalling 240 fits
```

```
CPU times: total: 15min 3s
```

```
Wall time: 21min 3s
```

```
2024/03/28 09:54:32 INFO mlflow.utils.autologging_utils: Created MLflow autologging run with ID 'b96e24da48b74943a25ffd7e0a11f9c2', which will track hyperparameters, performance metrics, model artifacts, and lineage information for the current sklearn workflow
```

```
2024/03/28 09:54:32 WARNING mlflow.sklearn: Failed to log training dataset information to MLflow Tracking. Reason: 'numpy.ndarray' object has no attribute 'toarray'
```

```
Accuracy on Test Data: 0.9111842105263158
```

```
F1 Score (Positive Class) on Test Data: 0.9493704795070989
```

```
F1 Score (Negative Class) on Test Data: 0.638623326959847
```

```
Classification Report:
```

	precision	recall	f1-score	support
Negative	0.62	0.65	0.64	255
Positive	0.95	0.95	0.95	1873
accuracy			0.91	2128
macro avg	0.79	0.80	0.79	2128
weighted avg	0.91	0.91	0.91	2128

Logistic Regression

```
***** logistic_regression *****
2024/03/28 09:55:03 WARNING mlflow.sklearn: Unrecognized dataset type <class 'pandas.core.series.Series'>. Dataset logging skipped.
Fitting 5 folds for each of 72 candidates, totalling 360 fits
CPU times: total: 21min 9s
Wall time: 24min 44s
2024/03/28 10:19:49 INFO mlflow.utils.autologging_utils: Created MLflow autologging run with ID '3ed352f0c7714d34828f5988cdb1987c', which will track hyperparameters, performance metrics, model artifacts, and lineage information for the current sklearn workflow
2024/03/28 10:19:49 WARNING mlflow.sklearn: Failed to log training dataset information to MLflow Tracking. Reason: 'numpy.ndarray' object has no attribute 'toarray'
Accuracy on Test Data: 0.8952067669172933
F1 Score (Positive Class) on Test Data: 0.9394186362401521
F1 Score (Negative Class) on Test Data: 0.6121739130434782
Classification Report:
              precision    recall  f1-score   support

   Negative         0.55         0.69         0.61         255
   Positive         0.96         0.92         0.94        1873

   accuracy                   0.90         2128
  macro avg         0.75         0.81         0.78         2128
 weighted avg         0.91         0.90         0.90         2128
```

XGBoost

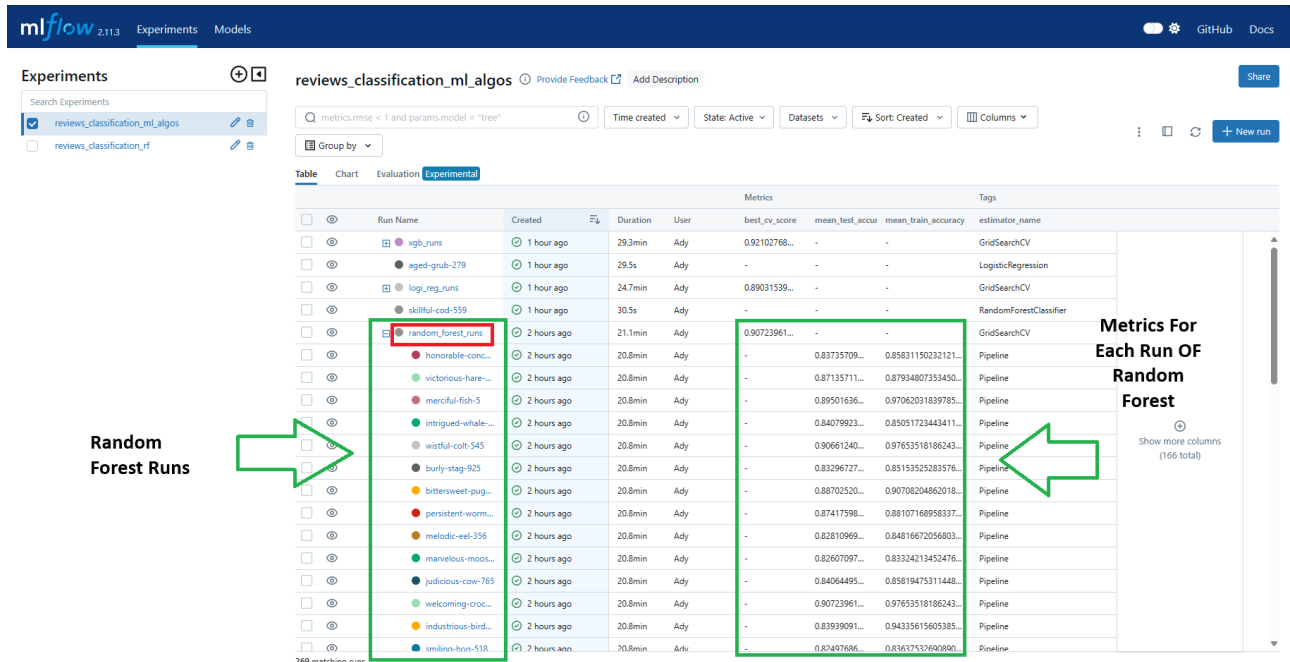
```
***** xgboost *****
2024/03/28 10:20:19 WARNING mlflow.sklearn: Unrecognized dataset type <class 'pandas.core.series.Series'>. Dataset logging skipped.
Fitting 5 folds for each of 144 candidates, totalling 720 fits
CPU times: total: 1h 22min
Wall time: 29min 16s
Accuracy on Test Data: 0.9172932330827067
F1 Score (Positive Class) on Test Data: 0.9534391534391534
F1 Score (Negative Class) on Test Data: 0.6302521008403361
Classification Report:
              precision    recall  f1-score   support

   Negative         0.68         0.59         0.63         255
   Positive         0.94         0.96         0.95        1873

