AWS ML Heroes in 15:

Detect Bias in ML Data, Models & Explain Predictions

Kesha Williams

AWS ML Hero & Program Director, AWS Cloud Residency
Slalom

© 2023, Amazon Web Services, Inc. or its affiliates.

Agenda

Understand bias in machine learning

Amazon SageMaker Clarify to the rescue

Uncover bias in your data

Detect bias in your model

Summary

© 2023, Amazon Web Services, Inc. or its affiliates.

Understand bias in machine learning

© 2023, Amazon Web Services, Inc. or its affiliates.

What is bias?

- Prediction behavior less favorable to an individual or group when there is no relevant difference that justifies the prediction
- Bias surfaces throughout the machine learning lifecycle

© 2023, Amazon Web Services, Inc. or its affiliates.

Where can bias appear in the ML lifecycle?

Bias can appear throughout the lifecycle

- Bias can be found in the training data
- ML algorithms can introduce bias
- Model can become biased because of drift

© 2023, Amazon Web Services, Inc. or its affiliates.

Amazon SageMaker Clarify to the rescue

© 2023, Amazon Web Services, Inc. or its affiliates.

ML Lifecycle

Prepare training data

Clarify in Data Wrangler

- Measure bias in your data

Train and tune model

Clarify in Studio

- Measure bias in your model
- Explain model predictions

Deploy model to production

Monitor your model for drift

Clarify in Model Monitor

- Explain model predictions in real time
- Monitor model for drift

© 2023, Amazon Web Services, Inc. or its affiliates.

Amazon SageMaker Clarify

Detect bias in ML models and understand model predictions

Detect bias during preparation

Identify data imbalances

Check trained model for bias

Evaluate different type of bias in your model

Explain model behavior

Determine relative importance of features

Monitor model once in production

Detect drift in bias and model behavior

© 2023, Amazon Web Services, Inc. or its affiliates.

SageMaker Clarify Metrics

8 Pretraining bias metrics

- Class Imbalance (CI)

  Measures if facet values have equal representation.

- Difference in Proportions of Labels (DPL)

  Measures the distribution of positive outcomes

13 Post-training bias metrics

- Accuracy Difference (AD)

  Measures accuracy across the facets

- Difference in Acceptance Rate (DAR)

