#AlwaysTrainingWeights&Biases

Group 5
Team Batak

"You can't spell MUSCLE without MLE."

- ssob Marc

apper.ph

Table of Contents

Meet the Team

Other SageMaker Capabilities
- ☐ Ground Truth
- ☐ Canvas

Amazon SageMaker Cookbook
- ☐ Chapter 4 Pre-Processing
- ☐ Chapter 5 Debugger & Experiments
- ☐ Chapter 6 Autopilot

Summary

apper.ph

Meet the Team

“What’s a perfect day for you?”

Edward Vincent **“VINCE”** Duero

A peaceful day with the fam

Rosiel Jazmine **“ROSE”** Villareal

Coffee convos & gig / chillnuman nights with friends

Jericho Carlo **“ECHO”** Agudo

Rest all day

apper.ph

Other SageMaker Capabilities

Ground Truth

Canvas

SageMaker Ground Truth

apper.ph

SageMaker Data Labeling

To train a machine learning model, you need a large, high-quality, **LABELED** dataset

Amazon SageMaker offers two options:

Ground Truth Plus

Allows you to create high-quality training datasets **without** having to build labeling or manage labeling workforces on your own

Ground Truth

Provides **flexibility** to build and manage your **own** data labeling workflows and workforce

SageMaker Ground Truth

SageMaker Ground Truth is a fully managed data labeling service that makes it easy to build highly accurate training datasets for machine learning

You can also generate labeled synthetic data without manually collecting or labeling real-world data

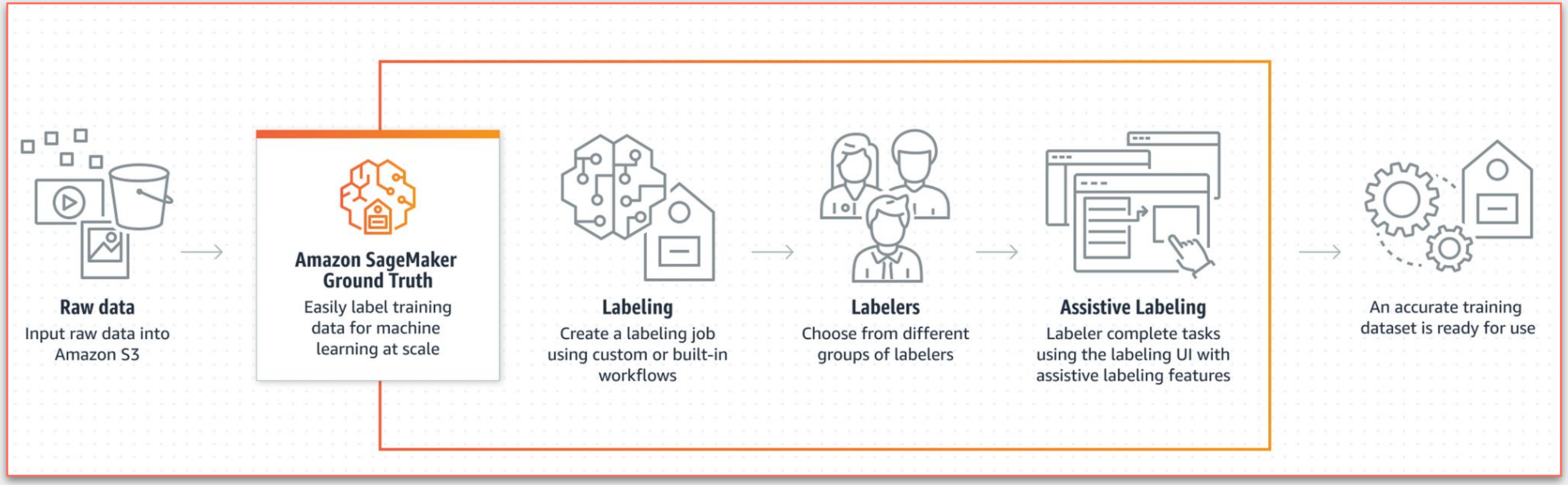
amazon mechanical turk

3rd Party Vendor

Own private workforce

SageMaker Ground Truth: How it Works

Labeling Data with SageMaker Ground Truth: Helps you build and manage your own data labeling workflows and data labeling workforce

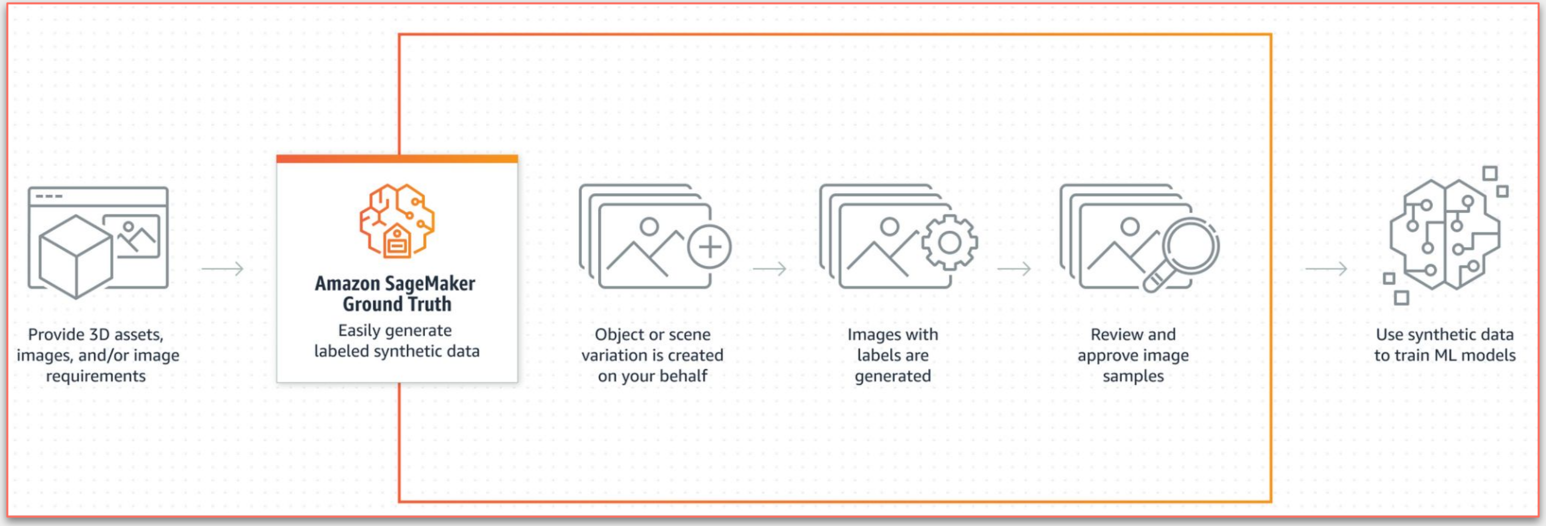


Raw data
Input raw data into Amazon S3
Amazon SageMaker Ground Truth
Easily label training data for machine learning at scale
Labeling
Create a labeling job using custom or built-in workflows
Labelers
Choose from different groups of labelers
Assistive Labeling
Labeler complete tasks using the labeling UI with assistive labeling features
An accurate training dataset is ready for use

apper.ph

SageMaker Ground Truth: How it Works

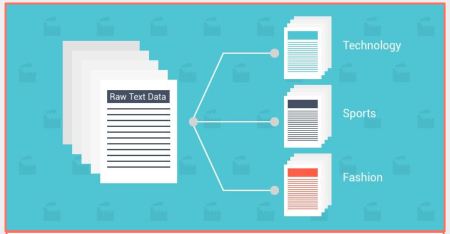
Generate Labeled Synthetic Data: Amazon SageMaker Ground Truth helps you generate labeled synthetic data


