

# MLD-5-Sagemaker\_Hillary

EU13\_AWS\_Sagemaker  
Training Clarusway  
Pear Deck - May 3, 2023 at 7:10PM

## Part 1 - Summary

Use this space to summarize your thoughts on the lesson

## Part 2 - Responses

Slide 1



Use this space to take notes:

## Slide 2

### ► Table of Contents

- ▶ Cloud Computing Market Size
- ▶ What is SageMaker?
- ▶ SageMaker Process
- ▶ Algorithms
- ▶ Notebook Instance
- ▶ S3
- ▶ Billing



2

Use this space to take notes:

## Slide 3

### Your Response

I've completed the pre-class content?

Students choose an option

Pear Deck Interactive Slide  
Do not remove this bar

You Chose

- Yes

Other Choices

- No

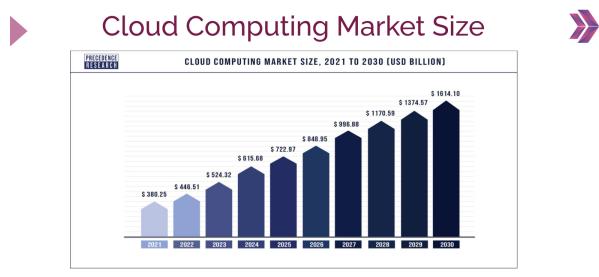
Use this space to take notes:

## Slide 4



Use this space to take notes:

## Slide 5



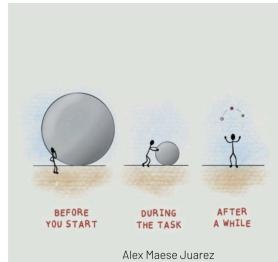
The rising popularity of the latest novel technologies like **artificial intelligence** and **machine learning** and its rapid adoption in the cloud computing is empowering the growth of the global cloud computing market.

5

Use this space to take notes:

## Slide 6

### ▶ SageMaker ➤



Use this space to take notes:

## Slide 7

### ▶ What is SageMaker ➤

Amazon SageMaker is a platform service that simplifies the process of **building, training, and deploying ML models** by providing everything organizations need to connect to their training data, select and optimize the **best algorithm and framework**, and **deploy their model** on auto-scaling clusters of Amazon EC2.



Use this space to take notes:

## Slide 8

### ▶ WHY Amazon SageMaker? ➤

- ▶ Accelerating machine learning innovation through security
- ▶ Security features from Amazon SageMaker and the AWS Cloud can help organizations go from idea to production faster
- ▶ <https://amer.resources.awscloud.com/ai-ml/accelerating-machine-learning-innovation-through-security>

8

Link(s) on this slide:

- <https://amer.resources.awscloud.com/ai-ml/accelerating-machine-learning-innovation-through-security>
- <https://amer.resources.awscloud.com/ai-ml/accelerating-machine-learning-innovation-through-security>

Use this space to take notes:

## Slide 9

### ▶ ML with AWS, by the numbers ➤

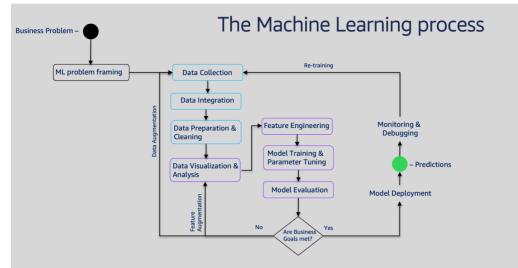
- ▶ Reduce training time by 50%
- ▶ Provide 90% scaling efficiency
- ▶ Deliver 3x faster network throughput
- ▶ Improve price and performance by 25%
- ▶ 91% of cloud-based PyTorch runs on AWS
- ▶ 92% of cloud-based TensorFlow runs on AWS

9

Use this space to take notes:

## Slide 10

### ► ML process



Use this space to take notes:

## Slide 11



CLARUSWAY®  
WAY TO REINVENT YOURSELF

Use this space to take notes:

## Slide 12

### ► SageMaker Free Tier (2 months) ➤

Two months [free tier](#) – starts from the first month you create a SageMaker resource

Development – 250 Hours/Month t2.medium or t3.medium

Train – 50 Hours/Month m4.xlarge or m5.xlarge

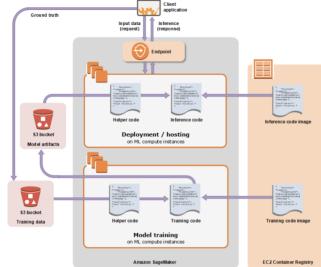
Deploy – 125 Hours/Month m4.xlarge or m5.xlarge

12

Use this space to take notes:

## Slide 13

### ► Model Deployment Architecture ➤



13

Use this space to take notes:

## Slide 14

### ► SageMaker Process



- Get your data



- Explore and refine models in a single Notebook Instance



- Train on the full dataset in a cluster of GPU instances..