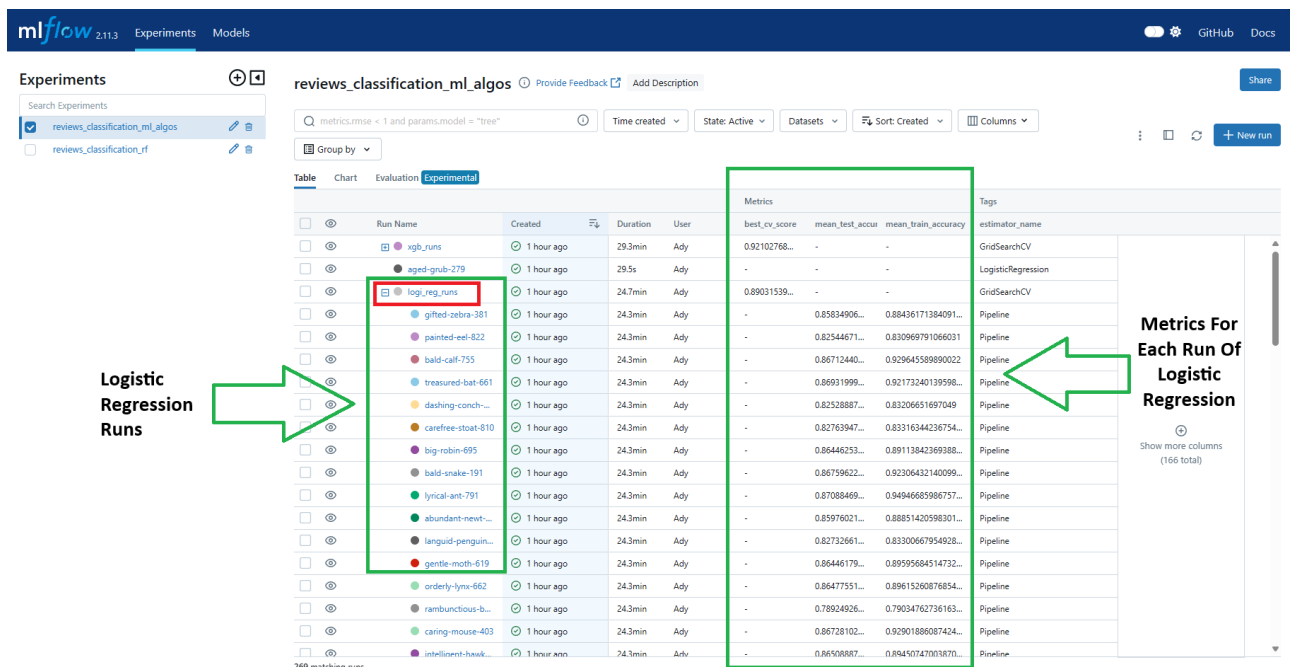
   accuracy                   0.92         2128
  macro avg         0.81         0.78         0.79         2128
 weighted avg         0.91         0.92         0.91         2128
```

Task 3: Demonstrate how to log parameters, metrics, and artifacts using MLflow tracking APIs.

Random Forest



Logistic Regression



XGBoost

The screenshot shows the MLflow Experiments page for the experiment 'reviews_classification_ml_algos'. The 'Run Name' column is highlighted with a green box, and a green arrow points to it with the text 'XGBoost Runs'. The 'Metrics' columns are also highlighted with a green box, and a green arrow points to it with the text 'Metrics For Each Run Of XGBoost'.

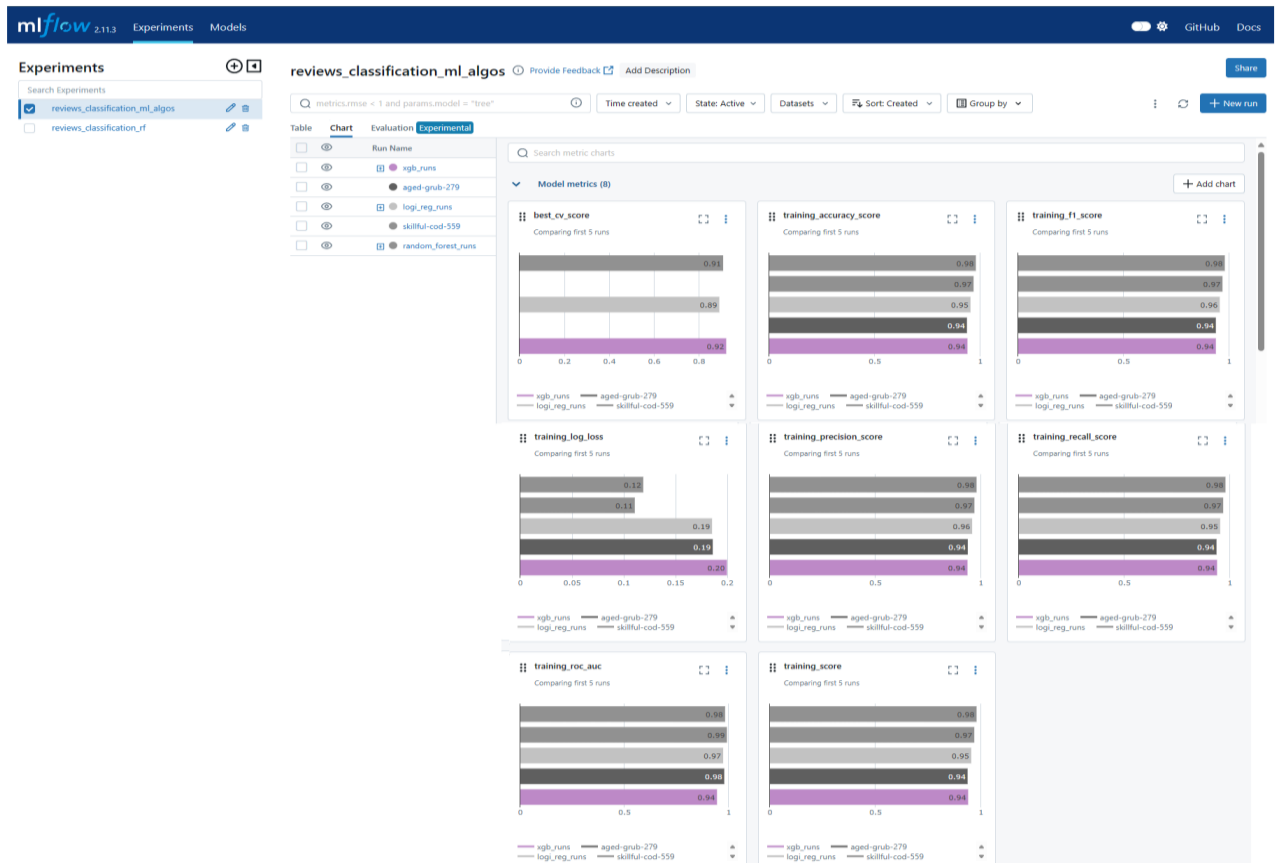
Run Name	Created	Duration	User	best_cv_score	mean_test_accu	mean_train_accuracy	estimator_name
xgb_runs	1 hour ago	29.3min	Ady	0.92102768...	-	-	GridSearchCV
trusting-smelt-3...	1 hour ago	28.3min	Ady	-	0.89877639...	0.90249927396203...	Pipeline
likeable-snake-4...	1 hour ago	28.3min	Ady	-	0.89517211...	0.90492798892662...	Pipeline
exultant-finch-492	1 hour ago	28.3min	Ady	-	0.89783632...	0.89854272958816...	Pipeline
