02_model_training_and_deployment

September 1, 2025

[ ]:
```
# Cell 1: Environment Setup & Load Prepared Data
import sagemaker
import boto3
import pandas as pd

# Initialize SageMaker session and get execution role
sagemaker_session = sagemaker.Session()
role = sagemaker.get_execution_role()
bucket = sagemaker_session.default_bucket()
prefix = 'iot-intrusion-detection' # The main project folder

print("--- Loading prepared feature set from S3 ---")
print(bucket)
# Define the path to the feature set created by the first notebook
feature_path = f"s3://{bucket}/{prefix}/features/combined_features.parquet"

# Load the data into a DataFrame
combined_features_df = pd.read_parquet(feature_path)

print(f"Successfully loaded {len(combined_features_df)} records.")
display(combined_features_df.head())
```

```
sagemaker.config INFO - Not applying SDK defaults from location:
/etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location:
/home/sagemaker-user/.config/sagemaker/config.yaml
--- Loading prepared feature set from S3 ---
sagemaker-us-east-2-696680564117
Successfully loaded 149 records.
```

	device_ip	timestamp	orig_bytes_sum	resp_bytes_sum	\
0	172.31.0.1	2025-08-30 21:01:00+00:00	1792.0	0.0	
1	172.31.0.1	2025-08-30 21:02:00+00:00	0.0	0.0	
2	172.31.0.1	2025-08-30 21:03:00+00:00	1792.0	0.0	
3	172.31.0.1	2025-08-30 21:04:00+00:00	0.0	0.0	
4	172.31.0.1	2025-08-30 21:05:00+00:00	1792.0	0.0	

	orig_pkts_sum	resp_pkts_sum	duration_mean	unique_dest_ips	\

```
0              12              0       0.795472                1
1               0              0       0.000000                0
2              12              0       0.793615                1
3               0              0       0.000000                0
4              12              0       0.792963                1

   unique_dest_ports  conn_count  alert_count  unique_alert_signatures
0                  1           2          0.0                      0.0
1                  0           0          0.0                      0.0
2                  1           2          0.0                      0.0
3                  0           0          0.0                      0.0
4                  1           2          0.0                      0.0
```