- Deploy to production



14

Use this space to take notes:

## Slide 15

### ► Built-in algorithms



- XGBoost, FM, Linear, k-NN, and Forecasting for supervised learning



- k-Means, PCA, and Random Cut Forest for unsupervised learning



- Image classification and object detection for computer vision



- LDA, Neural Topic Model, Seq2Seq, and Word2Vec for text and NLP



15

Use this space to take notes:

## Slide 16

### ► Built-in algorithms-Supervised ➤

- ▶ Linear Learner: regression
- ▶ K-Nearest Neighbors: non-parametric regression and classification
- ▶ XGBoost: regression, classification
- ▶ Factorization Machines: regression, classification, recommendation
- ▶ Semantic Segmentation: Deep Learning
- ▶ Image Classification: Deep Learning (ResNet)
- ▶ Object Detection (SSD): Deep Learning  
(VGG or ResNet)
- ▶ Sequence to Sequence: machine translation, speech to text and more
- ▶ DeepAR: time-series forecasting (RNN)

18

Use this space to take notes:

## Slide 17

### ► Built-in algorithms-Unsupervised ➤

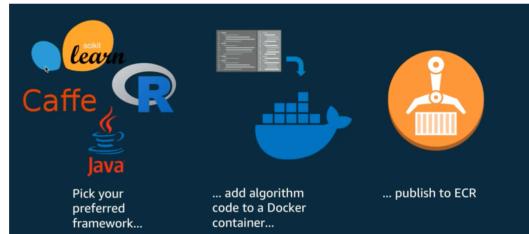
- ▶ K-Means: clustering
- ▶ Principal Component Analysis: dimensionality reduction
- ▶ Random Cut Forest: anomaly detection
- ▶ Object2Vec: general-purpose embedding
- ▶ Neural Topic Model: topic modeling
- ▶ Latent Dirichlet Allocation: topic modeling (mostly)
- ▶ Blazing Text: GPU-based Word2 Vec, and text classification
- ▶ IP Insights: usage patterns for IP addresses

19

Use this space to take notes:

## Slide 18

► Bring your own algorithm ➤



Use this space to take notes:

## Slide 19

► Let's jump to the AWS SageMaker Service! ➤

Use this space to take notes:

## Slide 20

### ► S3 BUCKET and OBJECTS

```
▶ iden-ml-sagemaker/
  ○ bikerental/
    ■ train/
      • train.csv
    ■ validation/
      • validation.csv
    ■ test/
      • test.csv
    ■ model/
      • xgboost-bikerental-v1-2022-04-21-10-23-10-964/ (Job_name+datetime)
        ○ output
        ■ modelTargz
```

Amazon S3>Buckets>iden-ml-sagemaker>bikerental>output>xgboostbikerental-v1-2022-04-21-10-23-10-964>output>modelTargz



20

Use this space to take notes:

## Slide 21

### ► Notebook Instance

An *Amazon SageMaker notebook instance* is a machine learning (ML) compute instance running the Jupyter Notebook App. SageMaker manages creating the instance and related resources.



21

Use this space to take notes:

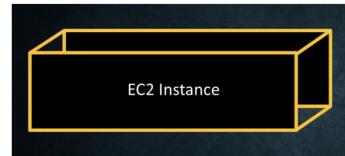
## Slide 22

### ▶ Notebook Instance



Use Jupyter notebooks in your notebook instance

- To prepare and process data,
- Write code to train models,
- Deploy models to SageMaker hosting,
- Test or validate your models.

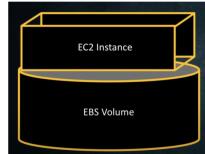


22

Use this space to take notes:

## Slide 23

### ▶ Notebook Instance



- Anaconda Packages,
- Tensorflow and Apache MXnet,
- Storage volume,
- Sample notebooks that contain complete code walkthroughs

23

Use this space to take notes:

## Slide 24

► You are using CSV formatted files to train on SageMaker's built-in XGBoost algorithm. SageMaker expects your training and validation to follow this convention:

- ▶ CSV must have column headers and target variable must be the last column
- ▶ CSV must have column headers with the target variable in the first column
- ▶ CSV must not have a column header record. Target variable must be the last column
- ▶ CSV must not have a column header record. Target variable must be the first column



24

Use this space to take notes:

## Slide 25

► How does SageMaker built-in know the target variable?