funny-sow-933	1 hour ago	28.3min	Ady	-	0.90566951...	0.954403004972353	Pipeline
gifted-shrimp-494	1 hour ago	28.3min	Ady	-	0.90269342...	0.91648395062865...	Pipeline
burly-mewt-81	1 hour ago	28.3min	Ady	-	0.90864855...	0.95314933210261...	Pipeline
blushing-seal-59	1 hour ago	28.3min	Ady	-	0.88545866...	0.88914126492885...	Pipeline
funny-goose-744	1 hour ago	28.3min	Ady	-	0.91084119...	0.91801168796441...	Pipeline
honorable-elk-24	1 hour ago	28.3min	Ady	-	0.91413284...	0.92741299916788...	Pipeline
omniscient-ant-...	1 hour ago	28.3min	Ady	-	0.91037097...	0.92329983008601...	Pipeline
adaptable-rook-...	1 hour ago	28.3min	Ady	-	0.91601311...	0.943082122278988	Pipeline
unruly-ape-433	1 hour ago	28.3min	Ady	-	0.89752272...	0.90763074533887...	Pipeline
wistful-turtle-25	1 hour ago	28.3min	Ady	-	0.90316449...	0.96756466725210...	Pipeline
numbing-pig-727	1 hour ago	28.3min	Ady	-	0.90159624...	0.91899086669175...	Pipeline
whimsical-sloth...	1 hour ago	28.3min	Ady	-	0.91742278...	0.95267932754114...	Pipeline
youthful-mule-9...	1 hour ago	28.3min	Ady	-	0.91773590...	0.94801777632506...	Pipeline
rebellious-toad...	1 hour ago	28.3min	Ady	-	0.89846377...	0.90402698827942...	Pipeline
