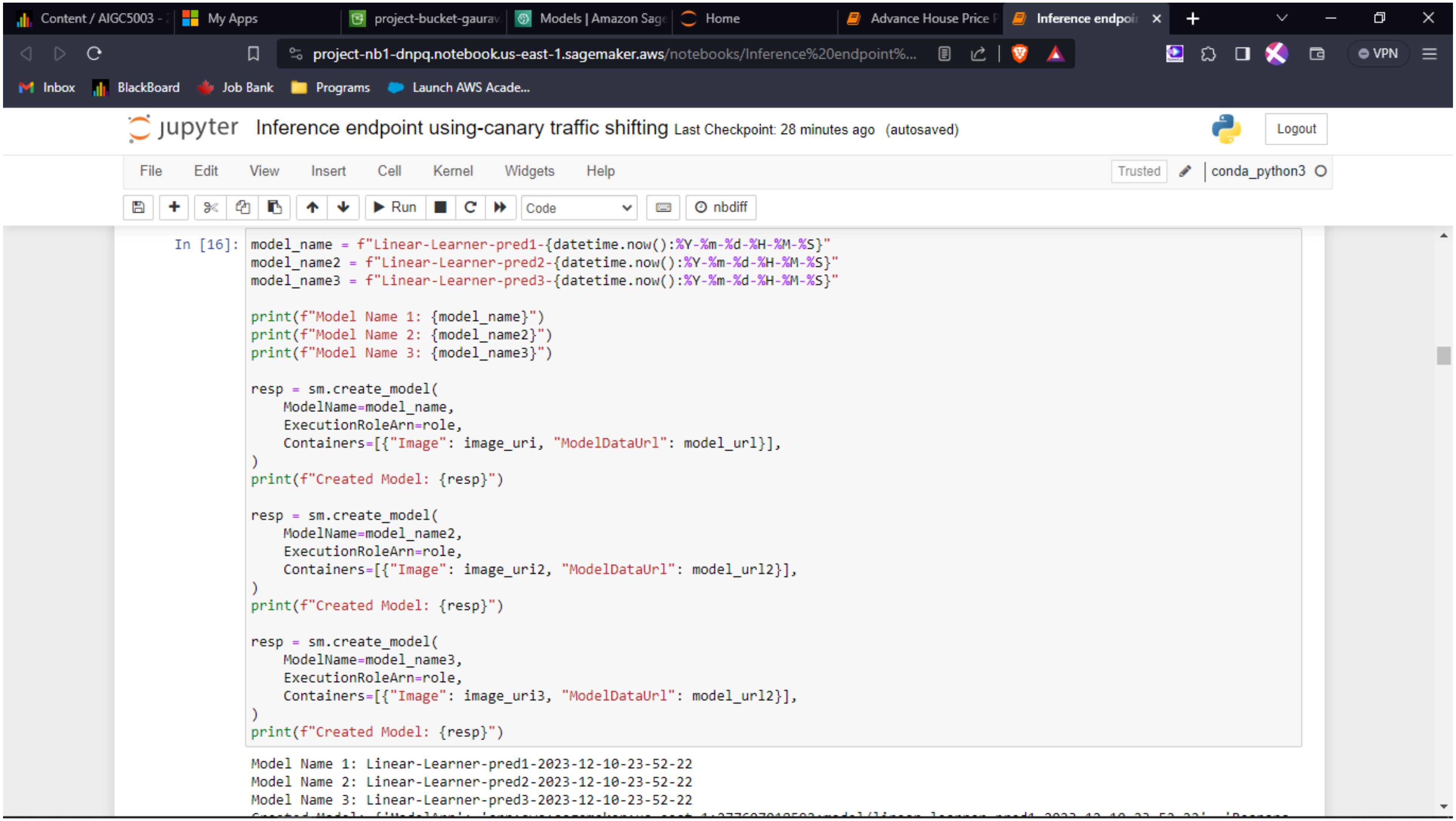
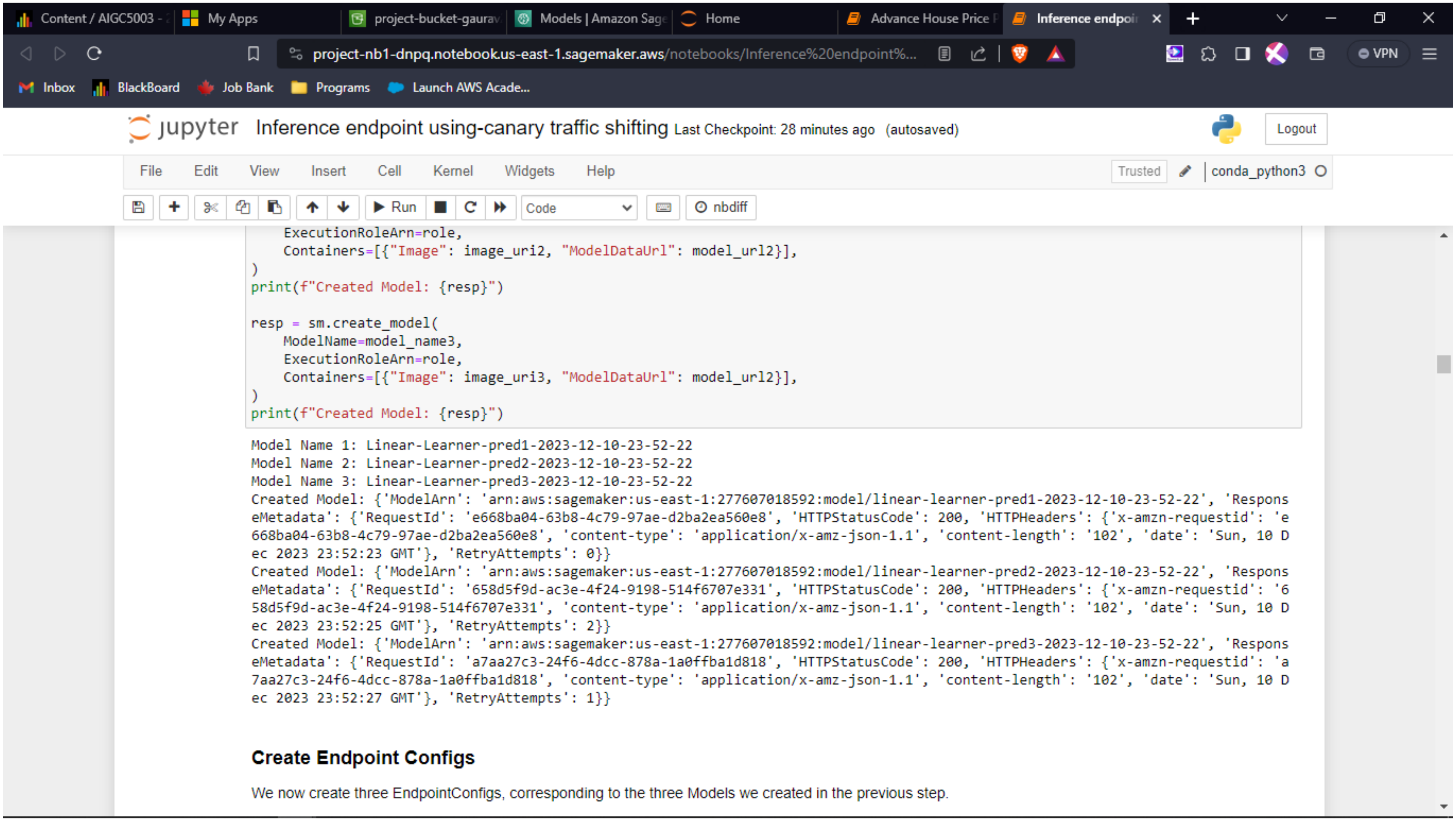
GuardRail



```
In [16]:
model_name = f"Linear-Learner-pred1-{datetime.now():%Y-%m-%d-%H-%M-%S}"
model_name2 = f"Linear-Learner-pred2-{datetime.now():%Y-%m-%d-%H-%M-%S}"
model_name3 = f"Linear-Learner-pred3-{datetime.now():%Y-%m-%d-%H-%M-%S}"

print(f"Model Name 1: {model_name}")
print(f"Model Name 2: {model_name2}")
print(f"Model Name 3: {model_name3}")

resp = sm.create_model(
    ModelName=model_name,
    ExecutionRoleArn=role,
    Containers=[{"Image": image_uri, "ModelDataUrl": model_url}],
)
print(f"Created Model: {resp}")

resp = sm.create_model(
    ModelName=model_name2,
    ExecutionRoleArn=role,
    Containers=[{"Image": image_uri2, "ModelDataUrl": model_url2}],
)
print(f"Created Model: {resp}")

resp = sm.create_model(
    ModelName=model_name3,
    ExecutionRoleArn=role,
    Containers=[{"Image": image_uri3, "ModelDataUrl": model_url2}],
)
print(f"Created Model: {resp}")
```

```
Model Name 1: Linear-Learner-pred1-2023-12-10-23-52-22
Model Name 2: Linear-Learner-pred2-2023-12-10-23-52-22
Model Name 3: Linear-Learner-pred3-2023-12-10-23-52-22
Created Model: {'ModelArn': 'arn:aws:sagemaker:us-east-1:277607910502:model/linear-learner-pred1-2023-12-10-23-52-22', 'Respons
```

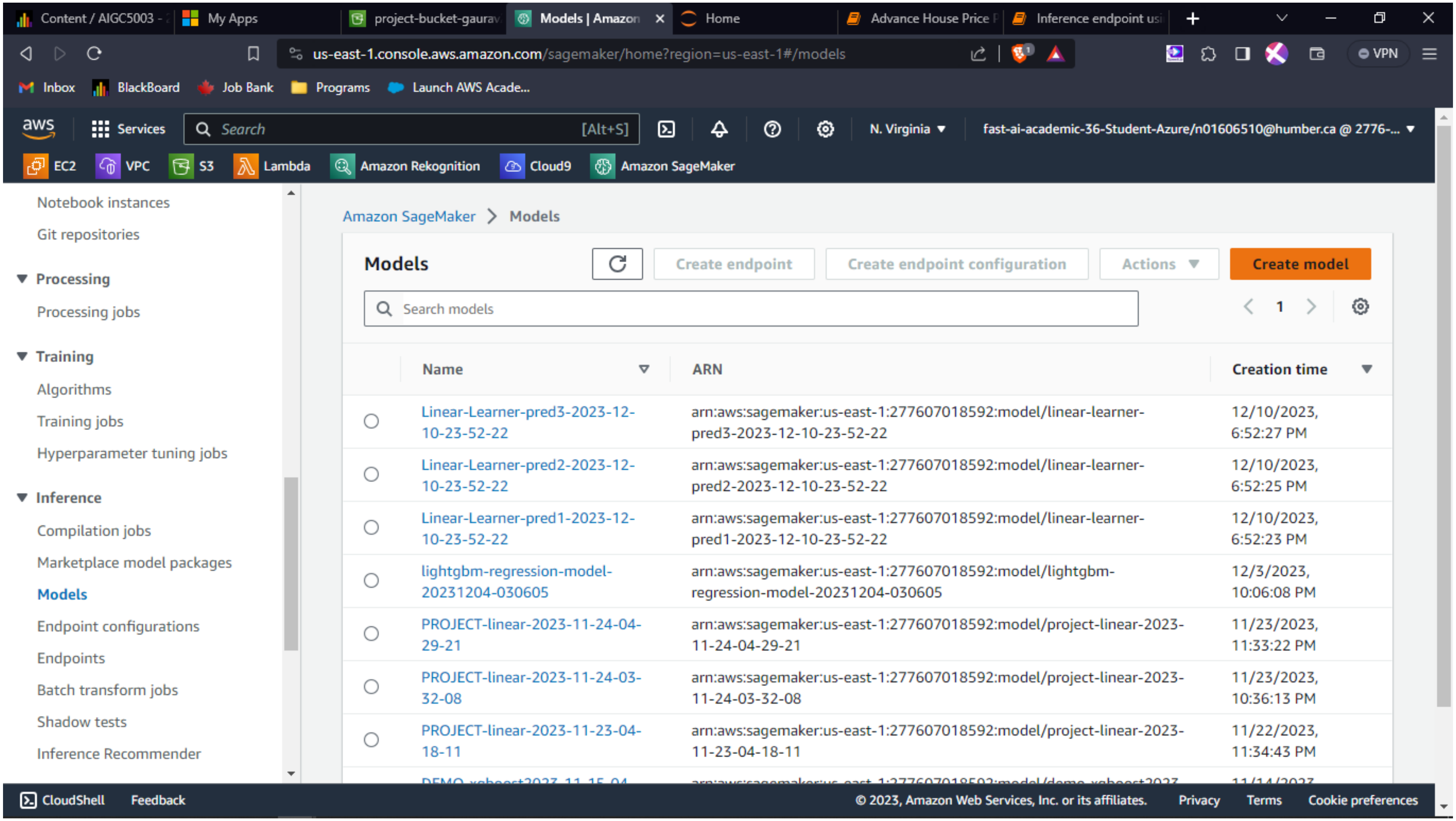


```
    ExecutionRoleArn=role,
    Containers=[{"Image": image_uri2, "ModelDataUrl": model_url2}],
)
print(f"Created Model: {resp}")

resp = sm.create_model(
    ModelName=model_name3,
    ExecutionRoleArn=role,
    Containers=[{"Image": image_uri3, "ModelDataUrl": model_url2}],
)
print(f"Created Model: {resp}")
```

```
Model Name 1: Linear-Learner-pred1-2023-12-10-23-52-22
Model Name 2: Linear-Learner-pred2-2023-12-10-23-52-22
Model Name 3: Linear-Learner-pred3-2023-12-10-23-52-22
Created Model: {'ModelArn': 'arn:aws:sagemaker:us-east-1:277607018592:model/linear-learner-pred1-2023-12-10-23-52-22', 'Respons
eMetadata': {'RequestId': 'e668ba04-63b8-4c79-97ae-d2ba2ea560e8', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-amzn-requestid': 'e
668ba04-63b8-4c79-97ae-d2ba2ea560e8', 'content-type': 'application/x-amz-json-1.1', 'content-length': '102', 'date': 'Sun, 10 D
ec 2023 23:52:23 GMT'}, 'RetryAttempts': 0}}
Created Model: {'ModelArn': 'arn:aws:sagemaker:us-east-1:277607018592:model/linear-learner-pred2-2023-12-10-23-52-22', 'Respons
eMetadata': {'RequestId': '658d5f9d-ac3e-4f24-9198-514f6707e331', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-amzn-requestid': '6
58d5f9d-ac3e-4f24-9198-514f6707e331', 'content-type': 'application/x-amz-json-1.1', 'content-length': '102', 'date': 'Sun, 10 D
ec 2023 23:52:25 GMT'}, 'RetryAttempts': 2}}
Created Model: {'ModelArn': 'arn:aws:sagemaker:us-east-1:277607018592:model/linear-learner-pred3-2023-12-10-23-52-22', 'Respons
eMetadata': {'RequestId': 'a7aa27c3-24f6-4dcc-878a-1a0ffba1d818', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-amzn-requestid': 'a
7aa27c3-24f6-4dcc-878a-1a0ffba1d818', 'content-type': 'application/x-amz-json-1.1', 'content-length': '102', 'date': 'Sun, 10 D
ec 2023 23:52:27 GMT'}, 'RetryAttempts': 1}}
```

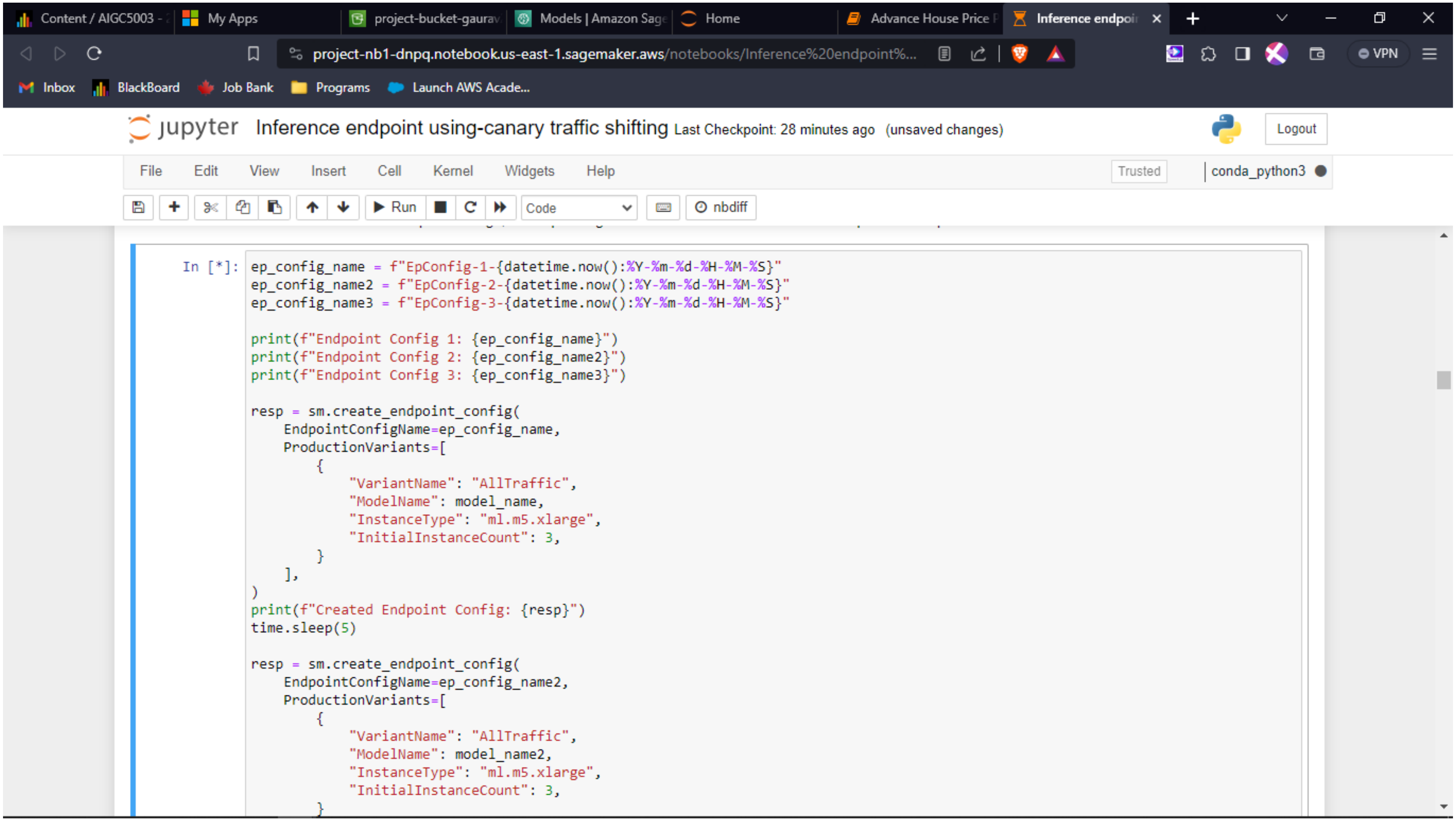
Create Endpoint Configs

We now create three EndpointConfigs, corresponding to the three Models we created in the previous step.