- ▶ For CSV training, the algorithm assumes that the target variable is in the first column and that the CSV does not have a header record.(Train.Validation)
- ▶ For CSV inference, the algorithm assumes that CSV input does not have the label column.

<https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html>

25

Link(s) on this slide:

- <https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost.html>

Use this space to take notes:

## Slide 26

### ► SageMaker Billing Dashboard

Cost Management	EC2 Container Registry (ECR)	\$0.00
Cost Explorer	Elastic Compute Cloud	\$0.00
Budgets	Key Management Service	\$0.30
Budgets Reports	SageMaker	\$0.00
Savings Plans	- US East (N. Virginia)	\$0.00
Promotions	Amazon SageMaker CreateVolume-Gp2	\$0.00
Billing preferences	\$0.00 for SageMaker Debugger Built-in Rule Volume	0.025 GB-Mo
	\$0.14 per GB-Mo of Training job ML storage	0.00004 GB-Mo
	\$0.14 per GB-Mo of Notebook Instance ML storage	0.00004 GB-Mo
Payment methods	\$0.14 per GB-Mo of Training job ML storage	0.0000274 GB-Mo
Consolidated billing	Amazon SageMaker Endpoint	\$0.00
Tax settings	\$0.016 per GB for Endpoint Data IN	0.001 GB
	\$0.016 per GB for Endpoint Data OUT	0.0002039 GB
	Amazon SageMaker RunInstance	\$0.00
	\$0.0 for SageMaker Debugger Built-in Rule Instance	\$0.00
	\$0.00 for Host ml.m4.xlarge per hour under monthly free tier	0.169 Hrs
	\$0.00 for ml.m4.xlarge per hour under monthly free tier	2.000 Hrs
	\$0.00 for Notebook.ml.s3.medium per hour under monthly free tier	1.153 Hrs
	\$0.23 per Training ml.c4.xlarge hour in US East (N. Virginia)	0.230 Hrs
	\$0.478 per Training ml.c4.2xlarge hour in US East (N. Virginia)	0.016 Hrs
	Management of Training Job ml.m4.xlarge hour in US East (N. Virginia)	\$0.003 Hrs
		\$0.00
	<b>US East (N. Virginia)</b>	\$0.00
	Simple Queue Service	\$0.00
	Simple Notification Service	\$0.00
	Simple Storage Service	\$0.00

28

Use this space to take notes:

## Slide 27

### ► SageMaker Use Case Example

#### ZAPPOS

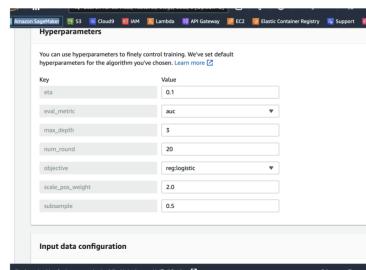
*"We are...using analytics and machine learning solutions to personalize sizing and search results for individual users. AWS services (including Amazon SageMaker) allow (our) engineers to focus on improving performance and results rather than DevOps overhead."*

29

Use this space to take notes:

## Slide 28

### ▶ Hyperparameters



The screenshot shows the 'Hyperparameters' configuration page in the AWS SageMaker console. It displays a list of hyperparameters with their current values:

Key	Value
eta	0.1
eval_metric	auc
max_depth	5
num_round	20
objective	reg:logistic
scale_pos_weight	2.0
subsample	0.5

Below the table, there is a section titled 'Input data configuration'.



Use this space to take notes:

## Slide 29

**THANKS!**  
Any questions?



CLARUSWAY®  
WAY TO REINVENT YOURSELF

29

Use this space to take notes:

## Slide 30

### ► S3 Settings

The screenshot shows the 'S3 Settings' page for a bucket named 'richard-new'. The 'Bucket name' field contains 'richard-new'. The 'AWS Region' dropdown is set to 'US East (N. Virginia) us-east-1'. Below these fields is a 'Copy settings from existing bucket - optional' section with a 'Choose bucket' button. The 'Object Ownership' section contains two radio button options: 'ACLs disabled (recommended)' (selected) and 'ACLs enabled'. A note below states: 'Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can modify access to objects.' The 'Object Ownership' section also includes a 'Bucket owner enforced' link.



30

Use this space to take notes:

## Slide 31

### ► S3 Settings

The screenshot shows the 'S3 Settings' page with the 'Block Public Access settings for this bucket' section expanded. It contains four checkboxes under the heading 'Block all public access': 'Block all public access' (selected), 'Block public access to buckets and objects granted through new access control lists (ACLS)', 'Block public access to buckets and objects granted through any public bucket or access point policies', and 'Block public and cross-account access to buckets and objects through any public bucket or access point policies'. A note at the top of this section states: 'Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure your account's security, AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to this bucket or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more'.