racious-moth-970	1 hour ago	28.3min	Ady	-	0.90833433...	0.91037300895836...	Pipeline

Task 4: Customizing MLflow UI with run names.

The screenshot shows the MLflow Experiments page for the experiment 'reviews_classification_ml_algos'. The 'Run Name' column is highlighted with a green box, and a green arrow points to it with the text 'XGBoost Runs'. The 'Metrics' columns are also highlighted with a green box, and a green arrow points to it with the text 'Metrics For Each Run Of XGBoost'.

Run Name	Created	Duration	User	best_cv_score	mean_test_accu	mean_train_accuracy	estimator_name
xgb_runs	1 hour ago	29.3min	Ady	0.92102768...	-	-	GridSearchCV
aged-grub-279	1 hour ago	29.5s	Ady	-	-	-	LogisticRegression
logi_reg_runs	1 hour ago	24.7min	Ady	0.89031539...	-	-	GridSearchCV
skillful-cod-559	1 hour ago	30.5s	Ady	-	-	-	RandomForestClassifier
random_forest_runs	2 hours ago	21.1min	Ady	0.90723961...	-	-	GridSearchCV

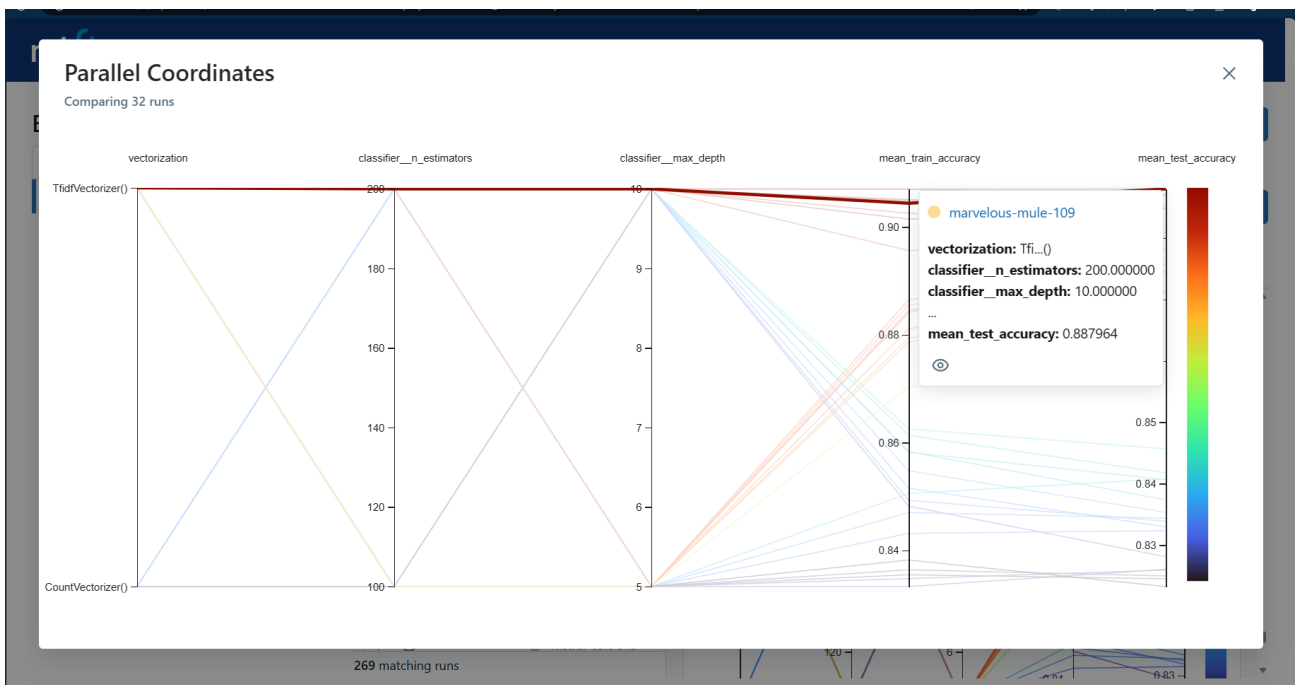
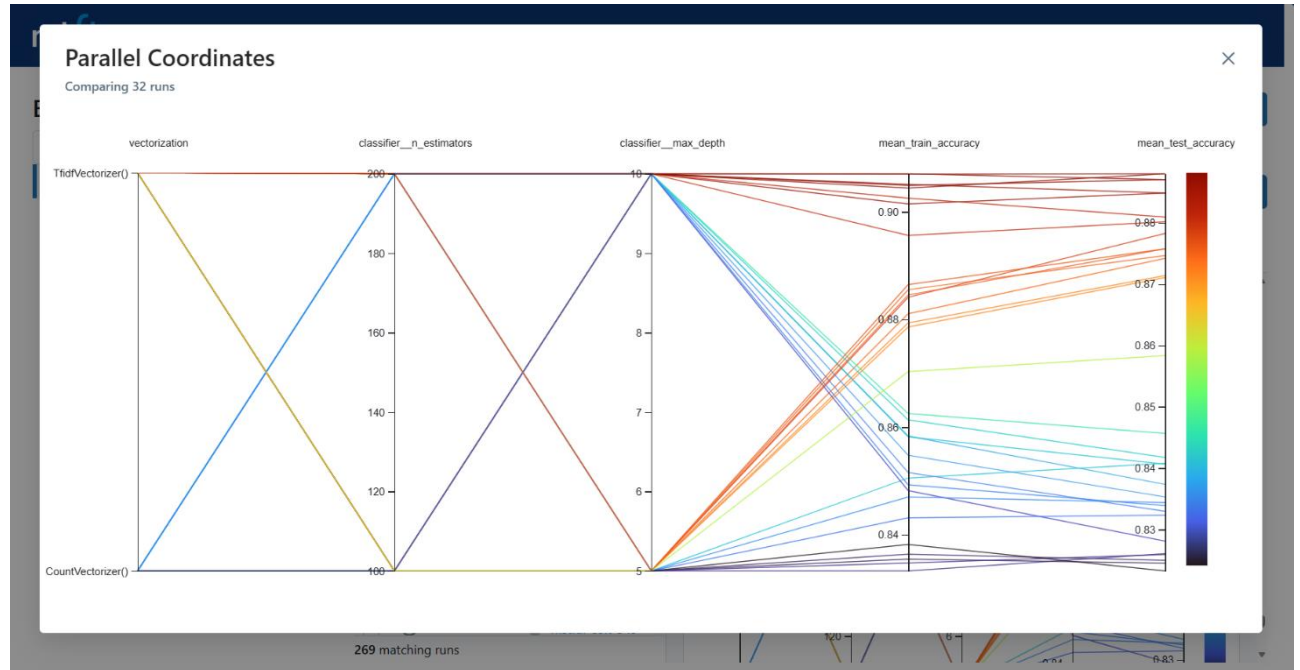
Task 5: Demonstrate metric plots.



Task 6: Demonstrate hyperparameter plots.

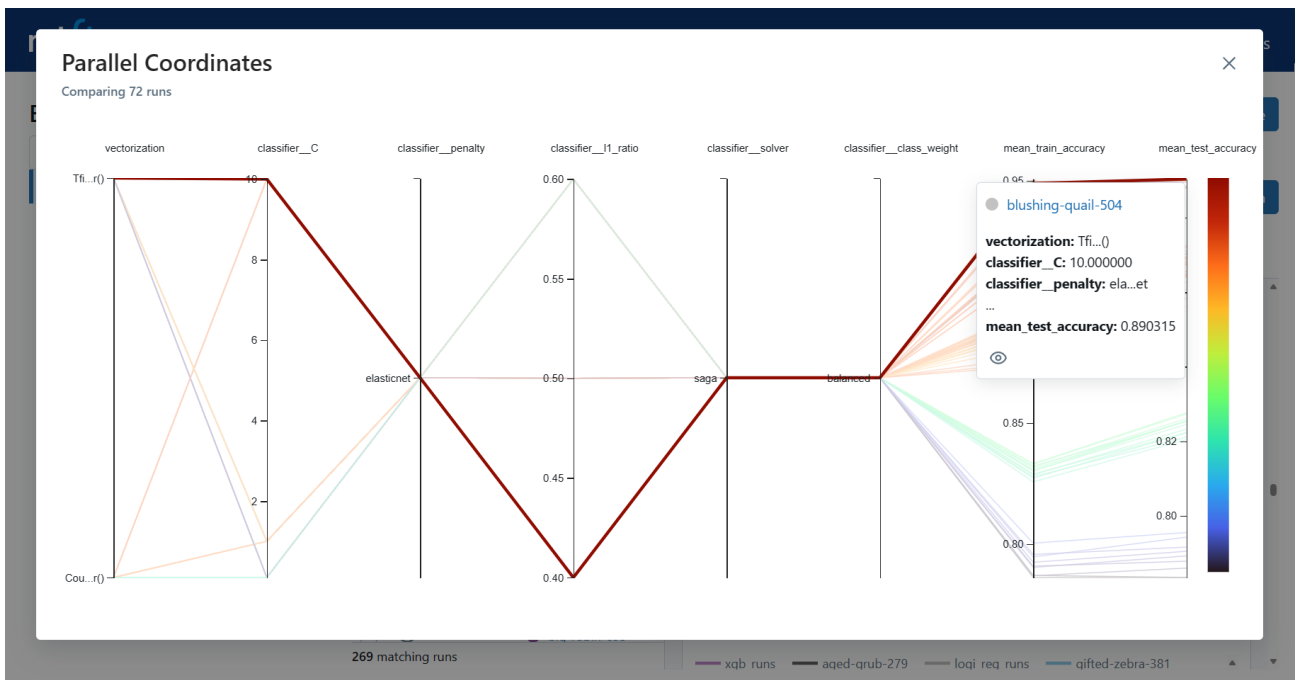
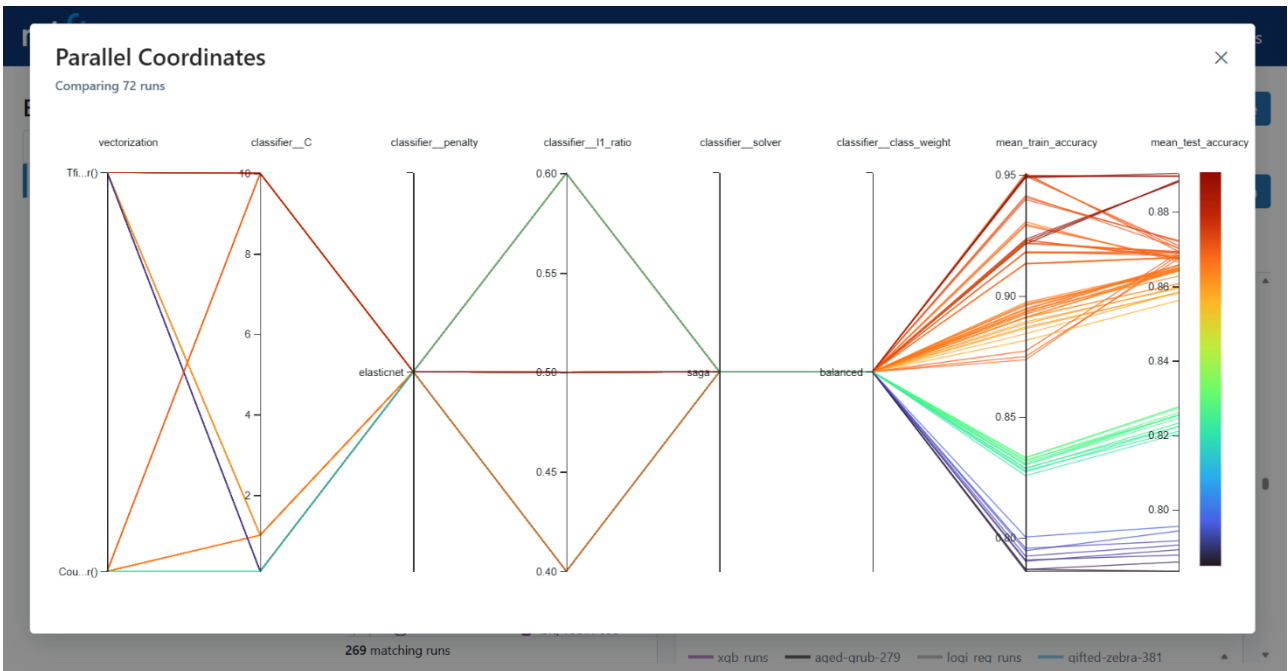
Random Forest