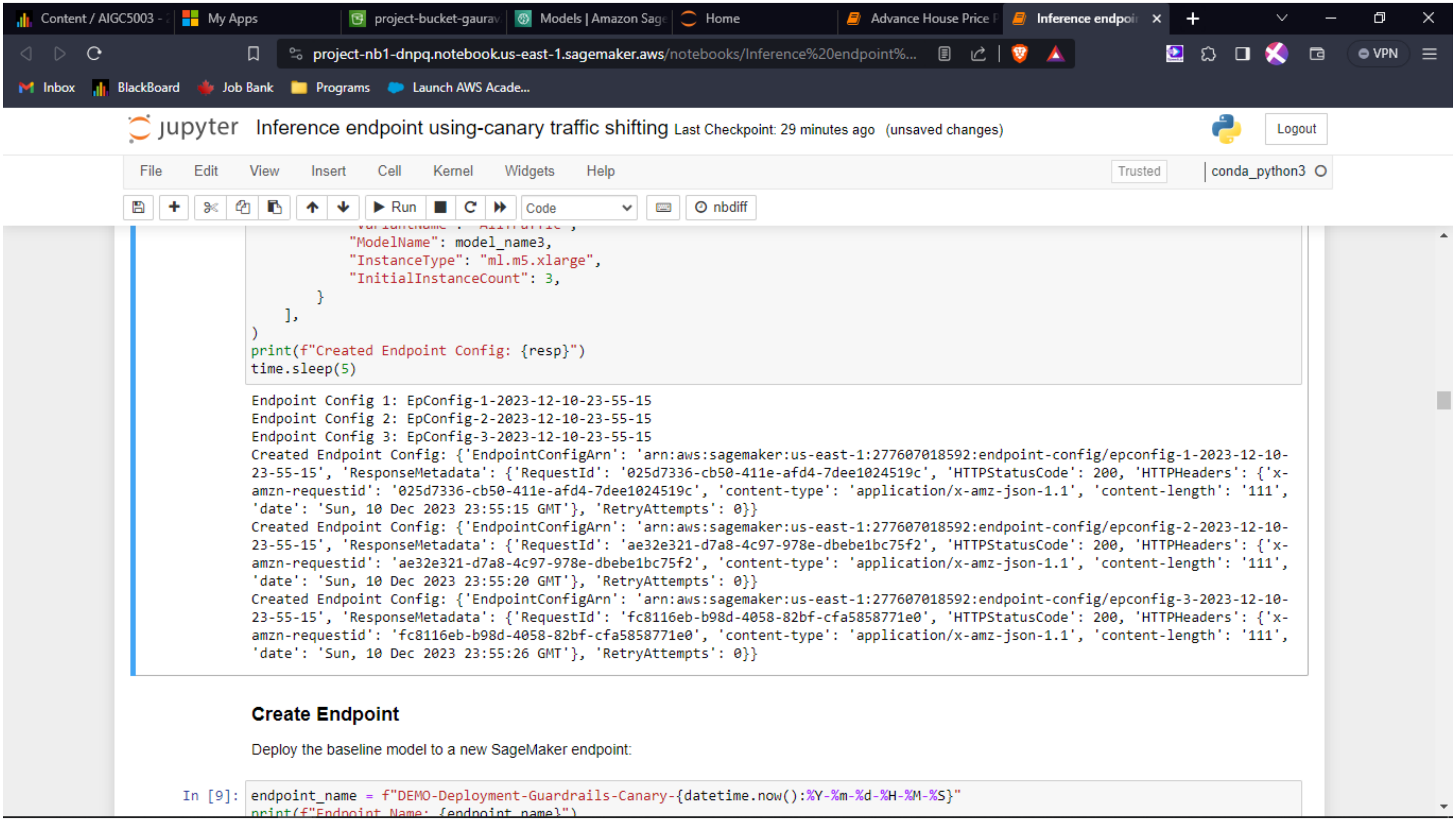


Models

	Name	ARN	Creation time
○	Linear-Learner-pred3-2023-12-10-23-52-22	arn:aws:sagemaker:us-east-1:277607018592:model/linear-learner-pred3-2023-12-10-23-52-22	12/10/2023, 6:52:27 PM
○	Linear-Learner-pred2-2023-12-10-23-52-22	arn:aws:sagemaker:us-east-1:277607018592:model/linear-learner-pred2-2023-12-10-23-52-22	12/10/2023, 6:52:25 PM
○	Linear-Learner-pred1-2023-12-10-23-52-22	arn:aws:sagemaker:us-east-1:277607018592:model/linear-learner-pred1-2023-12-10-23-52-22	12/10/2023, 6:52:23 PM
○	lightgbm-regression-model-20231204-030605	arn:aws:sagemaker:us-east-1:277607018592:model/lightgbm-regression-model-20231204-030605	12/3/2023, 10:06:08 PM
○	PROJECT-linear-2023-11-24-04-29-21	arn:aws:sagemaker:us-east-1:277607018592:model/project-linear-2023-11-24-04-29-21	11/23/2023, 11:33:22 PM
○	PROJECT-linear-2023-11-24-03-32-08	arn:aws:sagemaker:us-east-1:277607018592:model/project-linear-2023-11-24-03-32-08	11/23/2023, 10:36:13 PM
○	PROJECT-linear-2023-11-23-04-18-11	arn:aws:sagemaker:us-east-1:277607018592:model/project-linear-2023-11-23-04-18-11	11/22/2023, 11:34:43 PM
	DEMO-xgboost2023-11-15-04	arn:aws:sagemaker:us-east-1:277607018592:model/demo-xgboost2023-	11/14/2023



```
In [*]: ep_config_name = f"EpConfig-1-{datetime.now():%Y-%m-%d-%H-%M-%S}"
        ep_config_name2 = f"EpConfig-2-{datetime.now():%Y-%m-%d-%H-%M-%S}"
        ep_config_name3 = f"EpConfig-3-{datetime.now():%Y-%m-%d-%H-%M-%S}"

        print(f"Endpoint Config 1: {ep_config_name}")
        print(f"Endpoint Config 2: {ep_config_name2}")
        print(f"Endpoint Config 3: {ep_config_name3}")

        resp = sm.create_endpoint_config(
            EndpointConfigName=ep_config_name,
            ProductionVariants=[
                {
                    "VariantName": "AllTraffic",
                    "ModelName": model_name,
                    "InstanceType": "ml.m5.xlarge",
                    "InitialInstanceCount": 3,
                }
            ],
        )
        print(f"Created Endpoint Config: {resp}")
        time.sleep(5)

        resp = sm.create_endpoint_config(
            EndpointConfigName=ep_config_name2,
            ProductionVariants=[
                {
                    "VariantName": "AllTraffic",
                    "ModelName": model_name2,
                    "InstanceType": "ml.m5.xlarge",
                    "InitialInstanceCount": 3,
                }
```



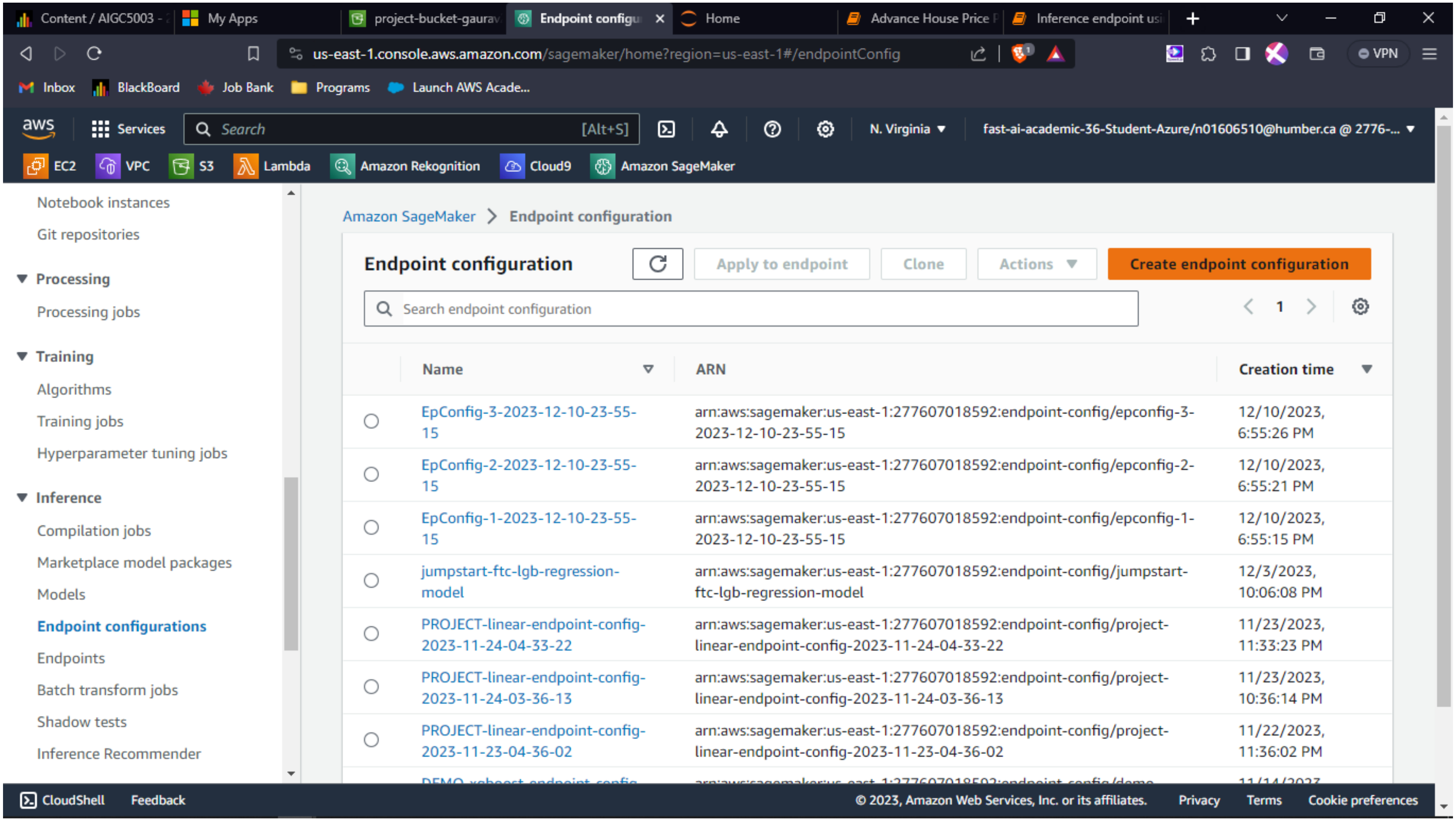
```
                "VariantName": "AllTraffic",
                "ModelName": model_name3,
                "InstanceType": "ml.m5.xlarge",
                "InitialInstanceCount": 3,
            }
        ],
)
print(f"Created Endpoint Config: {resp}")
time.sleep(5)
```

```
Endpoint Config 1: EpConfig-1-2023-12-10-23-55-15
Endpoint Config 2: EpConfig-2-2023-12-10-23-55-15
Endpoint Config 3: EpConfig-3-2023-12-10-23-55-15
Created Endpoint Config: {'EndpointConfigArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/epconfig-1-2023-12-10-
23-55-15', 'ResponseMetadata': {'RequestId': '025d7336-cb50-411e-afd4-7dee1024519c', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-
amzn-requestid': '025d7336-cb50-411e-afd4-7dee1024519c', 'content-type': 'application/x-amz-json-1.1', 'content-length': '111',
'date': 'Sun, 10 Dec 2023 23:55:15 GMT'}, 'RetryAttempts': 0}}
Created Endpoint Config: {'EndpointConfigArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/epconfig-2-2023-12-10-
23-55-15', 'ResponseMetadata': {'RequestId': 'ae32e321-d7a8-4c97-978e-dbebe1bc75f2', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-
amzn-requestid': 'ae32e321-d7a8-4c97-978e-dbebe1bc75f2', 'content-type': 'application/x-amz-json-1.1', 'content-length': '111',
'date': 'Sun, 10 Dec 2023 23:55:20 GMT'}, 'RetryAttempts': 0}}
Created Endpoint Config: {'EndpointConfigArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/epconfig-3-2023-12-10-
23-55-15', 'ResponseMetadata': {'RequestId': 'fc8116eb-b98d-4058-82bf-cfa5858771e0', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-
amzn-requestid': 'fc8116eb-b98d-4058-82bf-cfa5858771e0', 'content-type': 'application/x-amz-json-1.1', 'content-length': '111',
'date': 'Sun, 10 Dec 2023 23:55:26 GMT'}, 'RetryAttempts': 0}}
```

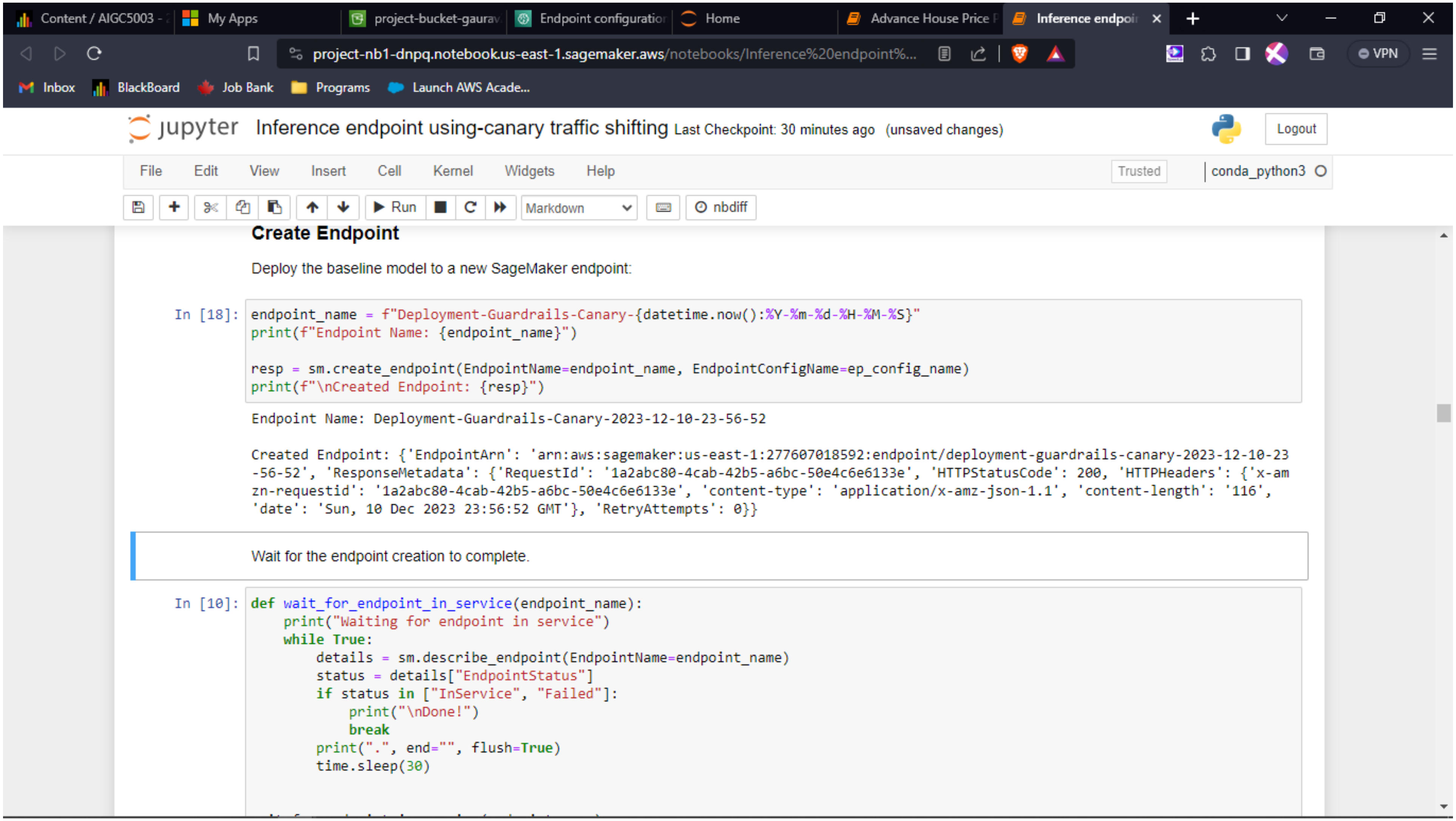
Create Endpoint

Deploy the baseline model to a new SageMaker endpoint:

```
In [9]: endpoint_name = f"DEMO-Deployment-Guardrails-Canary-{datetime.now():%Y-%m-%d-%H-%M-%S}"
        print(f"Endpoint Name: {endpoint_name}")
```



	Name	ARN	Creation time
○	EpConfig-3-2023-12-10-23-55-15	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/epconfig-3-2023-12-10-23-55-15	12/10/2023, 6:55:26 PM
○	EpConfig-2-2023-12-10-23-55-15	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/epconfig-2-2023-12-10-23-55-15	12/10/2023, 6:55:21 PM
○	EpConfig-1-2023-12-10-23-55-15	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/epconfig-1-2023-12-10-23-55-15	12/10/2023, 6:55:15 PM
○	jumpstart-ftc-lgb-regression-model	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/jumpstart-ftc-lgb-regression-model	12/3/2023, 10:06:08 PM
○	PROJECT-linear-endpoint-config-2023-11-24-04-33-22	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/project-linear-endpoint-config-2023-11-24-04-33-22	11/23/2023, 11:33:23 PM
○	PROJECT-linear-endpoint-config-2023-11-24-03-36-13	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/project-linear-endpoint-config-2023-11-24-03-36-13	11/23/2023, 10:36:14 PM
○	PROJECT-linear-endpoint-config-2023-11-23-04-36-02	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/project-linear-endpoint-config-2023-11-23-04-36-02	11/22/2023, 11:36:02 PM
	DEMO-xgboost-endpoint-config	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/demo	11/14/2023



Inference endpoint using-canary traffic shifting Last Checkpoint: 30 minutes ago (unsaved changes)

Create Endpoint

Deploy the baseline model to a new SageMaker endpoint:

In [18]:

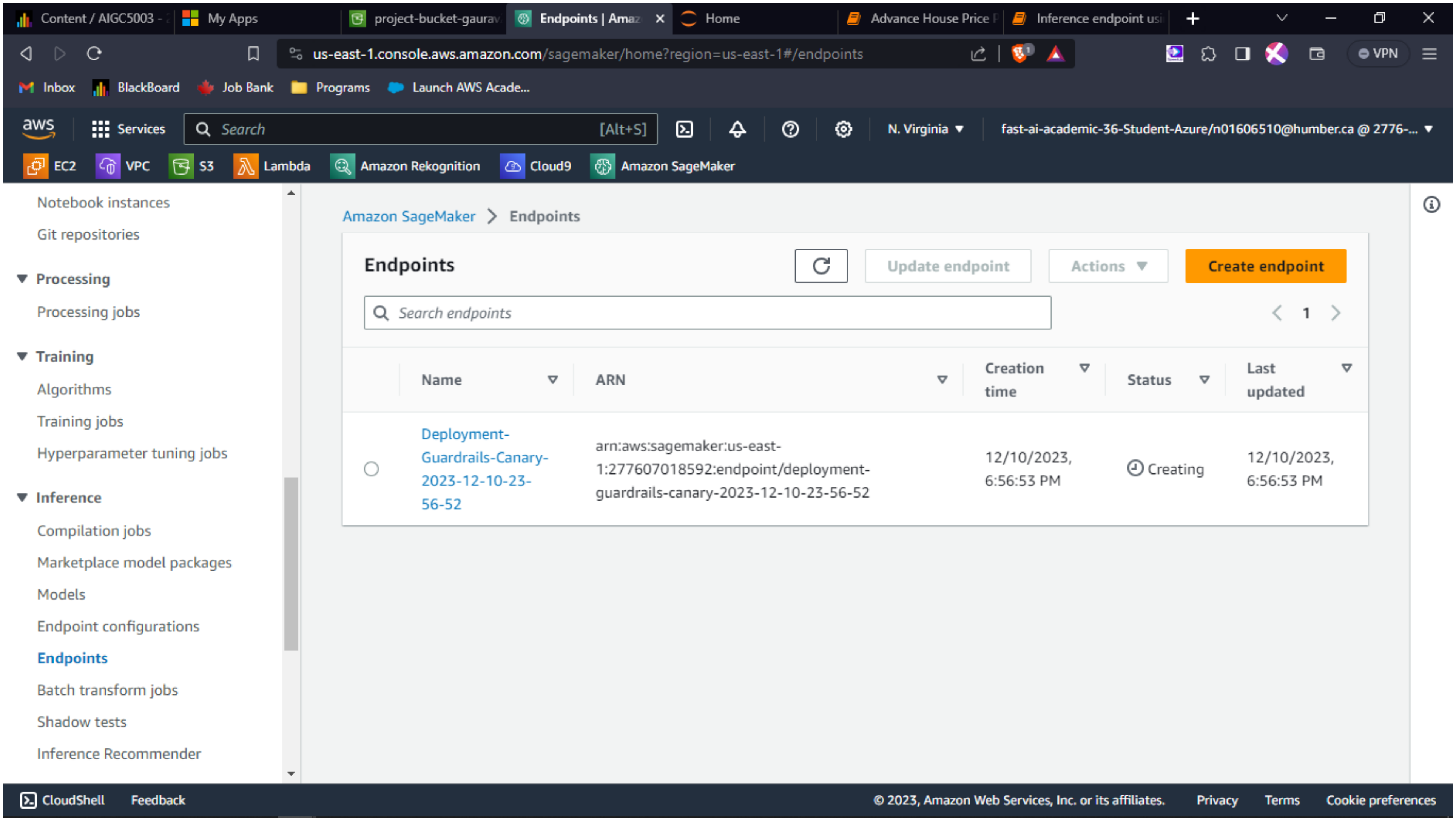
```
endpoint_name = f"Deployment-Guardrails-Canary-{datetime.now():%Y-%m-%d-%H-%M-%S}"
print(f"Endpoint Name: {endpoint_name}")

resp = sm.create_endpoint(EndpointName=endpoint_name, EndpointConfigName=ep_config_name)
print(f"\nCreated Endpoint: {resp}")
```

```
Endpoint Name: Deployment-Guardrails-Canary-2023-12-10-23-56-52

Created Endpoint: {'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23
-56-52', 'ResponseMetadata': {'RequestId': '1a2abc80-4cab-42b5-a6bc-50e4c6e6133e', 'HTTPStatusCode': 200, 'HTTPHeaders': {'x-am
zn-requestid': '1a2abc80-4cab-42b5-a6bc-50e4c6e6133e', 'content-type': 'application/x-amz-json-1.1', 'content-length': '116',
'date': 'Sun, 10 Dec 2023 23:56:52 GMT'}, 'RetryAttempts': 0}}
```

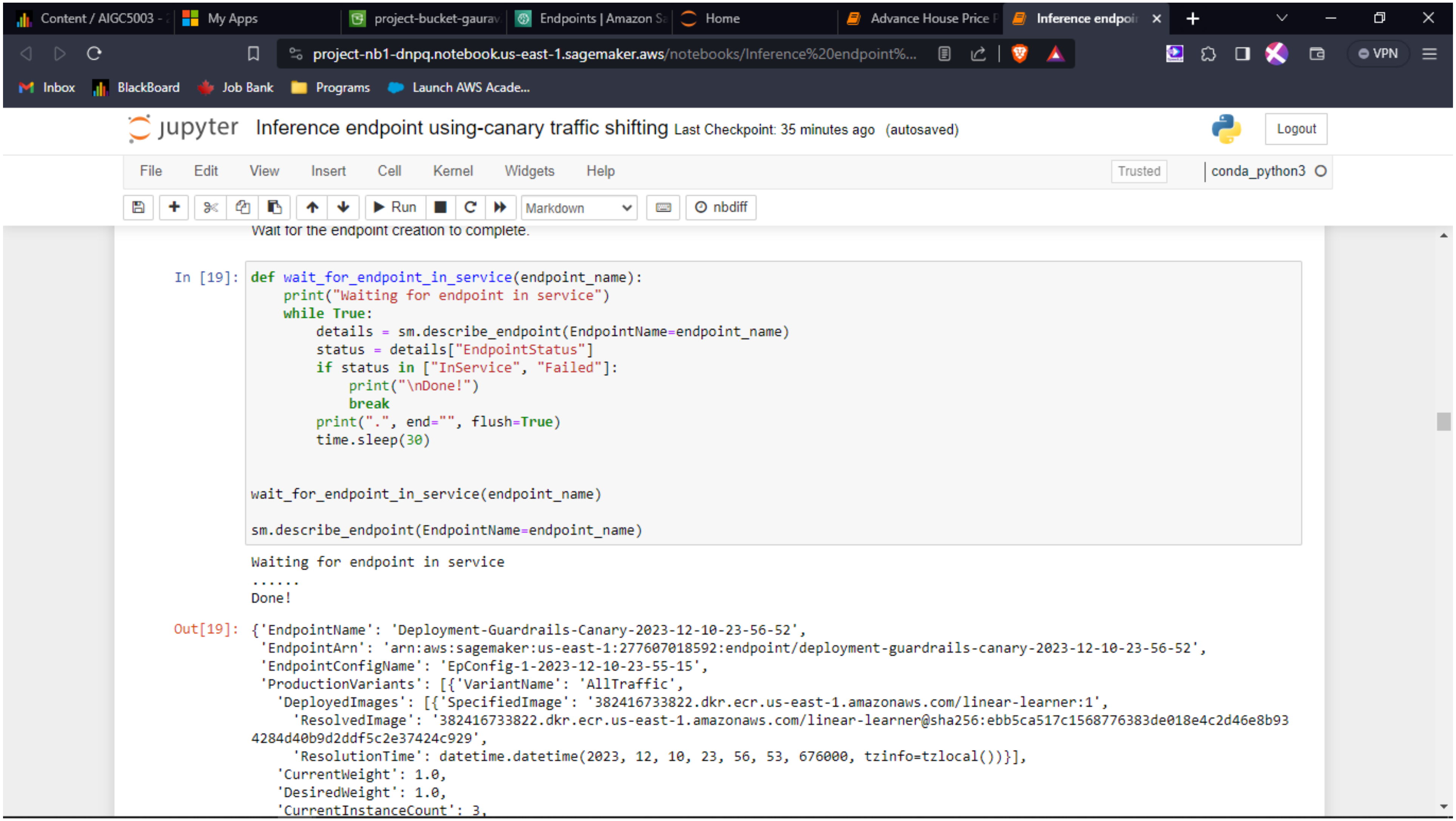
Wait for the endpoint creation to complete.