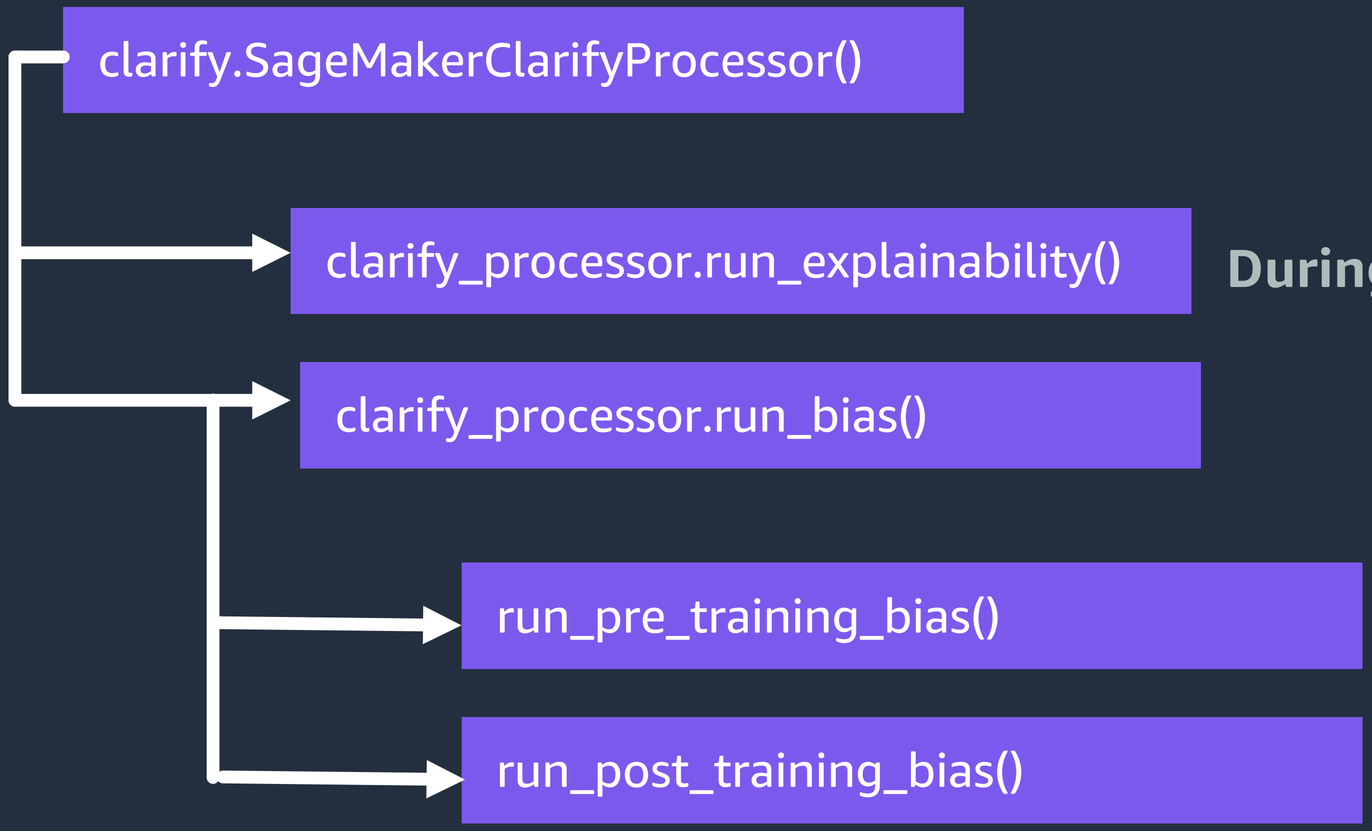
  Measures whether acceptance rate is the same across two facets

- Recall Difference (RD)

  Measures recall difference between the facets

© 2023, Amazon Web Services, Inc. or its affiliates.

SageMaker Clarify APIs



clarify.SageMakerClarifyProcessor()

clarify_processor.run_explainability() During model testing

clarify_processor.run_bias()

run_pre_training_bias() During data preparation

run_post_training_bias() After model training

© 2023, Amazon Web Services, Inc. or its affiliates.

Uncover bias in your data

© 2023, Amazon Web Services, Inc. or its affiliates.

Public Safety Model

Predicts if a stop will lead to an arrest.

Data: Stop and search data from

https://data.police.uk/data/

Problem Type: Binary classification

Algorithm: XGBoost

Photo by Krzysztof Hepner on Unsplash

© 2023, Amazon Web Service

The Data

Outcome	Date	Lat	Long	Gender	Age Range	Ethnicity
No action	2020-01-12T17:28:00+00:00	50.372336	-4.184631	Male	10-17	Black
No action	2020-01-23T00:10:00+00:00	50.720318	-3.610748	Female	18-24	White
Arrest	2020-01-24T01:08:00+00:00	50.372387	-4.143638	Male	25-34	White
No action	2020-01-24T11:40:00+00:00	50.391127	-3.534389	Male	Over 34	Mixed
Arrest	2020-01-24T11:45:00+00:00	50.528687	-3.770389	Male	10-17	Asian