```
random_forest
Pipeline(memory=Memory(location=.cache\joblib),
        steps=[('vectorization', TfidfVectorizer(max_features=5000)),
                ('smote', SMOTE(random_state=42)),
                ('classifier', RandomForestClassifier())])
```



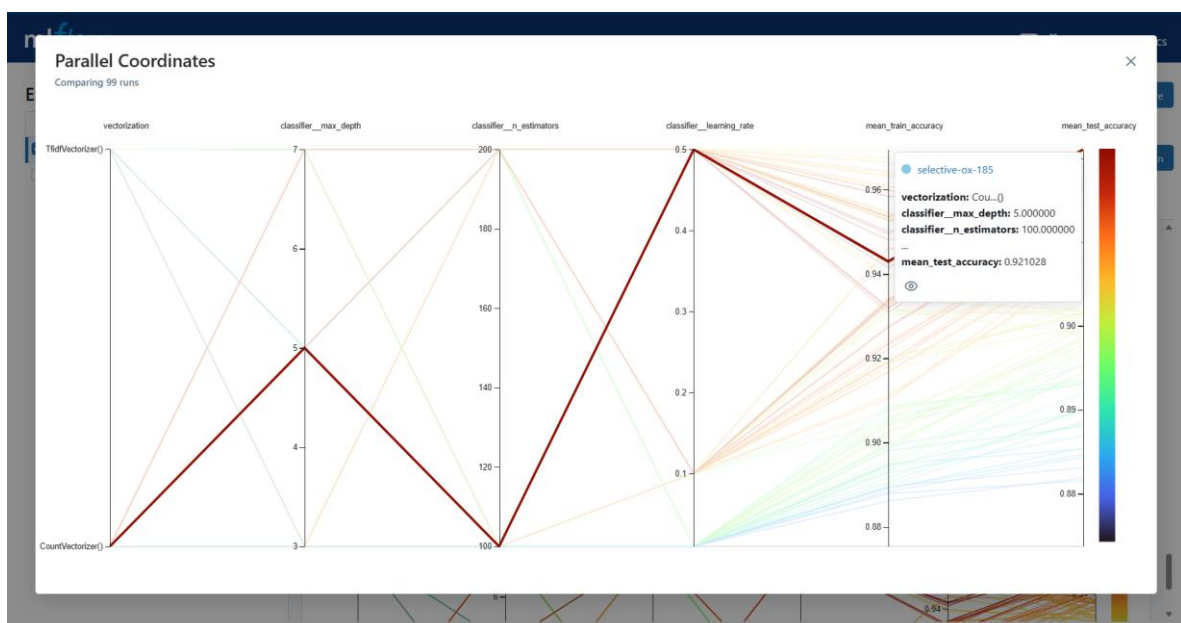
Logistic Regression

