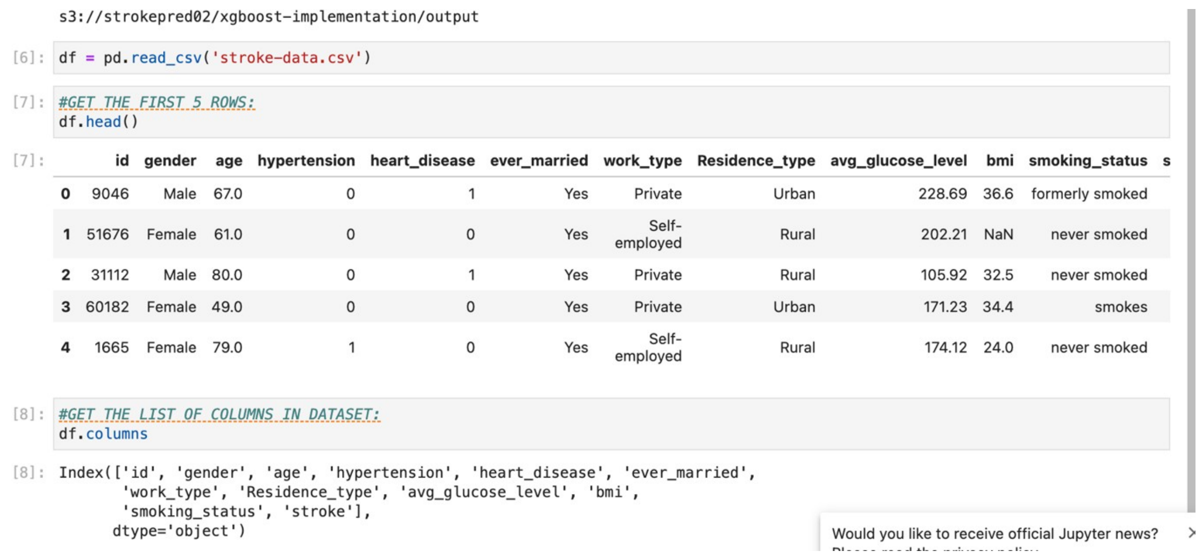
AWS FINAL PROJECT
SHARMI DAS **n01639206**

STROKE PREDICTION –

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

LOADING DATA



```
s3://strokepred02/xgboost-implementation/output
```

```
[6]: df = pd.read_csv('stroke-data.csv')
```

```
[7]: #GET THE FIRST 5 ROWS:
     df.head()
```

[7]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	s
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	

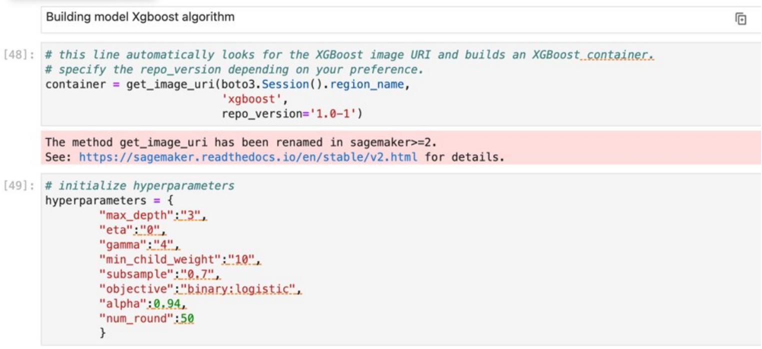
```
[8]: #GET THE LIST OF COLUMNS IN DATASET:
     df.columns
```

```
[8]: Index(['id', 'gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
            'work_type', 'Residence_type', 'avg_glucose_level', 'bmi',
            'smoking_status', 'stroke'],
           dtype='object')
```

Would you like to receive official Jupyter news?

COMPLETED TRAINING JOB-MODEL USED IS XGBOOST ALGORITHM

CONTAINER FOR THE MODEL

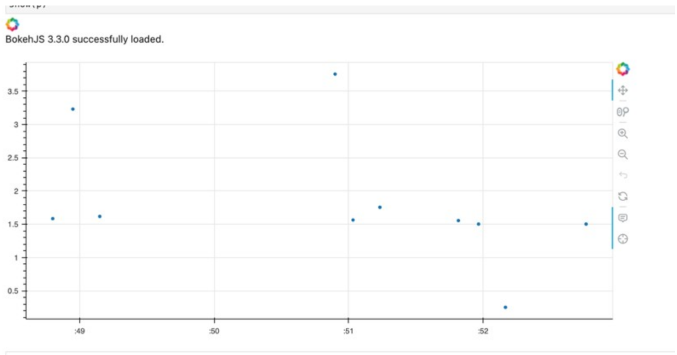


Building model Xgboost algorithm

```
[48]: # this line automatically looks for the XGBoost image URI and builds an XGBoost container.
      # specify the repo_version depending on your preference.
      container = get_image_uri(boto3.Session().region_name,
                                'xgboost',
                                repo_version='1.0-1')
```

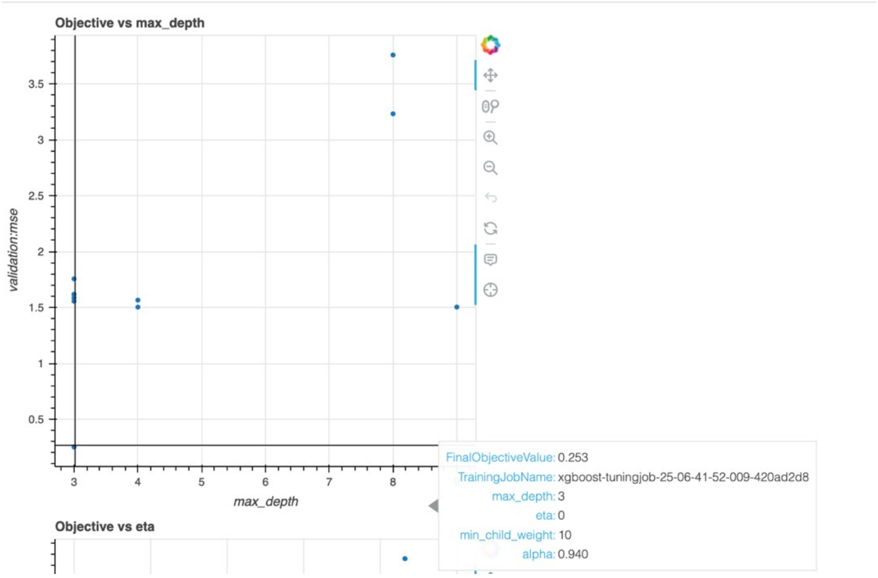
```
The method get_image_uri has been renamed in sagemaker>=2.
See: https://sagemaker.readthedocs.io/en/stable/v2.html for details.
```

```
[49]: # initialize hyperparameters
      hyperparameters = {
              "max_depth":"3",
              "eta":"0",
              "gamma":"4",
              "min_child_weight":"10",
              "subsample":"0.7",
              "objective":"binary:logistic",
              "alpha":0.94,
              "num_round":50
              }
```



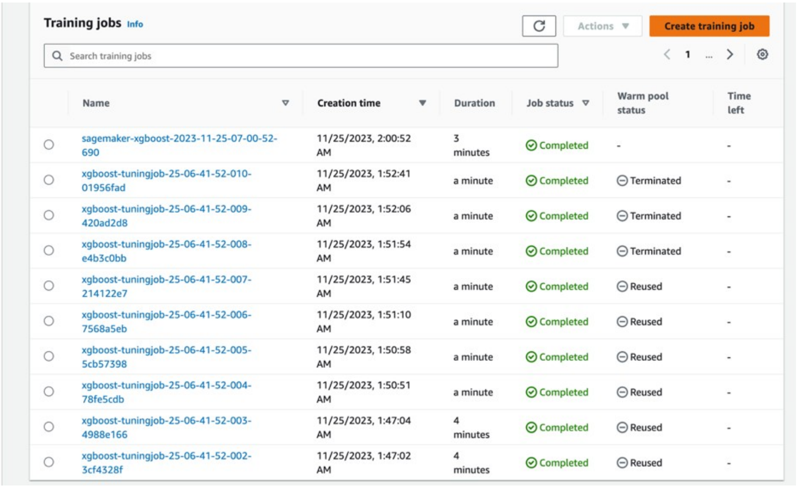
BokehJS 3.3.0 successfully loaded.

Graph Showing all the hyperparameter tuning jobs



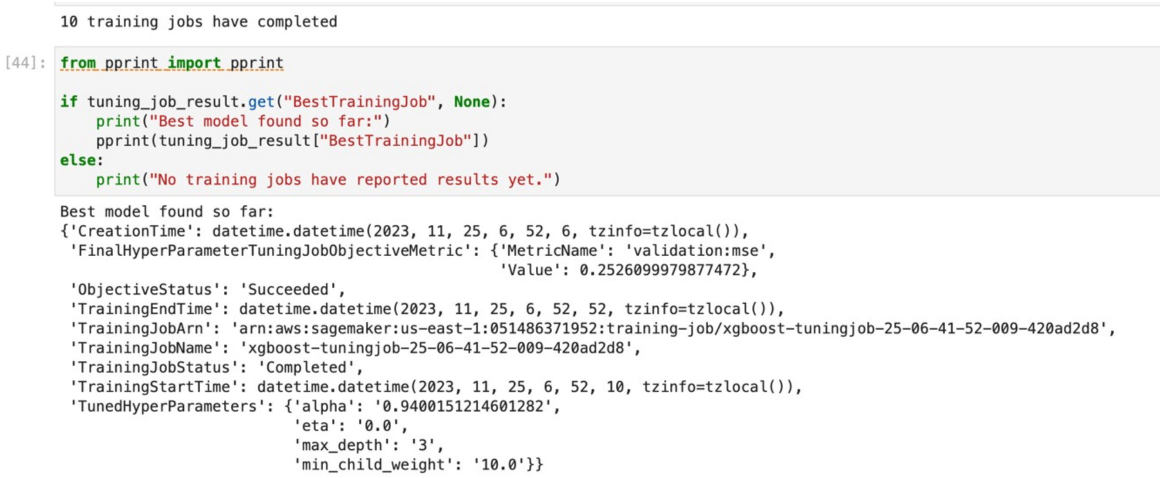
FinalObjectiveValue: 0.253
TrainingJobName: xgboost-tuningjob-25-06-41-52-009-420ad2d8
max_depth: 3
eta: 0
min_child_weight: 10
alpha: 0.940

Hyperparameter values of the best tuning job- later deployed successfully as Model 3 in Final Assignment



	Name	Creation time	Duration	Job status	Warm pool status	Time left
○	sagemaker-xgboost-2023-11-25-07-00-52-690	11/25/2023, 2:00:52 AM	3 minutes	✓ Completed	-	-
○	xgboost-tuningjob-25-06-41-52-010-01956fad	11/25/2023, 1:52:41 AM	a minute	✓ Completed	⊖ Terminated	-
○	xgboost-tuningjob-25-06-41-52-009-420ad2d8	11/25/2023, 1:52:06 AM	a minute	✓ Completed	⊖ Terminated	-
○	xgboost-tuningjob-25-06-41-52-008-e4b3c0bb	11/25/2023, 1:51:54 AM	a minute	✓ Completed	⊖ Terminated	-
○	xgboost-tuningjob-25-06-41-52-007-214122e7	11/25/2023, 1:51:45 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-006-7568a5eb	11/25/2023, 1:51:10 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-005-5cb57398	11/25/2023, 1:50:58 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-004-78fe5cdb	11/25/2023, 1:50:51 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-003-4988e166	11/25/2023, 1:47:04 AM	4 minutes	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-002-3cf4328f	11/25/2023, 1:47:02 AM	4 minutes	✓ Completed	⊖ Reused	-