Provide 3D assets, images, and/or image requirements
Amazon SageMaker Ground Truth
Easily generate labeled synthetic data
Object or scene variation is created on your behalf
Images with labels are generated
Review and approve image samples
Use synthetic data to train ML models

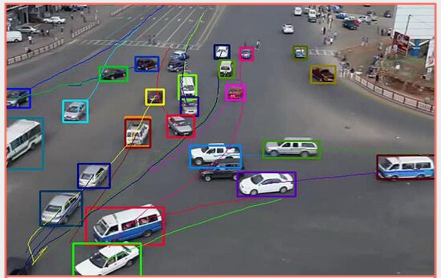
apper.ph

SageMaker Ground Truth: Built-in Tasks

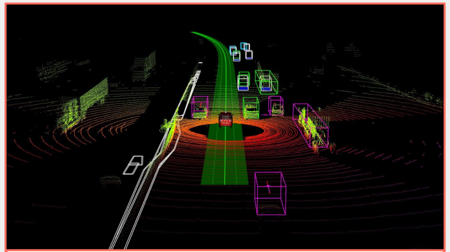
Label Images: Bounding Box



Label Text: Text Classification



Label Videos and Video Frames: Video Frame Object Tracking



Label 3D Point Clouds: 3D Point Cloud Object Detection

apper.ph

SageMaker Ground Truth: Benefits

Improve quality of training datasets

Choose your data labeling workforce

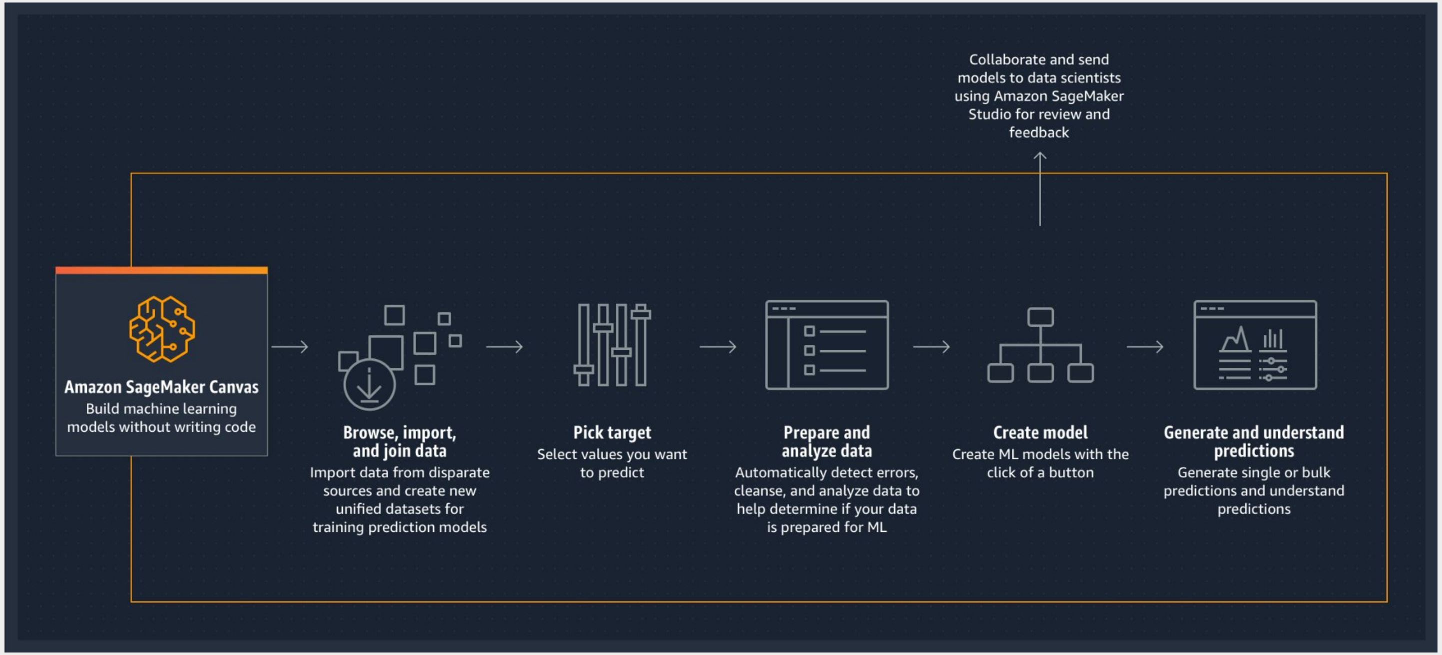
Increase visibility of data labeling operations

Receive high-quality labeled data quickly

SageMaker Canvas

SageMaker Canvas

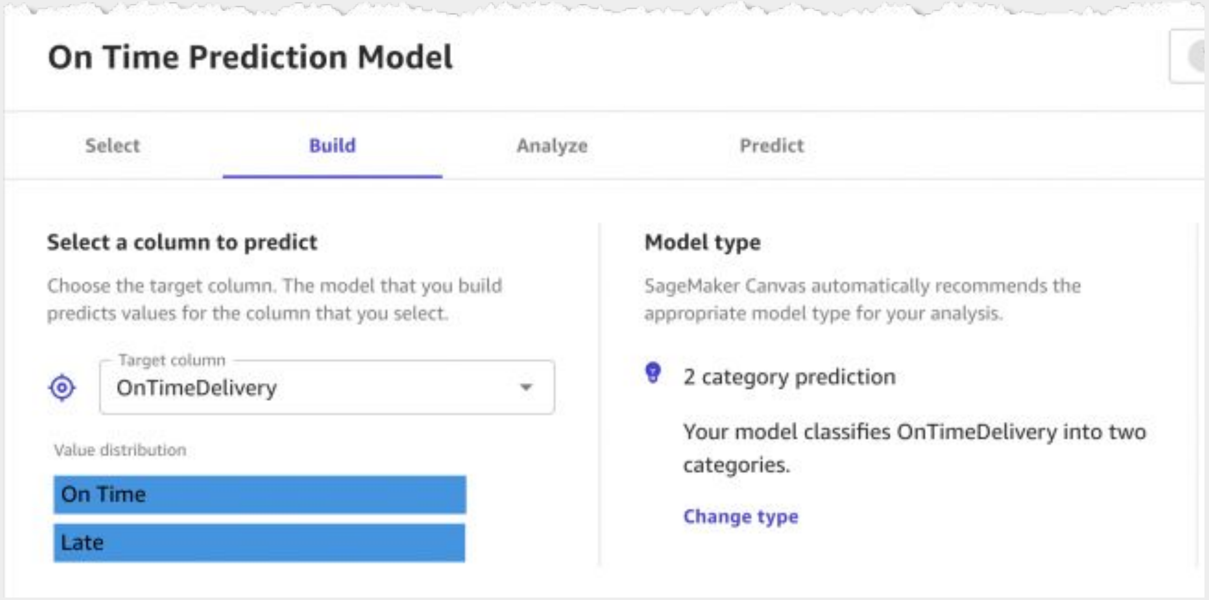
SageMaker Canvas is a visual, point-and-click service that allows analysts to generate accurate machine learning predictions without writing any code


Collaborate and send models to data scientists using Amazon SageMaker Studio for review and feedback
Amazon SageMaker Canvas
Build machine learning models without writing code
Browse, import, and join data
Import data from disparate sources and create new unified datasets for training prediction models
Pick target
Select values you want to predict
Prepare and analyze data
Automatically detect errors, cleanse, and analyze data to help determine if your data is prepared for ML
Create model
Create ML models with the click of a button
Generate and understand predictions
Generate single or bulk predictions and understand predictions

apper.ph

SageMaker Canvas: How it works

Pick target: Select values you want to predict


On Time Prediction Model

Select | Build | Analyze | Predict

Select a column to predict

Choose the target column. The model that you build predicts values for the column that you select.