[ ]:
```python
# Cell 2: Configure and Launch SageMaker Training Job (Corrected)
from io import StringIO
from sagemaker.amazon.amazon_estimator import get_image_uri
import sagemaker
import boto3

# 1. Select the numeric feature columns for the model
features_for_model = [
    'orig_bytes_sum', 'resp_bytes_sum', 'orig_pkts_sum', 'resp_pkts_sum',
    'duration_mean', 'unique_dest_ips', 'unique_dest_ports', 'conn_count',
    'alert_count', 'unique_alert_signatures'
]
training_data = combined_features_df[features_for_model].astype('float32')

# 2. Convert to CSV
csv_buffer = StringIO()
training_data.to_csv(csv_buffer, header=False, index=False)
csv_content = csv_buffer.getvalue()

# --- FIX IS HERE ---
# First, define the full S3 path where the data will be uploaded
training_s3_path = f"s3://{bucket}/{prefix}/rcf-training-data/train.csv"

# Now, upload the CSV string to that single, full path
sagemaker.s3.S3Uploader.upload_string_as_file_body(csv_content,␣
 ↪training_s3_path)
print(f"Training data uploaded to: {training_s3_path}")

# 3. Configure the SageMaker Estimator
rcf_image = get_image_uri(boto3.Session().region_name, "randomcutforest")
rcf = sagemaker.estimator.Estimator(
    image_uri=rcf_image,
    role=role, # <-- FIX: Was 'role-role'
    instance_count=1, # <-- FIX: Was 'instance_count := 1'
```

```
    instance_type='ml.m5.xlarge',
    output_path=f"s3://{bucket}/{prefix}/rcf-output",
    sagemaker_session=sagemaker_session # <-- FIX: Was missing '='
)

# 4. Set hyperparameters
rcf.set_hyperparameters(
    num_samples_per_tree=256,
    num_trees=100,
    feature_dim=len(features_for_model)
)

# 5. Launch the training job
s3_input_train = sagemaker.inputs.TrainingInput(s3_data=training_s3_path,␣
 ↪content_type='text/csv;label_size=0', distribution='ShardedByS3Key')

rcf.fit({'train': s3_input_train})
```

```
The method get_image_uri has been renamed in sagemaker>=2.
See: https://sagemaker.readthedocs.io/en/stable/v2.html for details.
INFO:sagemaker:Creating training-job with name:
randomcutforest-2025-09-01-22-01-10-444

Training data uploaded to: s3://sagemaker-us-east-2-696680564117/iot-intrusion-
detection/rcf-training-data/train.csv
2025-09-01 22:01:12 Starting - Starting the training job…
2025-09-01 22:01:27 Starting - Preparing the instances for training…
2025-09-01 22:02:07 Downloading - Downloading the training image…Docker
entrypoint called with argument(s): train
Running default environment configuration script
/opt/amazon/lib/python3.8/site-packages/mxnet/model.py:97: SyntaxWarning: "is"
with a literal. Did you mean "=="?
  if num_device is 1 and 'dist' not in kvstore:
/opt/amazon/lib/python3.8/site-packages/scipy/optimize/_shgo.py:495:
SyntaxWarning: "is" with a literal. Did you mean "=="?
  if cons['type'] is 'ineq':
/opt/amazon/lib/python3.8/site-packages/scipy/optimize/_shgo.py:743:
SyntaxWarning: "is not" with a literal. Did you mean "!="?
  if len(self.X_min) is not 0:
[09/01/2025 22:03:31 INFO 139630730504000] Reading default configuration from
/opt/amazon/lib/python3.8/site-packages/algorithm/resources/default-conf.json:
{'num_samples_per_tree': 256, 'num_trees': 100, 'force_dense': 'true',
'eval_metrics': ['accuracy', 'precision_recall_fscore'], 'epochs': 1,
'mini_batch_size': 1000, '_log_level': 'info', '_kvstore': 'dist_async',
'_num_kv_servers': 'auto', '_num_gpus': 'auto', '_tuning_objective_metric': '',
'_ftp_port': 8999}
[09/01/2025 22:03:31 INFO 139630730504000] Merging with provided configuration
from /opt/ml/input/config/hyperparameters.json: {'feature_dim': '10',
```