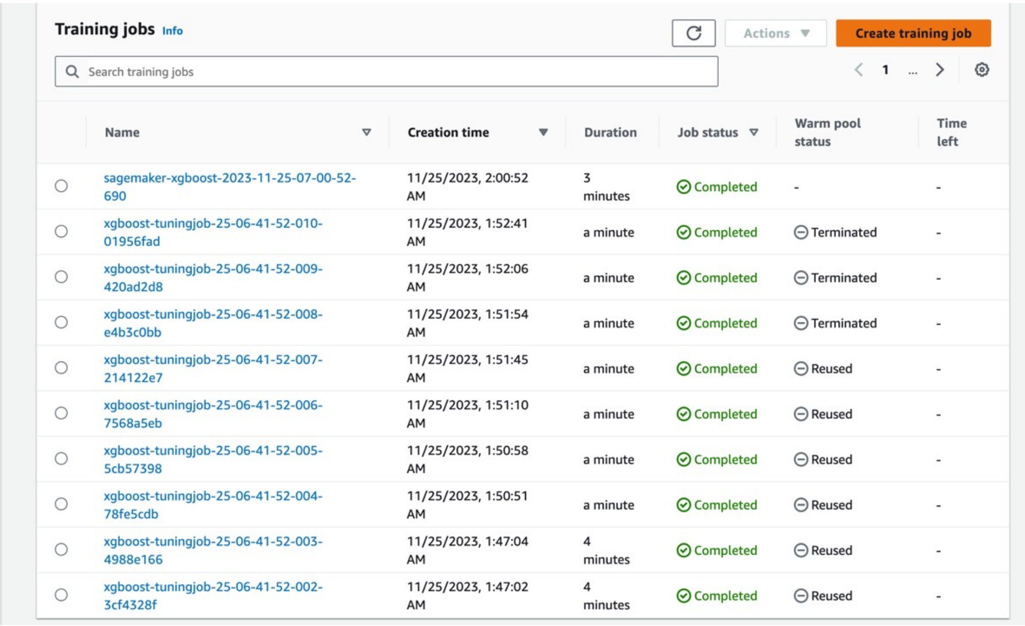
COMPLETED HYPERPARAMETER TUNING JOB



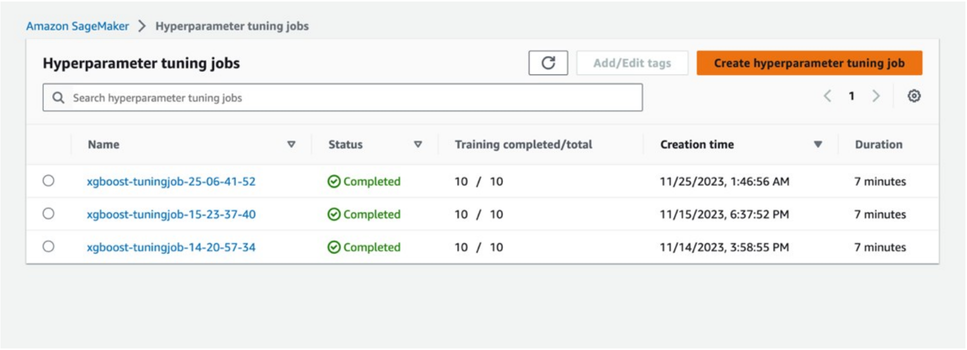
```
10 training jobs have completed
```

```
[44]: from pprint import pprint

      if tuning_job_result.get("BestTrainingJob", None):
          print("Best model found so far:")
          pprint(tuning_job_result["BestTrainingJob"])
      else:
          print("No training jobs have reported results yet.")
```

```
Best model found so far:
{'CreationTime': datetime.datetime(2023, 11, 25, 6, 52, 6, tzinfo=tzlocal()),
 'FinalHyperParameterTuningJobObjectiveMetric': {'MetricName': 'validation:mse',
                                                  'Value': 0.2526099979877472},
 'ObjectiveStatus': 'Succeeded',
 'TrainingEndTime': datetime.datetime(2023, 11, 25, 6, 52, 52, tzinfo=tzlocal()),
 'TrainingJobArn': 'arn:aws:sagemaker:us-east-1:051486371952:training-job/xgboost-tuningjob-25-06-41-52-009-420ad2d8',
 'TrainingJobName': 'xgboost-tuningjob-25-06-41-52-009-420ad2d8',
 'TrainingJobStatus': 'Completed',
 'TrainingStartTime': datetime.datetime(2023, 11, 25, 6, 52, 10, tzinfo=tzlocal()),
 'TunedHyperParameters': {'alpha': '0.9400151214601282',
                          'eta': '0.0',
                          'max_depth': '3',
                          'min_child_weight': '10.0'}}
```



	Name	Creation time	Duration	Job status	Warm pool status	Time left
○	sagemaker-xgboost-2023-11-25-07-00-52-690	11/25/2023, 2:00:52 AM	3 minutes	✓ Completed	-	-
○	xgboost-tuningjob-25-06-41-52-010-01956fad	11/25/2023, 1:52:41 AM	a minute	✓ Completed	⊖ Terminated	-
○	xgboost-tuningjob-25-06-41-52-009-420ad2d8	11/25/2023, 1:52:06 AM	a minute	✓ Completed	⊖ Terminated	-
○	xgboost-tuningjob-25-06-41-52-008-e4b3c0bb	11/25/2023, 1:51:54 AM	a minute	✓ Completed	⊖ Terminated	-
○	xgboost-tuningjob-25-06-41-52-007-214122e7	11/25/2023, 1:51:45 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-006-7568a5eb	11/25/2023, 1:51:10 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-005-5cb57398	11/25/2023, 1:50:58 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-004-78fe5cdb	11/25/2023, 1:50:51 AM	a minute	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-003-4988e166	11/25/2023, 1:47:04 AM	4 minutes	✓ Completed	⊖ Reused	-
○	xgboost-tuningjob-25-06-41-52-002-3cf4328f	11/25/2023, 1:47:02 AM	4 minutes	✓ Completed	⊖ Reused	-



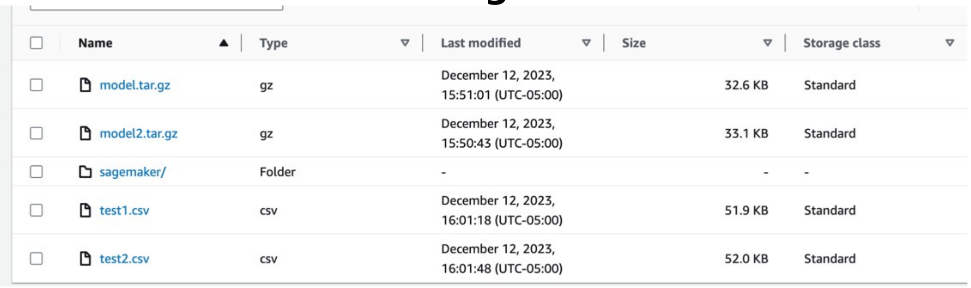
Amazon SageMaker > Hyperparameter tuning jobs
Hyperparameter tuning jobs
Add/Edit tags
Create hyperparameter tuning job
Search hyperparameter tuning jobs
< 1 >

	Name	Status	Training completed/total	Creation time	Duration
○	xgboost-tuningjob-25-06-41-52	✓ Completed	10 / 10	11/25/2023, 1:46:56 AM	7 minutes
○	xgboost-tuningjob-15-23-37-40	✓ Completed	10 / 10	11/15/2023, 6:37:52 PM	7 minutes
○	xgboost-tuningjob-14-20-57-34	✓ Completed	10 / 10	11/14/2023, 3:58:55 PM	7 minutes

0.25 best model
xgboost-tuningjob-25-06-41-52-009-420ad2d8
taken as Model 3 for Final Project

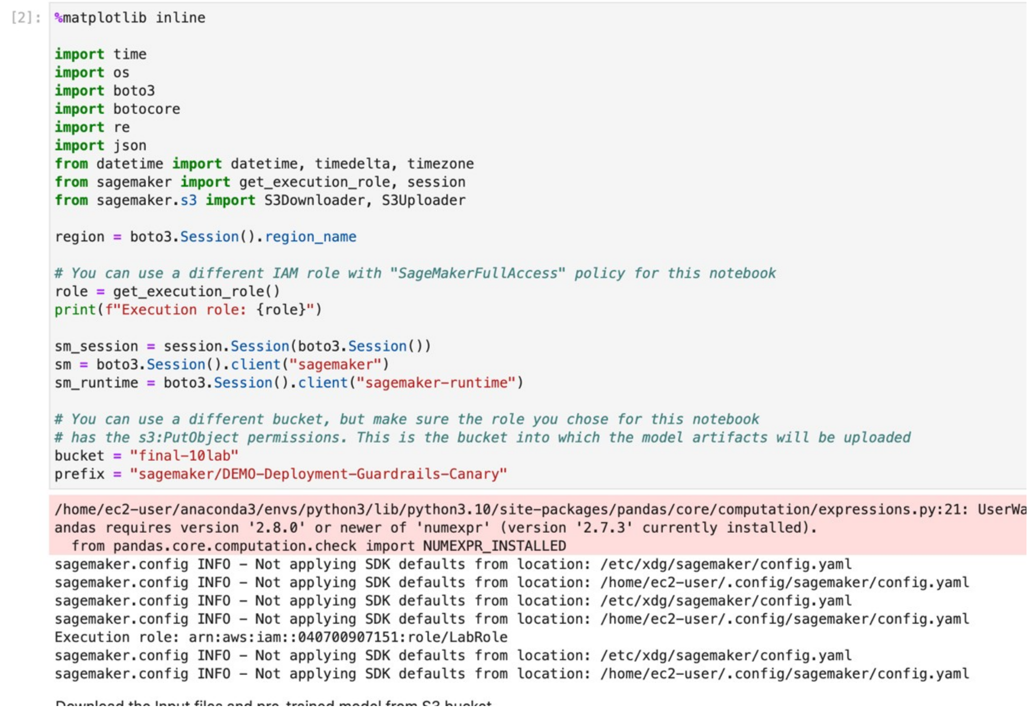
1.503 2nd best model
xgboost-tuningjob-25-06-41-52-010-01956fad
model 1 for Final Project

The model.tar.gz for these were downloaded and uploaded in bucket "final-10lab" along with test.csv files



□	Name ▲	Type ▽	Last modified ▽	Size ▽	Storage class ▽
□	model.tar.gz	gz	December 12, 2023, 15:51:01 (UTC-05:00)	32.6 KB	Standard
□	model2.tar.gz	gz	December 12, 2023, 15:50:43 (UTC-05:00)	33.1 KB	Standard
□	sagemaker/	Folder	-	-	-
□	test1.csv	csv	December 12, 2023, 16:01:18 (UTC-05:00)	51.9 KB	Standard
□	test2.csv	csv	December 12, 2023, 16:01:48 (UTC-05:00)	52.0 KB	Standard

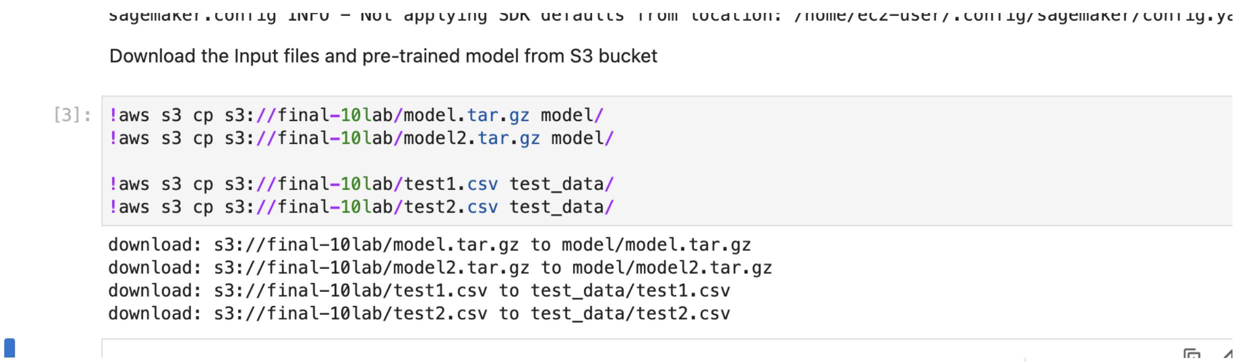
Deploying the models to S3



```
[2]: %matplotlib inline

     import time
     import os
     import boto3
     import botocore
     import re
     import json
     from datetime import datetime, timedelta, timezone
     from sagemaker import get_execution_role, session
     from sagemaker.s3 import S3Downloader, S3Uploader

     region = boto3.Session().region_name

     # You can use a different IAM role with "SageMakerFullAccess" policy for this notebook
     role = get_execution_role()
     print(f"Execution role: {role}")

     sm_session = session.Session(boto3.Session())
     sm = boto3.Session().client("sagemaker")
     sm_runtime = boto3.Session().client("sagemaker-runtime")

     # You can use a different bucket, but make sure the role you chose for this notebook
     # has the s3:PutObject permissions. This is the bucket into which the model artifacts will be uploaded
     bucket = "final-10lab"
     prefix = "sagemaker/DEMO-Deployment-Guardrails-Canary"
```

```
/home/ec2-user/anaconda3/envs/python3/lib/python3.10/site-packages/pandas/core/computation/expressions.py:21: UserWa
andas requires version '2.8.0' or newer of 'numexpr' (version '2.7.3' currently installed).
  from pandas.core.computation.check import NUMEXPR_INSTALLED
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
Execution role: arn:aws:iam::040700907151:role/LabRole
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
```