```
logistic_regression
Pipeline(memory=Memory(location=.cache\joblib),
  steps=[('vectorization', TfidfVectorizer(max_features=5000)),
        ('smote', SMOTE(random_state=42)),
        ('classifier',
         LogisticRegression(C=10, class_weight='balanced', l1_ratio=0.4,
                           penalty='elasticnet', solver='saga'))])
```



XGBoost

```
xgboost
Pipeline(memory=Memory(location=.cache\joblib),
  steps=[('vectorization', CountVectorizer(max_features=5000)),
        ('smote', SMOTE(random_state=42)),
        ('classifier',
         XGBClassifier(base_score=None, booster=None, callbacks=None,
                        colsample_bylevel=None, colsample_bynode=None,
                        colsample_bytree=None, device=None,
                        early_stopping_rounds=None,
                        enable_categorical=False, eval_me...
                        feature_types=None, gamma=None, grow_policy=None,
                        importance_type=None,
                        interaction_constraints=None, learning_rate=0.5,
                        max_bin=None, max_cat_threshold=None,
                        max_cat_to_onehot=None, max_delta_step=None,
                        max_depth=5, max_leaves=None,
                        min_child_weight=None, missing=nan,
                        monotone_constraints=None, multi_strategy=None,
                        n_estimators=100, n_jobs=None,
                        num_parallel_tree=None, random_state=None, ...)))]
```



Task 7: Demonstrate how to register models and manage by tagging them.

You Register Your Best Model (best_estimator)

Register model

- Unregistered models
- best_estimator View model
- model View model

Details

Created at	2024-03-28 09:15:39
Created by	Ady
Status	Finished
Run ID	43d2ece7043241d8b1f2b02159bd8c7c
Duration	7.3min
Datasets used	—