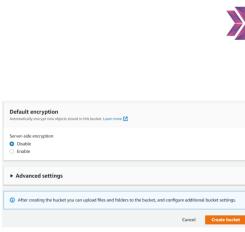


31

Use this space to take notes:

## Slide 32

### ▶ S3 Settings

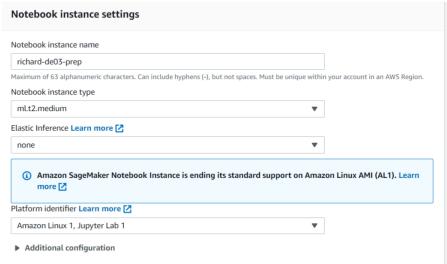


32

Use this space to take notes:

## Slide 33

### ▶ Notebook Settings



33

Use this space to take notes:

## Slide 34

### Notebook Settings

The screenshot shows the 'Permissions and encryption' section of the Notebook Settings page. It includes fields for creating an IAM role and selecting an encryption key. A prominent green success message box states: 'Success! You created an IAM role.' followed by the ARN 'AmazonSageMaker-ExecutionRole-20230101T194235'. Below this, there are options for 'Root access - optional' (with 'Enable' selected) and 'Encryption key - optional' (set to 'No Custom Encryption').



34

Use this space to take notes:

## Slide 35

### Notebook Settings

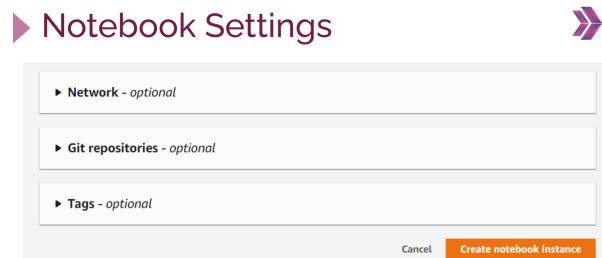
The screenshot shows the 'Create an IAM role' dialog box. It has sections for specifying S3 buckets (with 'Any S3 bucket' selected), object tagging, and S3 bucket policies. At the bottom, there are 'Cancel' and 'Create role' buttons.



35

Use this space to take notes:

## Slide 36



38

Use this space to take notes:

## Slide 37



39

*Now open the notebook from Sagemaker and send the data S3 via notebook*

*Sending data come back to the sagemaker console and do the jobs:*

*Training Job Creation-Model Creation- End Point Creation*

Use this space to take notes:

## Slide 38

### ▶ Train Job Settings



Job settings

Job name: richard-de-prep  
Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in each Region.

IAM role: AmazonSageMakerExecutionRole-20230101194255  
AmazonSageMaker requires permissions to call other services on your behalf. Choose a role or let us create a role that has the required permissions for this job.

Create role using the role creation wizard [\[?\]](#)

Algorithm options: Use an Amazon SageMaker built-in algorithm, your own algorithm, or a third-party algorithm from AWS Marketplace.

Algorithm source:

- Amazon SageMaker built-in algorithm [Learn more \[?\]](#)
- Your own algorithm resource
- Your own algorithm container in ECR [Learn more \[?\]](#)
- An algorithm subscription from AWS Marketplace

Choose an algorithm

38

Use this space to take notes:

## Slide 39

### ▶ Train Job Settings



Choose an algorithm

Tabular - XGBoost : v1.3

Container: The entry point where the training image is stored in Amazon ECR. [Learn more](#)  
68315268778.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.3-1

Input mode: You can provide your training data as a file or pipe.

File

Metrics

The algorithm you selected will publish the following metrics to CloudWatch metrics.

Metric name	Regex
trainrmse	.*\{[0-9]+\}.*#011train-rmse\{[+]\{[0-9]+\}\{[0-9]+\}\{[eE]\}\{+\}\{[0-9]+\}\{[eE]\}.*
trainmae	.*\{[0-9]+\}.*#011train-mae\{[+]\{[0-9]+\}\{[0-9]+\}\{[eE]\}\{+\}\{[0-9]+\}\{[eE]\}.*
trainlogloss	.*\{[0-9]+\}.*#011train-logloss\{[+]\{[0-9]+\}\{[0-9]+\}\{[!][eE]\}\{+\}\{[0-9]+\}\{[!][eE]\}.*
trainerror	.*\{[0-9]+\}.*#011train-error\{[+]\{[0-9]+\}\{[0-9]+\}\{[eE]\}\{+\}\{[0-9]+\}\{[eE]\}.*