Target column: OnTimeDelivery

Value distribution

- On Time
- Late

Model type

SageMaker Canvas automatically recommends the appropriate model type for your analysis.

2 category prediction

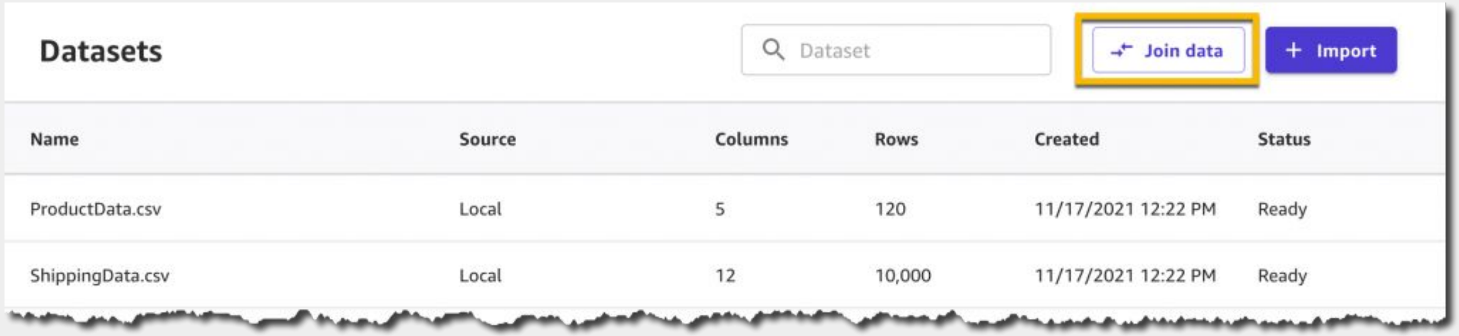
Your model classifies OnTimeDelivery into two categories.

Change type

apper.ph

SageMaker Canvas: How it works

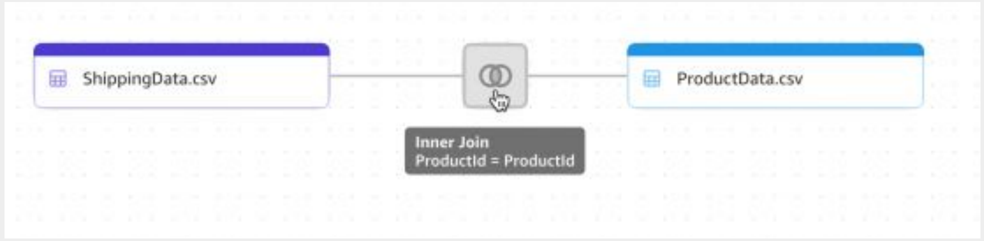
Browse, import, and join data: Import data from disparate sources & create unified datasets for training prediction models



Datasets

Dataset | Join data | + Import

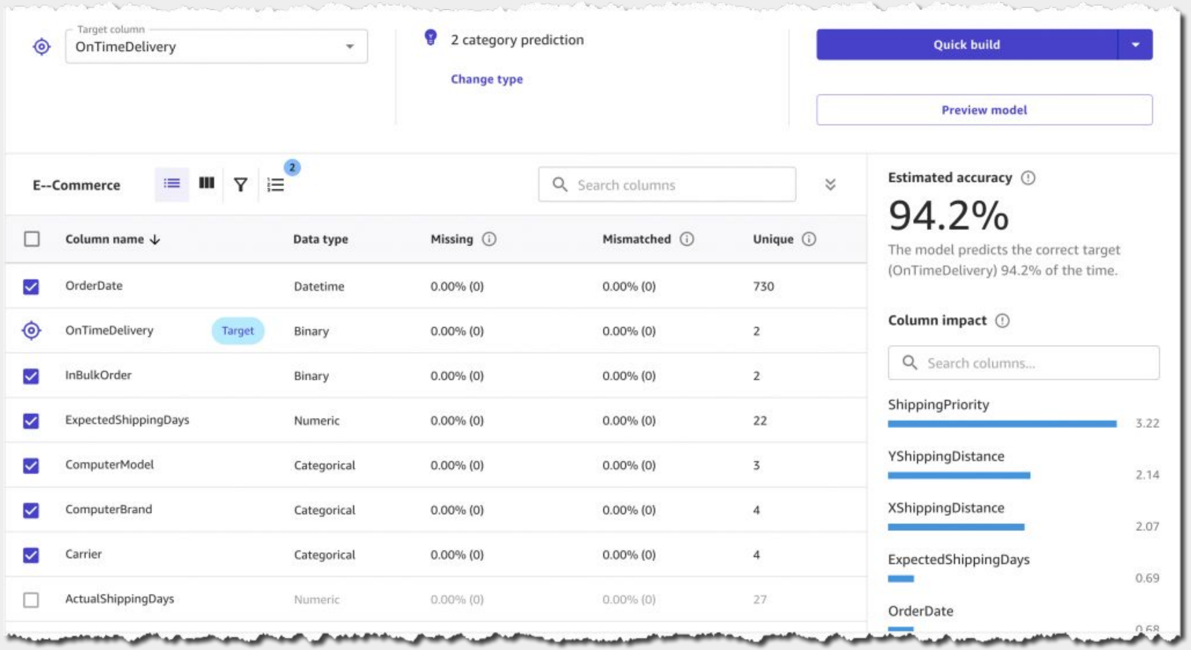
Name	Source	Columns	Rows	Created	Status
ProductData.csv	Local	5	120	11/17/2021 12:22 PM	Ready
ShippingData.csv	Local	12	10,000	11/17/2021 12:22 PM	Ready



apper.ph

SageMaker Canvas: How it works

Prepare and analyze data: Automatically detect errors, cleanse, analyze data to check if your data is ready for ML



Target column: OnTimeDelivery

2 category prediction

Change type

Quick build

Preview model

E--Commerce

Search columns

Column name		Data type	Missing	Mismatched	Unique
OrderDate		Datetime	0.00% (0)	0.00% (0)	730
OnTimeDelivery	Target	Binary	0.00% (0)	0.00% (0)	2
InBulkOrder		Binary	0.00% (0)	0.00% (0)	2
ExpectedShippingDays		Numeric	0.00% (0)	0.00% (0)	22
ComputerModel		Categorical	0.00% (0)	0.00% (0)	3
ComputerBrand		Categorical	0.00% (0)	0.00% (0)	4
Carrier		Categorical	0.00% (0)	0.00% (0)	4
ActualShippingDays		Numeric	0.00% (0)	0.00% (0)	27

Estimated accuracy

94.2%

The model predicts the correct target (OnTimeDelivery) 94.2% of the time.