In [10]:

```
def wait_for_endpoint_in_service(endpoint_name):
    print("Waiting for endpoint in service")
    while True:
        details = sm.describe_endpoint(EndpointName=endpoint_name)
        status = details["EndpointStatus"]
        if status in ["InService", "Failed"]:
            print("\nDone!")
            break
        print(".", end="", flush=True)
        time.sleep(30)
```



	Name	ARN	Creation time	Status	Last updated
○	Deployment-Guardrails-Canary-2023-12-10-23-56-52	arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23-56-52	12/10/2023, 6:56:53 PM	Creating	12/10/2023, 6:56:53 PM



Inference endpoint using-canary traffic shifting Last Checkpoint: 35 minutes ago (autosaved)

Wait for the endpoint creation to complete.

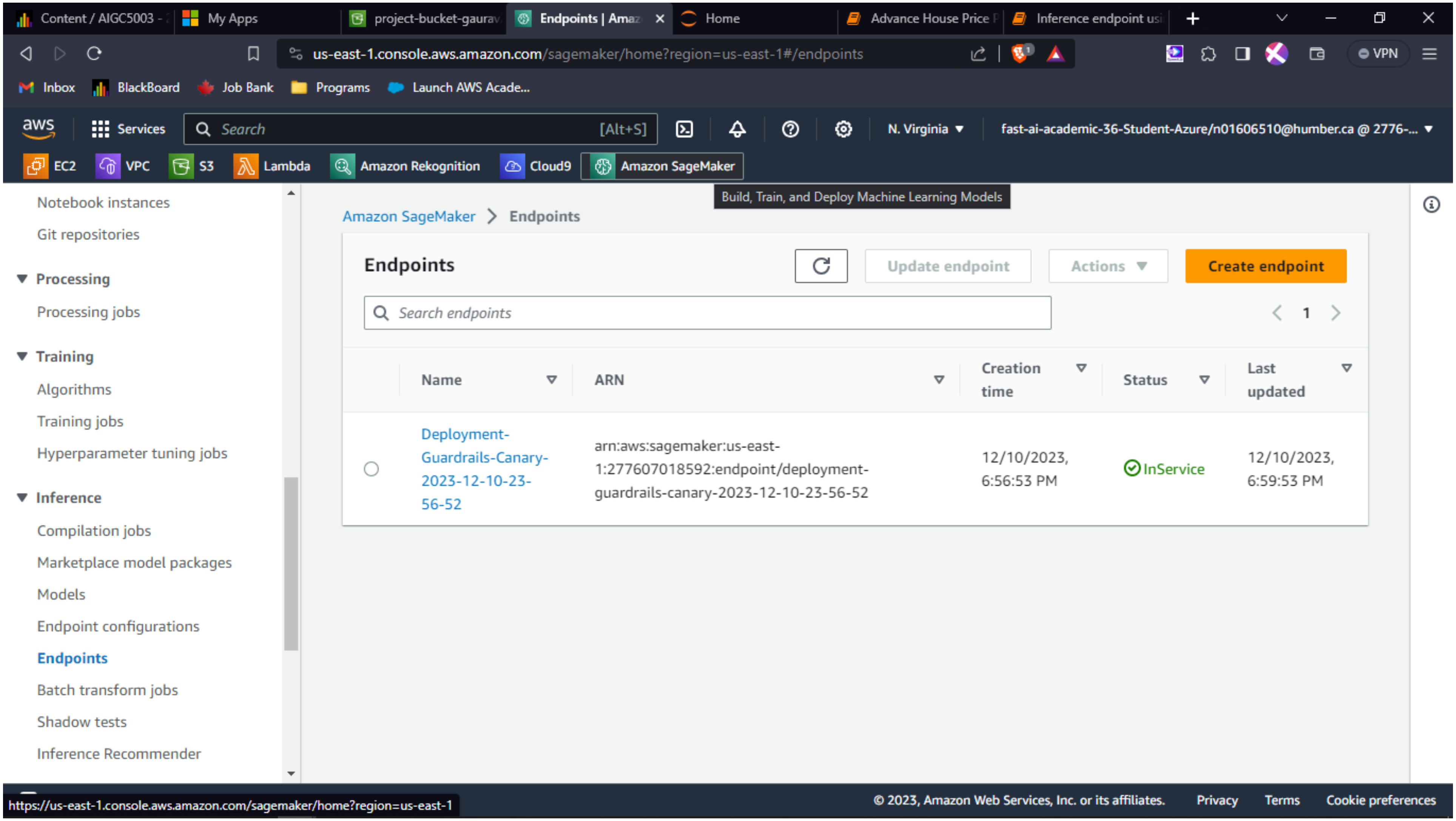
In [19]:

```
def wait_for_endpoint_in_service(endpoint_name):
    print("Waiting for endpoint in service")
    while True:
        details = sm.describe_endpoint(EndpointName=endpoint_name)
        status = details["EndpointStatus"]
        if status in ["InService", "Failed"]:
            print("\nDone!")
            break
        print(".", end="", flush=True)
        time.sleep(30)


wait_for_endpoint_in_service(endpoint_name)

sm.describe_endpoint(EndpointName=endpoint_name)
```

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

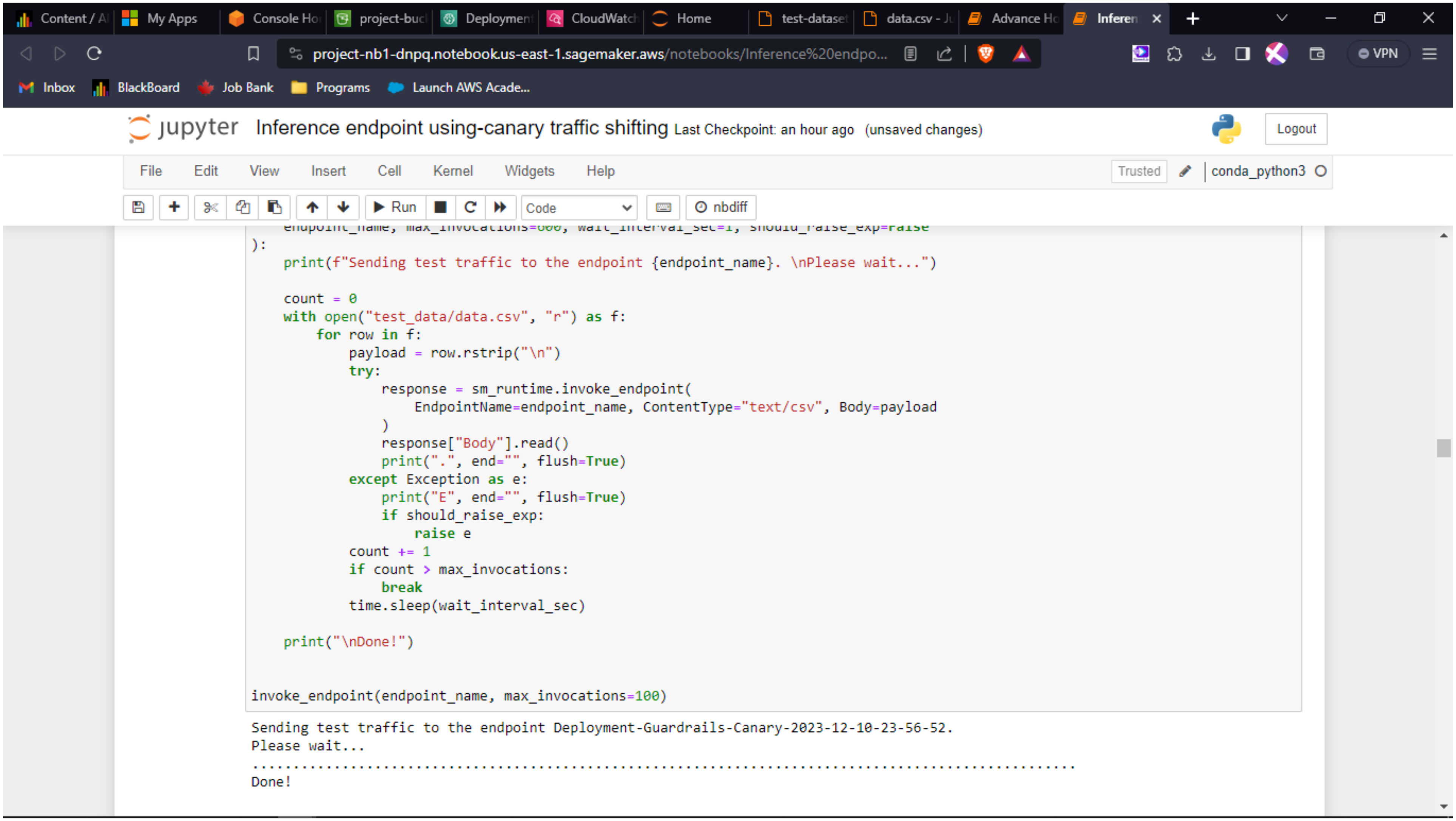
Out[19]:

```
{'EndpointName': 'Deployment-Guardrails-Canary-2023-12-10-23-56-52',
 'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23-56-52',
 'EndpointConfigName': 'EpConfig-1-2023-12-10-23-55-15',
 'ProductionVariants': [{'VariantName': 'AllTraffic',
   'DeployedImages': [{'SpecifiedImage': '382416733822.dkr.ecr.us-east-1.amazonaws.com/linear-learner:1',
     'ResolvedImage': '382416733822.dkr.ecr.us-east-1.amazonaws.com/linear-learner@sha256:ebb5ca517c1568776383de018e4c2d46e8b93
4284d40b9d2ddf5c2e37424c929',
     'ResolutionTime': datetime.datetime(2023, 12, 10, 23, 56, 53, 676000, tzinfo=tzlocal())}],
   'CurrentWeight': 1.0,
   'DesiredWeight': 1.0,
   'CurrentInstanceCount': 3,
```



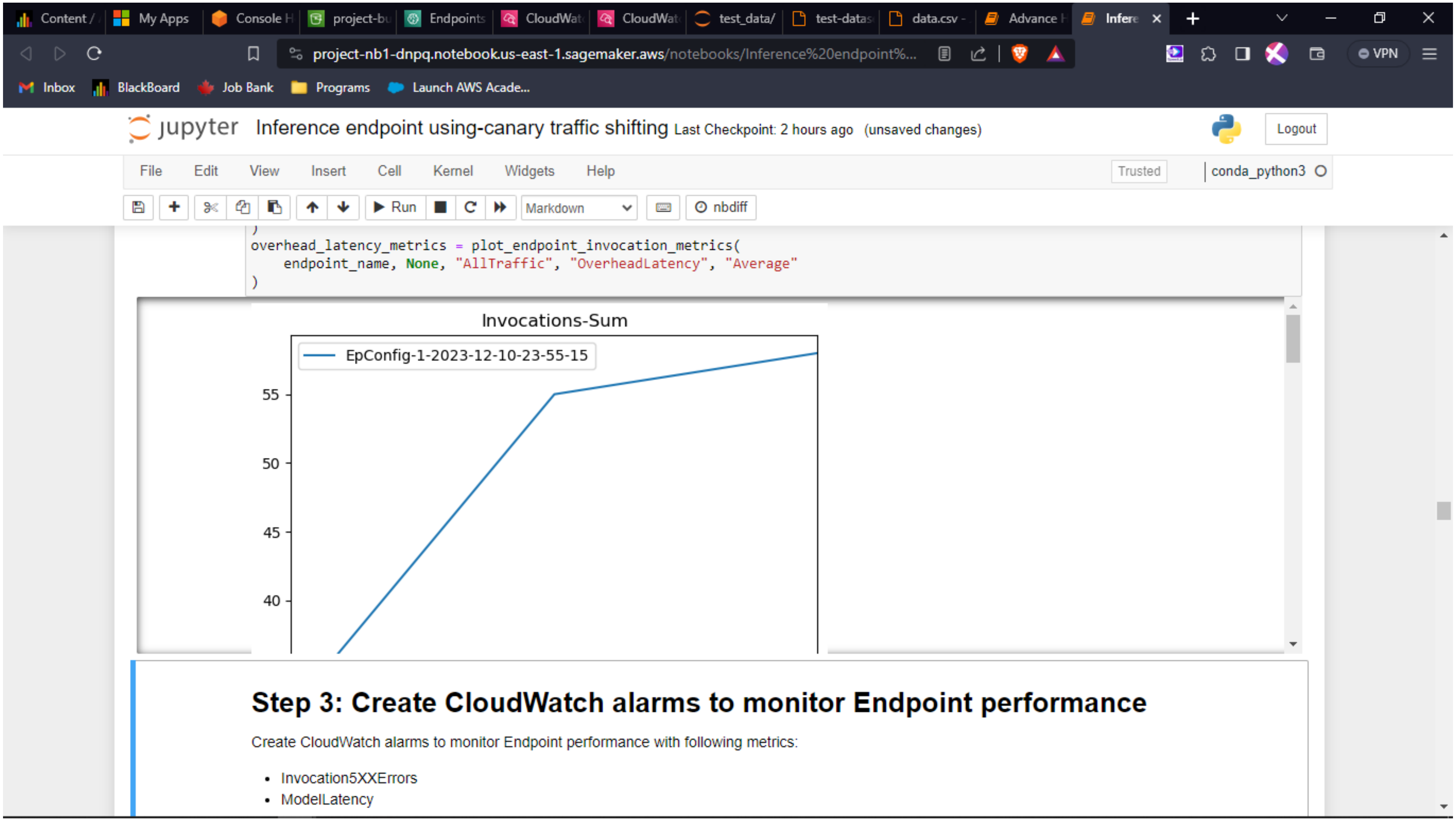
Endpoints

	Name	ARN	Creation time	Status	Last updated
○	Deployment-Guardrails-Canary-2023-12-10-23-56-52	arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23-56-52	12/10/2023, 6:56:53 PM	✓ InService	12/10/2023, 6:59:53 PM



```
    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/data.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 Deployment-Guardrails-Canary-2023-12-10-23-56-52.
Please wait...
.....................................................................................................
Done!
```



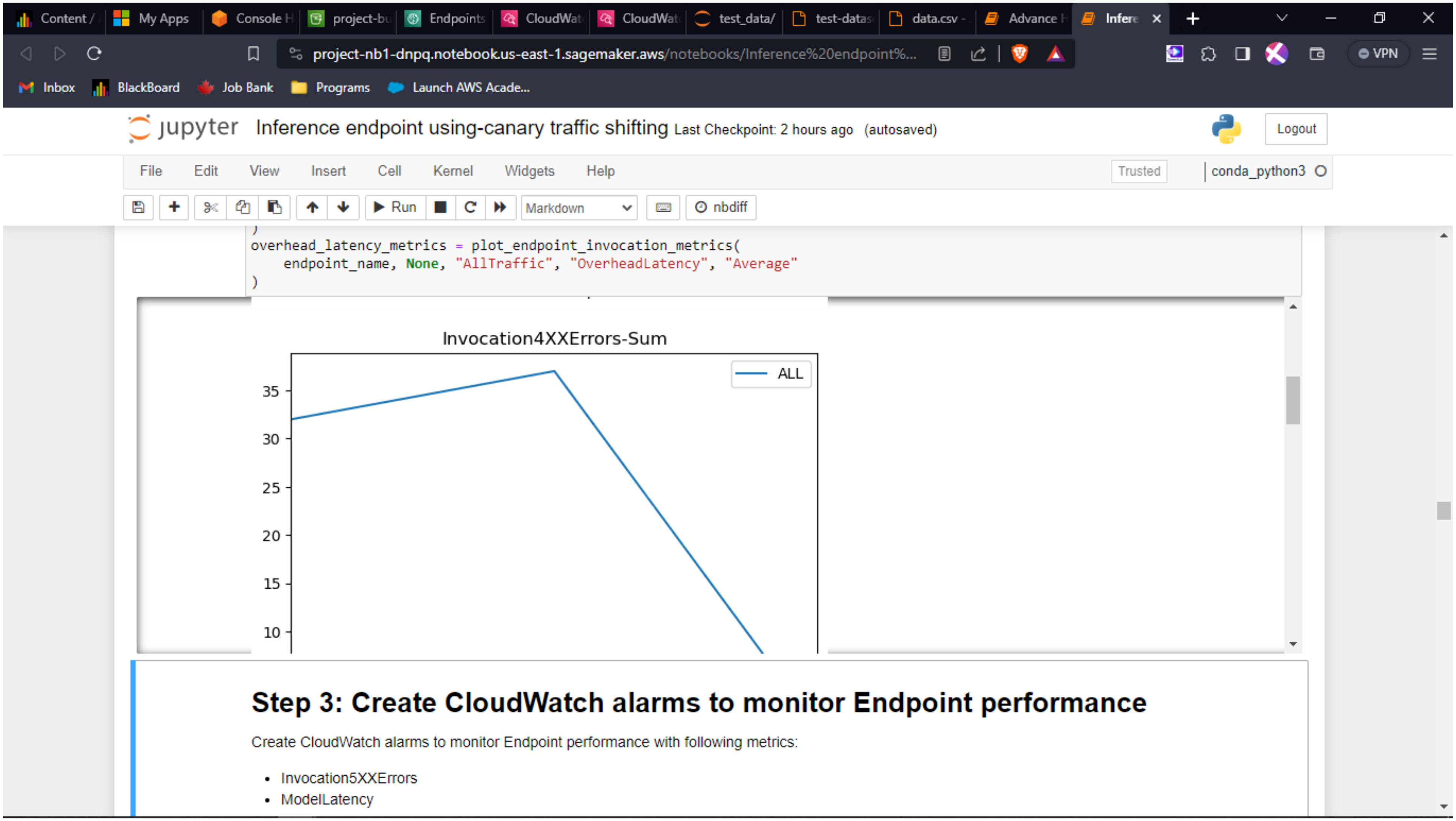
Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (unsaved changes)

```
)
overhead_latency_metrics = plot_endpoint_invocation_metrics(
    endpoint_name, None, "AllTraffic", "OverheadLatency", "Average"
)
```

Step 3: Create CloudWatch alarms to monitor Endpoint performance

Create CloudWatch alarms to monitor Endpoint performance with following metrics:

- Invocation5XXErrors
- ModelLatency



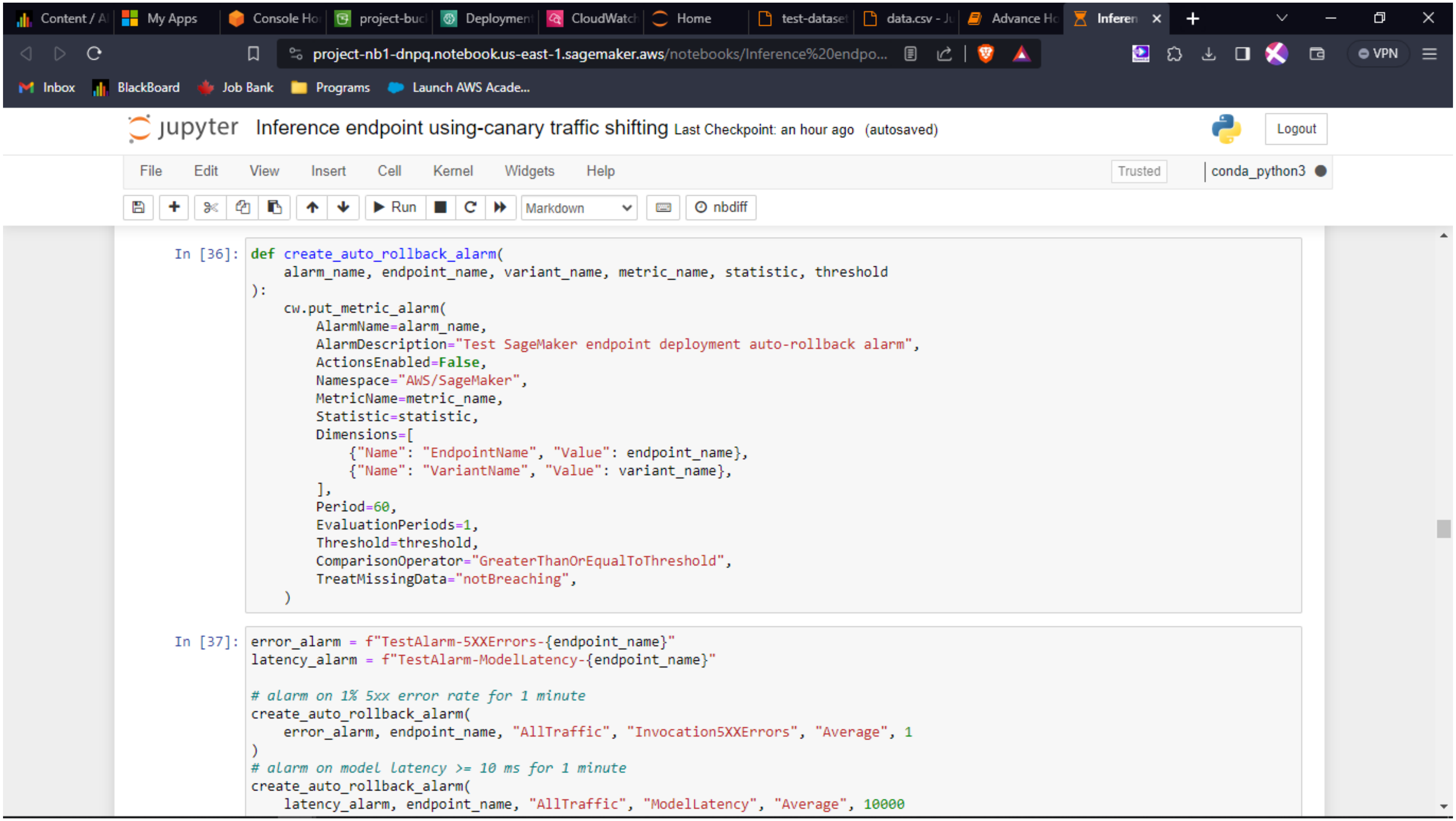
Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (autosaved)

```
)
overhead_latency_metrics = plot_endpoint_invocation_metrics(
    endpoint_name, None, "AllTraffic", "OverheadLatency", "Average"
)
```

Step 3: Create CloudWatch alarms to monitor Endpoint performance

Create CloudWatch alarms to monitor Endpoint performance with following metrics:

- Invocation5XXErrors
- ModelLatency

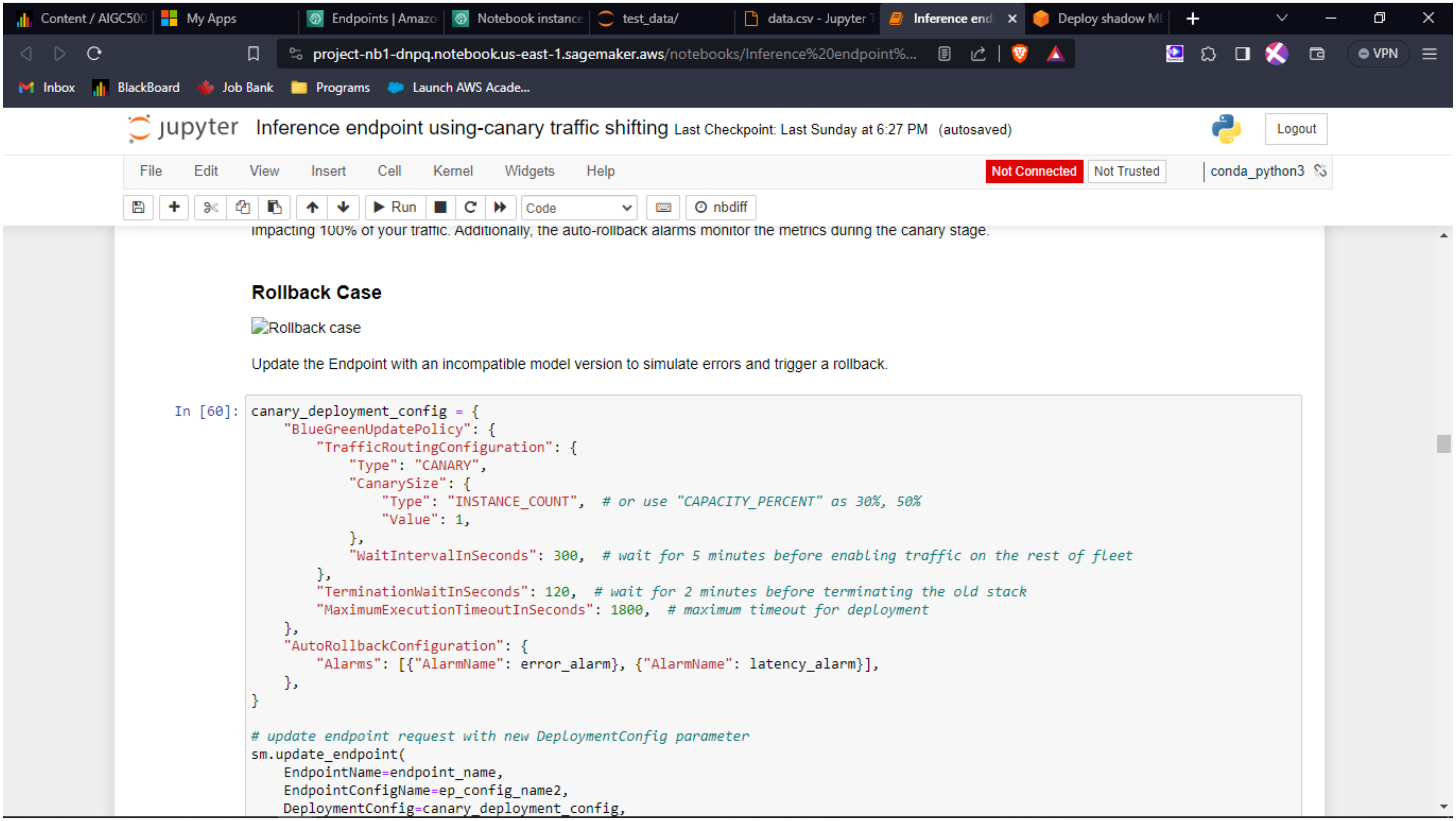


In [36]:

```python
def create_auto_rollback_alarm(
    alarm_name, endpoint_name, variant_name, metric_name, statistic, threshold
):
    cw.put_metric_alarm(
        AlarmName=alarm_name,
        AlarmDescription="Test SageMaker endpoint deployment auto-rollback alarm",
        ActionsEnabled=False,
        Namespace="AWS/SageMaker",
        MetricName=metric_name,
        Statistic=statistic,
        Dimensions=[
            {"Name": "EndpointName", "Value": endpoint_name},
            {"Name": "VariantName", "Value": variant_name},
        ],
        Period=60,
        EvaluationPeriods=1,
        Threshold=threshold,
        ComparisonOperator="GreaterThanOrEqualToThreshold",
        TreatMissingData="notBreaching",
    )
```

In [37]:

```python
error_alarm = f"TestAlarm-5XXErrors-{endpoint_name}"
latency_alarm = f"TestAlarm-ModelLatency-{endpoint_name}"

# alarm on 1% 5xx error rate for 1 minute
create_auto_rollback_alarm(
    error_alarm, endpoint_name, "AllTraffic", "Invocation5XXErrors", "Average", 1
)
# alarm on model latency >= 10 ms for 1 minute
create_auto_rollback_alarm(
    latency_alarm, endpoint_name, "AllTraffic", "ModelLatency", "Average", 10000
```



Inference endpoint using-canary traffic shifting Last Checkpoint: Last Sunday at 6:27 PM (autosaved)

impacting 100% of your traffic. Additionally, the auto-rollback alarms monitor the metrics during the canary stage.

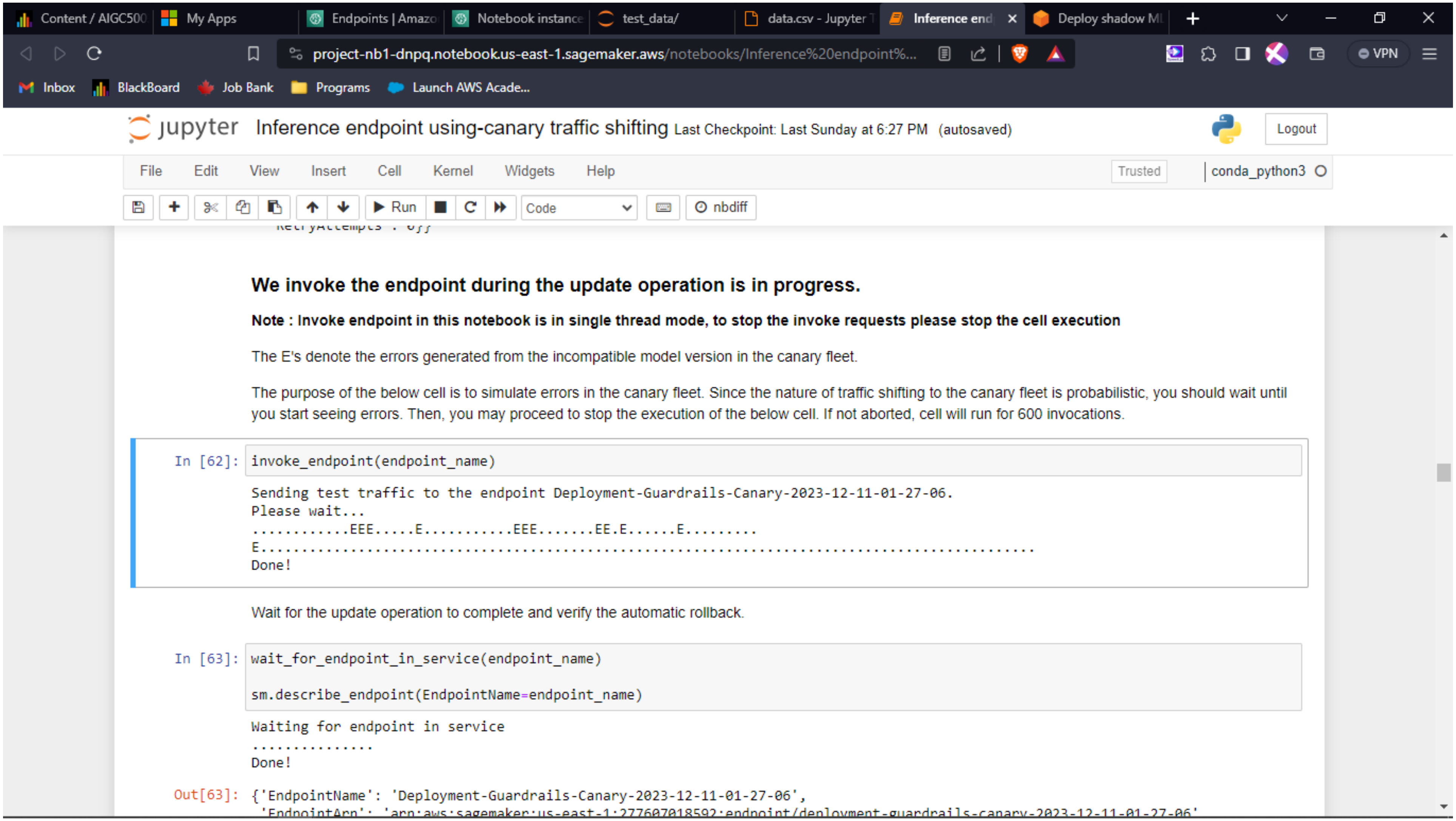
Rollback Case

Rollback case

Update the Endpoint with an incompatible model version to simulate errors and trigger a rollback.

In [60]:

```
canary_deployment_config = {
    "BlueGreenUpdatePolicy": {
        "TrafficRoutingConfiguration": {
            "Type": "CANARY",
            "CanarySize": {
                "Type": "INSTANCE_COUNT",  # or use "CAPACITY_PERCENT" as 30%, 50%
                "Value": 1,
            },
            "WaitIntervalInSeconds": 300,  # wait for 5 minutes before enabling traffic on the rest of fleet
        },
        "TerminationWaitInSeconds": 120,  # wait for 2 minutes before terminating the old stack
        "MaximumExecutionTimeoutInSeconds": 1800,  # maximum timeout for deployment
    },
    "AutoRollbackConfiguration": {
        "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,
```



```
  RetryAttempts': 0}}
```

We invoke the endpoint during the update operation is in progress.

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

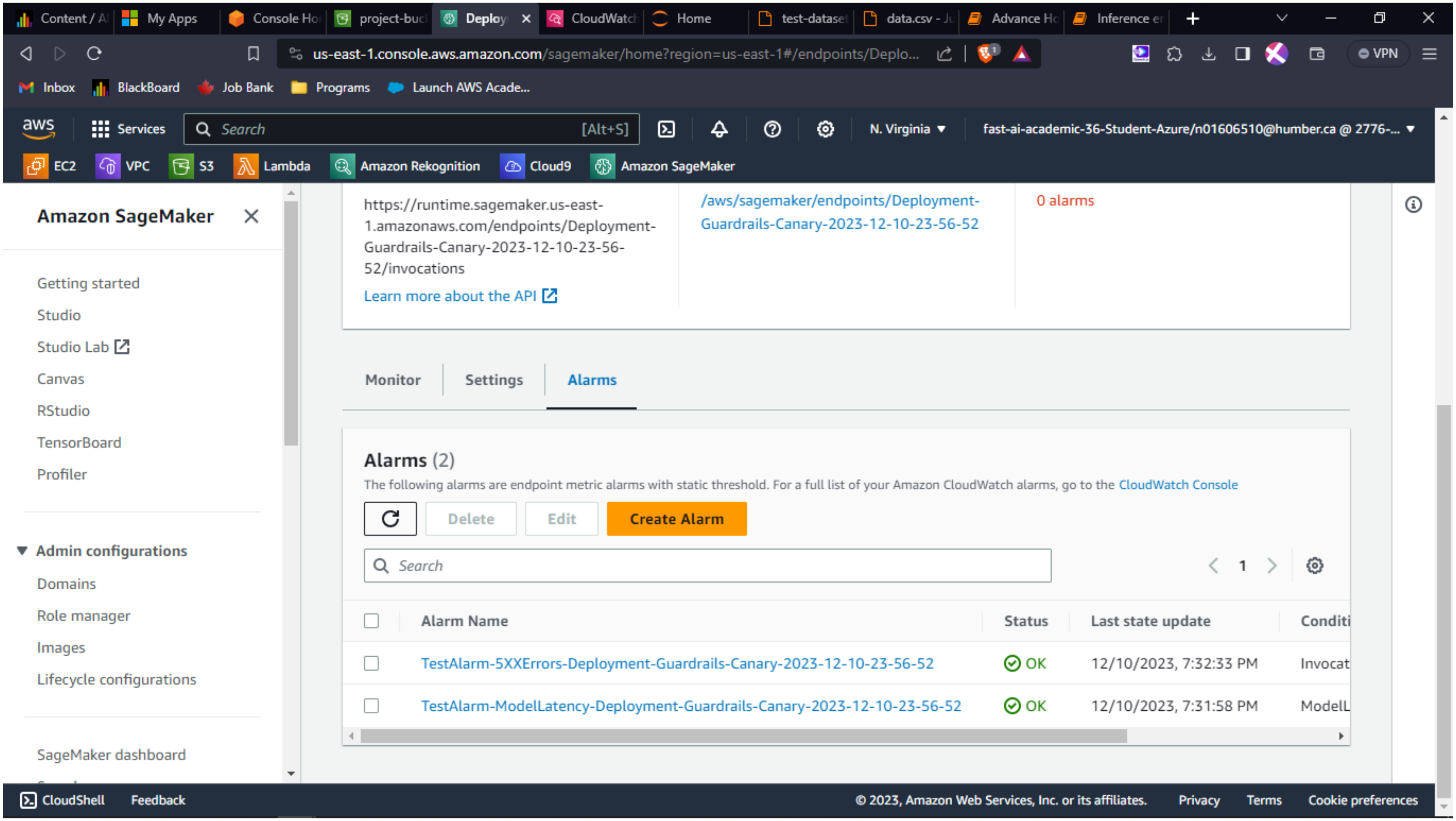
The E's denote the errors generated from the incompatible model version in the canary fleet.

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.

```
In [62]: invoke_endpoint(endpoint_name)

         Sending test traffic to the endpoint Deployment-Guardrails-Canary-2023-12-11-01-27-06.
         Please wait...
         ............EEE.....E...........EEE.......EE.E......E.........
         E...............................................................................................
         Done!
```

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

```
In [63]: wait_for_endpoint_in_service(endpoint_name)

         sm.describe_endpoint(EndpointName=endpoint_name)

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

Out[63]: {'EndpointName': 'Deployment-Guardrails-Canary-2023-12-11-01-27-06',
          'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-11-01-27-06'
```



https://runtime.sagemaker.us-east-1.amazonaws.com/endpoints/Deployment-Guardrails-Canary-2023-12-10-23-56-52/invocations

Learn more about the API

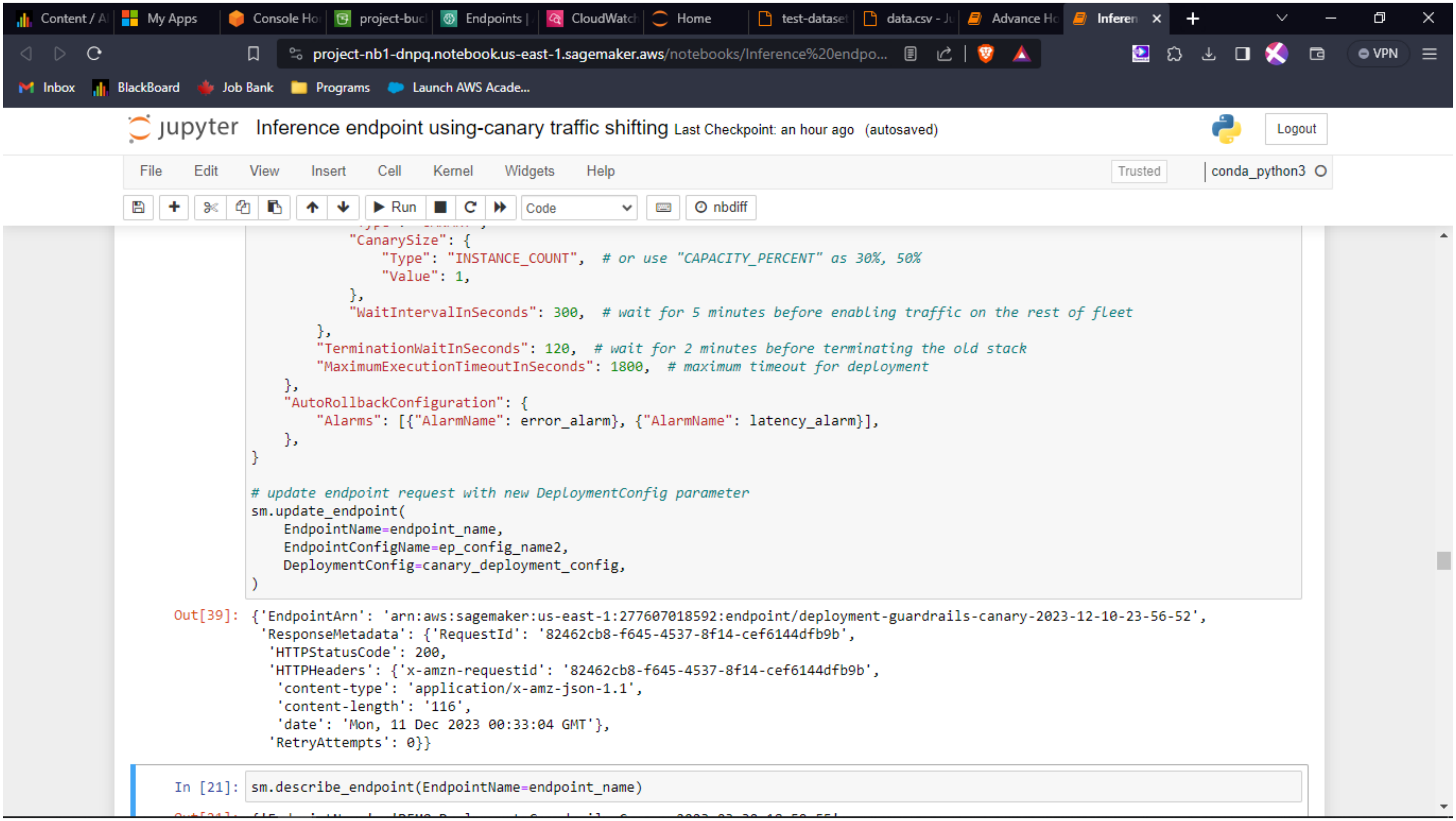
/aws/sagemaker/endpoints/Deployment-Guardrails-Canary-2023-12-10-23-56-52