Download the Input files and pre-trained model from S3 bucket

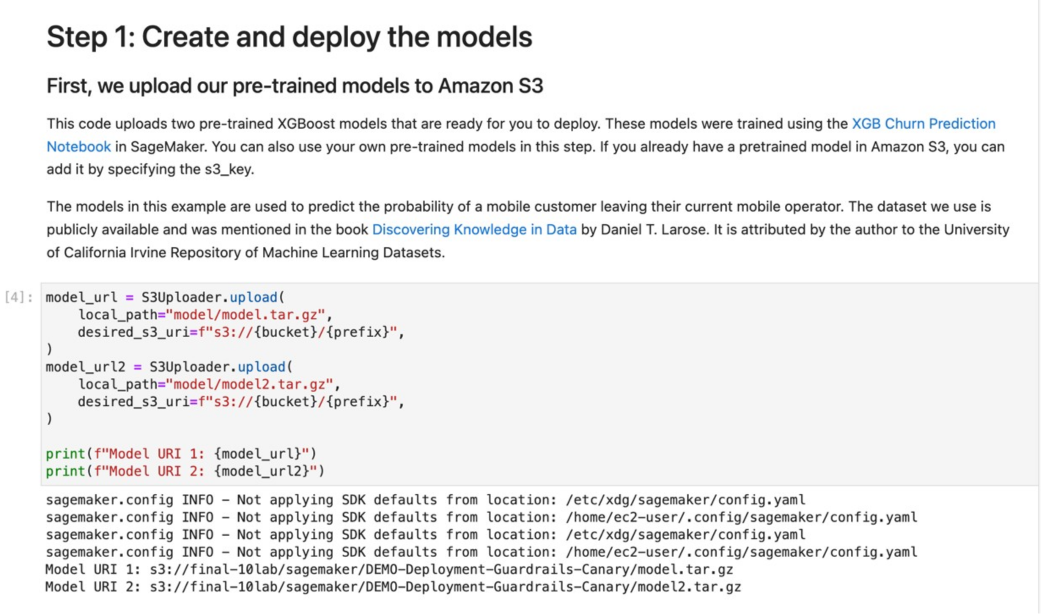


```
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.ya
```

Download the Input files and pre-trained model from S3 bucket

```
[3]: !aws s3 cp s3://final-10lab/model.tar.gz model/
     !aws s3 cp s3://final-10lab/model2.tar.gz model/

     !aws s3 cp s3://final-10lab/test1.csv test_data/
     !aws s3 cp s3://final-10lab/test2.csv test_data/
```

```
download: s3://final-10lab/model.tar.gz to model/model.tar.gz
download: s3://final-10lab/model2.tar.gz to model/model2.tar.gz
download: s3://final-10lab/test1.csv to test_data/test1.csv
download: s3://final-10lab/test2.csv to test_data/test2.csv
```



Step 1: Create and deploy the models

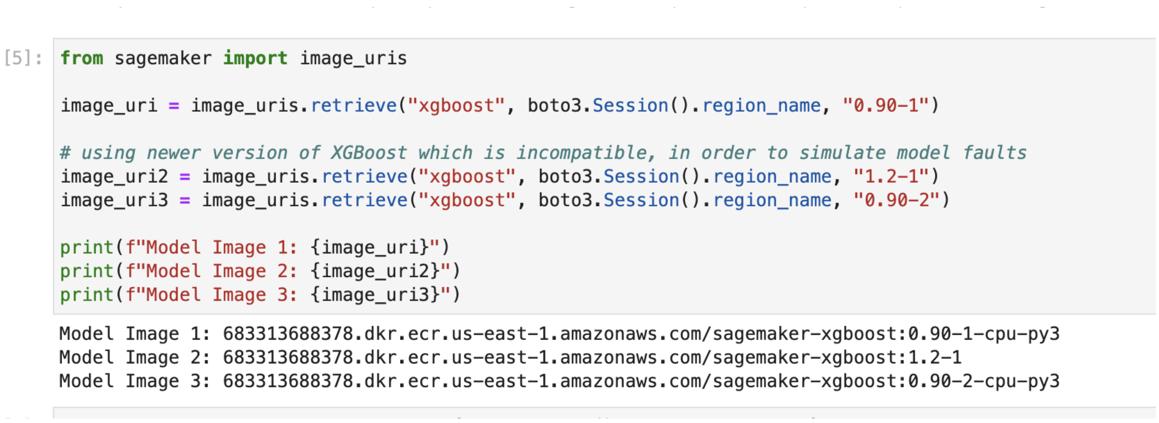
First, we upload our pre-trained models to Amazon S3

This code uploads two pre-trained XGBoost models that are ready for you to deploy. These models were trained using the XGB Churn Prediction Notebook in SageMaker. You can also use your own pre-trained models in this step. If you already have a pretrained model in Amazon S3, you can add it by specifying the s3_key.

The models in this example are used to predict the probability of a mobile customer leaving their current mobile operator. The dataset we use is publicly available and was mentioned in the book Discovering Knowledge in Data by Daniel T. Larose. It is attributed by the author to the University of California Irvine Repository of Machine Learning Datasets.

```
[4]: model_url = S3Uploader.upload(
         local_path="model/model.tar.gz",
         desired_s3_uri=f"s3://{bucket}/{prefix}",
     )
     model_url2 = S3Uploader.upload(
         local_path="model/model2.tar.gz",
         desired_s3_uri=f"s3://{bucket}/{prefix}",
     )

     print(f"Model URI 1: {model_url}")
     print(f"Model URI 2: {model_url2}")
```

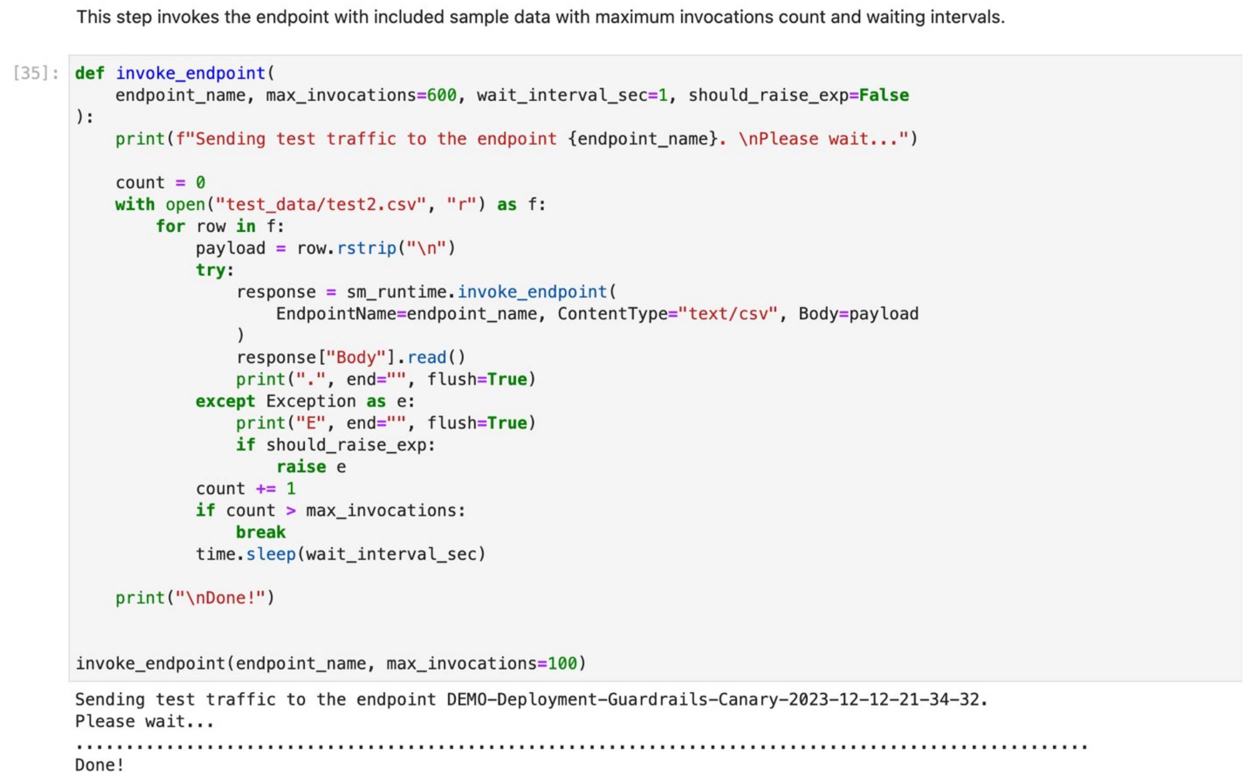
```
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
Model URI 1: s3://final-10lab/sagemaker/DEMO-Deployment-Guardrails-Canary/model.tar.gz
Model URI 2: s3://final-10lab/sagemaker/DEMO-Deployment-Guardrails-Canary/model2.tar.gz
```



```
[5]: from sagemaker import image_uris

     image_uri = image_uris.retrieve("xgboost", boto3.Session().region_name, "0.90-1")

     # using newer version of XGBoost which is incompatible, in order to simulate model faults
     image_uri2 = image_uris.retrieve("xgboost", boto3.Session().region_name, "1.2-1")
     image_uri3 = image_uris.retrieve("xgboost", boto3.Session().region_name, "0.90-2")

     print(f"Model Image 1: {image_uri}")
     print(f"Model Image 2: {image_uri2}")
     print(f"Model Image 3: {image_uri3}")
```

```
Model Image 1: 683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:0.90-1-cpu-py3
Model Image 2: 683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.2-1
Model Image 3: 683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:0.90-2-cpu-py3
```

CREATING MODEL OBJECTS WITH IMAGE AND MODEL DATA

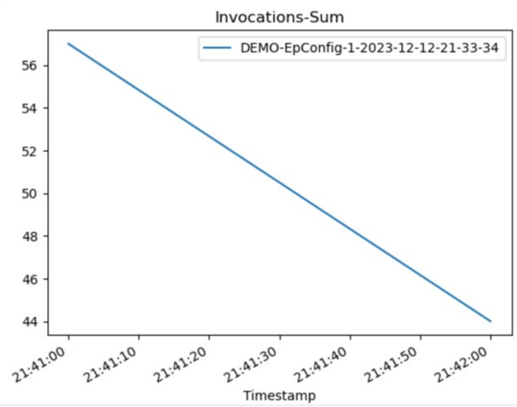


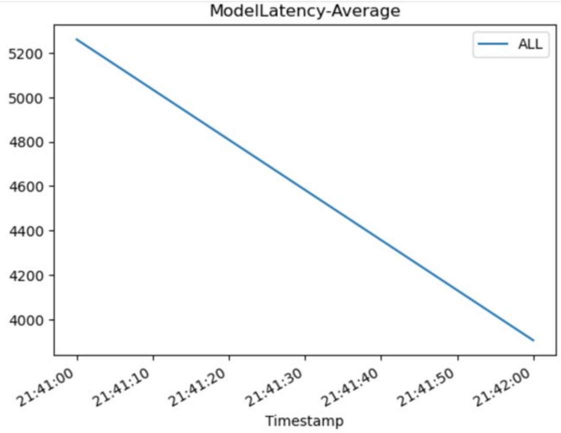
This step invokes the endpoint with included sample data with maximum invocations count and waiting intervals.

```
[35]: def invoke_endpoint(
          endpoint_name, max_invocations=600, wait_interval_sec=1, should_raise_exp=False
      ):
          print(f"Sending test traffic to the endpoint {endpoint_name}. \nPlease wait...")

          count = 0
          with open("test_data/test2.csv", "r") as f:
              for row in f:
                  payload = row.rstrip("\n")
                  try:
                      response = sm_runtime.invoke_endpoint(
                          EndpointName=endpoint_name, ContentType="text/csv", Body=payload
                      )
                      response["Body"].read()
                      print(".", end="", flush=True)
                  except Exception as e:
                      print("E", end="", flush=True)
                      if should_raise_exp:
                          raise e
                  count += 1
                  if count > max_invocations:
                      break
                  time.sleep(wait_interval_sec)

          print("\nDone!")


      invoke_endpoint(endpoint_name, max_invocations=100)
```

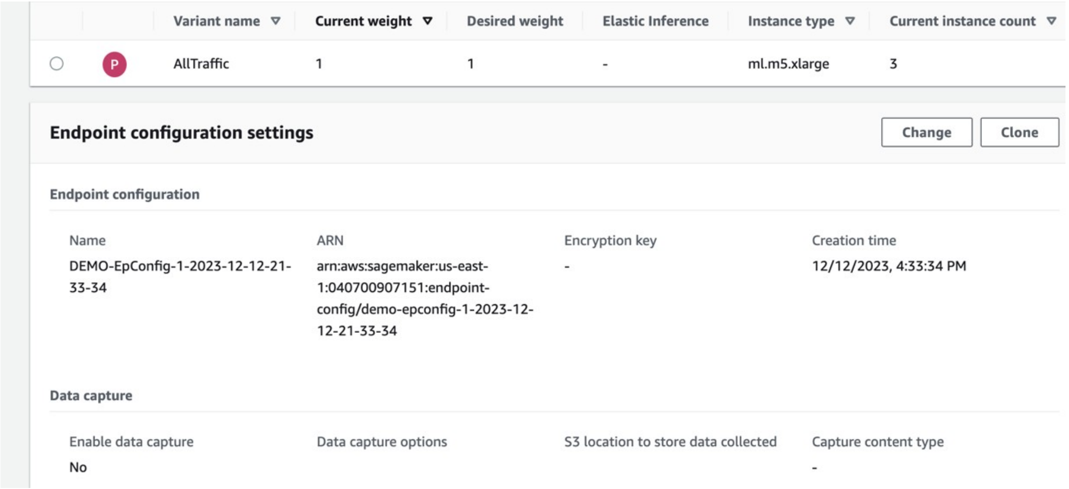
```
Sending test traffic to the endpoint DEMO-Deployment-Guardrails-Canary-2023-12-12-21-34-32.
Please wait...
....................................................................................................
Done!
```

SENDING TRAFFIC TO THE ENDPOINT WITH CONFIGURATION -1 (2^{nd} best model)





ModelLatency and OverheadLatency will start decreasing over time.