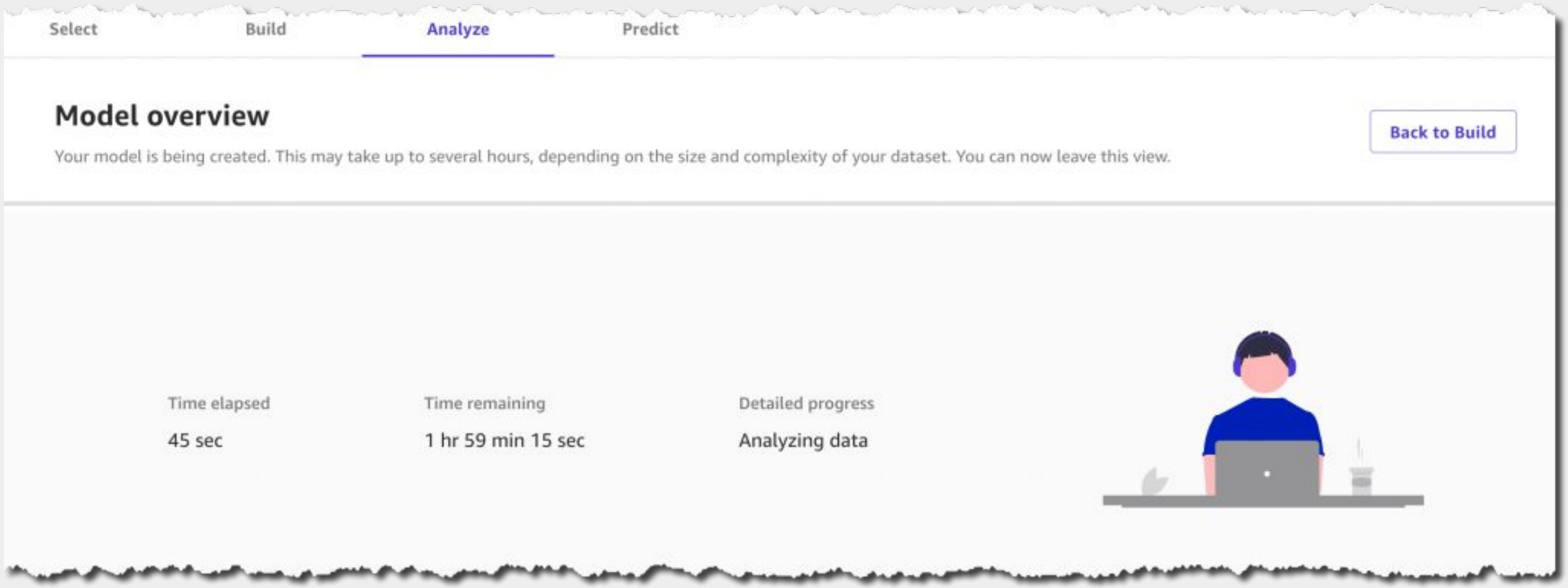
Column impact

Search columns...

ShippingPriority	3.22
YShippingDistance	2.14
XShippingDistance	2.07
ExpectedShippingDays	0.69
OrderDate	0.68

SageMaker Canvas: How it works

Create model: Create ML models with the click of a button



Model overview

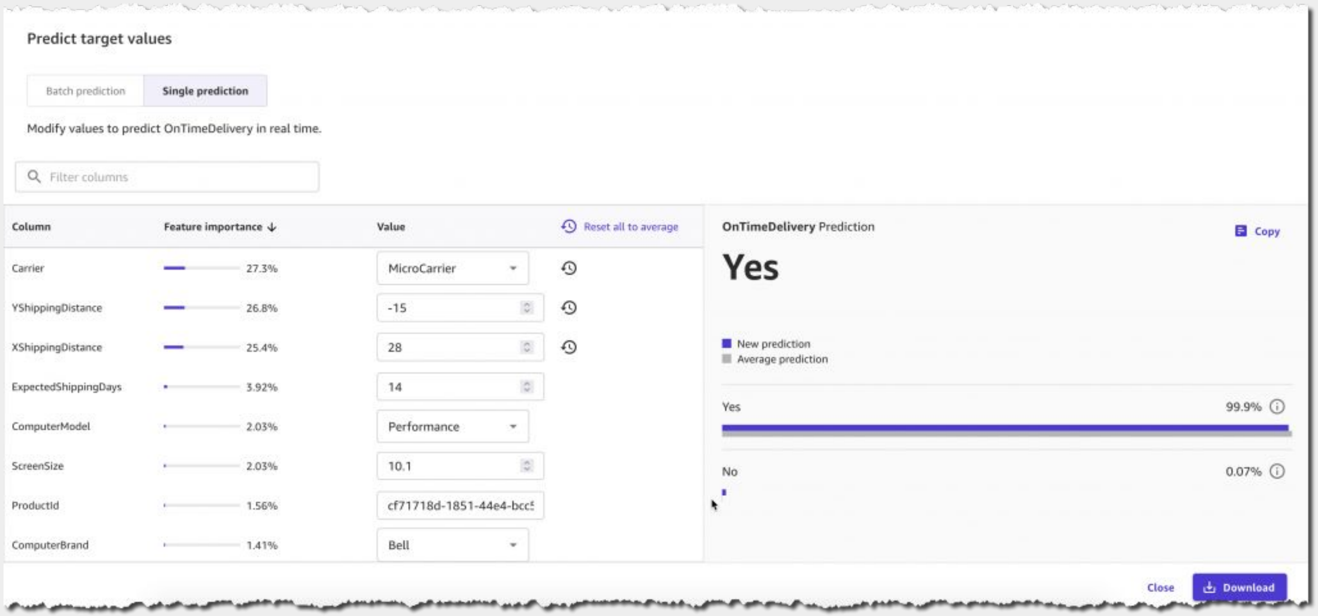
Your model is being created. This may take up to several hours, depending on the size and complexity of your dataset. You can now leave this view.

Time elapsed	Time remaining	Detailed progress
45 sec	1 hr 59 min 15 sec	Analyzing data

apper.ph

SageMaker Canvas: How it works

Generate and understand predictions: Generate single or bulk predictions and understand predictions



Predict target values

Modify values to predict OnTimeDelivery in real time.