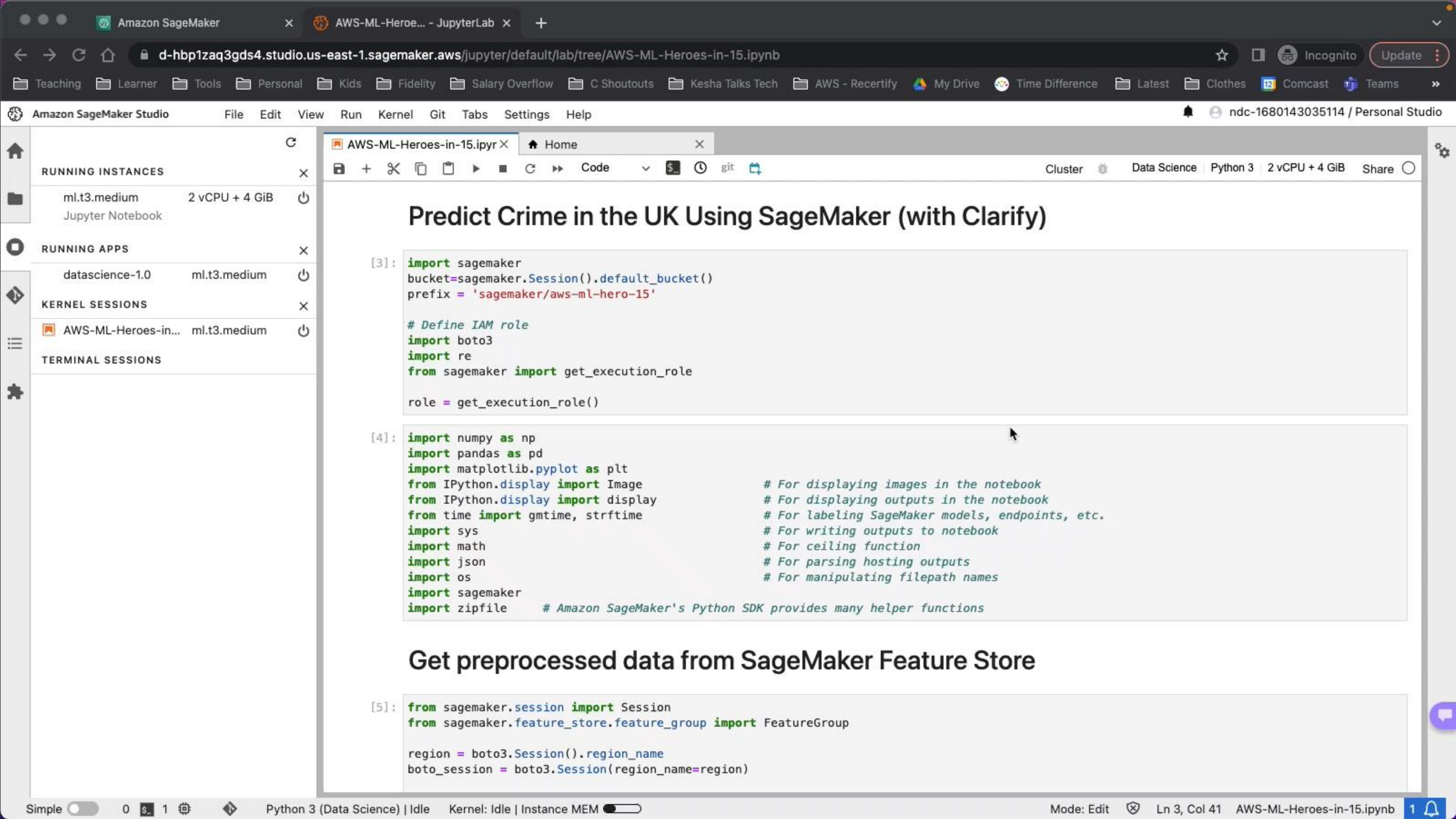
© 2023, Amazon Web Services, Inc. or its affiliates.

Pretraining Bias Metrics

- Determines if facet values have equal (or similar) representation in the data
- Indicates imbalances in the data

- Class Imbalance (CI)
  - Measures if facet values have equal representation.
- Difference in Proportions of Labels (DPL)
  - Measures the distribution of positive outcomes

© 2023, Amazon Web Services, Inc. or its affiliates.