'num_samples_per_tree': '256', 'num_trees': '100'}
[09/01/2025 22:03:31 INFO 139630730504000] Final configuration:
{'num_samples_per_tree': '256', 'num_trees': '100', 'force_dense': 'true',
'eval_metrics': ['accuracy', 'precision_recall_fscore'], 'epochs': 1,
'mini_batch_size': 1000, '_log_level': 'info', '_kvstore': 'dist_async',
'_num_kv_servers': 'auto', '_num_gpus': 'auto', '_tuning_objective_metric': '',
'_ftp_port': 8999, 'feature_dim': '10'}
[09/01/2025 22:03:31 WARNING 139630730504000] Loggers have already been setup.
[09/01/2025 22:03:31 INFO 139630730504000] Launching parameter server for role
scheduler
[09/01/2025 22:03:31 INFO 139630730504000] {'ENVROOT': '/opt/amazon',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION': 'cpp', 'HOSTNAME': 'ip-10-0-246-34.us-
east-2.compute.internal', 'TRAINING_JOB_NAME':
'randomcutforest-2025-09-01-22-01-10-444', 'NVIDIA_REQUIRE_CUDA': 'cuda>=9.0',
'TRAINING_JOB_ARN': 'arn:aws:sagemaker:us-east-2:696680564117:training-
job/randomcutforest-2025-09-01-22-01-10-444',
'AWS_CONTAINER_CREDENTIALS_RELATIVE_URI': '/v2/credentials/proxy-
bae370e84fba2bea1738f799b00fa19f9a7235c5fe8e1c7c002be6b6ee3f061a-customer',
'CANONICAL_ENVROOT': '/opt/amazon', 'PYTHONUNBUFFERED': 'TRUE',
'NVIDIA_VISIBLE_DEVICES': 'all', 'LD_LIBRARY_PATH':
'/opt/amazon/lib/python3.8/site-
packages/cv2/../../../../lib:/usr/local/nvidia/lib64:/opt/amazon/lib',
'MXNET_KVSTORE_BIGARRAY_BOUND': '400000000', 'NVIDIA_DRIVER_CAPABILITIES':
'compute,utility', 'SAGEMAKER_MANAGED_WARMPOOL_CACHE_DIRECTORY':
'/opt/ml/sagemaker/warmpoolcache', 'PATH': '/opt/amazon/bin:/usr/local/nvidia/bi
n:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin', 'PWD': '/',
'LANG': 'en_US.utf8', 'AWS_REGION': 'us-east-2', 'SAGEMAKER_METRICS_DIRECTORY':
'/opt/ml/output/metrics/sagemaker', 'CUDA_VERSION': '11.1', 'HOME': '/root',
'SHLVL': '1', 'CUDA_COMPAT_NDRIVER_SUPPORTED_VERSION': '455.32.00',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION_VERSION': '2', 'OMP_NUM_THREADS': '2',
'DMLC_INTERFACE': 'eth0', 'SAGEMAKER_HTTP_PORT': '8080', 'SAGEMAKER_DATA_PATH':
'/opt/ml', 'KMP_DUPLICATE_LIB_OK': 'True', 'KMP_INIT_AT_FORK': 'FALSE'}
[09/01/2025 22:03:31 INFO 139630730504000] envs={'ENVROOT': '/opt/amazon',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION': 'cpp', 'HOSTNAME': 'ip-10-0-246-34.us-
east-2.compute.internal', 'TRAINING_JOB_NAME':
'randomcutforest-2025-09-01-22-01-10-444', 'NVIDIA_REQUIRE_CUDA': 'cuda>=9.0',
'TRAINING_JOB_ARN': 'arn:aws:sagemaker:us-east-2:696680564117:training-
job/randomcutforest-2025-09-01-22-01-10-444',
'AWS_CONTAINER_CREDENTIALS_RELATIVE_URI': '/v2/credentials/proxy-
bae370e84fba2bea1738f799b00fa19f9a7235c5fe8e1c7c002be6b6ee3f061a-customer',
'CANONICAL_ENVROOT': '/opt/amazon', 'PYTHONUNBUFFERED': 'TRUE',
'NVIDIA_VISIBLE_DEVICES': 'all', 'LD_LIBRARY_PATH':
'/opt/amazon/lib/python3.8/site-
packages/cv2/../../../../lib:/usr/local/nvidia/lib64:/opt/amazon/lib',
'MXNET_KVSTORE_BIGARRAY_BOUND': '400000000', 'NVIDIA_DRIVER_CAPABILITIES':
'compute,utility', 'SAGEMAKER_MANAGED_WARMPOOL_CACHE_DIRECTORY':
'/opt/ml/sagemaker/warmpoolcache', 'PATH': '/opt/amazon/bin:/usr/local/nvidia/bi
n:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin', 'PWD': '/',