Column	Feature importance	Value
Carrier	27.3%	MicroCarrier
YShippingDistance	26.8%	-15
XShippingDistance	25.4%	28
ExpectedShippingDays	3.92%	14
ComputerModel	2.03%	Performance
ScreenSize	2.03%	10.1
ProductId	1.56%	cf71718d-1851-44e4-bcc5
ComputerBrand	1.41%	Bell

OnTimeDelivery Prediction

Yes

■ New prediction
■ Average prediction

Yes	99.9%
No	0.07%

apper.ph

SageMaker Canvas: Benefits

Generate ML predictions without writing code

Quickly access and prepare data for ML

Use built-in AutoML to generate predictions

Validate ML models with data scientists

SageMaker Canvas: Pricing

Session charges:

$1.9 per hour

Training charges:

First 10M cells = $30 per million cells

Next 90M cells = $15 per million cells

Over 100M cells = $7 per million cells

Example:

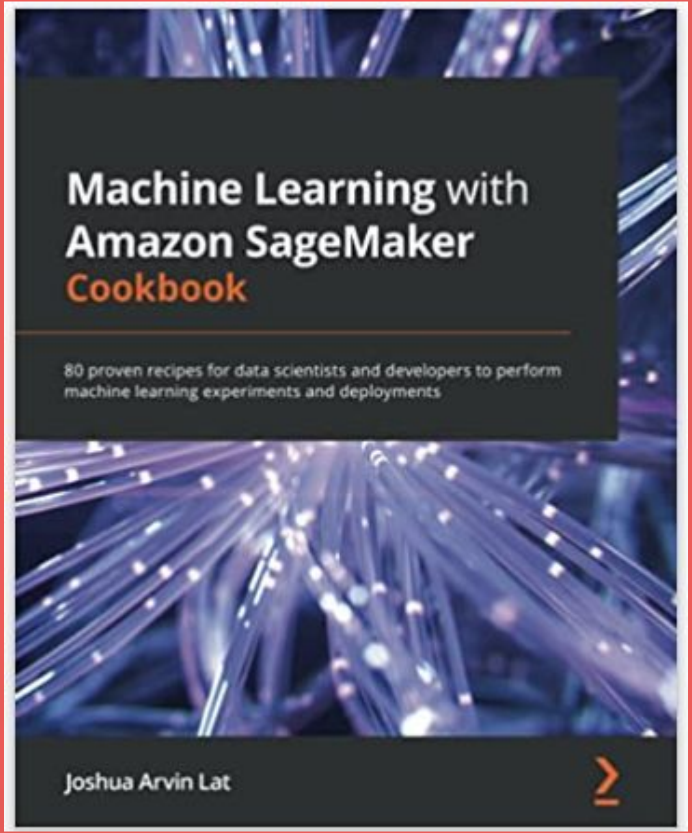
- 500,000 customers with 26 attributes, which translates to 13 million cells
- 40 hours logged into the SageMaker Canvas

($1.9 x 40 hours) = **$76** and ($30 x 10M cells + $15 x 3M cells) = **$345**

Total: $421

apper.ph

Amazon SageMaker Cookbook


Machine Learning with Amazon SageMaker Cookbook

80 proven recipes for data scientists and developers to perform machine learning experiments and deployments

Joshua Arvin Lat

apper.ph

CHAPTER 4

Preparing, Processing, and Analyzing the Data

Chapter 4 Summary

focuses on the key SageMaker **capabilities**, **algorithms**, and **features** to perform **data processing** and **analysis**

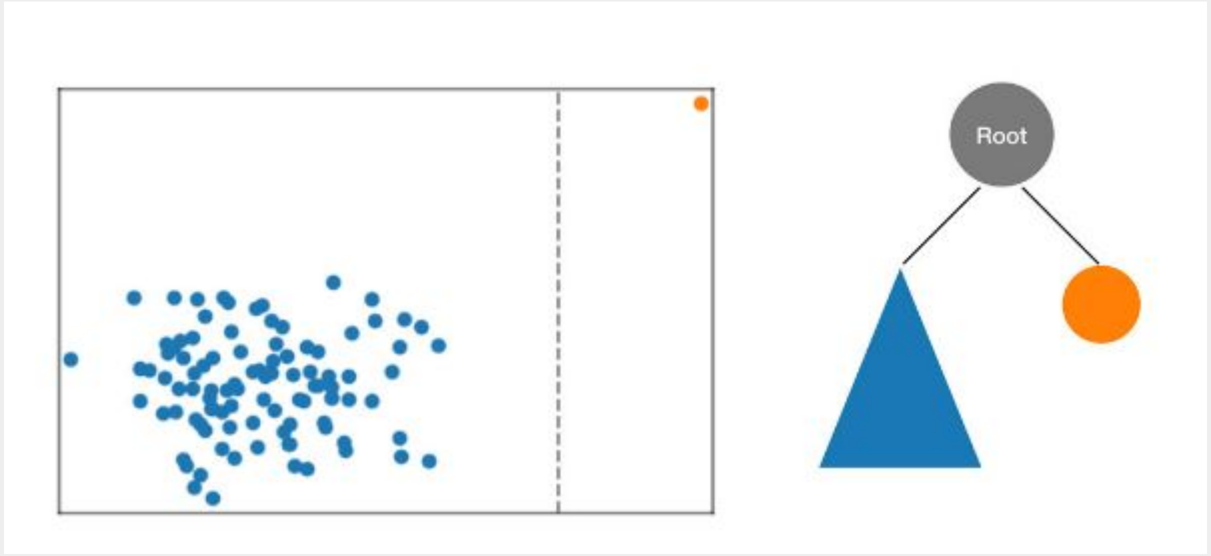
What we learned:

- RCF Model
- Amazon Athena
- PCA Algorithm
- KMeans Algorithm
- K-Nearest Neighbors
- SageMaker Processing

apper.ph

Chapter 4 RCF Model

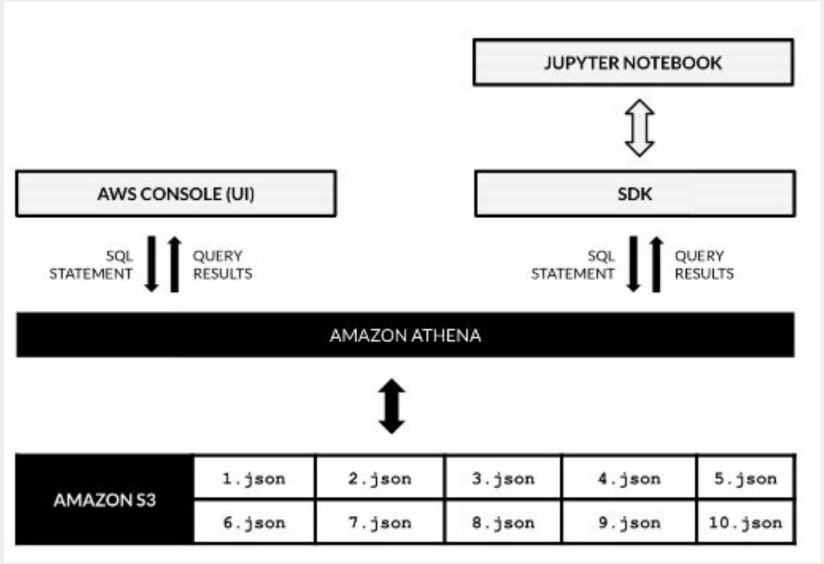
Useful for **detecting anomalies** in datasets. Data points are associated with an **anomaly score** and anomalies are associated with higher scores



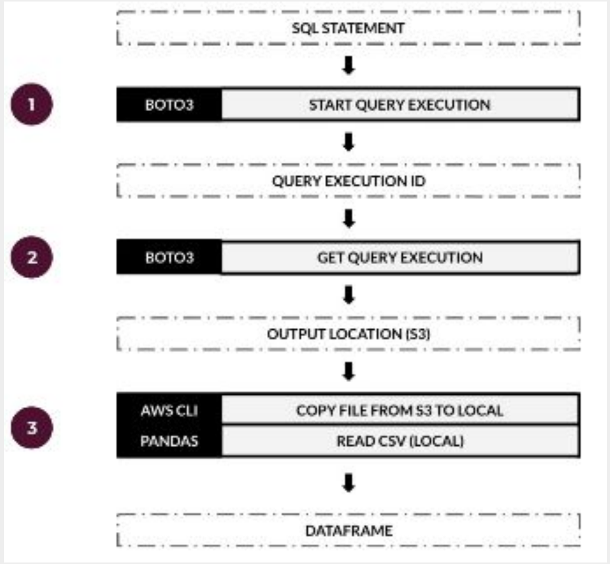
apper.ph

Chapter 4 Amazon Athena

Helps us **analyze** the **data** inside the files stored in our **S3 buckets**.