Tags	estimator_class: sklearn.model_selection_search.G... estimator_name: GridSearchCV
Source	C:\Users\adykh\AppData\Local\Programs\Python\Python312\Lib\site-packages\ipykernel_launcher.py
Logged models	sklearn +1
Registered models	—

Parameters (14)

Parameter	Value
best_classifier__max_depth	None
best_classifier__n_estimators	100
best_vectorization	TfidfVectorizer()
best_vectorization__max_features	2000
cv	2

Metrics (8)

Metric	Value
best_cv_score	0.9094327796928863
training_accuracy_score	0.9780633030397994
training_f1_score	0.9780747404724318
training_log_loss	0.12335042108047667
training_precision_score	0.9780865834243873

Register model

Model

Select a model

+ Create New Model

Models

- reviews_classification

Register New Model

Already Registered Model

Details

Created at	2024-03-28 09:15:39
Created by	Ady
Status	Finished
Run ID	43d2ece7043241d8b1f2b02159bd8c7c
Duration	7.3min
Datasets used	—
Tags	estimator_class: sklearn.model_selection_search.G... estimator_name: GridSearchCV
Source	C:\Users\adykh\AppData\Local\Programs\Python\Python312\Lib\site-packages\ipykernel_launcher.py
Logged models	sklearn +1
Registered models	—

Parameters (14)

Parameter	Value
best_classifier__max_depth	None
best_classifier__n_estimators	100
best_vectorization	TfidfVectorizer()
best_vectorization__max_features	2000
cv	2

Metrics (8)

Metric	Value
best_cv_score	0.9094327796928863
training_accuracy_score	0.9780633030397994
training_f1_score	0.9780747404724318
training_log_loss	0.12335042108047667
training_precision_score	0.9780865834243873

Registered Models

reviews_classification

Created Time: 2024-03-28 13:24:41

Last Modified: 2024-03-28 13:30:22

Description Edit

Tags

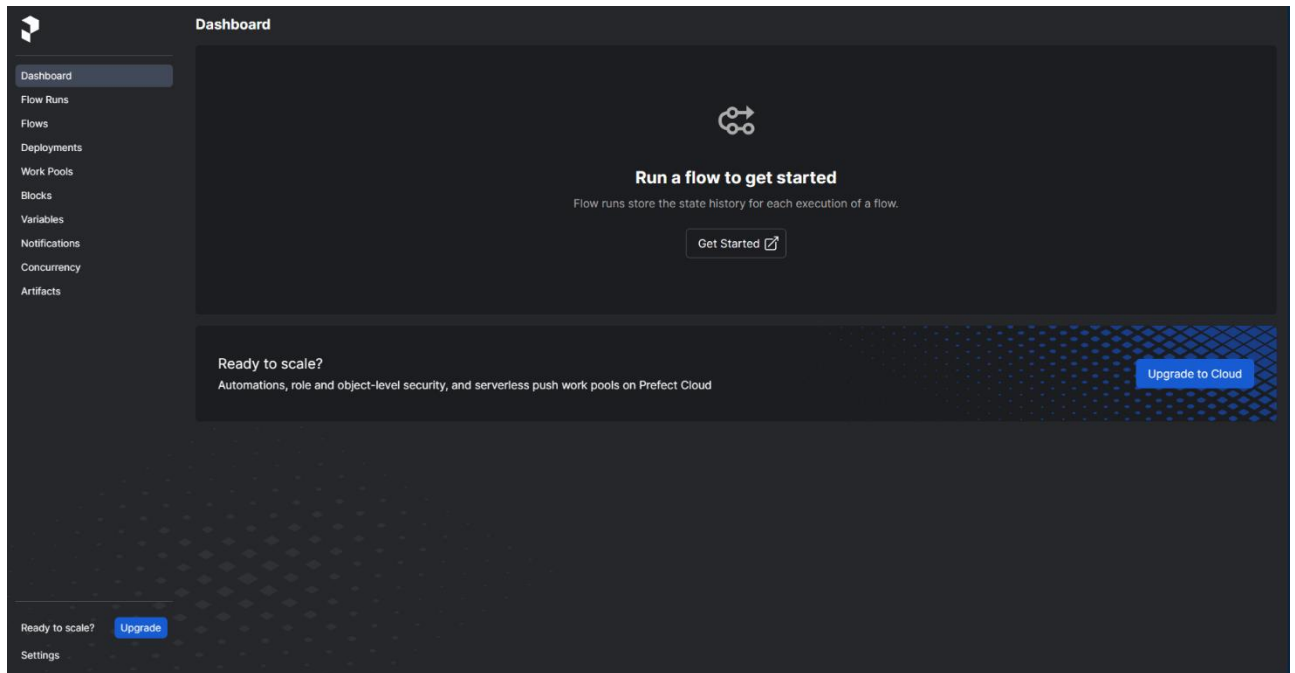
Versions All Active 2 Compare

New model registry UI

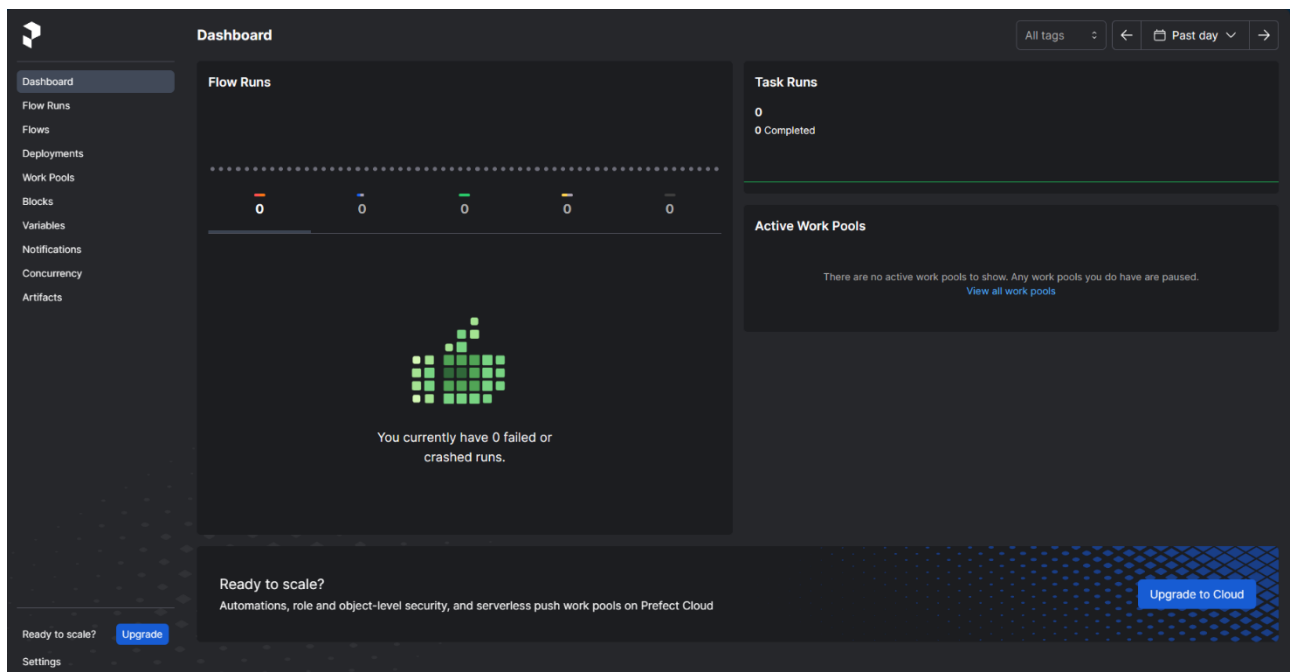
Version	Registered at	Created by	Stage	Description