Predict Crime in the UK Using SageMaker (with Clarify)

[3]:

```python
import sagemaker
bucket=sagemaker.Session().default_bucket()
prefix = 'sagemaker/aws-ml-hero-15'

# Define IAM role
import boto3
import re
from sagemaker import get_execution_role

role = get_execution_role()
```

[4]:

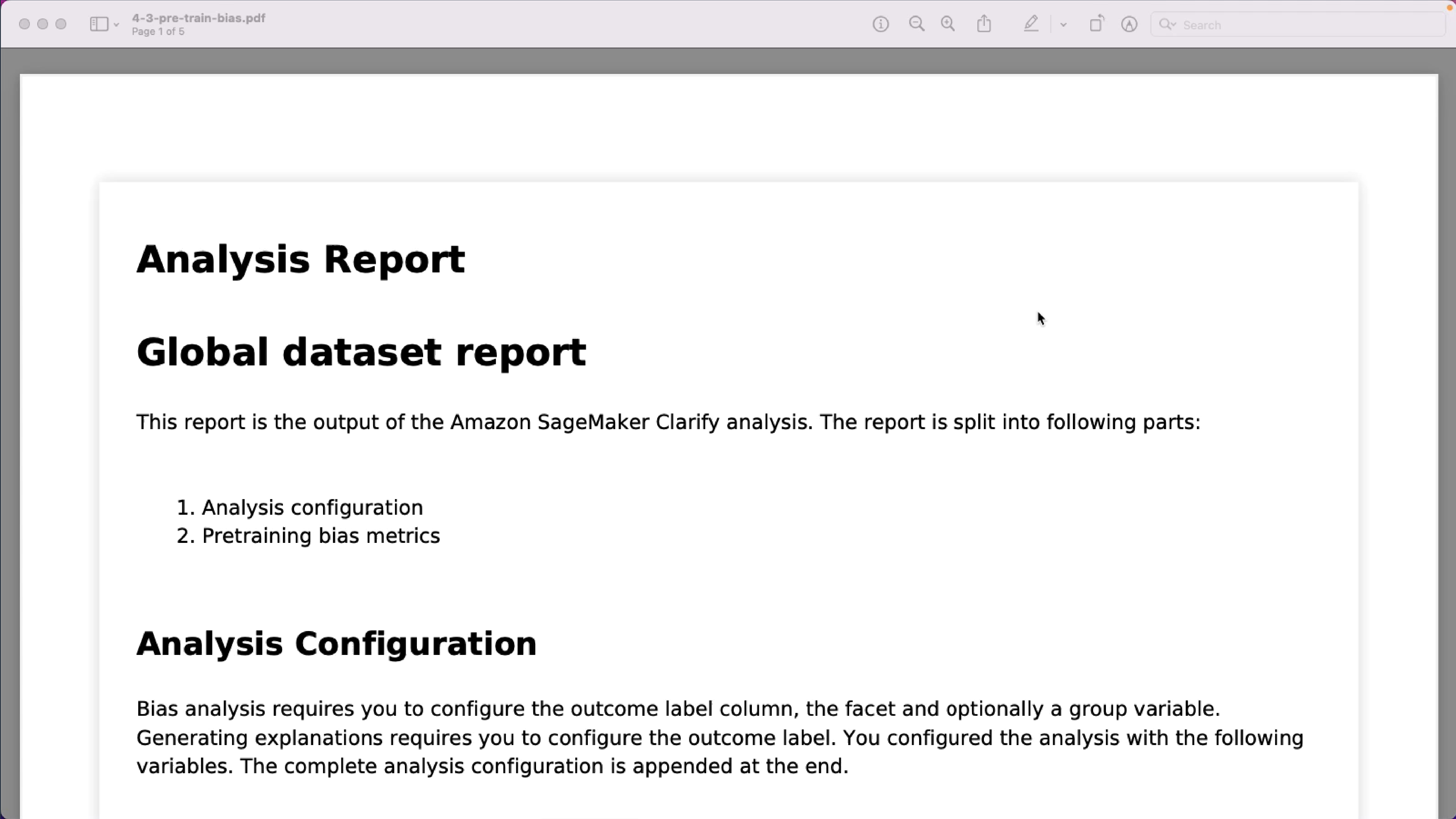
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from IPython.display import Image                 # For displaying images in the notebook
from IPython.display import display               # For displaying outputs in the notebook
from time import gmtime, strftime                 # For labeling SageMaker models, endpoints, etc.
import sys                                        # For writing outputs to notebook
import math                                       # For ceiling function
import json                                       # For parsing hosting outputs
import os                                         # For manipulating filepath names
import sagemaker
import zipfile     # Amazon SageMaker's Python SDK provides many helper functions
```

Get preprocessed data from SageMaker Feature Store

[5]:

```python
from sagemaker.session import Session
from sagemaker.feature_store.feature_group import FeatureGroup

region = boto3.Session().region_name
boto_session = boto3.Session(region_name=region)
```



Analysis Report

Global dataset report

This report is the output of the Amazon SageMaker Clarify analysis. The report is split into following parts:

1. Analysis configuration
2. Pretraining bias metrics

Analysis Configuration

Bias analysis requires you to configure the outcome label column, the facet and optionally a group variable. Generating explanations requires you to configure the outcome label. You configured the analysis with the following variables. The complete analysis configuration is appended at the end.

Detect bias in your model

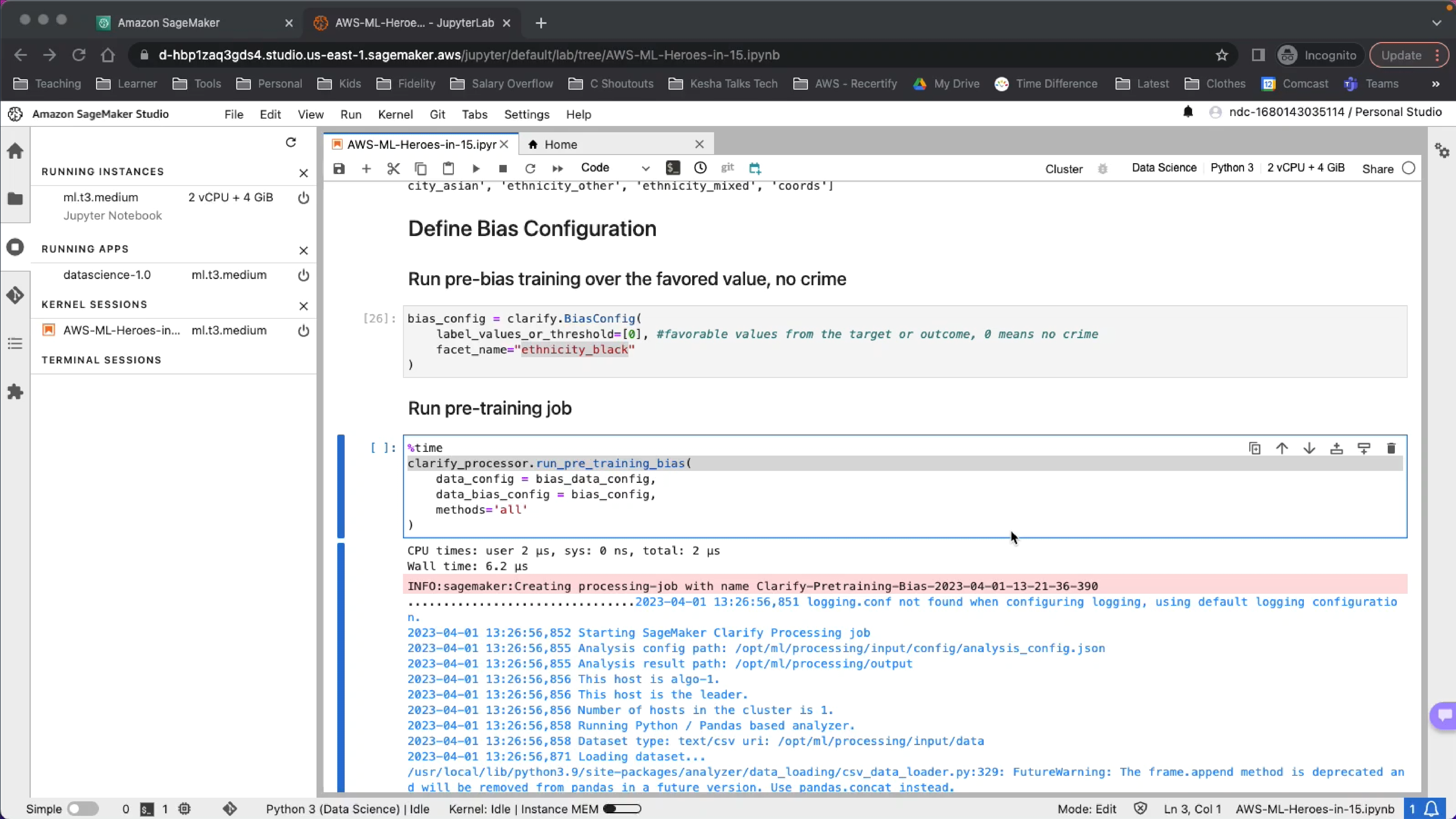
© 2023, Amazon Web Services, Inc. or its affiliates.

Posttraining Bias Metrics

- Determines if facet values represented at a similar rate in favorable model predictions
- Determines if the model has similar predictive power for all facet values
- Considers the predictions of the model

- Accuracy Difference (AD)
  - Measures accuracy across the facets
- Difference in Acceptance Rate (DAR)
  - Measures whether acceptance rate is the same across two facets
- Recall Difference (RD)
  - Measures recall difference between the facets

© 2023, Amazon Web Services, Inc. or its affiliates.