Using AWS UI

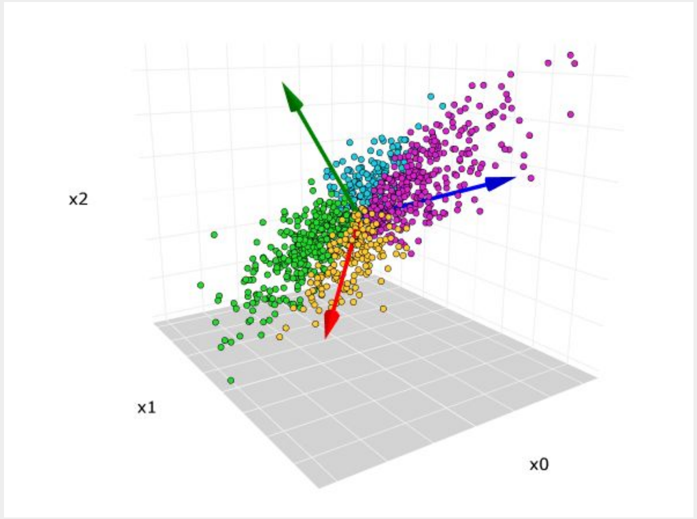


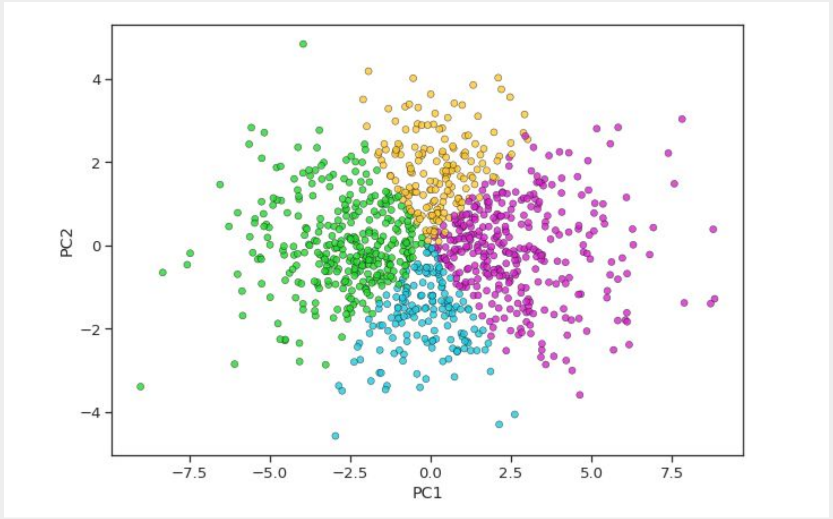
Using boto3

apper.ph

Chapter 4 PCA Algorithm

Built-in algorithm used to perform **dimensionality reduction** on a dataset

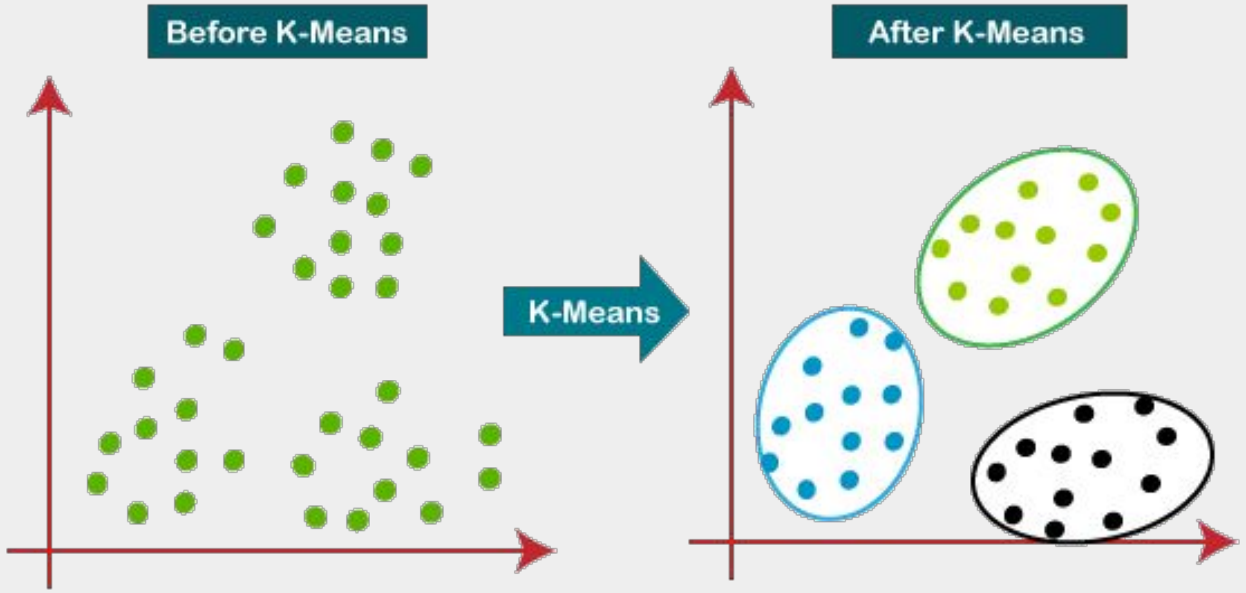




apper.ph

Chapter 4 KMeans Algorithm

Built-in algorithm used to perform **cluster analysis** on a dataset



apper.ph

Chapter 4 K-Nearest Neighbors (KNN)

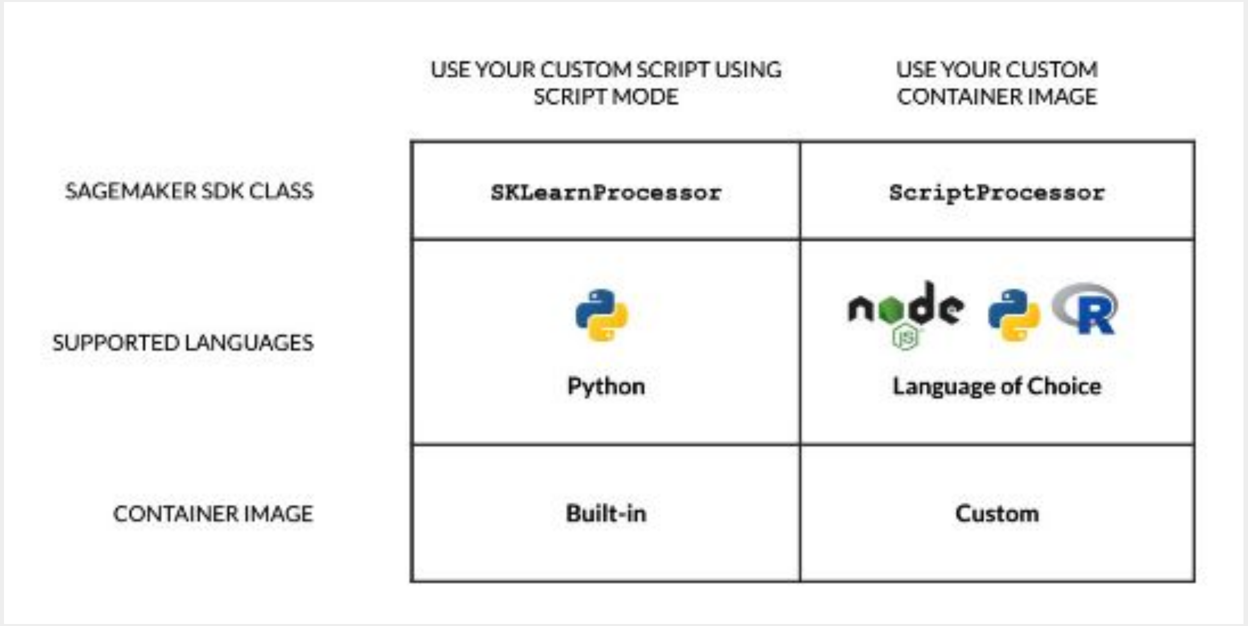
In **protobuf recordIO** format, training start times will be **faster** as the training job streams directly from the S3 bucket source