'LANG': 'en_US.utf8', 'AWS_REGION': 'us-east-2', 'SAGEMAKER_METRICS_DIRECTORY':
'/opt/ml/output/metrics/sagemaker', 'CUDA_VERSION': '11.1', 'HOME': '/root',
'SHLVL': '1', 'CUDA_COMPAT_NDRIVER_SUPPORTED_VERSION': '455.32.00',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION_VERSION': '2', 'OMP_NUM_THREADS': '2',
'DMLC_INTERFACE': 'eth0', 'SAGEMAKER_HTTP_PORT': '8080', 'SAGEMAKER_DATA_PATH':
'/opt/ml', 'KMP_DUPLICATE_LIB_OK': 'True', 'KMP_INIT_AT_FORK': 'FALSE',
'DMLC_ROLE': 'scheduler', 'DMLC_PS_ROOT_URI': '10.0.246.34',
'DMLC_PS_ROOT_PORT': '9000', 'DMLC_NUM_SERVER': '1', 'DMLC_NUM_WORKER': '1'}
[09/01/2025 22:03:31 INFO 139630730504000] Launching parameter server for role
server
[09/01/2025 22:03:31 INFO 139630730504000] {'ENVROOT': '/opt/amazon',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION': 'cpp', 'HOSTNAME': 'ip-10-0-246-34.us-
east-2.compute.internal', 'TRAINING_JOB_NAME':
'randomcutforest-2025-09-01-22-01-10-444', 'NVIDIA_REQUIRE_CUDA': 'cuda>=9.0',
'TRAINING_JOB_ARN': 'arn:aws:sagemaker:us-east-2:696680564117:training-
job/randomcutforest-2025-09-01-22-01-10-444',
'AWS_CONTAINER_CREDENTIALS_RELATIVE_URI': '/v2/credentials/proxy-
bae370e84fba2bea1738f799b00fa19f9a7235c5fe8e1c7c002be6b6ee3f061a-customer',
'CANONICAL_ENVROOT': '/opt/amazon', 'PYTHONUNBUFFERED': 'TRUE',
'NVIDIA_VISIBLE_DEVICES': 'all', 'LD_LIBRARY_PATH':
'/opt/amazon/lib/python3.8/site-
packages/cv2/../../../../lib:/usr/local/nvidia/lib64:/opt/amazon/lib',
'MXNET_KVSTORE_BIGARRAY_BOUND': '400000000', 'NVIDIA_DRIVER_CAPABILITIES':
'compute,utility', 'SAGEMAKER_MANAGED_WARMPOOL_CACHE_DIRECTORY':
'/opt/ml/sagemaker/warmpoolcache', 'PATH': '/opt/amazon/bin:/usr/local/nvidia/bi
n:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin', 'PWD': '/',
'LANG': 'en_US.utf8', 'AWS_REGION': 'us-east-2', 'SAGEMAKER_METRICS_DIRECTORY':
'/opt/ml/output/metrics/sagemaker', 'CUDA_VERSION': '11.1', 'HOME': '/root',
'SHLVL': '1', 'CUDA_COMPAT_NDRIVER_SUPPORTED_VERSION': '455.32.00',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION_VERSION': '2', 'OMP_NUM_THREADS': '2',
'DMLC_INTERFACE': 'eth0', 'SAGEMAKER_HTTP_PORT': '8080', 'SAGEMAKER_DATA_PATH':
'/opt/ml', 'KMP_DUPLICATE_LIB_OK': 'True', 'KMP_INIT_AT_FORK': 'FALSE'}
[09/01/2025 22:03:31 INFO 139630730504000] envs={'ENVROOT': '/opt/amazon',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION': 'cpp', 'HOSTNAME': 'ip-10-0-246-34.us-
east-2.compute.internal', 'TRAINING_JOB_NAME':
'randomcutforest-2025-09-01-22-01-10-444', 'NVIDIA_REQUIRE_CUDA': 'cuda>=9.0',
'TRAINING_JOB_ARN': 'arn:aws:sagemaker:us-east-2:696680564117:training-
job/randomcutforest-2025-09-01-22-01-10-444',
'AWS_CONTAINER_CREDENTIALS_RELATIVE_URI': '/v2/credentials/proxy-
bae370e84fba2bea1738f799b00fa19f9a7235c5fe8e1c7c002be6b6ee3f061a-customer',
'CANONICAL_ENVROOT': '/opt/amazon', 'PYTHONUNBUFFERED': 'TRUE',
'NVIDIA_VISIBLE_DEVICES': 'all', 'LD_LIBRARY_PATH':
'/opt/amazon/lib/python3.8/site-
packages/cv2/../../../../lib:/usr/local/nvidia/lib64:/opt/amazon/lib',
'MXNET_KVSTORE_BIGARRAY_BOUND': '400000000', 'NVIDIA_DRIVER_CAPABILITIES':
'compute,utility', 'SAGEMAKER_MANAGED_WARMPOOL_CACHE_DIRECTORY':
'/opt/ml/sagemaker/warmpoolcache', 'PATH': '/opt/amazon/bin:/usr/local/nvidia/bi
n:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin', 'PWD': '/',