0 alarms

Alarms (2)

The following alarms are endpoint metric alarms with static threshold. For a full list of your Amazon CloudWatch alarms, go to the CloudWatch Console

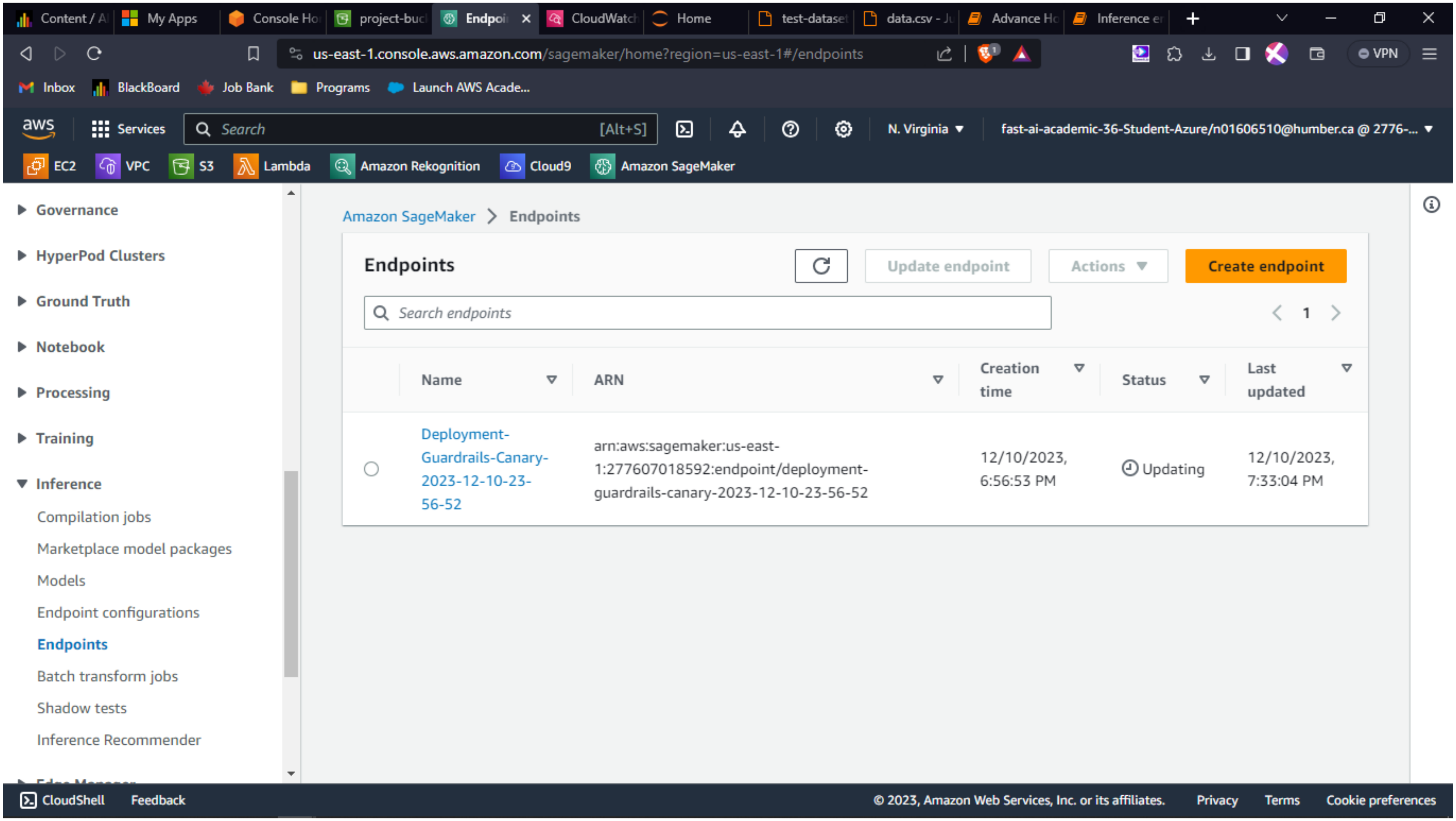
□	Alarm Name	Status	Last state update	Conditi
□	TestAlarm-5XXErrors-Deployment-Guardrails-Canary-2023-12-10-23-56-52	✓ OK	12/10/2023, 7:32:33 PM	Invocat
□	TestAlarm-ModelLatency-Deployment-Guardrails-Canary-2023-12-10-23-56-52	✓ OK	12/10/2023, 7:31:58 PM	ModelL



```
            "CanarySize": {
                "Type": "INSTANCE_COUNT",  # or use "CAPACITY_PERCENT" as 30%, 50%
                "Value": 1,
            },
            "WaitIntervalInSeconds": 300,  # wait for 5 minutes before enabling traffic on the rest of fleet
        },
        "TerminationWaitInSeconds": 120,  # wait for 2 minutes before terminating the old stack
        "MaximumExecutionTimeoutInSeconds": 1800,  # maximum timeout for deployment
    },
    "AutoRollbackConfiguration": {
        "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,
)
```

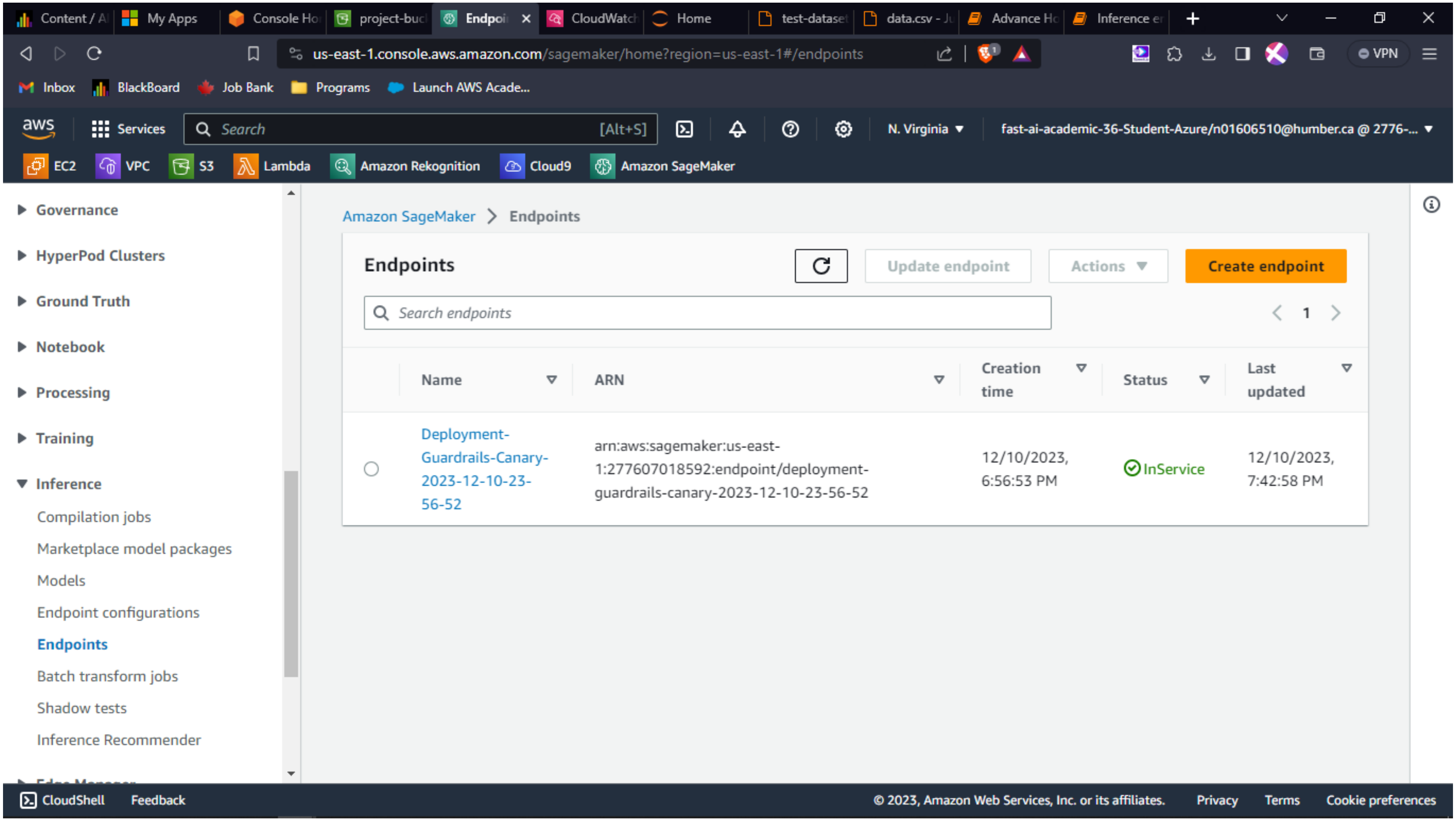
```
Out[39]: {'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23-56-52',
          'ResponseMetadata': {'RequestId': '82462cb8-f645-4537-8f14-cef6144dfb9b',
           'HTTPStatusCode': 200,
           'HTTPHeaders': {'x-amzn-requestid': '82462cb8-f645-4537-8f14-cef6144dfb9b',
            'content-type': 'application/x-amz-json-1.1',
            'content-length': '116',
            'date': 'Mon, 11 Dec 2023 00:33:04 GMT'},
           'RetryAttempts': 0}}
```

```
In [21]: sm.describe_endpoint(EndpointName=endpoint_name)
```



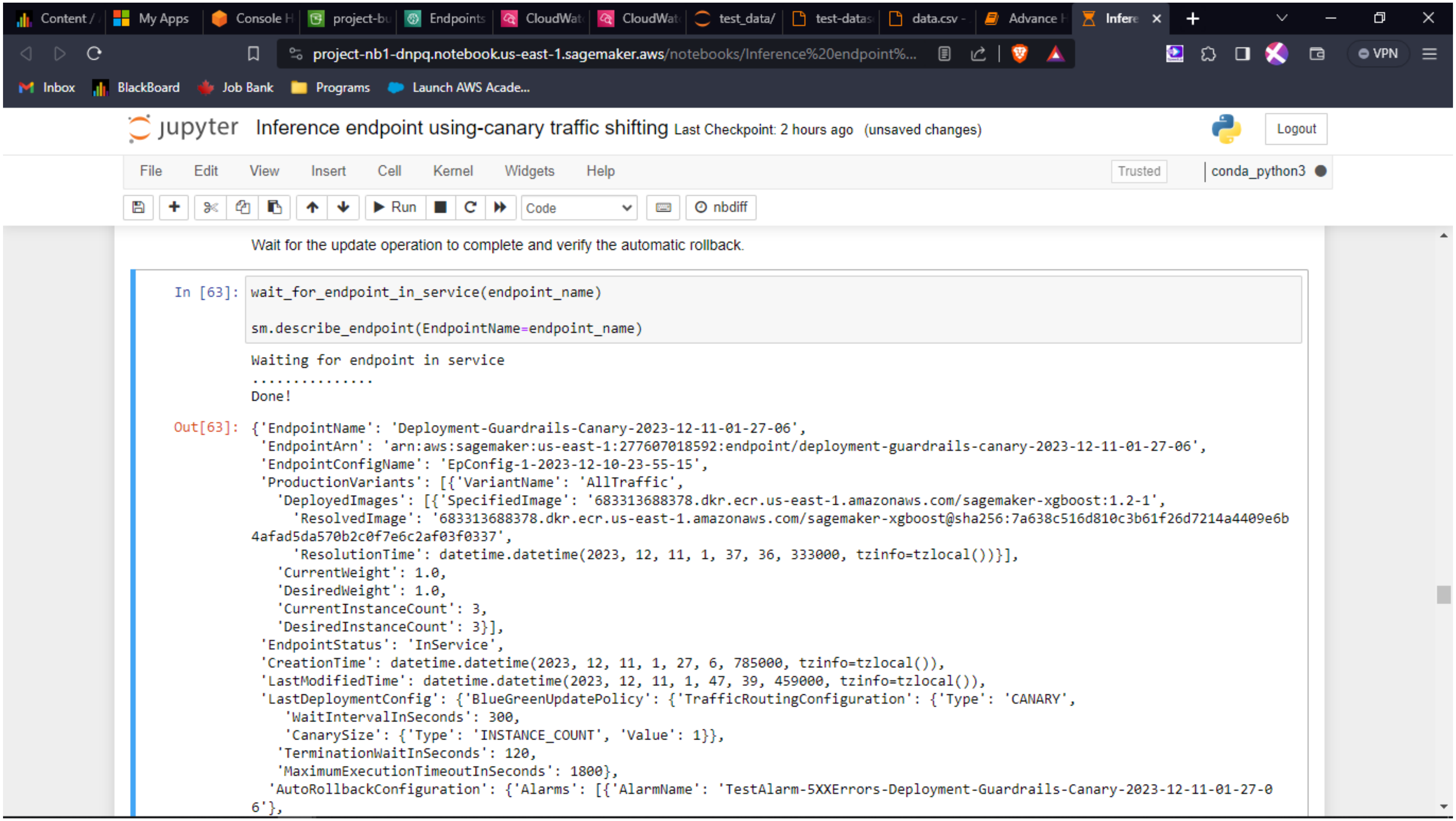
Endpoints

	Name	ARN	Creation time	Status	Last updated
○	Deployment-Guardrails-Canary-2023-12-10-23-56-52	arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23-56-52	12/10/2023, 6:56:53 PM	Updating	12/10/2023, 7:33:04 PM



Endpoints

	Name	ARN	Creation time	Status	Last updated
○	Deployment-Guardrails-Canary-2023-12-10-23-56-52	arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-10-23-56-52	12/10/2023, 6:56:53 PM	✓ InService	12/10/2023, 7:42:58 PM


Content / | My Apps | Console H | project-bu | Endpoints | CloudWat | CloudWat | test_data/ | test-datas | data.csv - | Advance I | Infere
project-nb1-dnpq.notebook.us-east-1.sagemaker.aws/notebooks/Inference%20endpoint%...
Inbox | BlackBoard | Job Bank | Programs | Launch AWS Acade...
jupyter
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
conda_python3
Run
Code
nbdiff

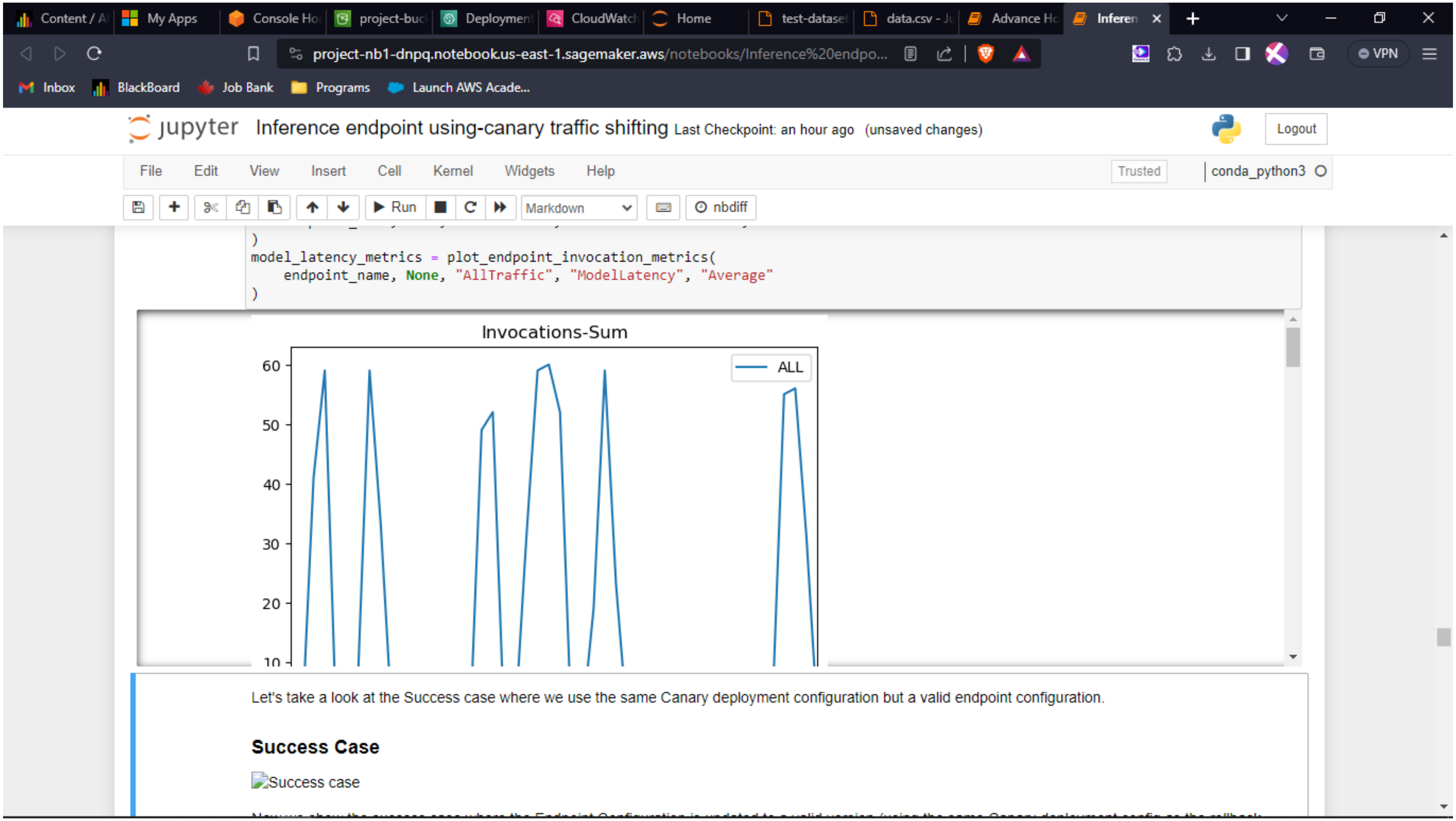
Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (unsaved changes)

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

```
In [63]: wait_for_endpoint_in_service(endpoint_name)

         sm.describe_endpoint(EndpointName=endpoint_name)
```

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

```
Out[63]: {'EndpointName': 'Deployment-Guardrails-Canary-2023-12-11-01-27-06',
          'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-11-01-27-06',
          'EndpointConfigName': 'EpConfig-1-2023-12-10-23-55-15',
          'ProductionVariants': [{'VariantName': 'AllTraffic',
            'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.2-1',
              'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:7a638c516d810c3b61f26d7214a4409e6b
         4afad5da570b2c0f7e6c2af03f0337',
              'ResolutionTime': datetime.datetime(2023, 12, 11, 1, 37, 36, 333000, tzinfo=tzlocal())}],
            'CurrentWeight': 1.0,
            'DesiredWeight': 1.0,
            'CurrentInstanceCount': 3,
            'DesiredInstanceCount': 3}],
          'EndpointStatus': 'InService',
          'CreationTime': datetime.datetime(2023, 12, 11, 1, 27, 6, 785000, tzinfo=tzlocal()),
          'LastModifiedTime': datetime.datetime(2023, 12, 11, 1, 47, 39, 459000, tzinfo=tzlocal()),
          'LastDeploymentConfig': {'BlueGreenUpdatePolicy': {'TrafficRoutingConfiguration': {'Type': 'CANARY',
             'WaitIntervalInSeconds': 300,
             'CanarySize': {'Type': 'INSTANCE_COUNT', 'Value': 1}},
            'TerminationWaitInSeconds': 120,
            'MaximumExecutionTimeoutInSeconds': 1800},
           'AutoRollbackConfiguration': {'Alarms': [{'AlarmName': 'TestAlarm-5XXErrors-Deployment-Guardrails-Canary-2023-12-11-01-27-0
         6'},
```



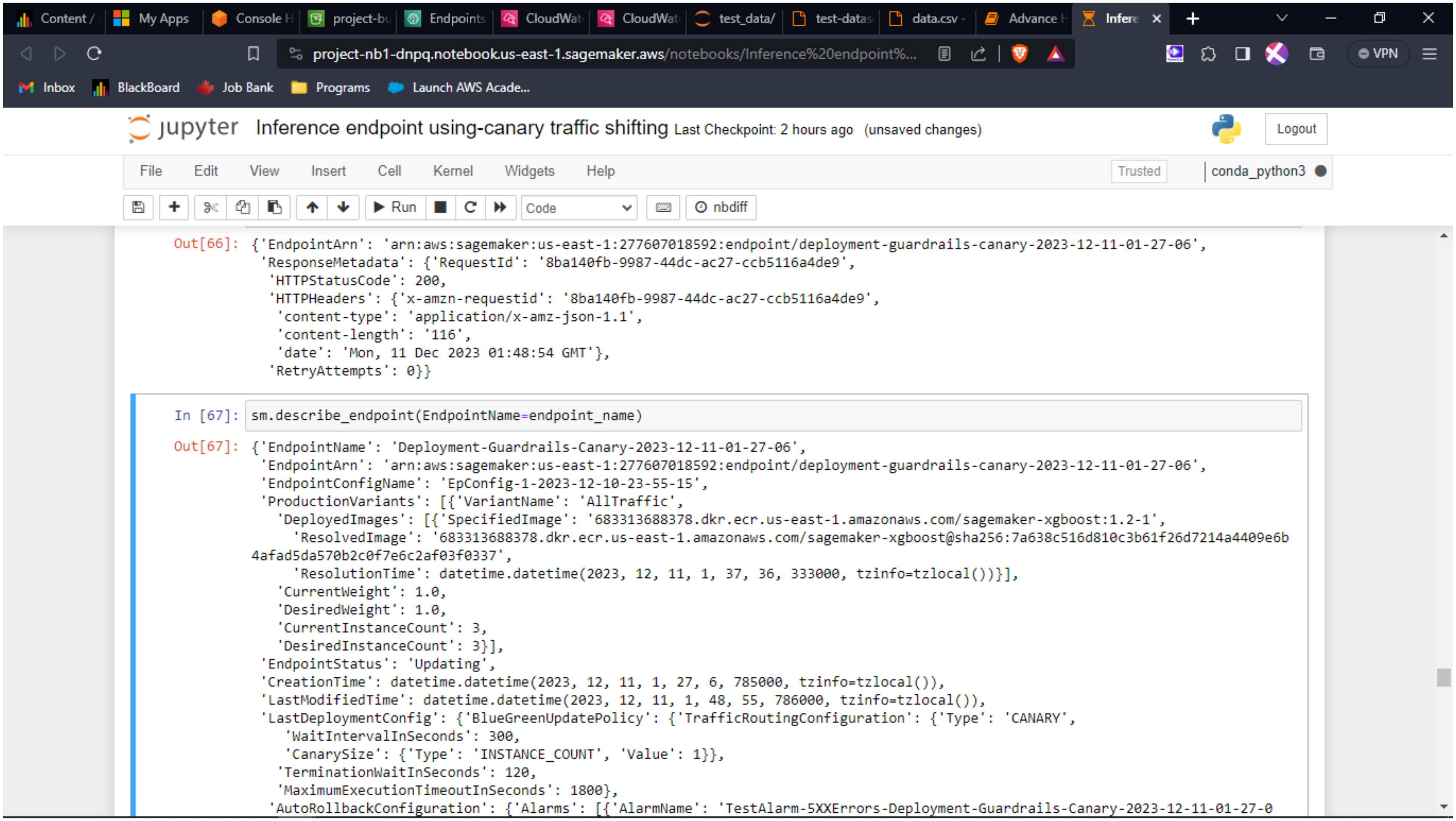
Inference endpoint using-canary traffic shifting Last Checkpoint: an hour ago (unsaved changes)

```
)
model_latency_metrics = plot_endpoint_invocation_metrics(
    endpoint_name, None, "AllTraffic", "ModelLatency", "Average"
)
```

Let's take a look at the Success case where we use the same Canary deployment configuration but a valid endpoint configuration.