```
city_asian', 'ethnicity_other', 'ethnicity_mixed', 'coords']
```

Define Bias Configuration

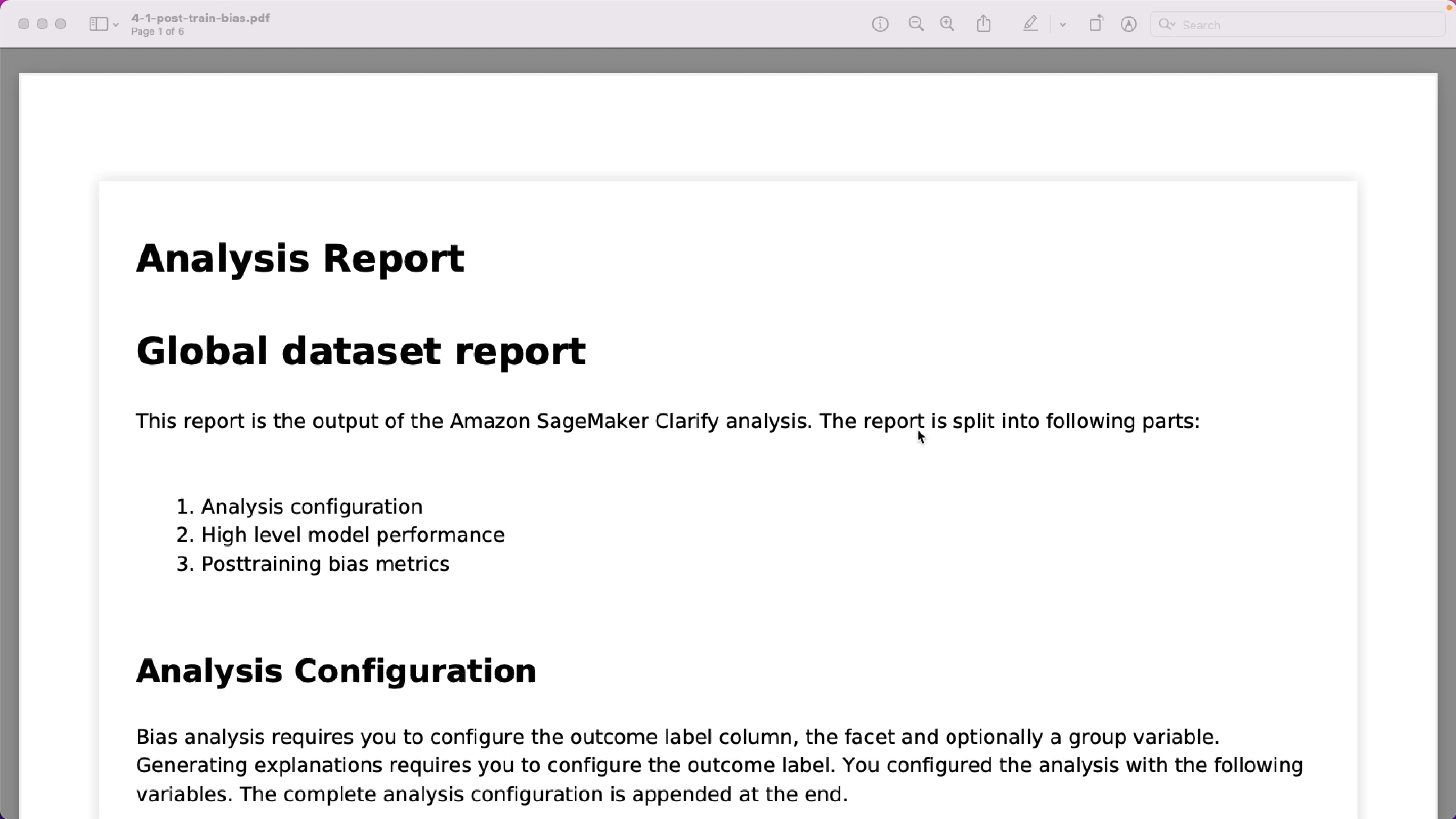
Run pre-bias training over the favored value, no crime