'LANG': 'en_US.utf8', 'AWS_REGION': 'us-east-2', 'SAGEMAKER_METRICS_DIRECTORY':
'/opt/ml/output/metrics/sagemaker', 'CUDA_VERSION': '11.1', 'HOME': '/root',
'SHLVL': '1', 'CUDA_COMPAT_NDRIVER_SUPPORTED_VERSION': '455.32.00',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION_VERSION': '2', 'OMP_NUM_THREADS': '2',
'DMLC_INTERFACE': 'eth0', 'SAGEMAKER_HTTP_PORT': '8080', 'SAGEMAKER_DATA_PATH':
'/opt/ml', 'KMP_DUPLICATE_LIB_OK': 'True', 'KMP_INIT_AT_FORK': 'FALSE',
'DMLC_ROLE': 'server', 'DMLC_PS_ROOT_URI': '10.0.246.34', 'DMLC_PS_ROOT_PORT':
'9000', 'DMLC_NUM_SERVER': '1', 'DMLC_NUM_WORKER': '1'}
[09/01/2025 22:03:31 INFO 139630730504000] Environment: {'ENVROOT':
'/opt/amazon', 'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION': 'cpp', 'HOSTNAME':
'ip-10-0-246-34.us-east-2.compute.internal', 'TRAINING_JOB_NAME':
'randomcutforest-2025-09-01-22-01-10-444', 'NVIDIA_REQUIRE_CUDA': 'cuda>=9.0',
'TRAINING_JOB_ARN': 'arn:aws:sagemaker:us-east-2:696680564117:training-
job/randomcutforest-2025-09-01-22-01-10-444',
'AWS_CONTAINER_CREDENTIALS_RELATIVE_URI': '/v2/credentials/proxy-
bae370e84fba2bea1738f799b00fa19f9a7235c5fe8e1c7c002be6b6ee3f061a-customer',
'CANONICAL_ENVROOT': '/opt/amazon', 'PYTHONUNBUFFERED': 'TRUE',
'NVIDIA_VISIBLE_DEVICES': 'all', 'LD_LIBRARY_PATH':
'/opt/amazon/lib/python3.8/site-
packages/cv2/../../../../lib:/usr/local/nvidia/lib64:/opt/amazon/lib',
'MXNET_KVSTORE_BIGARRAY_BOUND': '400000000', 'NVIDIA_DRIVER_CAPABILITIES':
'compute,utility', 'SAGEMAKER_MANAGED_WARMPOOL_CACHE_DIRECTORY':
'/opt/ml/sagemaker/warmpoolcache', 'PATH': '/opt/amazon/bin:/usr/local/nvidia/bi
n:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin', 'PWD': '/',
'LANG': 'en_US.utf8', 'AWS_REGION': 'us-east-2', 'SAGEMAKER_METRICS_DIRECTORY':
'/opt/ml/output/metrics/sagemaker', 'CUDA_VERSION': '11.1', 'HOME': '/root',
'SHLVL': '1', 'CUDA_COMPAT_NDRIVER_SUPPORTED_VERSION': '455.32.00',
'PROTOCOL_BUFFERS_PYTHON_IMPLEMENTATION_VERSION': '2', 'OMP_NUM_THREADS': '2',
'DMLC_INTERFACE': 'eth0', 'SAGEMAKER_HTTP_PORT': '8080', 'SAGEMAKER_DATA_PATH':
'/opt/ml', 'KMP_DUPLICATE_LIB_OK': 'True', 'KMP_INIT_AT_FORK': 'FALSE',
'DMLC_ROLE': 'worker', 'DMLC_PS_ROOT_URI': '10.0.246.34', 'DMLC_PS_ROOT_PORT':
'9000', 'DMLC_NUM_SERVER': '1', 'DMLC_NUM_WORKER': '1'}
Process 35 is a shell:scheduler.
Process 44 is a shell:server.
Process 8 is a worker.
[09/01/2025 22:03:31 INFO 139630730504000] Using default worker.
[09/01/2025 22:03:31 INFO 139630730504000] Loaded iterator creator
application/x-recordio-protobuf for content type ('application/x-recordio-
protobuf', '1.0')
[09/01/2025 22:03:31 INFO 139630730504000] Checkpoint loading and saving are
disabled.
[09/01/2025 22:03:31 INFO 139630730504000] Verifying hyperparamemters…
[09/01/2025 22:03:31 INFO 139630730504000] Hyperparameters are correct.
[09/01/2025 22:03:31 INFO 139630730504000] Validating that feature_dim agrees
with dimensions in training data…
[09/01/2025 22:03:31 INFO 139630730504000] feature_dim is correct.
[09/01/2025 22:03:31 INFO 139630730504000] Validating memory limits…
[09/01/2025 22:03:31 INFO 139630730504000] Available memory in bytes:

15004098560
[09/01/2025 22:03:31 INFO 139630730504000] Estimated sample size in bytes:
2048000
[09/01/2025 22:03:31 INFO 139630730504000] Estimated memory needed to build the
forest in bytes: 10240000
[09/01/2025 22:03:31 INFO 139630730504000] Memory limits validated.
[09/01/2025 22:03:31 INFO 139630730504000] Starting cluster sharing
facilities…
[09/01/2025 22:03:31 INFO 139628263417600] concurrency model: async
[09/01/2025 22:03:31 INFO 139630730504000] Create Store: dist_async
[09/01/2025 22:03:31 INFO 139628263417600] masquerade (NAT) address: None
[09/01/2025 22:03:31 INFO 139628263417600] passive ports: None
[09/01/2025 22:03:31 INFO 139628263417600] >>> starting FTP server on
0.0.0.0:8999, pid=8 <<<
[09/01/2025 22:03:32 INFO 139630730504000] Cluster sharing facilities started.
[09/01/2025 22:03:32 INFO 139630730504000] Verifying all workers are
accessible…
[09/01/2025 22:03:32 INFO 139630730504000] All workers accessible.
[09/01/2025 22:03:32 INFO 139630730504000] Initializing Sampler…
[09/01/2025 22:03:32 INFO 139630730504000] Sampler correctly initialized.
#metrics {"StartTime": 1756764211.4041913, "EndTime": 1756764212.4254599,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training"}, "Metrics": {"initialize.time": {"sum": 1020.9815502166748, "count":
1, "min": 1020.9815502166748, "max": 1020.9815502166748}}}
#metrics {"StartTime": 1756764212.4256122, "EndTime": 1756764212.4256432,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training", "Meta": "init_train_data_iter"}, "Metrics": {"Total Records Seen":
{"sum": 0.0, "count": 1, "min": 0, "max": 0}, "Total Batches Seen": {"sum": 0.0,
"count": 1, "min": 0, "max": 0}, "Max Records Seen Between Resets": {"sum": 0.0,
"count": 1, "min": 0, "max": 0}, "Max Batches Seen Between Resets": {"sum": 0.0,
"count": 1, "min": 0, "max": 0}, "Reset Count": {"sum": 0.0, "count": 1, "min":
0, "max": 0}, "Number of Records Since Last Reset": {"sum": 0.0, "count": 1,
"min": 0, "max": 0}, "Number of Batches Since Last Reset": {"sum": 0.0, "count":
1, "min": 0, "max": 0}}}
[09/01/2025 22:03:32 INFO 139630730504000] Sampling training data…
[09/01/2025 22:03:32 INFO 139630730504000] Sampling training data completed.
#metrics {"StartTime": 1756764212.42556, "EndTime": 1756764212.4402416,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training"}, "Metrics": {"epochs": {"sum": 1.0, "count": 1, "min": 1, "max": 1},
"update.time": {"sum": 14.405250549316406, "count": 1, "min":
14.405250549316406, "max": 14.405250549316406}}}
[09/01/2025 22:03:32 INFO 139630730504000] Early stop condition met. Stopping
training.
[09/01/2025 22:03:32 INFO 139630730504000] #progress_metric: host=algo-1,
completed 100 % epochs
#metrics {"StartTime": 1756764212.4258134, "EndTime": 1756764212.440445,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training", "epoch": 0, "Meta": "training_data_iter"}, "Metrics": {"Total