Success Case

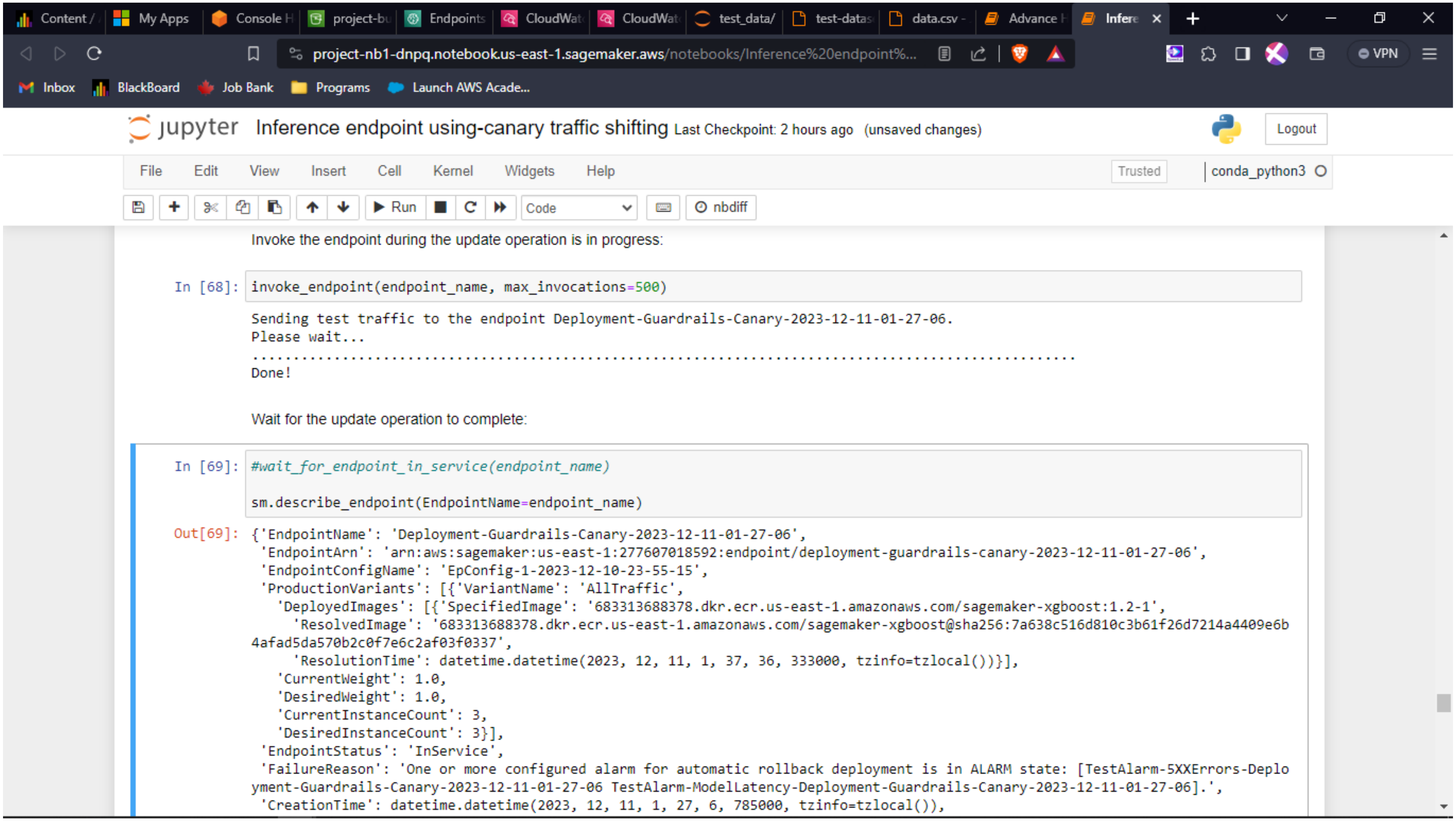
Success case



```
Out[66]: {'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-11-01-27-06',
           'ResponseMetadata': {'RequestId': '8ba140fb-9987-44dc-ac27-ccb5116a4de9',
            'HTTPStatusCode': 200,
            'HTTPHeaders': {'x-amzn-requestid': '8ba140fb-9987-44dc-ac27-ccb5116a4de9',
             'content-type': 'application/x-amz-json-1.1',
             'content-length': '116',
             'date': 'Mon, 11 Dec 2023 01:48:54 GMT'},
            'RetryAttempts': 0}}
```

```
In [67]: sm.describe_endpoint(EndpointName=endpoint_name)
```

```
Out[67]: {'EndpointName': 'Deployment-Guardrails-Canary-2023-12-11-01-27-06',
           'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-11-01-27-06',
           'EndpointConfigName': 'EpConfig-1-2023-12-10-23-55-15',
           'ProductionVariants': [{'VariantName': 'AllTraffic',
             'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.2-1',
               'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:7a638c516d810c3b61f26d7214a4409e6b
         4afad5da570b2c0f7e6c2af03f0337',
               'ResolutionTime': datetime.datetime(2023, 12, 11, 1, 37, 36, 333000, tzinfo=tzlocal())}],
             'CurrentWeight': 1.0,
             'DesiredWeight': 1.0,
             'CurrentInstanceCount': 3,
             'DesiredInstanceCount': 3}],
           'EndpointStatus': 'Updating',
           'CreationTime': datetime.datetime(2023, 12, 11, 1, 27, 6, 785000, tzinfo=tzlocal()),
           'LastModifiedTime': datetime.datetime(2023, 12, 11, 1, 48, 55, 786000, tzinfo=tzlocal()),
           'LastDeploymentConfig': {'BlueGreenUpdatePolicy': {'TrafficRoutingConfiguration': {'Type': 'CANARY',
              'WaitIntervalInSeconds': 300,
              'CanarySize': {'Type': 'INSTANCE_COUNT', 'Value': 1}},
             'TerminationWaitInSeconds': 120,
             'MaximumExecutionTimeoutInSeconds': 1800},
            'AutoRollbackConfiguration': {'Alarms': [{'AlarmName': 'TestAlarm-5XXErrors-Deployment-Guardrails-Canary-2023-12-11-01-27-0
```



Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (unsaved changes)

Invoke the endpoint during the update operation is in progress:

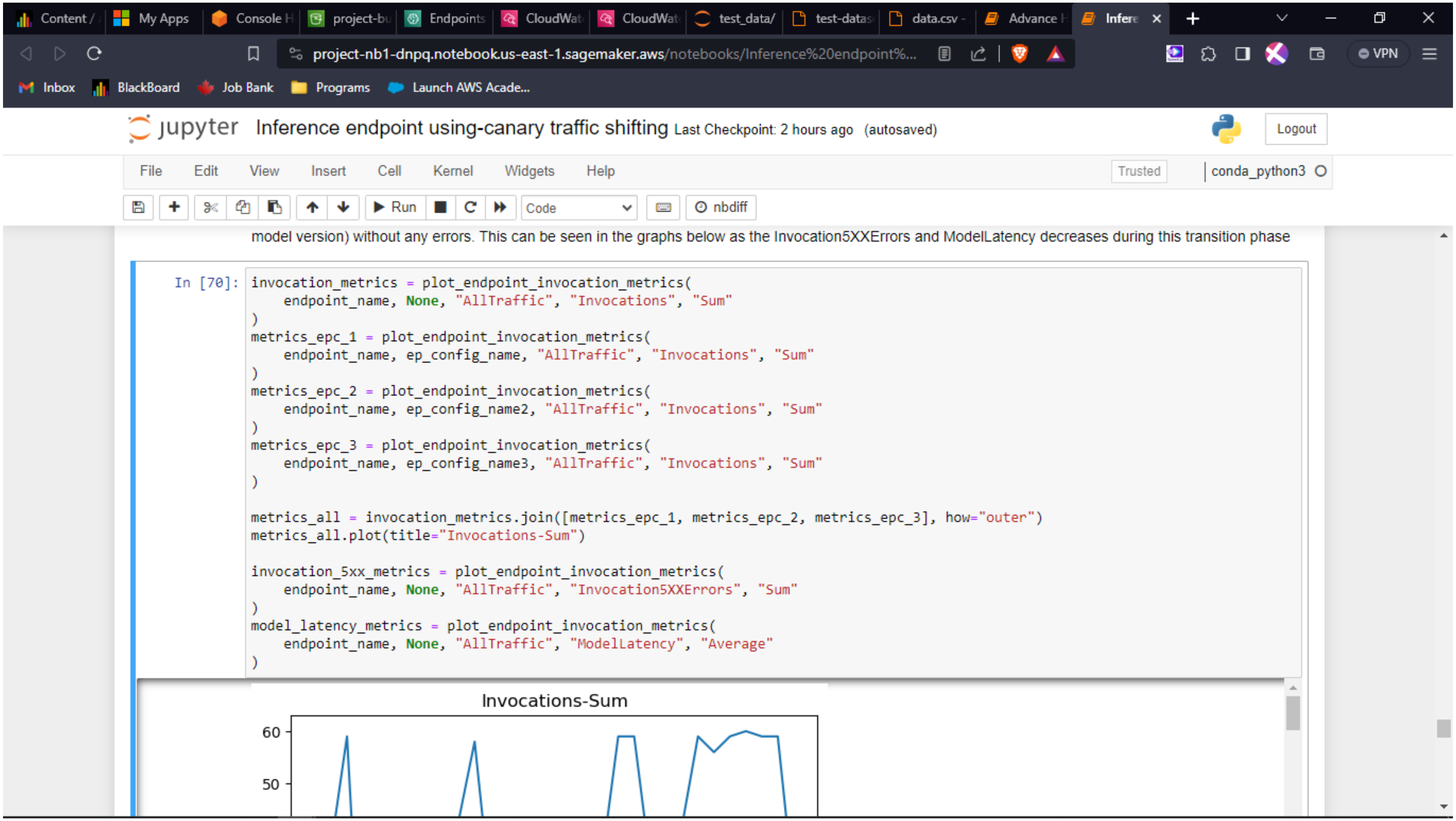
```
In [68]: invoke_endpoint(endpoint_name, max_invocations=500)
```

```
Sending test traffic to the endpoint Deployment-Guardrails-Canary-2023-12-11-01-27-06.
Please wait...
......................................................................................................
Done!
```

Wait for the update operation to complete:

```
In [69]: #wait_for_endpoint_in_service(endpoint_name)

         sm.describe_endpoint(EndpointName=endpoint_name)
```

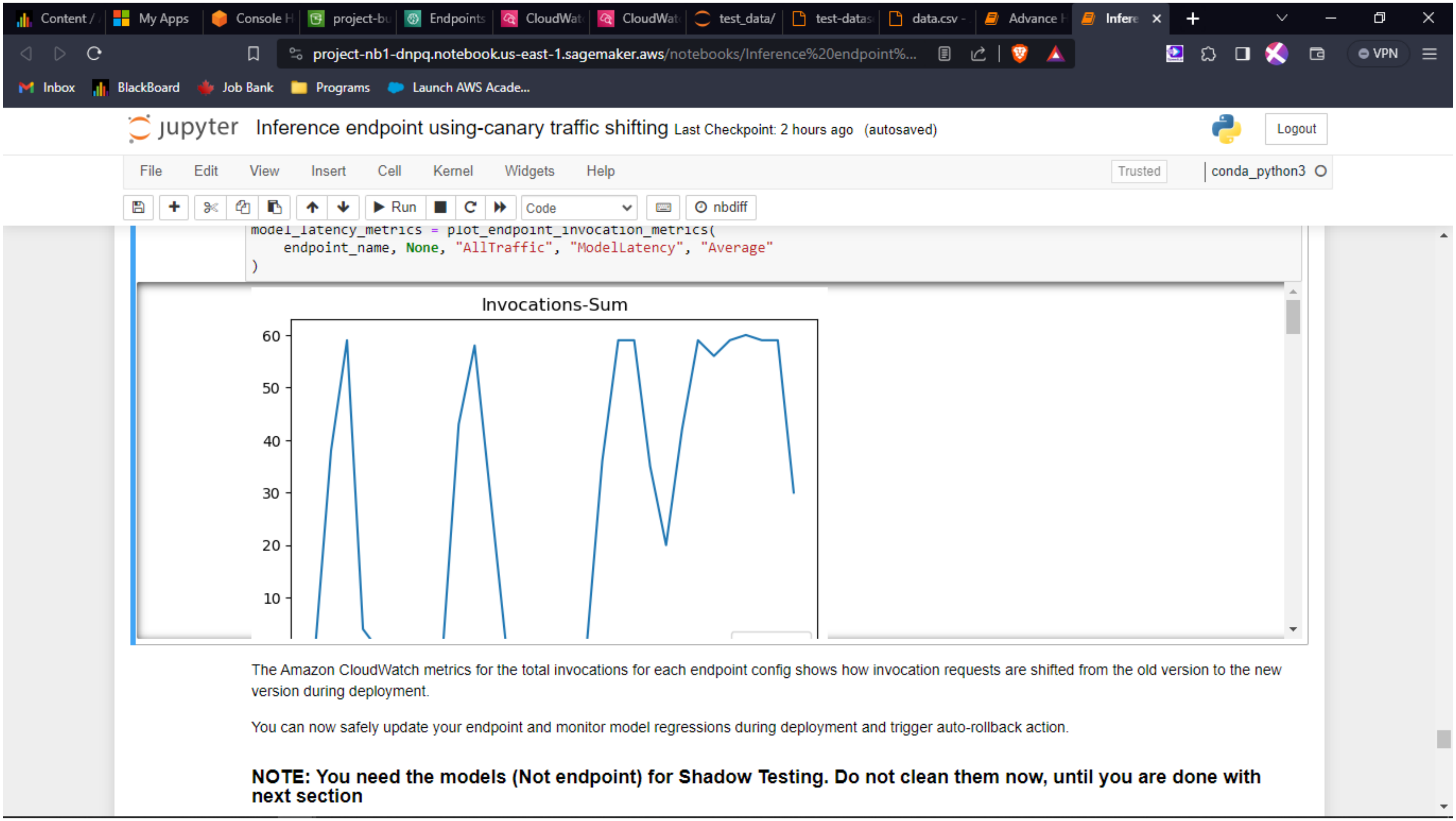
```
Out[69]: {'EndpointName': 'Deployment-Guardrails-Canary-2023-12-11-01-27-06',
          'EndpointArn': 'arn:aws:sagemaker:us-east-1:277607018592:endpoint/deployment-guardrails-canary-2023-12-11-01-27-06',
          'EndpointConfigName': 'EpConfig-1-2023-12-10-23-55-15',
          'ProductionVariants': [{'VariantName': 'AllTraffic',
            'DeployedImages': [{'SpecifiedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.2-1',
              'ResolvedImage': '683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost@sha256:7a638c516d810c3b61f26d7214a4409e6b
         4afad5da570b2c0f7e6c2af03f0337',
              'ResolutionTime': datetime.datetime(2023, 12, 11, 1, 37, 36, 333000, tzinfo=tzlocal())}],
            'CurrentWeight': 1.0,
            'DesiredWeight': 1.0,
            'CurrentInstanceCount': 3,
            'DesiredInstanceCount': 3}],
          'EndpointStatus': 'InService',
          'FailureReason': 'One or more configured alarm for automatic rollback deployment is in ALARM state: [TestAlarm-5XXErrors-Deplo
         yment-Guardrails-Canary-2023-12-11-01-27-06 TestAlarm-ModelLatency-Deployment-Guardrails-Canary-2023-12-11-01-27-06].',
          'CreationTime': datetime.datetime(2023, 12, 11, 1, 27, 6, 785000, tzinfo=tzlocal()),
```



Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (autosaved)

model version) without any errors. This can be seen in the graphs below as the Invocation5XXErrors and ModelLatency decreases during this transition phase

```
In [70]: invocation_metrics = plot_endpoint_invocation_metrics(
             endpoint_name, None, "AllTraffic", "Invocations", "Sum"
         )
         metrics_epc_1 = plot_endpoint_invocation_metrics(
             endpoint_name, ep_config_name, "AllTraffic", "Invocations", "Sum"
         )
         metrics_epc_2 = plot_endpoint_invocation_metrics(
             endpoint_name, ep_config_name2, "AllTraffic", "Invocations", "Sum"
         )
         metrics_epc_3 = plot_endpoint_invocation_metrics(
             endpoint_name, ep_config_name3, "AllTraffic", "Invocations", "Sum"
         )

         metrics_all = invocation_metrics.join([metrics_epc_1, metrics_epc_2, metrics_epc_3], how="outer")
         metrics_all.plot(title="Invocations-Sum")

         invocation_5xx_metrics = plot_endpoint_invocation_metrics(
             endpoint_name, None, "AllTraffic", "Invocation5XXErrors", "Sum"
         )
         model_latency_metrics = plot_endpoint_invocation_metrics(
             endpoint_name, None, "AllTraffic", "ModelLatency", "Average"
         )
```

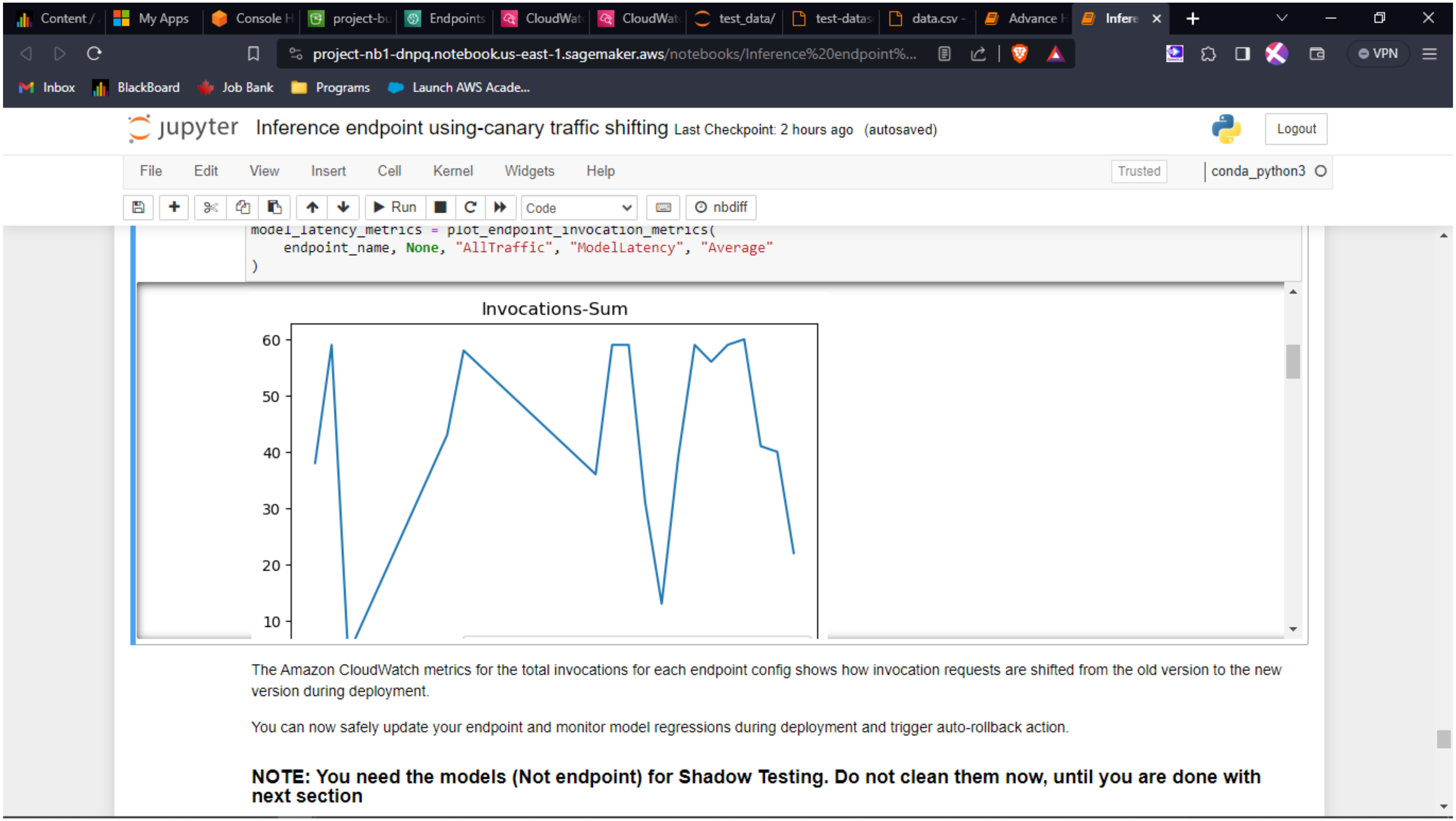


```
model_latency_metrics = plot_endpoint_invocation_metrics(
    endpoint_name, None, "AllTraffic", "ModelLatency", "Average"
)
```

The Amazon CloudWatch metrics for the total invocations for each endpoint config shows how invocation requests are shifted from the old version to the new version during deployment.