Using record_set as training input data

Using protobuf recordIO as training input data

Chapter 4 SageMaker Processing

Any **processing** that involves using a **managed service** to handle infrastructure component and a **custom script** to perform a certain action



	USE YOUR CUSTOM SCRIPT USING SCRIPT MODE	USE YOUR CUSTOM CONTAINER IMAGE
SAGEMAKER SDK CLASS	`SKLearnProcessor`	`ScriptProcessor`
SUPPORTED LANGUAGES	Python	Language of Choice
CONTAINER IMAGE	Built-in	Custom

apper.ph

CHAPTER 5
Effectively Managing
Machine Learning Experiments

apper.ph

Chapter 5 Introduction to Debugger & Experiments

When we build multiple ML experiments, we need to **detect** and **monitor changes** in the values of **parameters, metrics**, & other **variables**, and sometimes we want to automate an action to perform when specific **rules** and **conditions** are met.

SageMaker Debugger

Moreover, we need to keep **track** of **datasets**, **hyperparameters**, and other **inputs** and **outputs** of multiple ML experiments, so we can easily **reproduce** them.

SageMaker Experiments

apper.ph

Chapter 5 SageMaker Debugger

Detect issues and profile training jobs using **Debugger Hooks** to **capture debug data**…

…and check whether **certain conditions are met** using **Debugger rules**, e.g. detect if loss is not decreasing by 5% every 2 steps

Inspect the **debugger output artifacts** and check the **logs** as well to find where issues were detected during training using **`smdebug`** and **`awslogs`**

Chapter 5 SageMaker Debugger

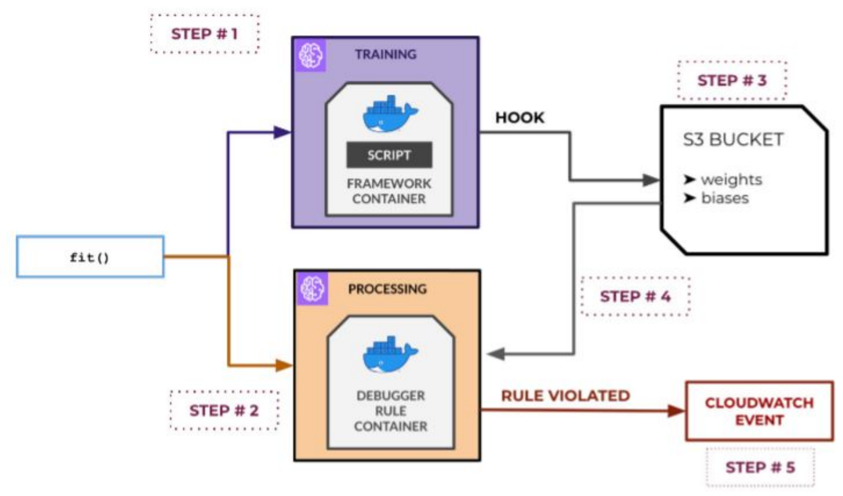


Figure 5.10 – What happens behind the scenes when using SageMaker Debugger

(1) **Training jobs** built on specific framework containers are run

(2) **Processing jobs** built on Debugger Rule containers **monitor training jobs**

(3) **Hooks capture debug data** and store them in an S3 bucket

(4) Processing jobs inspect debug data and **check whether a rule is violated**

(5) Once a rule is violated, a **CloudWatch event** can be **triggered**

apper.ph

Chapter 5 SageMaker Experiments

Set up an **Experiment** with multiple **Trials**, each with **Trial Components** and a **Tracker**

Analyze details of previous experiments we performed and tracked using **`ExperimentAnalytics`** from **`sagemaker.analytics`**

Inspect metadata of Experiments & Trials and **details** (parameters, metrics, artifacts, metadata) of Trial Components

apper.ph

Chapter 5 SageMaker Experiments

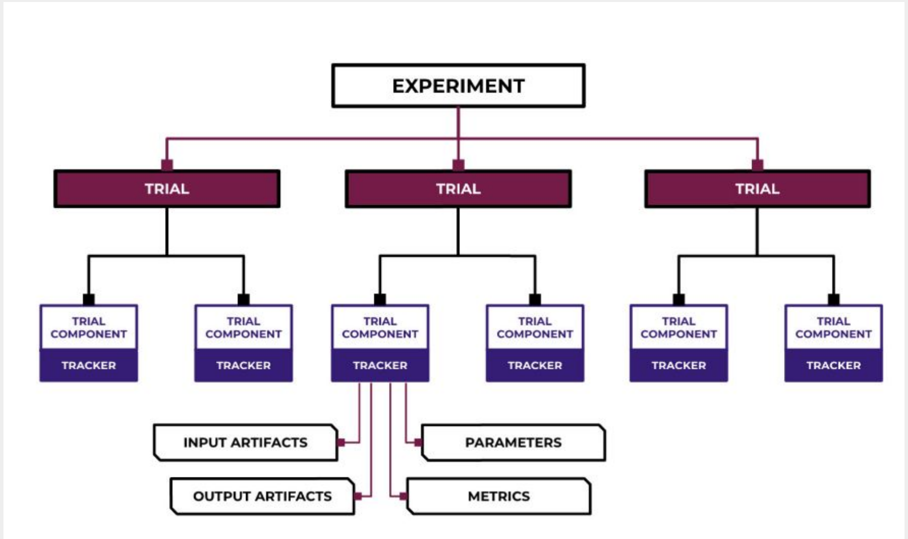


Figure 5.20 – How Experiment, Trial, TrialComponent, and Tracker resources are connected

We use…

- **Experiment**
- **Trial** - a training iteration or job as part of an experiment
- **Trial Components** - parameters, metrics, artifacts, and metadata of a trial
- **Tracker** - artifact logger