Records Seen": {"sum": 149.0, "count": 1, "min": 149, "max": 149}, "Total
Batches Seen": {"sum": 1.0, "count": 1, "min": 1, "max": 1}, "Max Records Seen
Between Resets": {"sum": 149.0, "count": 1, "min": 149, "max": 149}, "Max
Batches Seen Between Resets": {"sum": 1.0, "count": 1, "min": 1, "max": 1},
"Reset Count": {"sum": 1.0, "count": 1, "min": 1, "max": 1}, "Number of Records
Since Last Reset": {"sum": 149.0, "count": 1, "min": 149, "max": 149}, "Number
of Batches Since Last Reset": {"sum": 1.0, "count": 1, "min": 1, "max": 1}}}
[09/01/2025 22:03:32 INFO 139630730504000] #throughput_metric: host=algo-1,
train throughput=10125.916199487994 records/second
[09/01/2025 22:03:32 INFO 139630730504000] Master node: building Random Cut
Forest…
[09/01/2025 22:03:32 INFO 139630730504000] Gathering samples…
[09/01/2025 22:03:32 INFO 139630730504000] 149 samples gathered
[09/01/2025 22:03:32 INFO 139630730504000] Building Random Cut Forest…
[09/01/2025 22:03:32 INFO 139630730504000] Random Cut Forest built:
ForestInfo{num_trees: 100, num_samples_in_forest: 100, num_samples_per_tree: 1,
sample_dim: 10, shingle_size: 1, trees_num_nodes: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1,], trees_depth: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1,], max_num_nodes: 1, min_num_nodes: 1, avg_num_nodes: 1,
max_tree_depth: 1, min_tree_depth: 1, avg_tree_depth: 1, mem_size: 25648}
#metrics {"StartTime": 1756764212.4402997, "EndTime": 1756764212.4414628,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training"}, "Metrics": {"fit_model.time": {"sum": 0.25463104248046875, "count":
1, "min": 0.25463104248046875, "max": 0.25463104248046875}, "model.bytes":
{"sum": 25648.0, "count": 1, "min": 25648, "max": 25648}, "finalize.time":
{"sum": 0.7967948913574219, "count": 1, "min": 0.7967948913574219, "max":
0.7967948913574219}}}
[09/01/2025 22:03:32 INFO 139630730504000] Master node: Serializing the
RandomCutForest model
#metrics {"StartTime": 1756764212.441564, "EndTime": 1756764212.442215,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training"}, "Metrics": {"serialize_model.time": {"sum": 0.6248950958251953,
"count": 1, "min": 0.6248950958251953, "max": 0.6248950958251953}}}
[09/01/2025 22:03:32 INFO 139630730504000] Test data is not provided.
#metrics {"StartTime": 1756764212.4423044, "EndTime": 1756764212.4425132,
"Dimensions": {"Algorithm": "RandomCutForest", "Host": "algo-1", "Operation":
"training"}, "Metrics": {"setuptime": {"sum": 21.821975708007812, "count": 1,
"min": 21.821975708007812, "max": 21.821975708007812}, "totaltime": {"sum":
1064.8455619812012, "count": 1, "min": 1064.8455619812012, "max":
1064.8455619812012}}}

2025-09-01 22:03:46 Training - Training image download completed. Training in

```
progress.
2025-09-01 22:03:46 Uploading - Uploading generated training model
2025-09-01 22:03:46 Completed - Training job completed
Training seconds: 119
Billable seconds: 119
```

[ ]:
```python
# Cell 3: Deploy the Trained Model
rcf_predictor = rcf.deploy(
    initial_instance_count=1,
    instance_type='ml.t2.medium'
)
print("Model endpoint is now active.")
```

```
INFO:sagemaker:Creating model with name: randomcutforest-2025-09-01-22-03-57-575

INFO:sagemaker:Creating endpoint-config with name
randomcutforest-2025-09-01-22-03-57-575
INFO:sagemaker:Creating endpoint with name
randomcutforest-2025-09-01-22-03-57-575

-----
```

[ ]:
```python
# Cell 4: Test the Endpoint with Sample Data
import json
from sagemaker.serializers import CSVSerializer
from sagemaker.deserializers import JSONDeserializer

# Select a sample of data to test
sample_data = training_data.head(5).to_numpy()

# Set the correct serializer/deserializer for the RCF model
rcf_predictor.serializer = CSVSerializer()
rcf_predictor.deserializer = JSONDeserializer()

# Get predictions
results = rcf_predictor.predict(sample_data)
scores = [record['score'] for record in results['scores']]

print("Anomaly scores for sample data (higher is more anomalous):")
print(scores)
```

```
Anomaly scores for sample data (higher is more anomalous):
[0.0, 0.0, 0.0, 0.0, 0.0]
```