You can now safely update your endpoint and monitor model regressions during deployment and trigger auto-rollback action.

NOTE: You need the models (Not endpoint) for Shadow Testing. Do not clean them now, until you are done with next section

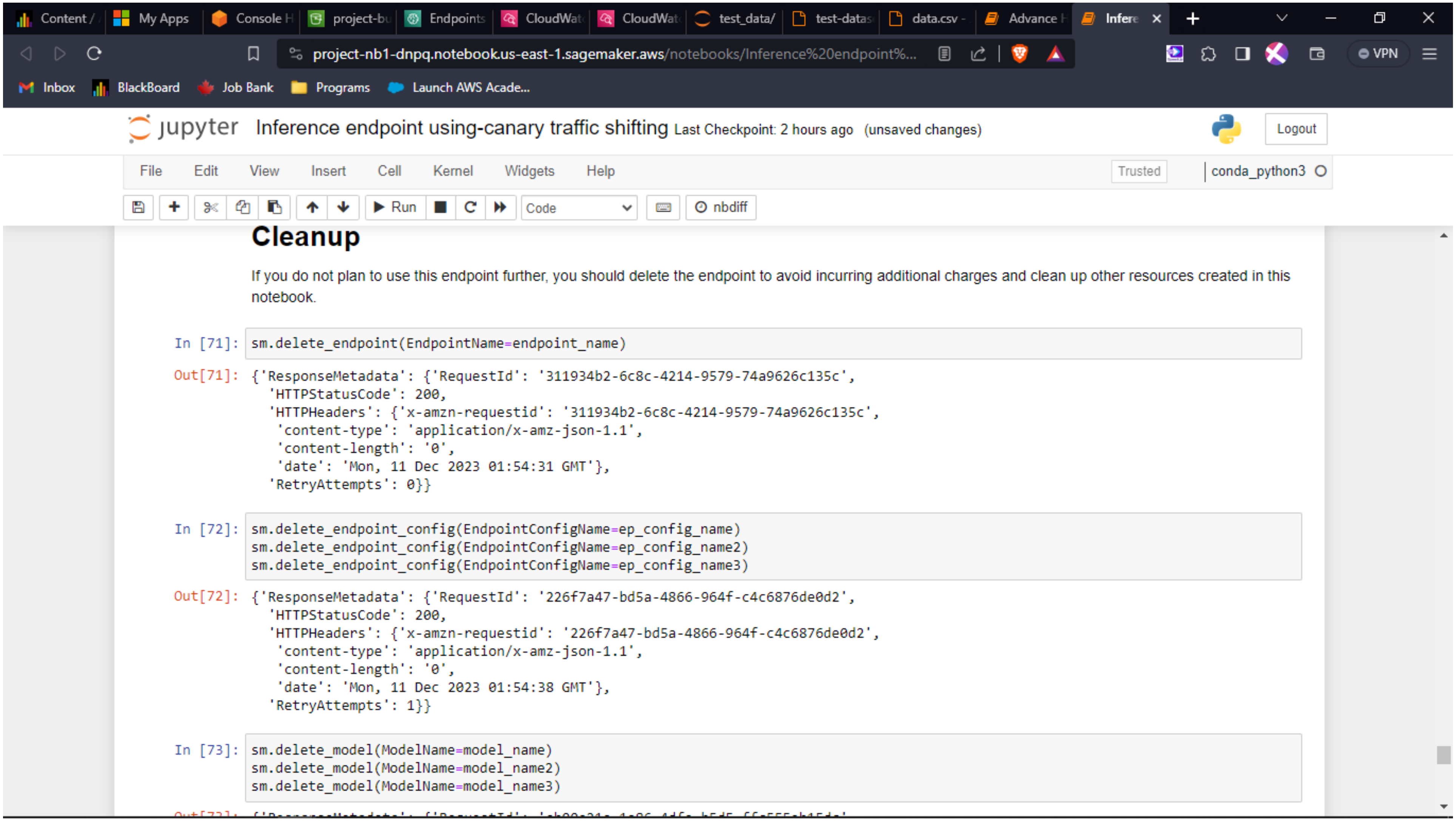


```
model_latency_metrics = plot_endpoint_invocation_metrics(
    endpoint_name, None, "AllTraffic", "ModelLatency", "Average"
)
```

The Amazon CloudWatch metrics for the total invocations for each endpoint config shows how invocation requests are shifted from the old version to the new version during deployment.

You can now safely update your endpoint and monitor model regressions during deployment and trigger auto-rollback action.

NOTE: You need the models (Not endpoint) for Shadow Testing. Do not clean them now, until you are done with next section



Content / · My Apps · Console H · project-bu · Endpoints · CloudWat · CloudWat · test_data/ · test-datas · data.csv - · Advance H · Infere
project-nb1-dnpq.notebook.us-east-1.sagemaker.aws/notebooks/Inference%20endpoint%...
Inbox · BlackBoard · Job Bank · Programs · Launch AWS Acade...
VPN
jupyter
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
conda_python3
Run
Code
nbdiff

Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (unsaved changes)

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.

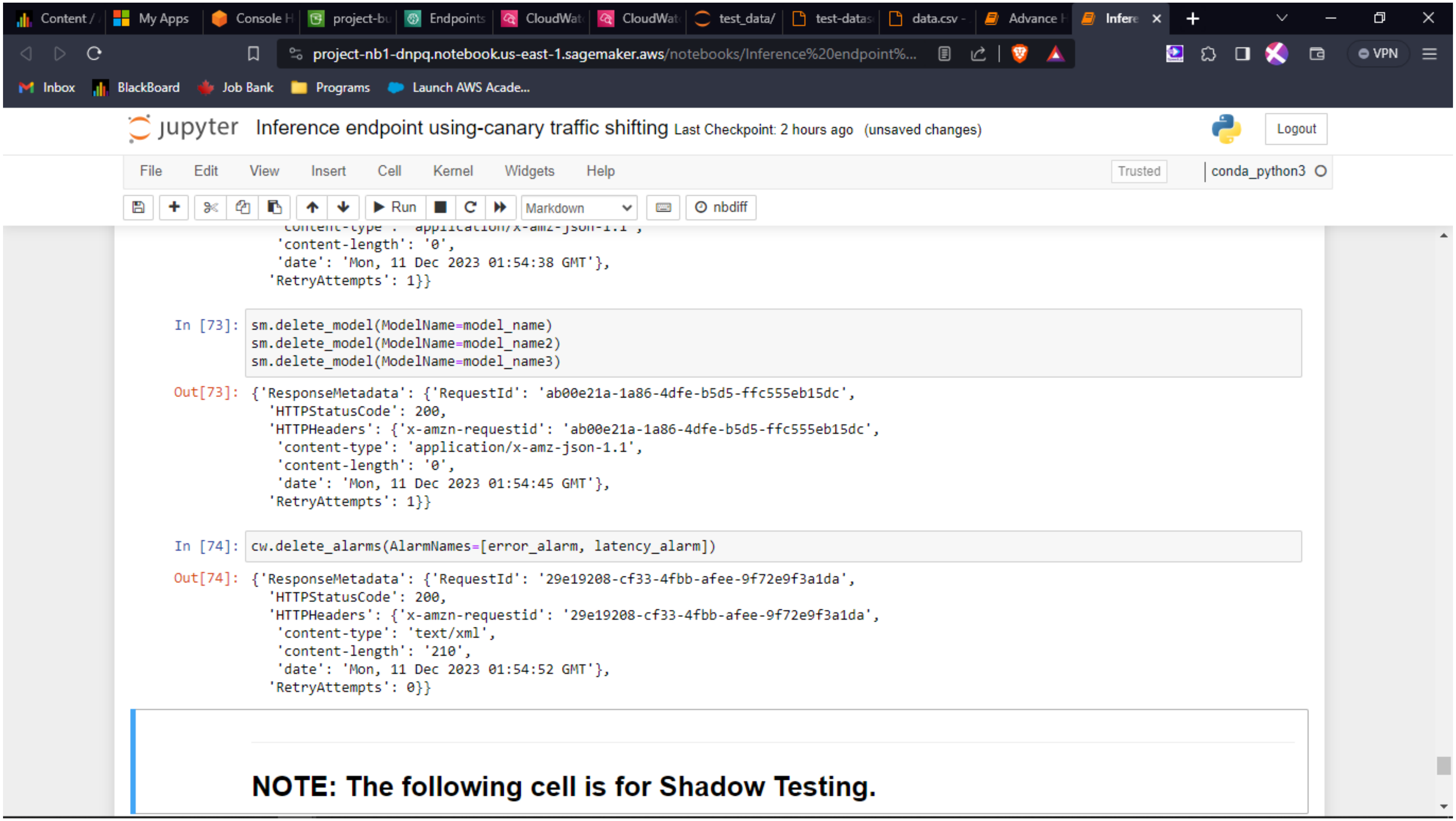
```
In [71]: sm.delete_endpoint(EndpointName=endpoint_name)
```

```
Out[71]: {'ResponseMetadata': {'RequestId': '311934b2-6c8c-4214-9579-74a9626c135c',
           'HTTPStatusCode': 200,
           'HTTPHeaders': {'x-amzn-requestid': '311934b2-6c8c-4214-9579-74a9626c135c',
            'content-type': 'application/x-amz-json-1.1',
            'content-length': '0',
            'date': 'Mon, 11 Dec 2023 01:54:31 GMT'},
           'RetryAttempts': 0}}
```

```
In [72]: sm.delete_endpoint_config(EndpointConfigName=ep_config_name)
         sm.delete_endpoint_config(EndpointConfigName=ep_config_name2)
         sm.delete_endpoint_config(EndpointConfigName=ep_config_name3)
```

```
Out[72]: {'ResponseMetadata': {'RequestId': '226f7a47-bd5a-4866-964f-c4c6876de0d2',
           'HTTPStatusCode': 200,
           'HTTPHeaders': {'x-amzn-requestid': '226f7a47-bd5a-4866-964f-c4c6876de0d2',
            'content-type': 'application/x-amz-json-1.1',
            'content-length': '0',
            'date': 'Mon, 11 Dec 2023 01:54:38 GMT'},
           'RetryAttempts': 1}}
```

```
In [73]: sm.delete_model(ModelName=model_name)
         sm.delete_model(ModelName=model_name2)
         sm.delete_model(ModelName=model_name3)
```



Inference endpoint using-canary traffic shifting Last Checkpoint: 2 hours ago (unsaved changes)

```
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '0',
   'date': 'Mon, 11 Dec 2023 01:54:38 GMT'},
  'RetryAttempts': 1}}
```

In [73]:

```
sm.delete_model(ModelName=model_name)
sm.delete_model(ModelName=model_name2)
sm.delete_model(ModelName=model_name3)
```

Out[73]:

```
{'ResponseMetadata': {'RequestId': 'ab00e21a-1a86-4dfe-b5d5-ffc555eb15dc',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': 'ab00e21a-1a86-4dfe-b5d5-ffc555eb15dc',
   'content-type': 'application/x-amz-json-1.1',
   'content-length': '0',
   'date': 'Mon, 11 Dec 2023 01:54:45 GMT'},
  'RetryAttempts': 1}}
```

In [74]:

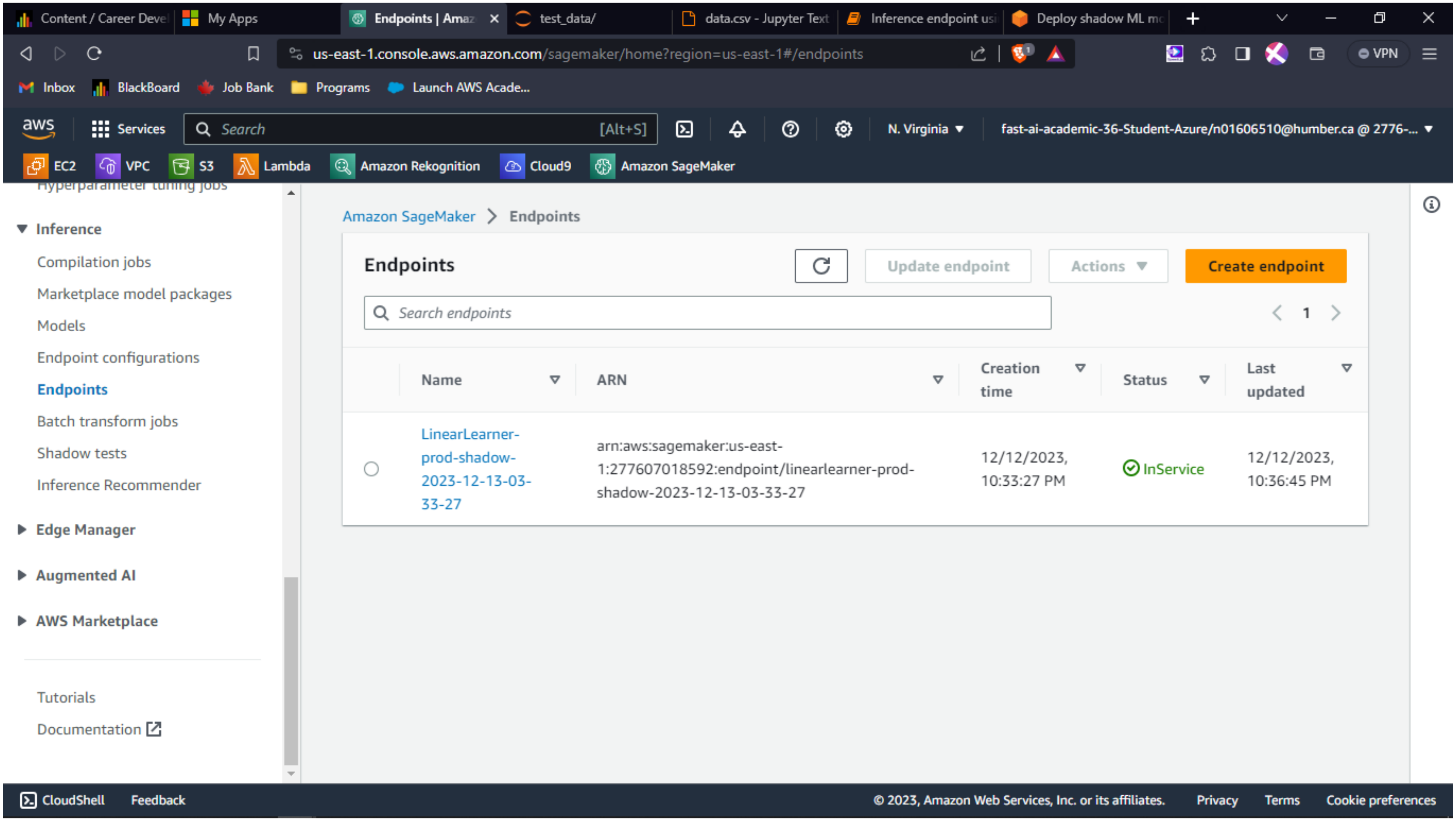
```
cw.delete_alarms(AlarmNames=[error_alarm, latency_alarm])
```

Out[74]:

```
{'ResponseMetadata': {'RequestId': '29e19208-cf33-4fbb-afee-9f72e9f3a1da',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'x-amzn-requestid': '29e19208-cf33-4fbb-afee-9f72e9f3a1da',
   'content-type': 'text/xml',
   'content-length': '210',
   'date': 'Mon, 11 Dec 2023 01:54:52 GMT'},
  'RetryAttempts': 0}}
```

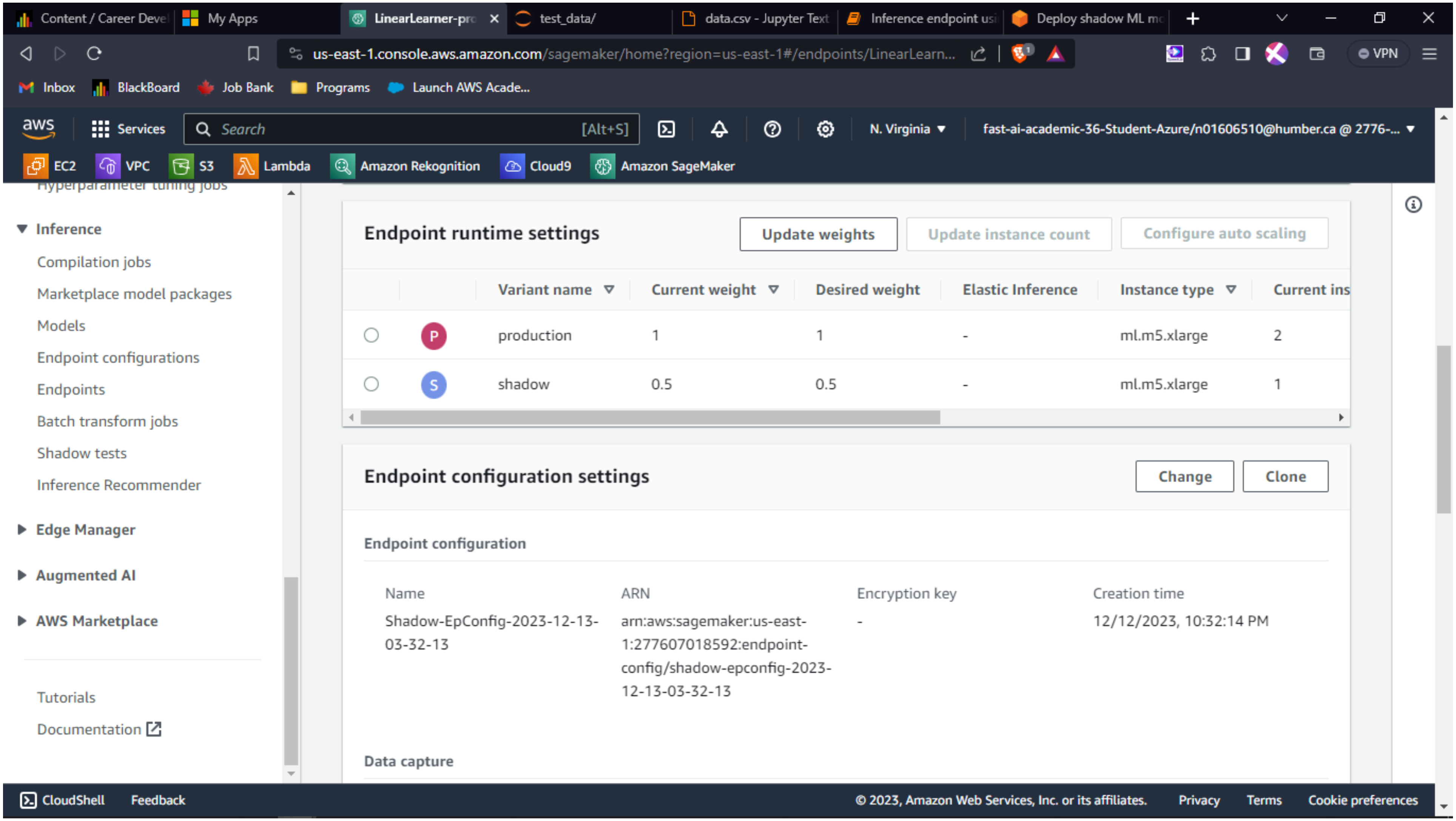
NOTE: The following cell is for Shadow Testing.