```
[26]: bias_config = clarify.BiasConfig(
          label_values_or_threshold=[0], #favorable values from the target or outcome, 0 means no crime
          facet_name="ethnicity_black"
      )
```

Run pre-training job

```
[ ]: %time
     clarify_processor.run_pre_training_bias(
         data_config = bias_data_config,
         data_bias_config = bias_config,
         methods='all'
     )
```

```
CPU times: user 2 µs, sys: 0 ns, total: 2 µs
Wall time: 6.2 µs
INFO:sagemaker:Creating processing-job with name Clarify-Pretraining-Bias-2023-04-01-13-21-36-390
................................2023-04-01 13:26:56,851 logging.conf not found when configuring logging, using default logging configuratio
n.
2023-04-01 13:26:56,852 Starting SageMaker Clarify Processing job
2023-04-01 13:26:56,855 Analysis config path: /opt/ml/processing/input/config/analysis_config.json
2023-04-01 13:26:56,855 Analysis result path: /opt/ml/processing/output
2023-04-01 13:26:56,856 This host is algo-1.
2023-04-01 13:26:56,856 This host is the leader.
2023-04-01 13:26:56,856 Number of hosts in the cluster is 1.
2023-04-01 13:26:56,858 Running Python / Pandas based analyzer.
2023-04-01 13:26:56,858 Dataset type: text/csv uri: /opt/ml/processing/input/data
2023-04-01 13:26:56,871 Loading dataset...
/usr/local/lib/python3.9/site-packages/analyzer/data_loading/csv_data_loader.py:329: FutureWarning: The frame.append method is deprecated an
d will be removed from pandas in a future version. Use pandas.concat instead.
```



Analysis Report

Global dataset report

This report is the output of the Amazon SageMaker Clarify analysis. The report is split into following parts:

1. Analysis configuration
2. High level model performance
3. Posttraining bias metrics

Analysis Configuration

Bias analysis requires you to configure the outcome label column, the facet and optionally a group variable. Generating explanations requires you to configure the outcome label. You configured the analysis with the following variables. The complete analysis configuration is appended at the end.

Summary

© 2023, Amazon Web Services, Inc. or its affiliates.

SageMaker Clarify

- Measure bias in your data
- Measure bias in your model
- Explain model predictions
- Monitor model for drift

© 2023, Amazon Web Services, Inc. or its affiliates.

Thank you!

Kesha Williams

@KeshaWillz

© 2023, Amazon Web Services, Inc. or its affiliates.