		Variant name	Current weight	Desired weight	Elastic Inference	Instance type	Current instance count
○	P	AllTraffic	1	1	-	ml.m5.xlarge	3

Endpoint configuration settings

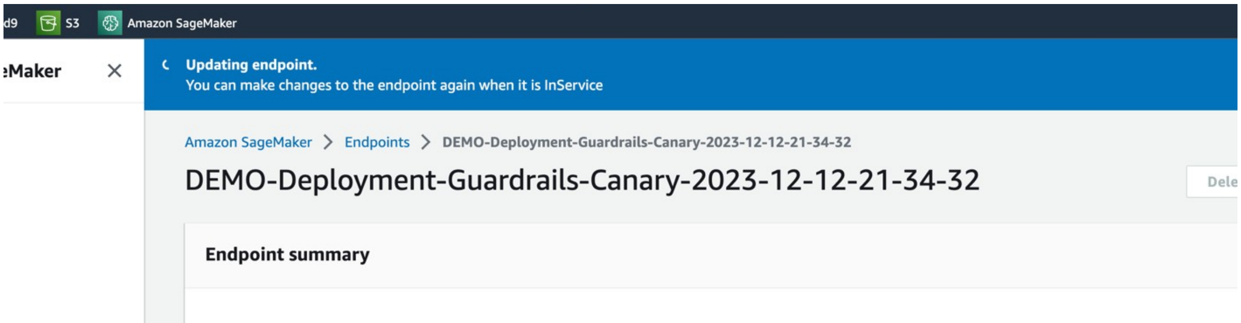
Endpoint configuration

Name	ARN	Encryption key	Creation time
DEMO-EpConfig-1-2023-12-12-21-33-34	arn:aws:sagemaker:us-east-1:040700907151:endpoint-config/demo-epconfig-1-2023-12-12-21-33-34	-	12/12/2023, 4:33:34 PM

Data capture

Enable data capture	Data capture options	S3 location to store data collected	Capture content type
No			-

METRICS FOR THE ENDPOINT WITH CONFIGURATION -1



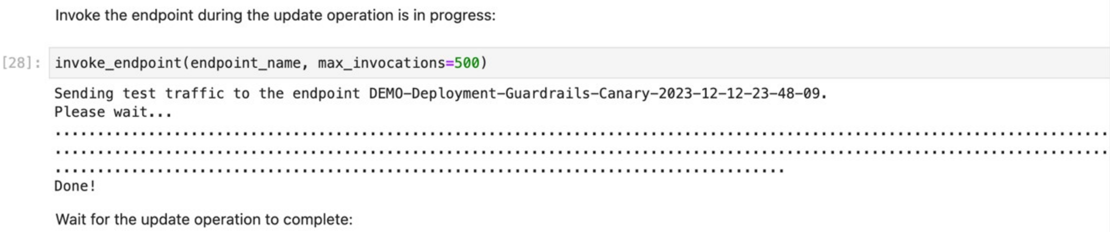
Updating endpoint.
You can make changes to the endpoint again when it is InService

Amazon SageMaker > Endpoints > DEMO-Deployment-Guardrails-Canary-2023-12-12-21-34-32

DEMO-Deployment-Guardrails-Canary-2023-12-12-21-34-32

Endpoint summary

SHOWS AN ENDPOINT BEING UPDATED



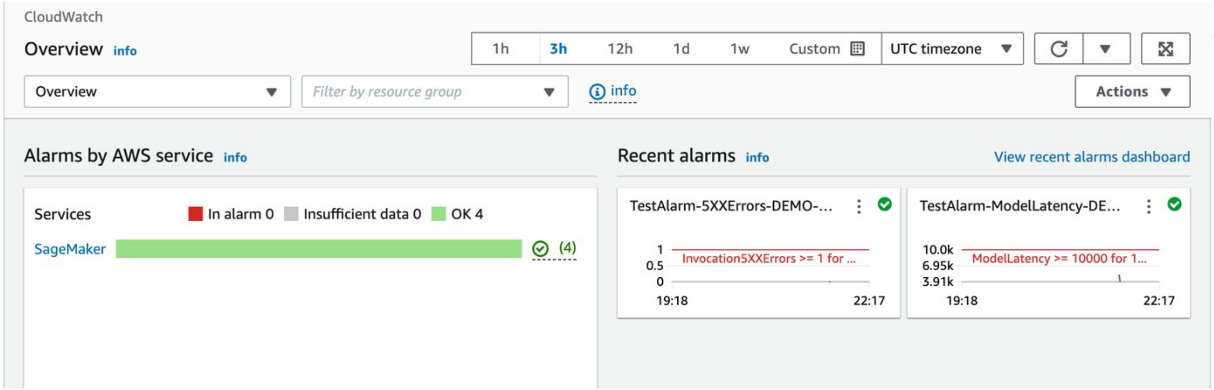
Invoke the endpoint during the update operation is in progress:

```
[28]: invoke_endpoint(endpoint_name, max_invocations=500)
```

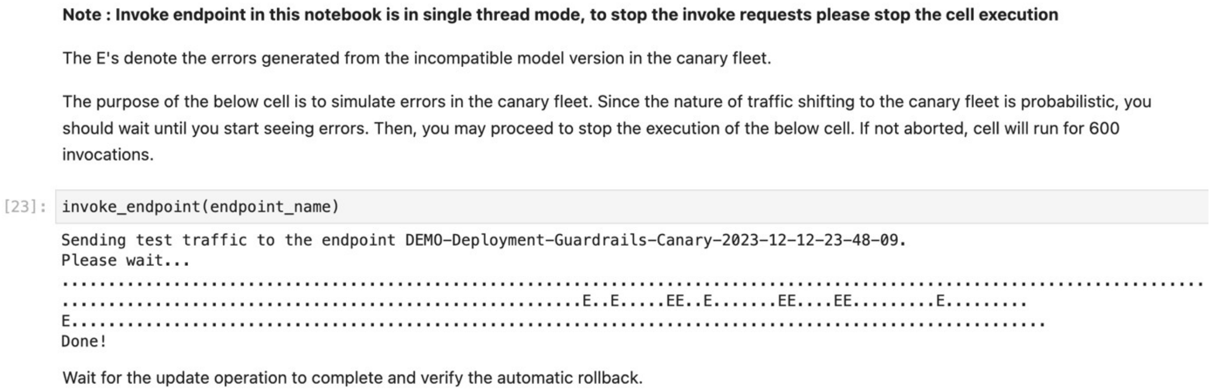
```
Sending test traffic to the endpoint DEMO-Deployment-Guardrails-Canary-2023-12-12-23-48-09.
Please wait...
...................................................................................................................
...................................................................................................................
...............................................................................
Done!
```

Wait for the update operation to complete:

This step invokes the endpoint with included sample data with maximum invocations count and waiting intervals



GREEN REGION FOR SUCCESSFUL DEPLOYMENT



Note : Invoke endpoint in this notebook is in single thread mode, to stop the invoke requests please stop the cell execution

The E's denote the errors generated from the incompatible model version in the canary fleet.

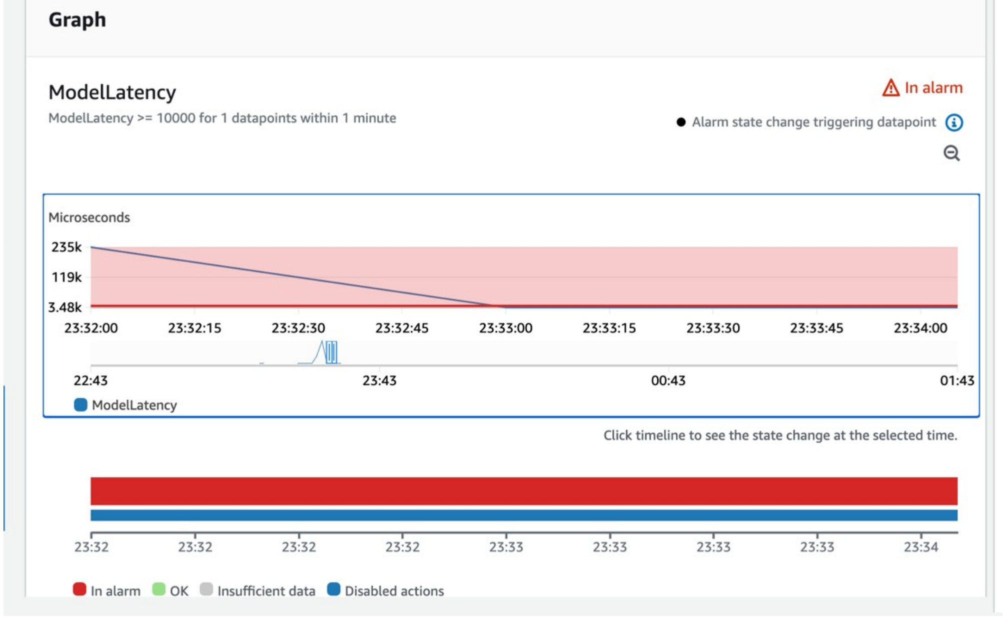
The purpose of the below cell is to simulate errors in the canary fleet. Since the nature of traffic shifting to the canary fleet is probabilistic, you should wait until you start seeing errors. Then, you may proceed to stop the execution of the below cell. If not aborted, cell will run for 600 invocations.

```
[23]: invoke_endpoint(endpoint_name)
```

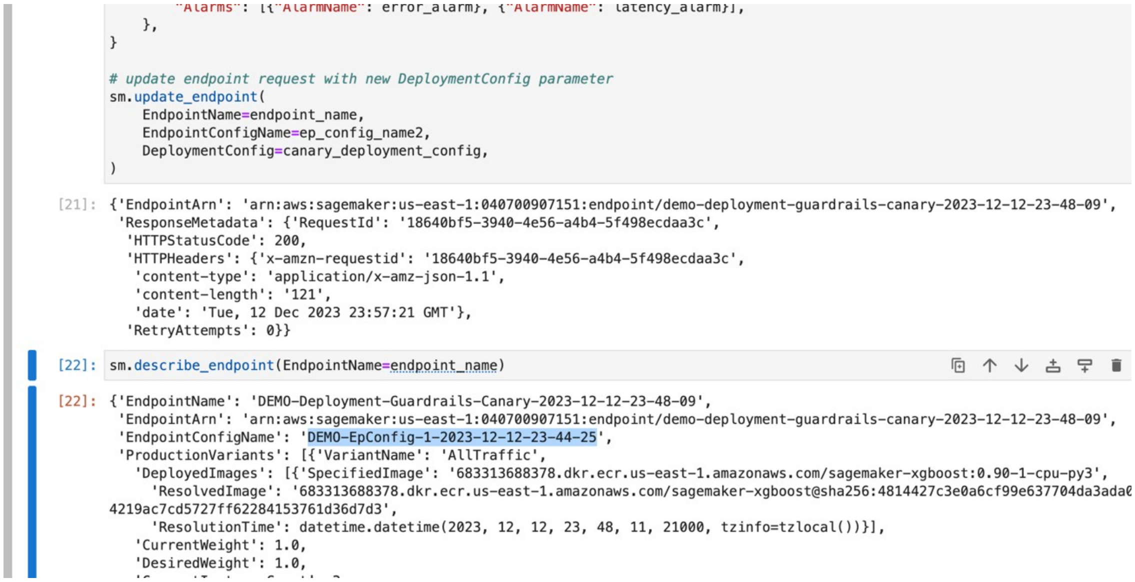
```
Sending test traffic to the endpoint DEMO-Deployment-Guardrails-Canary-2023-12-12-23-48-09.
Please wait...
..............................................................................................................................
.........................................................E..E.....EE..E.......EE....EE.........E.........
E...........................................................................................................
Done!
```

Wait for the update operation to complete and verify the automatic rollback.