Version 3	2024-03-28 13:28:33		Production	XGBOOST Model
Version 2	2024-03-28 13:27:51		Staging	Logistic Regression Model
Version 1	2024-03-28 13:24:41		Archived	Random Forest Model

Task 8: (BONUS) Build a Prefect Workflow and Auto Schedule it. Show the Prefect Dashboard with relevant outputs.

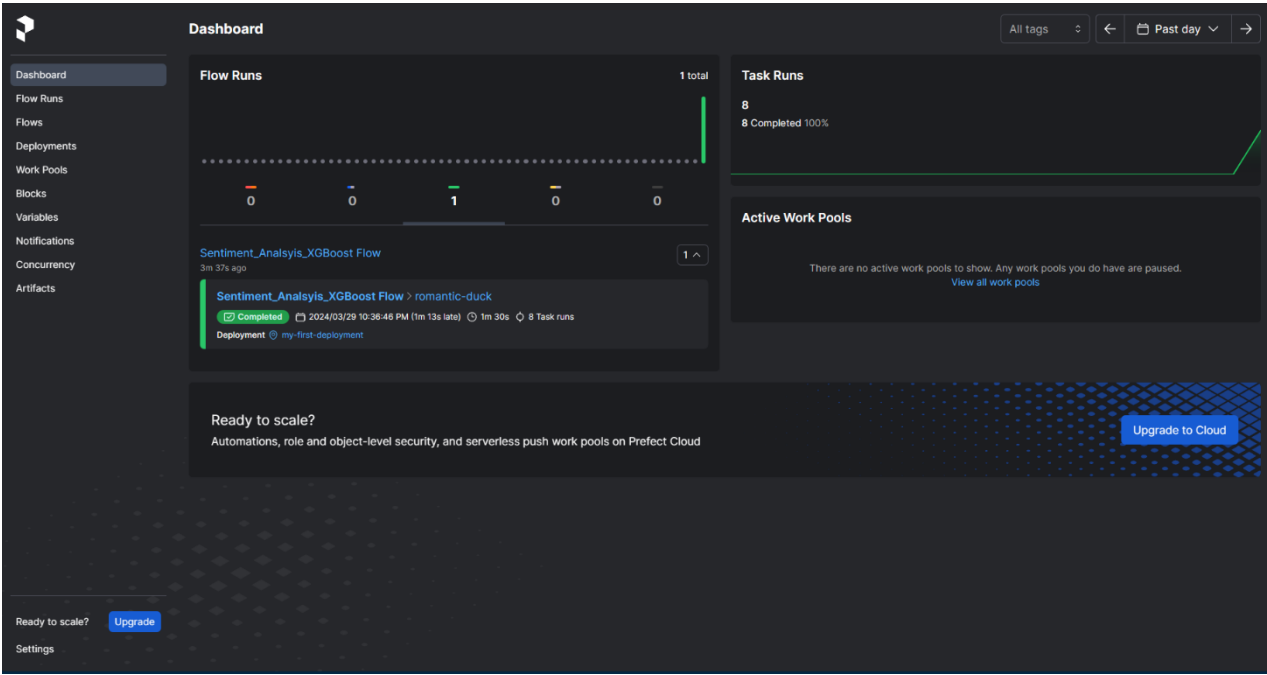
Prefect Dashboard



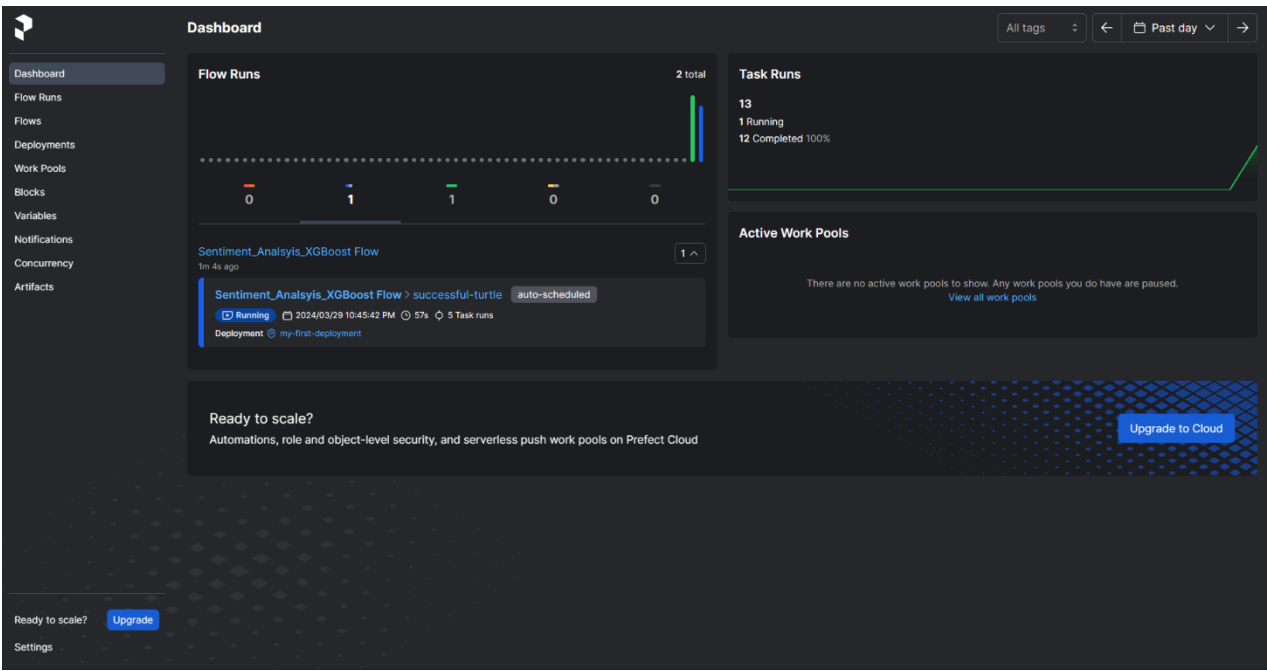
Prefect Dashboard After 1st Deployment



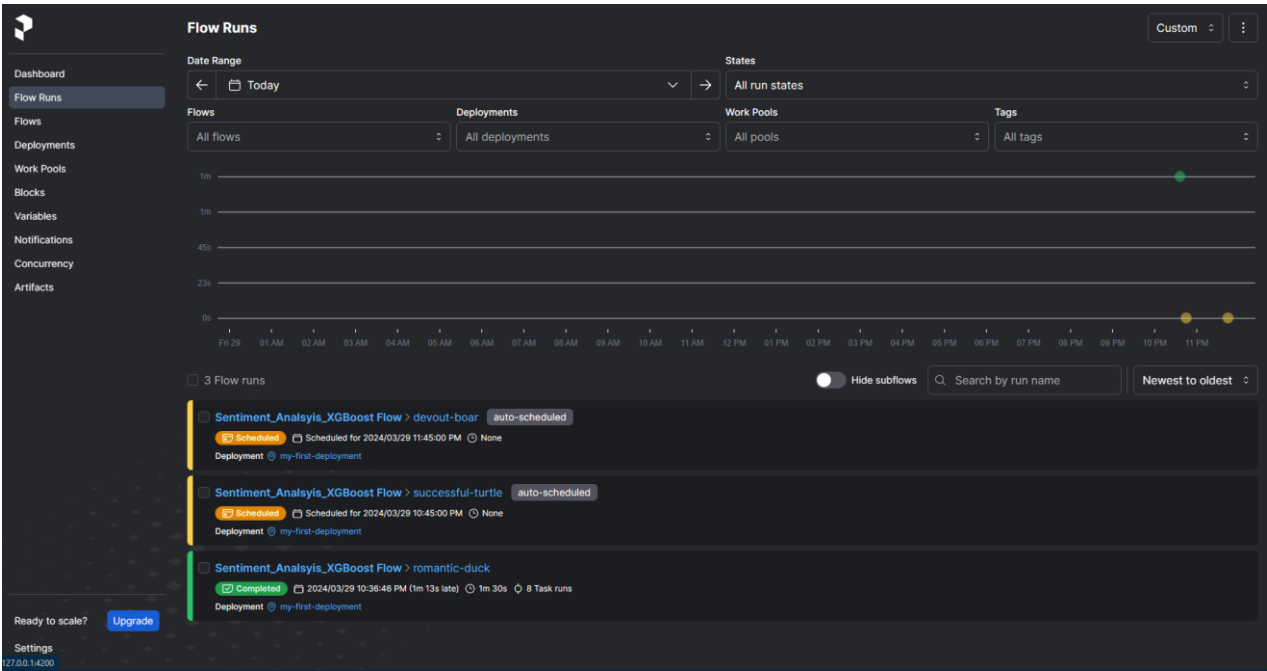
Prefect Dashboard (After Completing) 1st Task



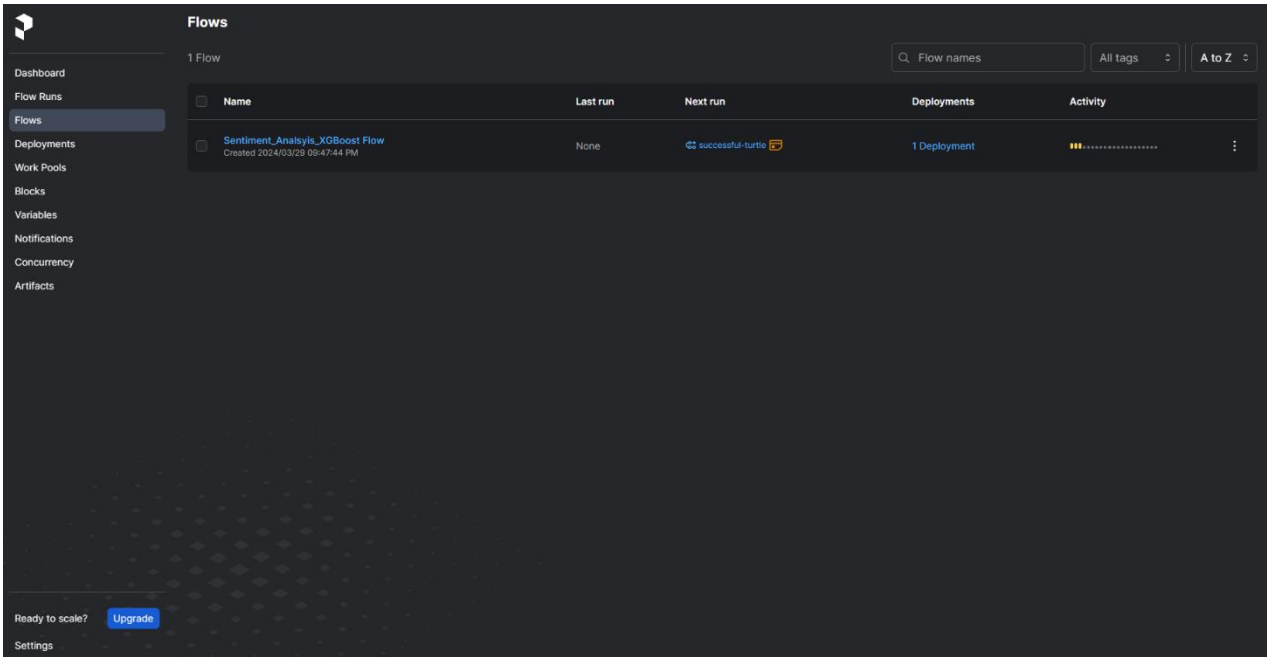
Prefect Dashboard (2nd Task Pending)



Prefect Flow Runs



Prefect Flows (Last Run: None)



Prefect Flows (With Last & Next Run)

Dashboard

Flow Runs

Flows

Deployments

Work Pools

Blocks

Variables

Notifications

Concurrency

Artifacts

Flows

1 Flow

Flow names

All tags

A to Z

<input type="checkbox"/>	Name	Last run	Next run	Deployments	Activity
<input type="checkbox"/>	<div>Sentiment_Analysis_XGBoost Flow</div> <div>Created 2024/03/29 09:47:44 PM</div>	<div>romantic-duck</div> <div></div>	<div>successful-turtle</div> <div></div>	1 Deployment	<div></div> <div></div>

Ready to scale?

Upgrade

Settings

Prefect Deployments (Quick Run) (I Use Quick Run(name): romantic-duck

Dashboard

Flow Runs

Flows

Deployments

Work Pools

Blocks

Variables

Notifications

Concurrency

Artifacts

Deployments

1 Deployment

Deployment names

All tags

A to Z

<input type="checkbox"/>	Deployment name	Flow name	Schedule	Tags	Activity
<input type="checkbox"/>	<div>my-first-deployment</div> <div>Created 2024/03/29 09:47:44 PM</div>	Sentiment_Analysis_XGBoost Flow	At 15 minutes past the hour every day		<div></div> <div></div>

Quick run

Custom run

Copy ID

Edit

Delete

Ready to scale?

Upgrade

Settings

Prefect Flows Runs (successful-turtle)(logs)