SHADOW TESTING



Endpoints

	Name	ARN	Creation time	Status	Last updated
○	LinearLearner-prod-shadow-2023-12-13-03-33-27	arn:aws:sagemaker:us-east-1:277607018592:endpoint/linearlearner-prod-shadow-2023-12-13-03-33-27	12/12/2023, 10:33:27 PM	✓ InService	12/12/2023, 10:36:45 PM



Endpoint runtime settings

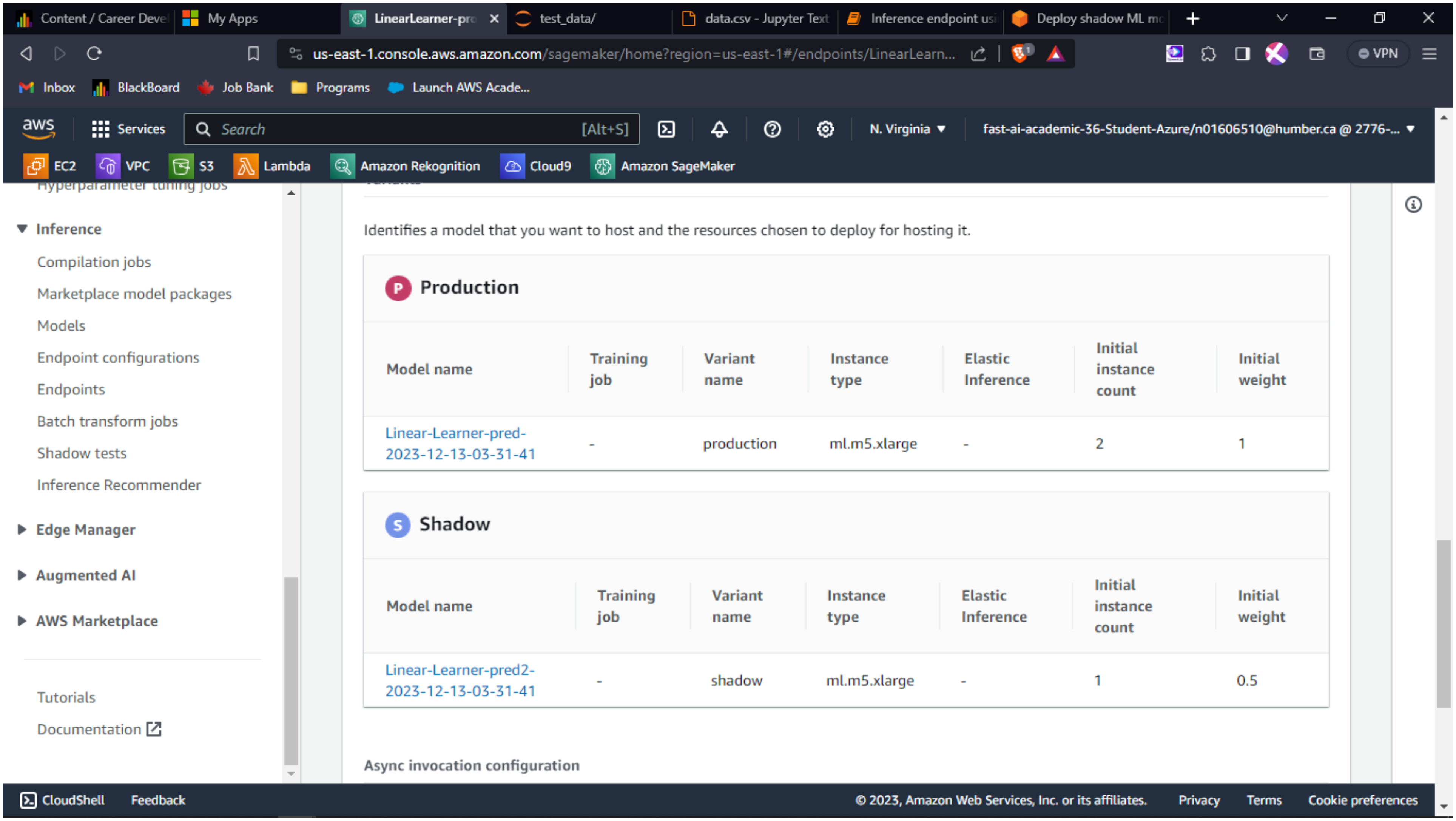
		Variant name	Current weight	Desired weight	Elastic Inference	Instance type	Current ins
○	P	production	1	1	-	ml.m5.xlarge	2
○	S	shadow	0.5	0.5	-	ml.m5.xlarge	1

Endpoint configuration settings

Endpoint configuration

Name	ARN	Encryption key	Creation time
Shadow-EpConfig-2023-12-13-03-32-13	arn:aws:sagemaker:us-east-1:277607018592:endpoint-config/shadow-epconfig-2023-12-13-03-32-13	-	12/12/2023, 10:32:14 PM

Data capture



Variants

Identifies a model that you want to host and the resources chosen to deploy for hosting it.

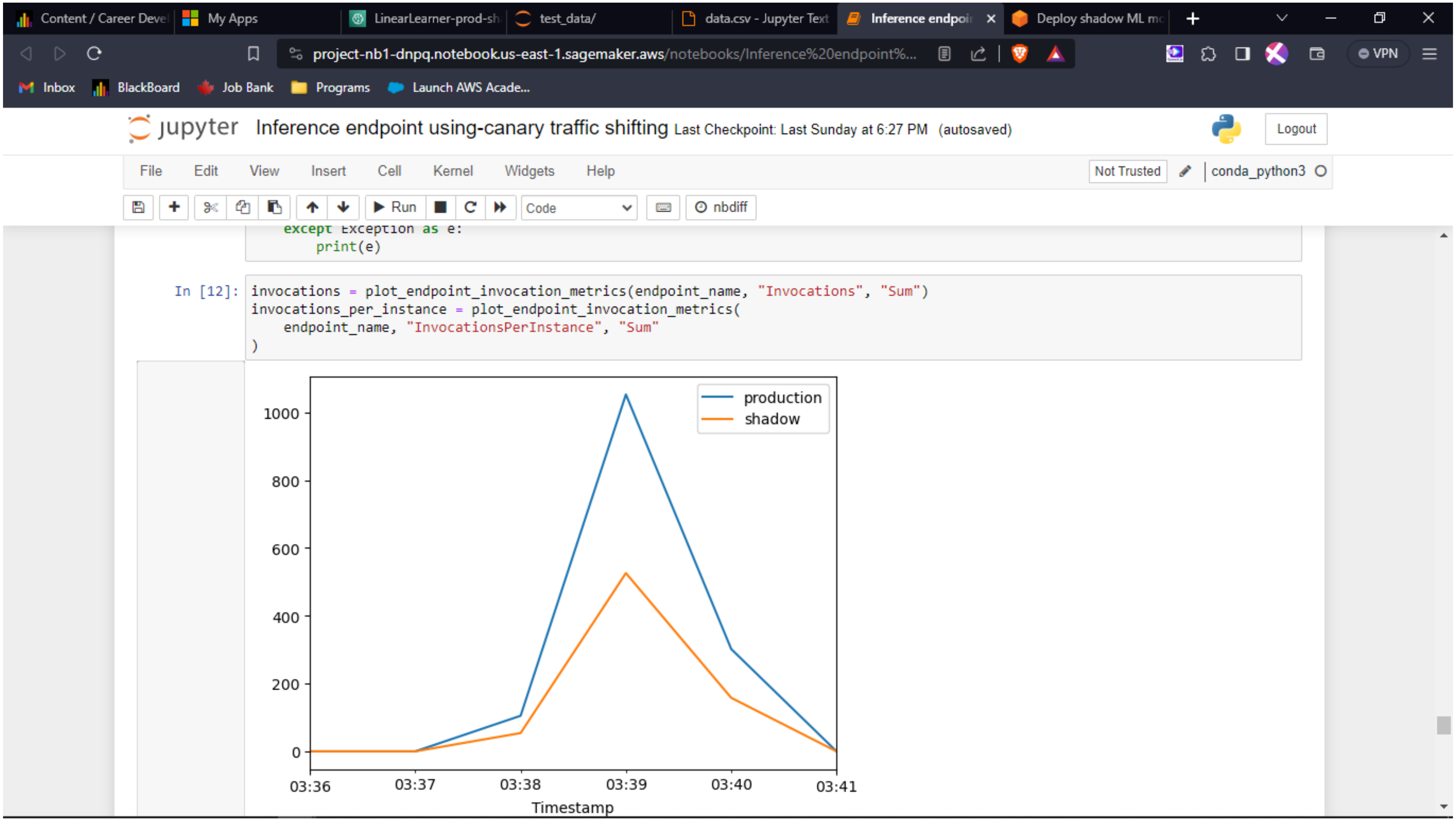
P Production

Model name	Training job	Variant name	Instance type	Elastic Inference	Initial instance count	Initial weight
Linear-Learner-pred-2023-12-13-03-31-41	-	production	ml.m5.xlarge	-	2	1

S Shadow

Model name	Training job	Variant name	Instance type	Elastic Inference	Initial instance count	Initial weight
Linear-Learner-pred2-2023-12-13-03-31-41	-	shadow	ml.m5.xlarge	-	1	0.5

Async invocation configuration

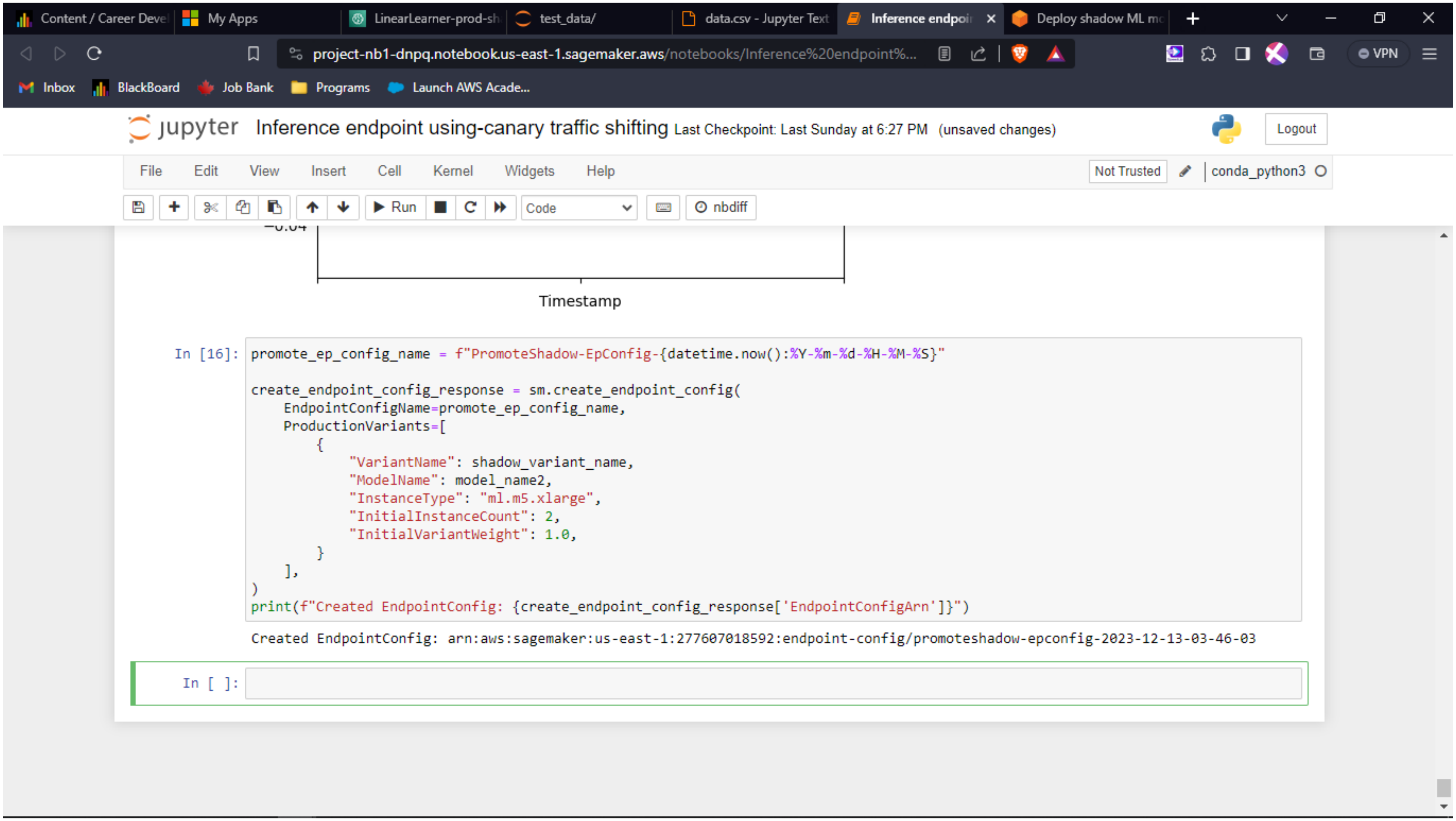


Inference endpoint using-canary traffic shifting Last Checkpoint: Last Sunday at 6:27 PM (autosaved)

```
    except Exception as e:
        print(e)
```

In [12]:

```
invocations = plot_endpoint_invocation_metrics(endpoint_name, "Invocations", "Sum")
invocations_per_instance = plot_endpoint_invocation_metrics(
    endpoint_name, "InvocationsPerInstance", "Sum"
)
```

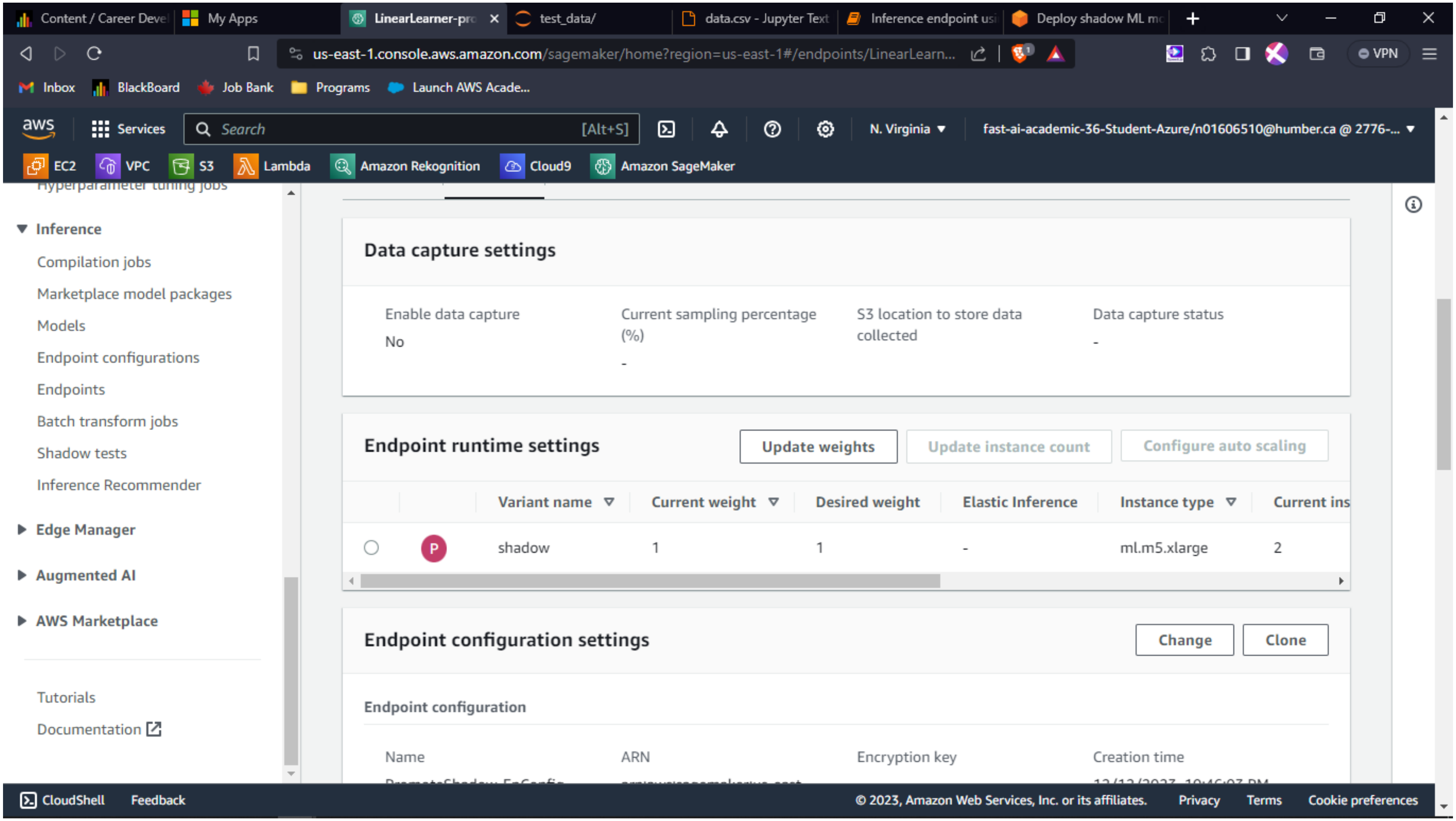


Inference endpoint using-canary traffic shifting Last Checkpoint: Last Sunday at 6:27 PM (unsaved changes)

In [16]:

```
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-1:277607018592:endpoint-config/promoteshadow-epconfig-2023-12-13-03-46-03
```



Data capture settings

Enable data capture	Current sampling percentage (%)	S3 location to store data collected	Data capture status
No	-		-

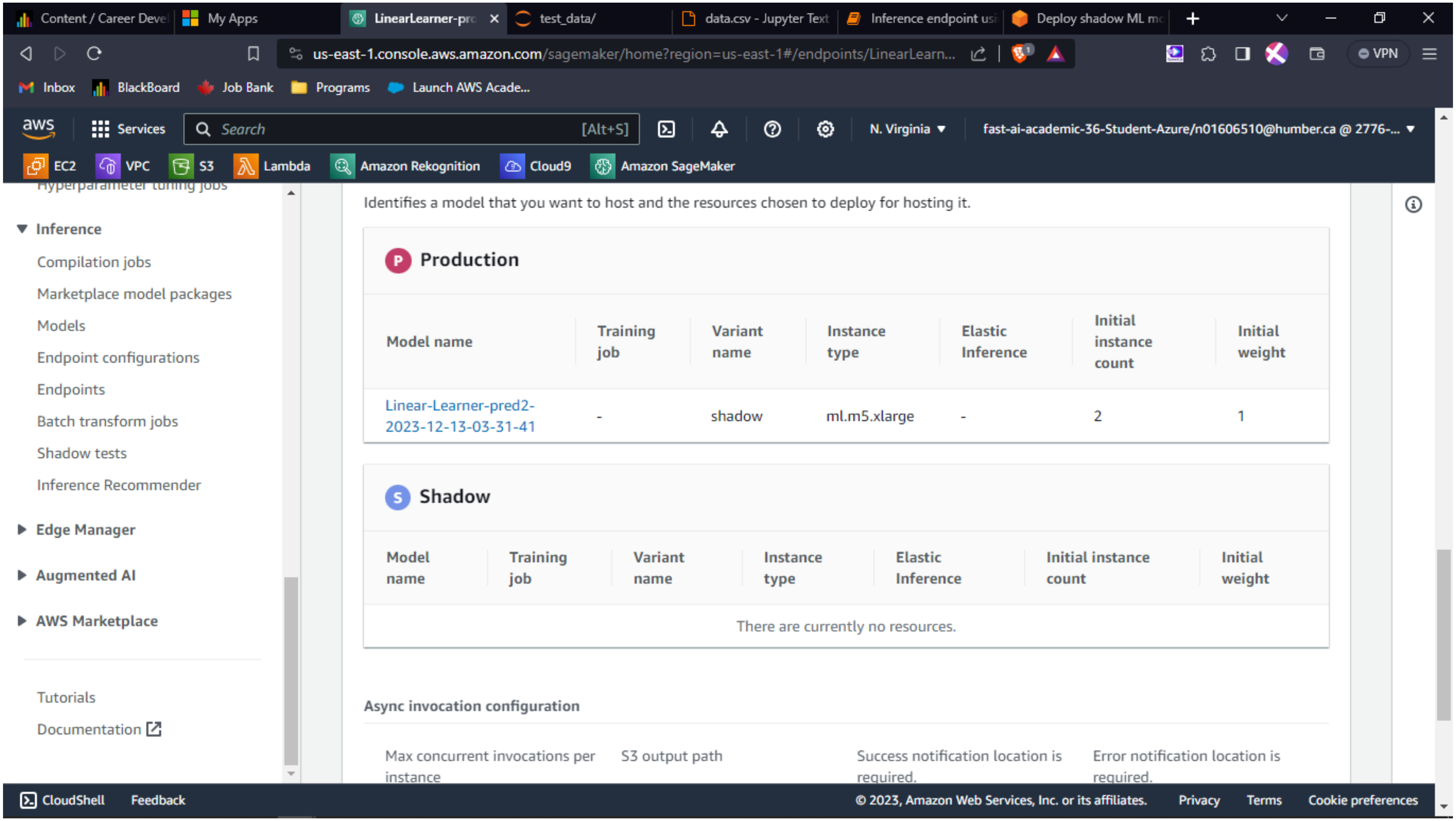
Endpoint runtime settings

		Variant name	Current weight	Desired weight	Elastic Inference	Instance type	Current ins
○	P	shadow	1	1	-	ml.m5.xlarge	2

Endpoint configuration settings

Endpoint configuration

Name	ARN	Encryption key	Creation time



Identifies a model that you want to host and the resources chosen to deploy for hosting it.

P Production

Model name	Training job	Variant name	Instance type	Elastic Inference	Initial instance count	Initial weight
Linear-Learner-pred2-2023-12-13-03-31-41	-	shadow	ml.m5.xlarge	-	2	1

S Shadow

Model name	Training job	Variant name	Instance type	Elastic Inference	Initial instance count	Initial weight
There are currently no resources.						

Async invocation configuration

Max concurrent invocations per instance	S3 output path	Success notification location is required.	Error notification location is required.