ERRORS SHOWN IN NOTEBOOK



ALARM SHOWING ERROR IN CLOUDWATCH LOGS



```
        "Alarms": [{"AlarmName": error_alarm}, {"AlarmName": latency_alarm}],
    },
}

# update endpoint request with new DeploymentConfig parameter
sm.update_endpoint(
    EndpointName=endpoint_name,
    EndpointConfigName=ep_config_name2,
    DeploymentConfig=canary_deployment_config,
)
```

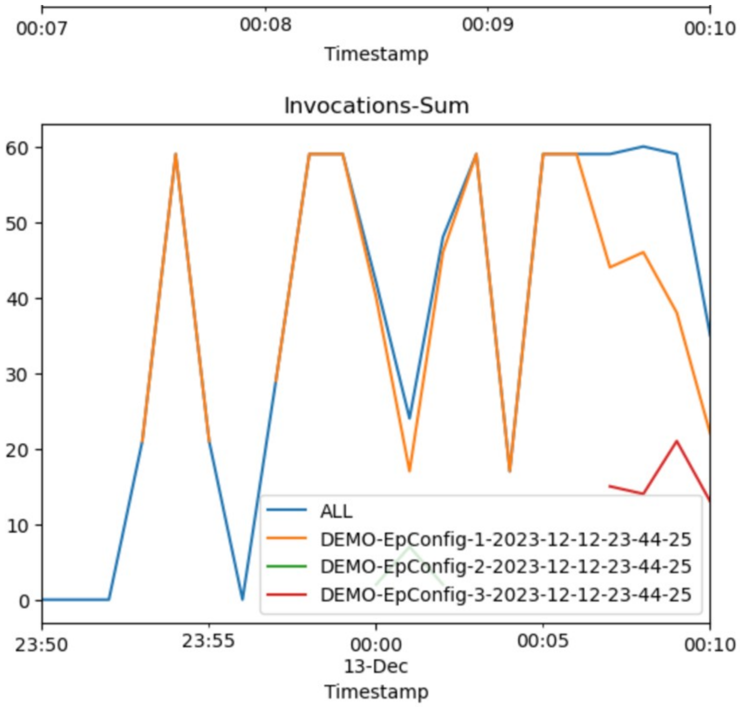
```
[21]: {'EndpointArn': 'arn:aws:sagemaker:us-east-1:040700907151:endpoint/demo-deployment-guardrails-canary-2023-12-12-23-48-09',
       'ResponseMetadata': {'RequestId': '18640bf5-3940-4e56-a4b4-5f498ecdaa3c',
        'HTTPStatusCode': 200,
        'HTTPHeaders': {'x-amzn-requestid': '18640bf5-3940-4e56-a4b4-5f498ecdaa3c',
         'content-type': 'application/x-amz-json-1.1',
         'content-length': '121',
         'date': 'Tue, 12 Dec 2023 23:57:21 GMT'},
        'RetryAttempts': 0}}
```

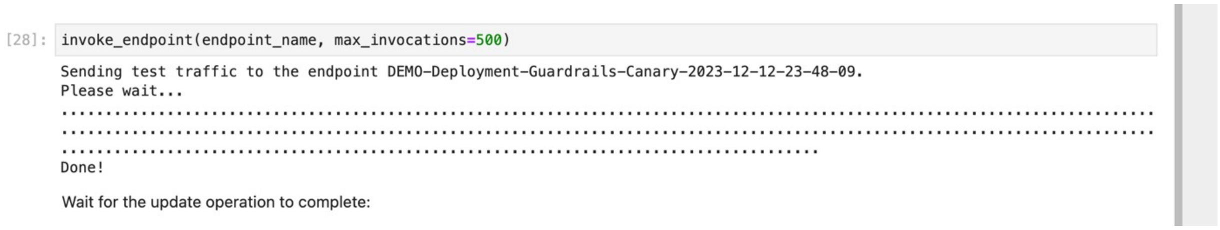
```
[22]: sm.describe_endpoint(EndpointName=endpoint_name)
```

```
[22]: {'EndpointName': 'DEMO-Deployment-Guardrails-Canary-2023-12-12-23-48-09',
       'EndpointArn': 'arn:aws:sagemaker:us-east-1:040700907151:endpoint/demo-deployment-guardrails-canary-2023-12-12-23-48-09',
       'EndpointConfigName': 'DEMO-EpConfig-1-2023-12-12-23-44-25',
       'ProductionVariants': [{'VariantName': 'AllTraffic',
         'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:0.90-1-cpu-py3',
           'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:4814427c3e0a6cf99e637704da3ada@
4219ac7cd5727ff62284153761d36d7d3',
           'ResolutionTime': datetime.datetime(2023, 12, 12, 23, 48, 11, 21000, tzinfo=tzlocal())}],
         'CurrentWeight': 1.0,
         'DesiredWeight': 1.0,
```

Back to Endpoint Configuration-1 after Rollback

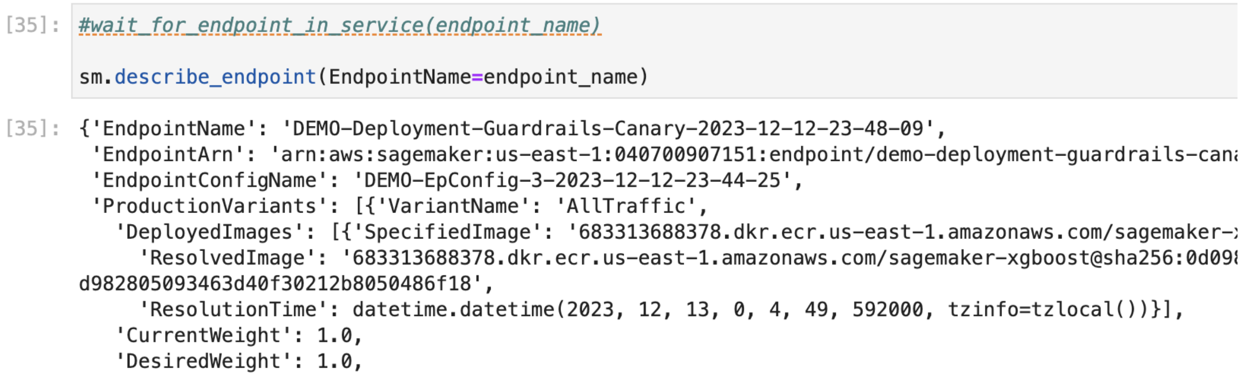
GRAPH FOR INVOCATIONS SUM FOR ALL THE ENDPOINT CONFIGURATIONS





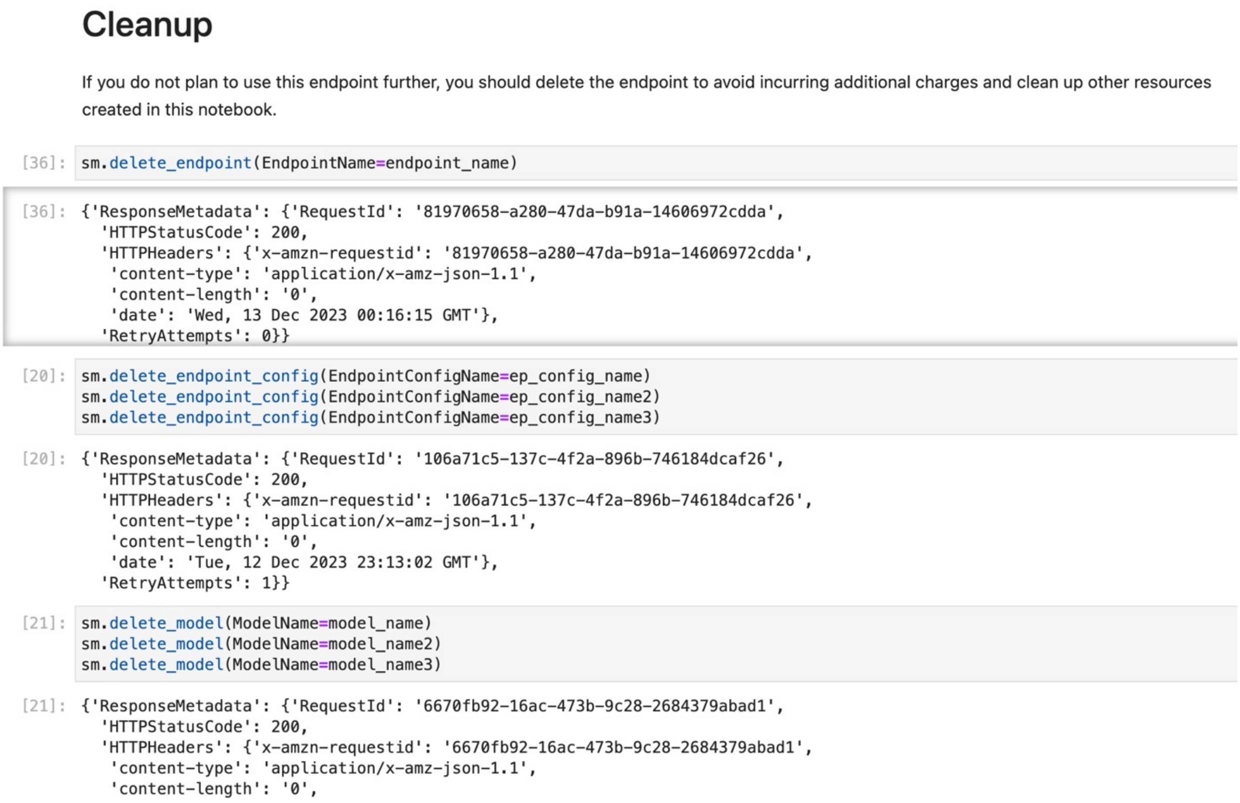
```
[28]: invoke_endpoint(endpoint_name, max_invocations=500)

Sending test traffic to the endpoint DEMO-Deployment-Guardrails-Canary-2023-12-12-23-48-09.
Please wait...
.....................................................................................................................
.....................................................................................................................
................................................................................
Done!
```

Wait for the update operation to complete:



```
[35]: #wait_for_endpoint_in_service(endpoint_name)

      sm.describe_endpoint(EndpointName=endpoint_name)

[35]: {'EndpointName': 'DEMO-Deployment-Guardrails-Canary-2023-12-12-23-48-09',
       'EndpointArn': 'arn:aws:sagemaker:us-east-1:040700907151:endpoint/demo-deployment-guardrails-cana
       'EndpointConfigName': 'DEMO-EpConfig-3-2023-12-12-23-44-25',
       'ProductionVariants': [{'VariantName': 'AllTraffic',
         'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-
           'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:0d09
      d982805093463d40f30212b8050486f18',
           'ResolutionTime': datetime.datetime(2023, 12, 13, 0, 4, 49, 592000, tzinfo=tzlocal())}],
         'CurrentWeight': 1.0,
         'DesiredWeight': 1.0,
```

CODE SHOWING SUCCESSFUL DEPLOYMENT TO MODEL WITH CONFIG3- BEST MODEL



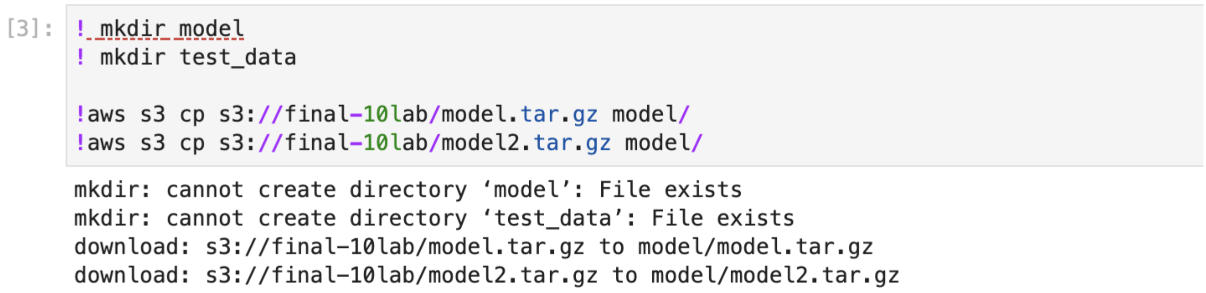
Cleanup

If you do not plan to use this endpoint further, you should delete the endpoint to avoid incurring additional charges and clean up other resources created in this notebook.

```
[36]: sm.delete_endpoint(EndpointName=endpoint_name)

[36]: {'ResponseMetadata': {'RequestId': '81970658-a280-47da-b91a-14606972cdda',
        'HTTPStatusCode': 200,
        'HTTPHeaders': {'x-amzn-requestid': '81970658-a280-47da-b91a-14606972cdda',
         'content-type': 'application/x-amz-json-1.1',
         'content-length': '0',
         'date': 'Wed, 13 Dec 2023 00:16:15 GMT'},
        'RetryAttempts': 0}}

[20]: sm.delete_endpoint_config(EndpointConfigName=ep_config_name)
      sm.delete_endpoint_config(EndpointConfigName=ep_config_name2)
      sm.delete_endpoint_config(EndpointConfigName=ep_config_name3)

[20]: {'ResponseMetadata': {'RequestId': '106a71c5-137c-4f2a-896b-746184dcaf26',
        'HTTPStatusCode': 200,
        'HTTPHeaders': {'x-amzn-requestid': '106a71c5-137c-4f2a-896b-746184dcaf26',
         'content-type': 'application/x-amz-json-1.1',
         'content-length': '0',
         'date': 'Tue, 12 Dec 2023 23:13:02 GMT'},
        'RetryAttempts': 1}}

[21]: sm.delete_model(ModelName=model_name)
      sm.delete_model(ModelName=model_name2)
      sm.delete_model(ModelName=model_name3)

[21]: {'ResponseMetadata': {'RequestId': '6670fb92-16ac-473b-9c28-2684379abad1',
        'HTTPStatusCode': 200,
        'HTTPHeaders': {'x-amzn-requestid': '6670fb92-16ac-473b-9c28-2684379abad1',
         'content-type': 'application/x-amz-json-1.1',
         'content-length': '0',
```

DELETING ENDPOINT,ENDPOINT CONFIGURATIONS AND MODELS.