We track…

- **Parameters** – e.g. max depth, gamma
- **Metrics** – e.g. accuracy, f1-score
- **Artifacts** – paths to training & validation inputs, container images, to outputs

apper.ph

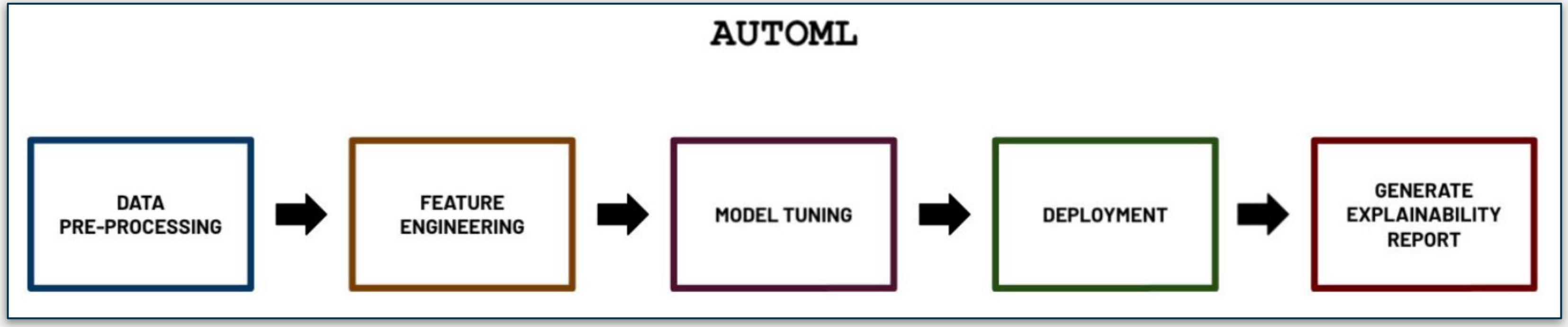
CHAPTER 6

Automated Machine Learning in Amazon SageMaker

apper.ph

Chapter 6 AutoML

AutoML is the process of automating aspects of the machine learning pipeline



In Amazon SageMaker, AutoML can be done using **SageMaker Autopilot**

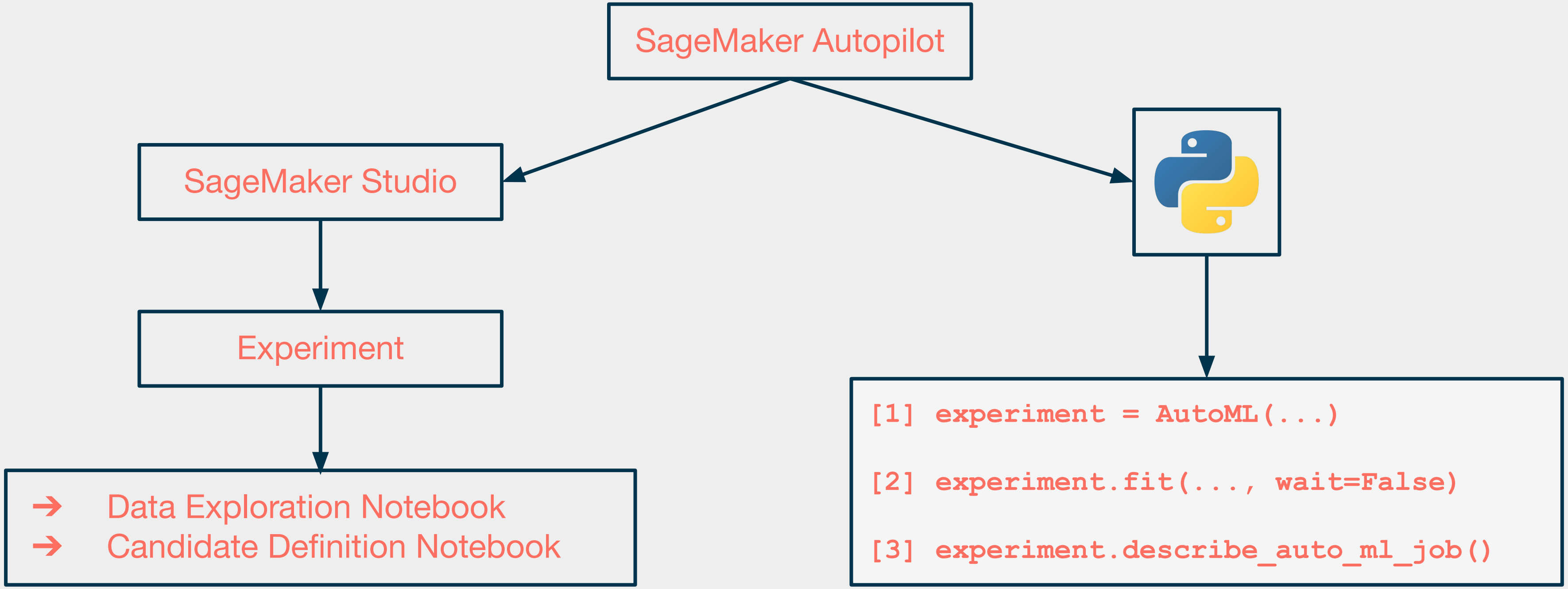
apper.ph

Chapter 6 SageMaker Autopilot

With SageMaker Autopilot, the different steps of the machine learning process are **performed automatically**

Data Analysis	Problem Definition	Database Schema Detection
Candidate Definitions Generation	Data Preprocessing and Feature Engineering	Algorithm Selection
Model Tuning	Deployment	Explainability Report Generation

Chapter 6 SageMaker Autopilot Implementation

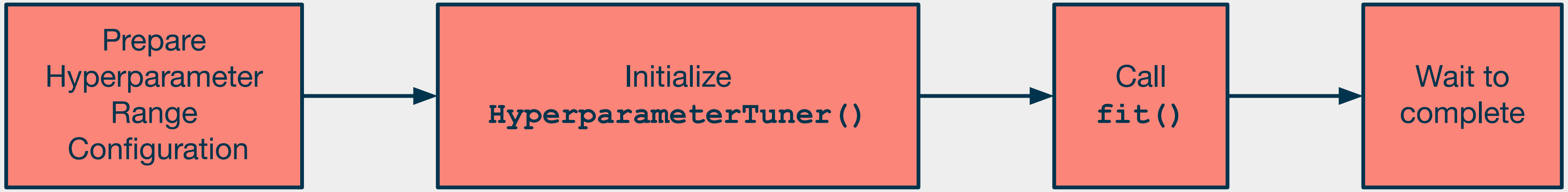


```
[1] experiment = AutoML(...)

[2] experiment.fit(..., wait=False)

[3] experiment.describe_auto_ml_job()
```

apper.ph

Chapter 6 Hyperparameter Optimization

Hyperparameter optimization is the process of looking for the best configuration and combination of hyperparameter values that produce the best mode


Prepare Hyperparameter Range Configuration
Initialize HyperparameterTuner()
Call fit()
Wait to complete

apper.ph

Chapter 6 Hyperparameter Tuning Job Analytics

Properties and details of the Automatic Model Tuning Job can be loaded using the `HyperparameterTuningJob` Analytics class

```
analytics = sagemaker.HyperparameterTuningJob(tuning_job_name)

full_df = analytics.dataframe()

full_df
```

Parameters, hyperparameters, and metric values associated with the training jobs can then be seen

Summary

Other SageMaker Capabilities

- ✅ Canvas
- ✅ Ground Truth

Amazon SageMaker Cookbook

- ✅ Chapter 4 Pre-Processing
- ✅ Chapter 5 Debugger & Experiments
- ✅ Chapter 6 Autopilot

apper.ph