SHADOW TESTING

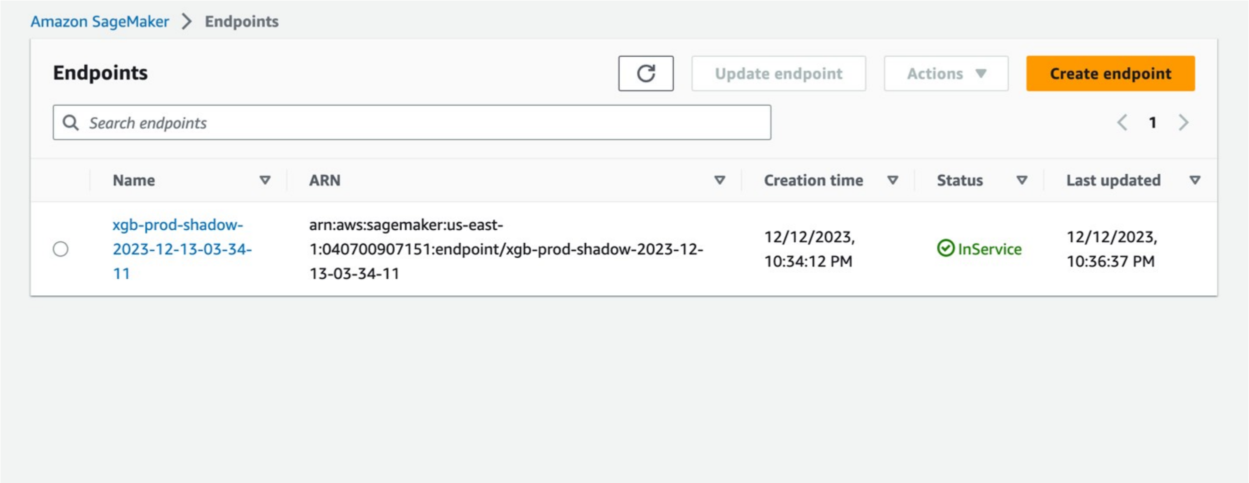


```
[3]: ! mkdir model
     ! mkdir test_data

     !aws s3 cp s3://final-10lab/model.tar.gz model/
     !aws s3 cp s3://final-10lab/model2.tar.gz model/
```

```
mkdir: cannot create directory ‘model’: File exists
mkdir: cannot create directory ‘test_data’: File exists
download: s3://final-10lab/model.tar.gz to model/model.tar.gz
download: s3://final-10lab/model2.tar.gz to model/model2.tar.gz
```

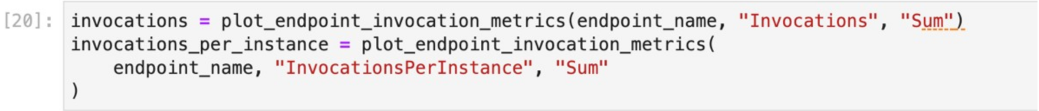
SAME MODELS USED AS THE GUARDRAIL

AN ENDPOINT CONFIGURATION WAS CREATED WITH SHADOW AND TEST VARIANTS,THEN AN ENDPOINT WAS CREATED,WHICH WAS THEN INVOKED.

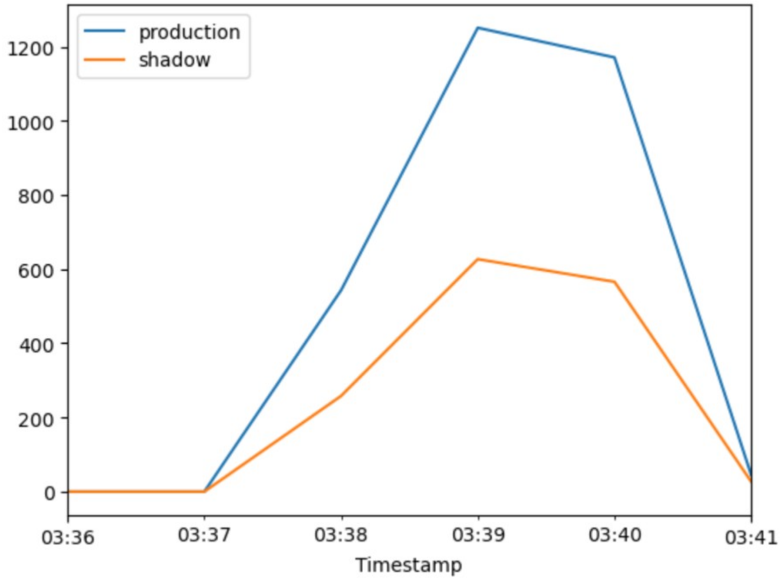


	Name	ARN	Creation time	Status	Last updated
○	xgb-prod-shadow-2023-12-13-03-34-11	arn:aws:sagemaker:us-east-1:040700907151:endpoint/xgb-prod-shadow-2023-12-13-03-34-11	12/12/2023, 10:34:12 PM	✓ InService	12/12/2023, 10:36:37 PM

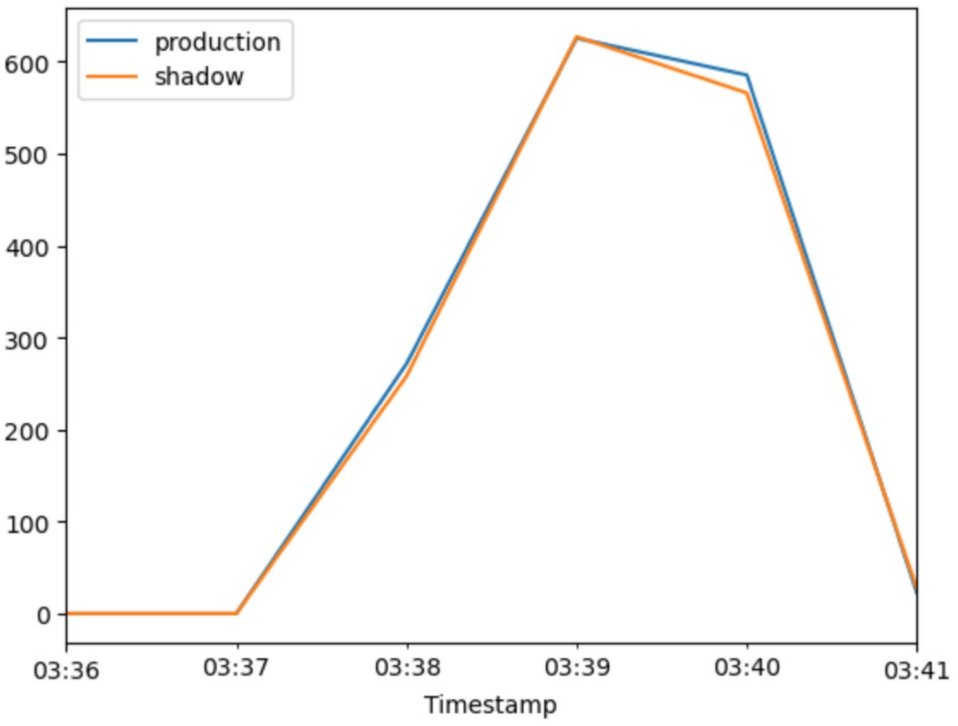
COMPARING INVOCATIONS BETWEEN PRODUCTION AND SHADOW VARIANTS

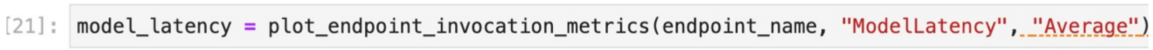


```
[20]: invocations = plot_endpoint_invocation_metrics(endpoint_name, "Invocations", "Sum")
      invocations_per_instance = plot_endpoint_invocation_metrics(
          endpoint_name, "InvocationsPerInstance", "Sum"
      )
```

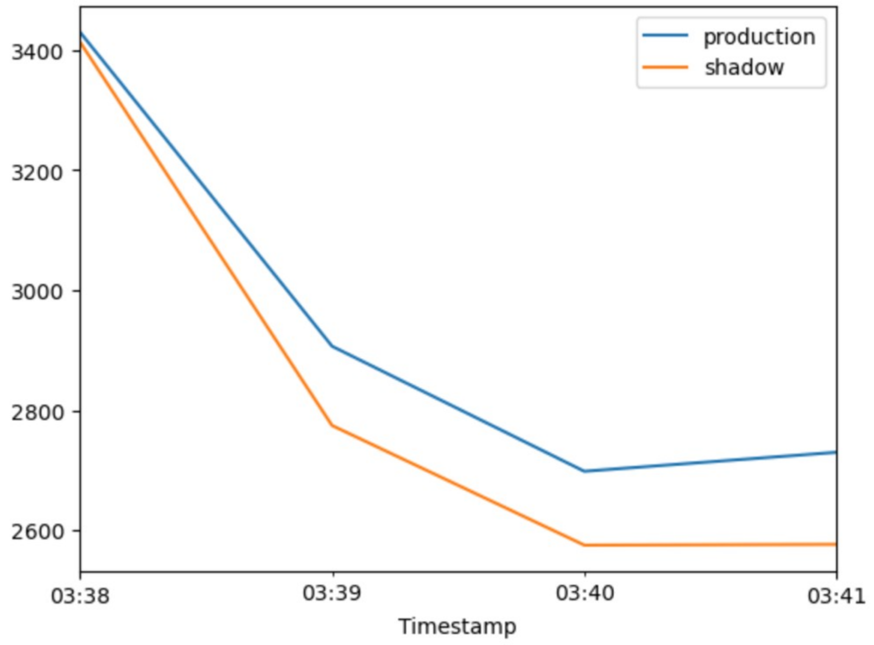


COMPARING INVOCATIONS BETWEEN PRODUCTION AND SHADOW VARIANTS

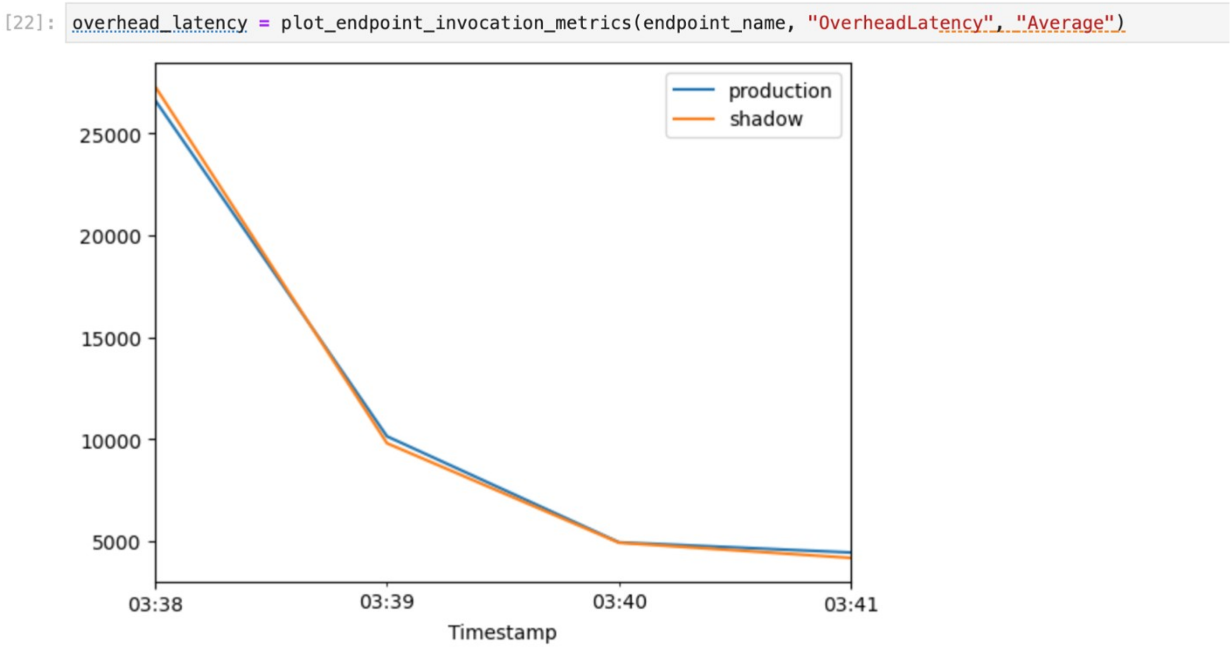




```
[21]: model_latency = plot_endpoint_invocation_metrics(endpoint_name, "ModelLatency", "Average")
```

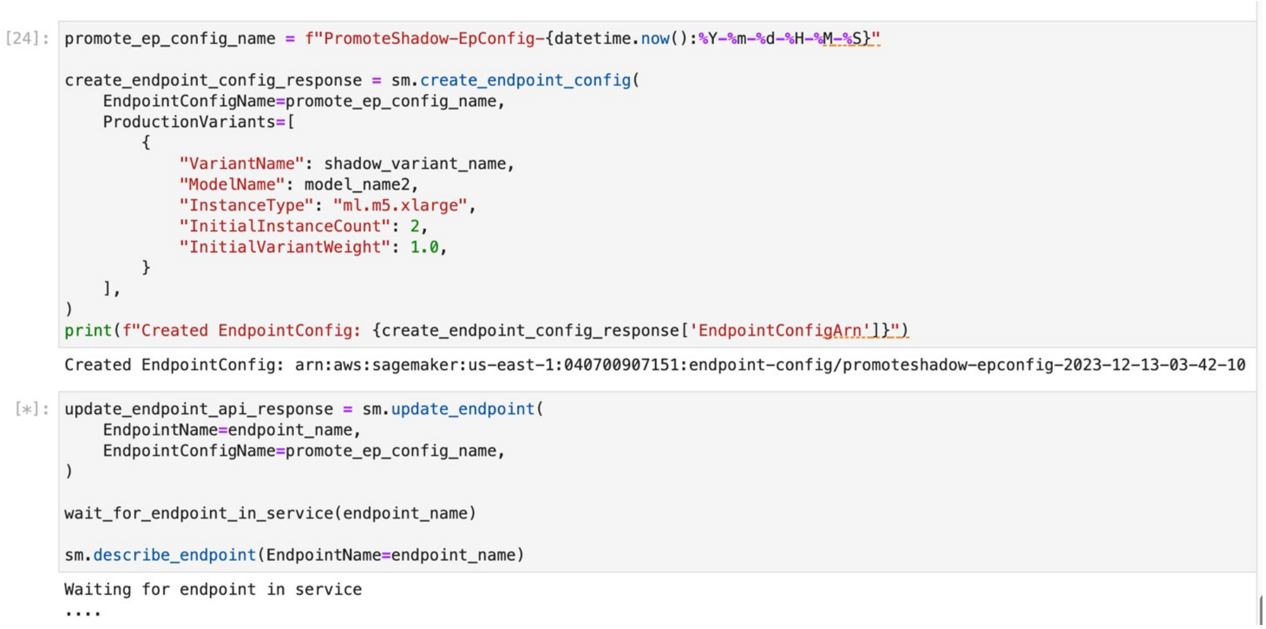


COMPARING INVOKATIONS BETWEEN PRODUCTION AND SHADOW VARIANTS



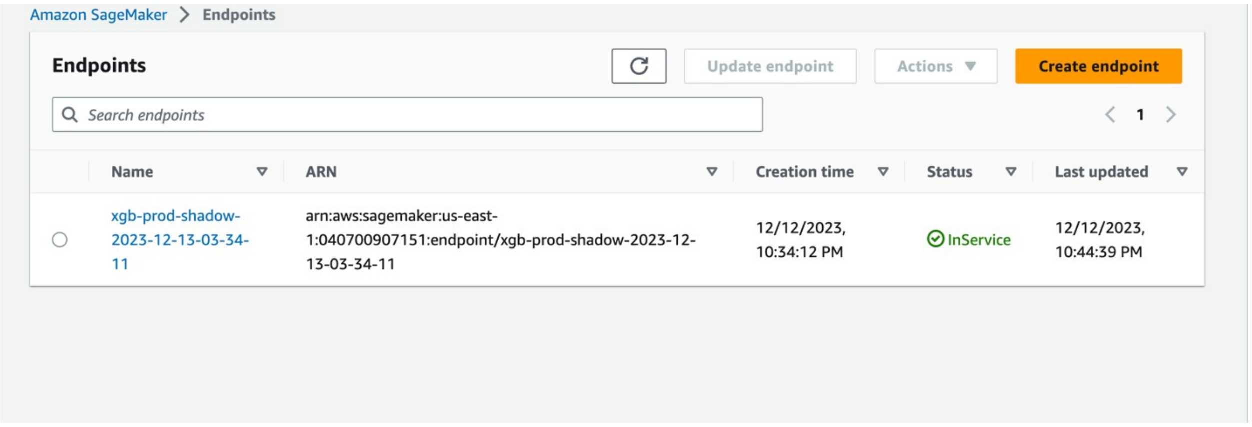
```
[22]: overhead_latency = plot_endpoint_invocation_metrics(endpoint_name, "OverheadLatency", "Average")
```

THIS SHOWS THE ENDPOINT BEING CREATED AND IN SERVICE

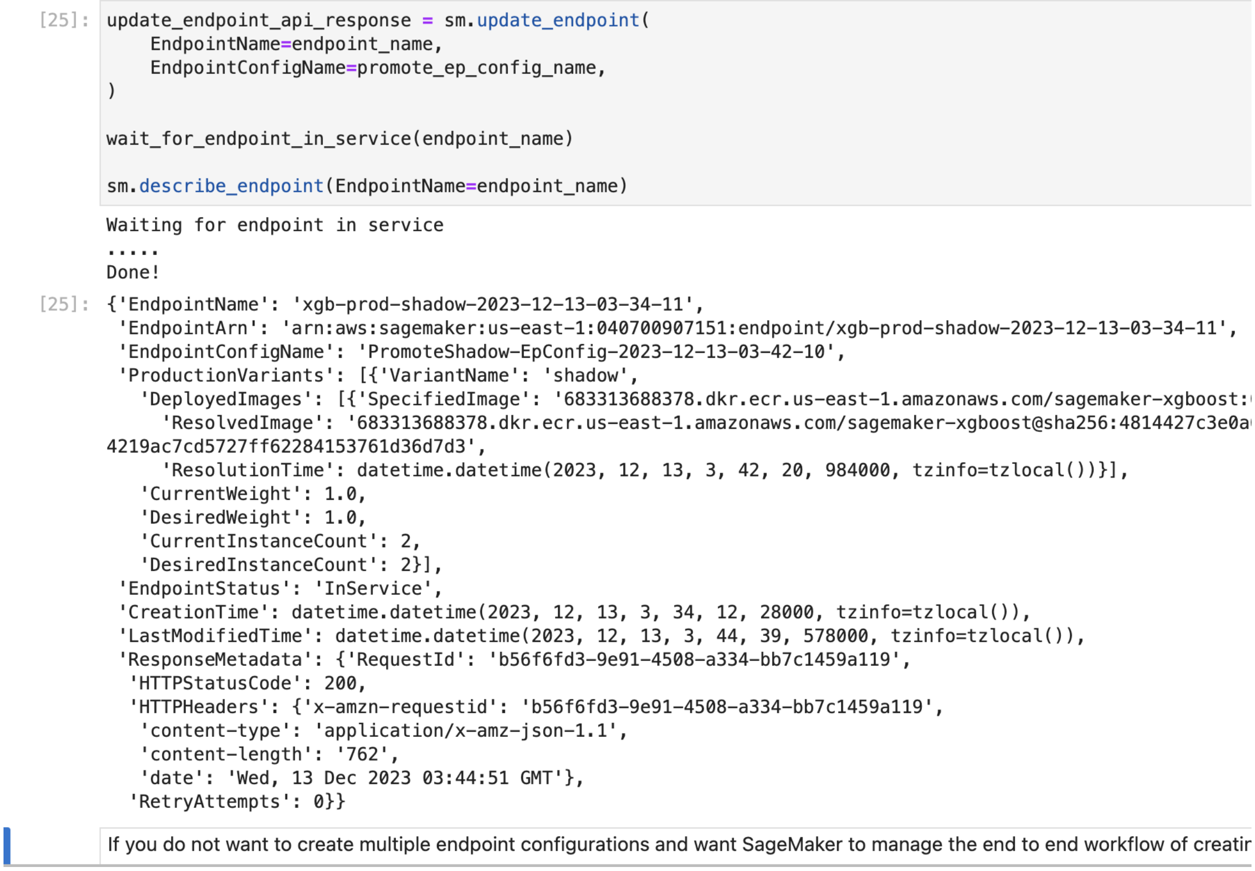


```
[24]: promote_ep_config_name = f"PromoteShadow-EpConfig-{datetime.now():%Y-%m-%d-%H-%M-%S}"

      create_endpoint_config_response = sm.create_endpoint_config(
          EndpointConfigName=promote_ep_config_name,
          ProductionVariants=[
              {
                  "VariantName": shadow_variant_name,
                  "ModelName": model_name2,
                  "InstanceType": "ml.m5.xlarge",
                  "InitialInstanceCount": 2,
                  "InitialVariantWeight": 1.0,
              }
          ],
      )
      print(f"Created EndpointConfig: {create_endpoint_config_response['EndpointConfigArn']}")

      Created EndpointConfig: arn:aws:sagemaker:us-east-1:040700907151:endpoint-config/promoteshadow-epconfig-2023-12-13-03-42-10

[*]:  update_endpoint_api_response = sm.update_endpoint(
          EndpointName=endpoint_name,
          EndpointConfigName=promote_ep_config_name,
      )

      wait_for_endpoint_in_service(endpoint_name)

      sm.describe_endpoint(EndpointName=endpoint_name)

      Waiting for endpoint in service
      ....
```

THIS SHOWS THE ENDPOINT BEING CREATED AND IN SERVICE



	Name	ARN	Creation time	Status	Last updated
○	xgb-prod-shadow-2023-12-13-03-34-11	arn:aws:sagemaker:us-east-1:040700907151:endpoint/xgb-prod-shadow-2023-12-13-03-34-11	12/12/2023, 10:34:12 PM	✓ InService	12/12/2023, 10:44:39 PM

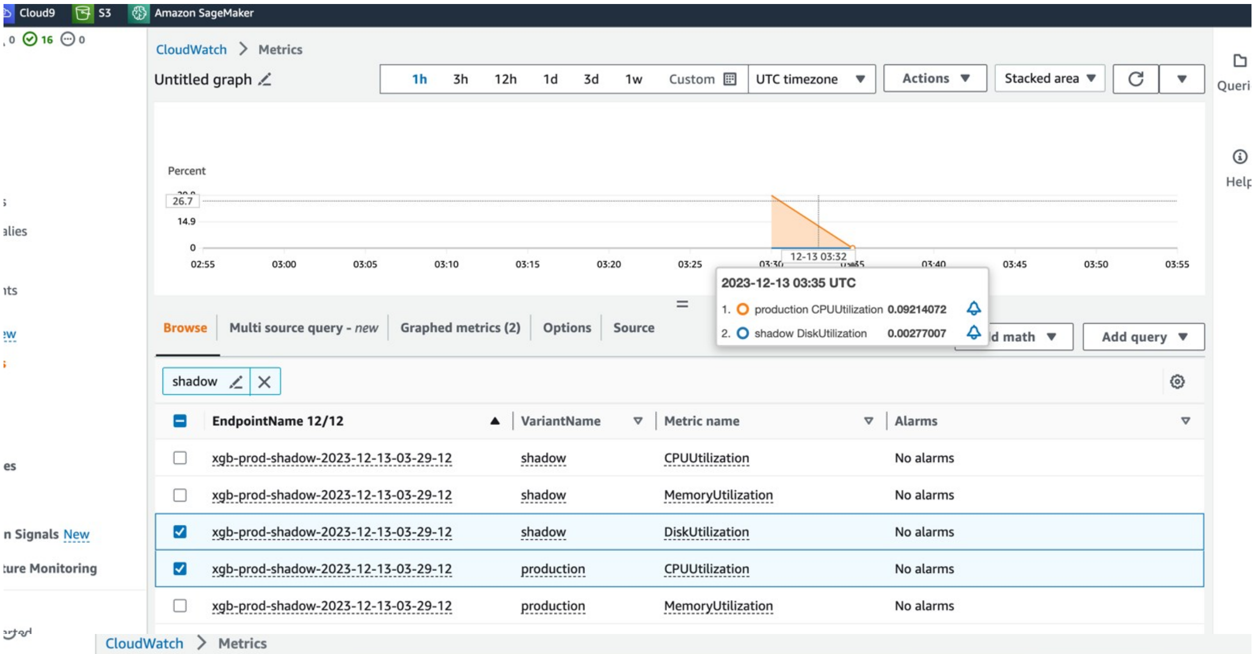


```
[25]: update_endpoint_api_response = sm.update_endpoint(
          EndpointName=endpoint_name,
          EndpointConfigName=promote_ep_config_name,
      )

      wait_for_endpoint_in_service(endpoint_name)

      sm.describe_endpoint(EndpointName=endpoint_name)
```

```
      Waiting for endpoint in service
      .....
      Done!
[25]: {'EndpointName': 'xgb-prod-shadow-2023-12-13-03-34-11',
       'EndpointArn': 'arn:aws:sagemaker:us-east-1:040700907151:endpoint/xgb-prod-shadow-2023-12-13-03-34-11',
       'EndpointConfigName': 'PromoteShadow-EpConfig-2023-12-13-03-42-10',
       'ProductionVariants': [{'VariantName': 'shadow',
         'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:
           'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:4814427c3e0a
      4219ac7cd5727ff62284153761d36d7d3',
           'ResolutionTime': datetime.datetime(2023, 12, 13, 3, 42, 20, 984000, tzinfo=tzlocal())}],
         'CurrentWeight': 1.0,
         'DesiredWeight': 1.0,
         'CurrentInstanceCount': 2,
         'DesiredInstanceCount': 2}],
       'EndpointStatus': 'InService',
       'CreationTime': datetime.datetime(2023, 12, 13, 3, 34, 12, 28000, tzinfo=tzlocal()),
       'LastModifiedTime': datetime.datetime(2023, 12, 13, 3, 44, 39, 578000, tzinfo=tzlocal()),
       'ResponseMetadata': {'RequestId': 'b56f6fd3-9e91-4508-a334-bb7c1459a119',
        'HTTPStatusCode': 200,
        'HTTPHeaders': {'x-amzn-requestid': 'b56f6fd3-9e91-4508-a334-bb7c1459a119',
         'content-type': 'application/x-amz-json-1.1',
         'content-length': '762',
         'date': 'Wed, 13 Dec 2023 03:44:51 GMT'},
        'RetryAttempts': 0}}
```

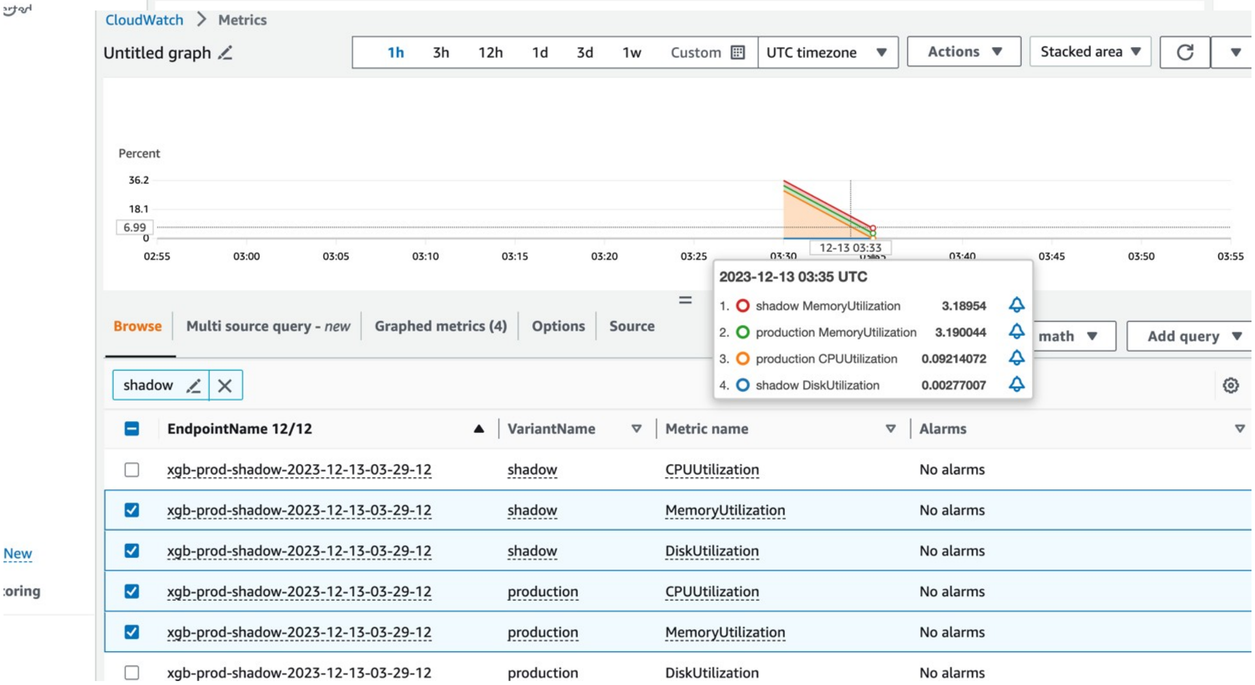
If you do not want to create multiple endpoint configurations and want SageMaker to manage the end to end workflow of creatir

THIS SHOWS THAT THE SHADOW VARIANT IS BETTER THAN THE PRODUCTION VARIANT WHEN COMPARED IN TERMS OF MEMORY UTILIZATION AND CPU UTILIZATION.



2023-12-13 03:35 UTC
1. production CPUUtilization 0.09214072
2. shadow DiskUtilization 0.00277007

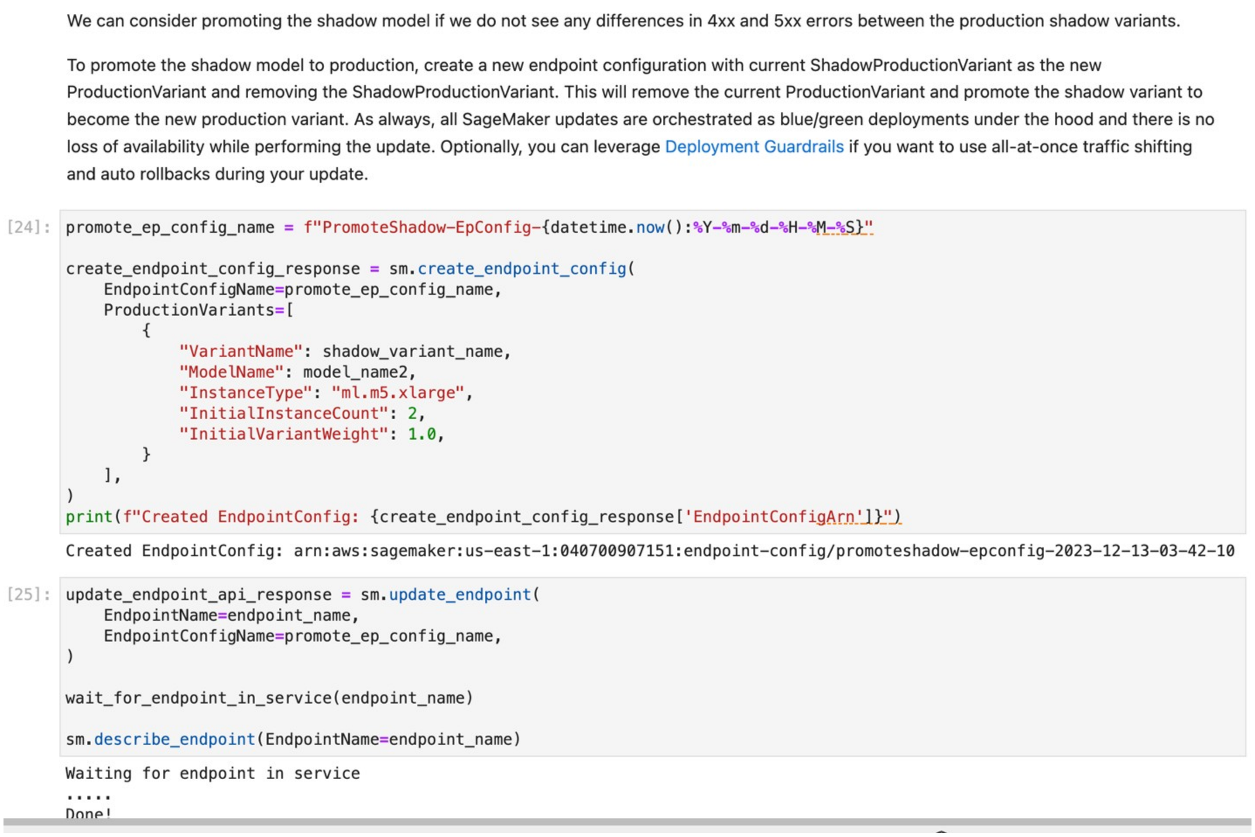
☐	EndpointName 12/12	VariantName	Metric name	Alarms
☐	xgb-prod-shadow-2023-12-13-03-29-12	shadow	CPUUtilization	No alarms
☐	xgb-prod-shadow-2023-12-13-03-29-12	shadow	MemoryUtilization	No alarms
☑	xgb-prod-shadow-2023-12-13-03-29-12	shadow	DiskUtilization	No alarms
☑	xgb-prod-shadow-2023-12-13-03-29-12	production	CPUUtilization	No alarms
☐	xgb-prod-shadow-2023-12-13-03-29-12	production	MemoryUtilization	No alarms



2023-12-13 03:35 UTC
1. shadow MemoryUtilization 3.18954
2. production MemoryUtilization 3.190044
3. production CPUUtilization 0.09214072
4. shadow DiskUtilization 0.00277007

☐	EndpointName 12/12	VariantName	Metric name	Alarms
☐	xgb-prod-shadow-2023-12-13-03-29-12	shadow	CPUUtilization	No alarms
☑	xgb-prod-shadow-2023-12-13-03-29-12	shadow	MemoryUtilization	No alarms
☑	xgb-prod-shadow-2023-12-13-03-29-12	shadow	DiskUtilization	No alarms
☑	xgb-prod-shadow-2023-12-13-03-29-12	production	CPUUtilization	No alarms
☑	xgb-prod-shadow-2023-12-13-03-29-12	production	MemoryUtilization	No alarms
☐	xgb-prod-shadow-2023-12-13-03-29-12	production	DiskUtilization	No alarms

THE SHADOW VARIANT LATER REPLACES THE PRODUCTION VARIANT



We can consider promoting the shadow model if we do not see any differences in 4xx and 5xx errors between the production shadow variants.

To promote the shadow model to production, create a new endpoint configuration with current ShadowProductionVariant as the new ProductionVariant and removing the ShadowProductionVariant. This will remove the current ProductionVariant and promote the shadow variant to become the new production variant. As always, all SageMaker updates are orchestrated as blue/green deployments under the hood and there is no loss of availability while performing the update. Optionally, you can leverage Deployment Guardrails if you want to use all-at-once traffic shifting and auto rollbacks during your update.

[24]:

```
promote_ep_config_name = f"PromoteShadow-EpConfig-{datetime.now():%Y-%m-%d-%H-%M-%S}"

create_endpoint_config_response = sm.create_endpoint_config(
    EndpointConfigName=promote_ep_config_name,
    ProductionVariants=[
        {
            "VariantName": shadow_variant_name,
            "ModelName": model_name2,
            "InstanceType": "ml.m5.xlarge",
            "InitialInstanceCount": 2,
            "InitialVariantWeight": 1.0,
        }
    ],
)
print(f"Created EndpointConfig: {create_endpoint_config_response['EndpointConfigArn']}")
```

```
Created EndpointConfig: arn:aws:sagemaker:us-east-1:040700907151:endpoint-config/promoteshadow-epconfig-2023-12-13-03-42-10
```

[25]:

```
update_endpoint_api_response = sm.update_endpoint(
    EndpointName=endpoint_name,
    EndpointConfigName=promote_ep_config_name,
)

wait_for_endpoint_in_service(endpoint_name)

sm.describe_endpoint(EndpointName=endpoint_name)
```

```
Waiting for endpoint in service
.....
Done!
```