39

Use this space to take notes:

## Slide 40

### ▶ Train Job Settings

Resource configuration

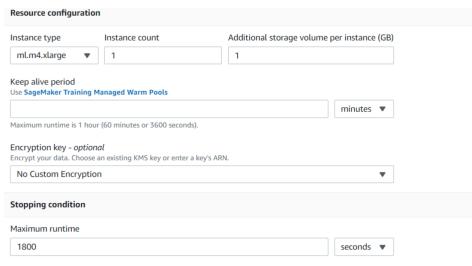
Instance type	ml.m4.xlarge	Instance count	1	Additional storage volume per instance (GB)	1
---------------	--------------	----------------	---	---	---

Keep alive period  
Use SageMaker Training Managed Warm Pools

Encryption key - optional  
Encrypt your data. Choose an existing KMS key or enter a key's ARN.  
No Custom Encryption

Stopping condition

Maximum runtime  
1800 seconds



40

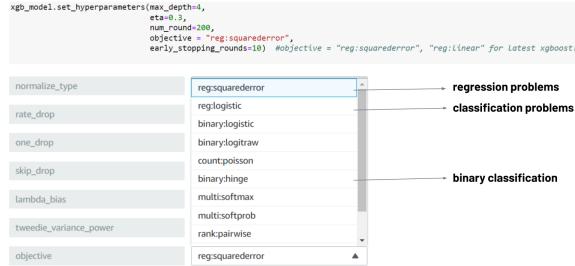
Use this space to take notes:

## Slide 41

### ▶ Train Job Settings

```
xgb_model.set_hyperparameters(max_depth=4,  
                               eta=0.3,  
                               n_estimators=380,  
                               objective = "reg:squarederror",  
                               evals=1, stopping_rounds=10) #objective = "reg:squarederror", "reg:linear" for latest xgboost!
```

normalize_type	reg:squarederror	→ regression problems
rate_drop	reg:logistic	→ classification problems
one_drop	binary:logitraw	
skip_drop	count:poisson	
lambda_bias	binary:hinge	→ binary classification
tweedie_variance_power	multi:softmax	
objective	multi:softprob	
	rank:pairwise	
	reg:squarederror	



41

Use this space to take notes:

## Slide 42

### ▶ Train Job Settings

Input data configuration

Create up to 10 channels of input sources. If the algorithm you chose supports multiple input channels here, See Algorithms Provided by Amazon SageMaker: Common Parameters

Channels

train

Channel name: train

Input mode - optional: File

Content type - optional: csv

Choose one of the formats below:

- json
- csv

Compression type: None

Record wrapper: None

Data source: S3

Add channel

S3 data type: S3Prefix

S3 data distribution type: FullyReplicated

S3 location: s3://richard-de-prep/sagemaker-autoscout/data/train.csv

Don't forget to add channel for validation data

Use this space to take notes:

## Slide 43

### ▶ Train Job Settings

validation

Channel name: validation

Input mode - optional: File

Content type - optional: csv

Choose one of the formats below:

- json
- csv

Compression type: None

Record wrapper: None

Data source: S3

S3 data type: S3Prefix

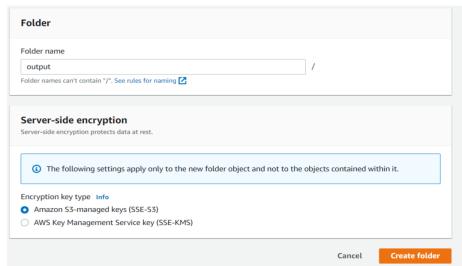
S3 data distribution type: FullyReplicated

S3 location: s3://richard-de-prep/sagemaker-autoscout/data/validation.csv

Use this space to take notes:

## Slide 44

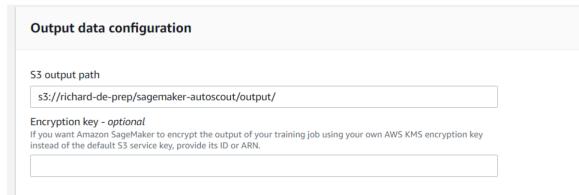
### ▶ Create Output Folder from S3



Use this space to take notes:

## Slide 45

### ▶ Train Job Settings



45

Use this space to take notes:

## Slide 46

### ▶ Train Job Settings

Managed spot training

Enable managed spot training - optional  
Save compute costs for jobs that have flexibility in start and end times. Amazon SageMaker will use spare capacity only to run this job. Learn more

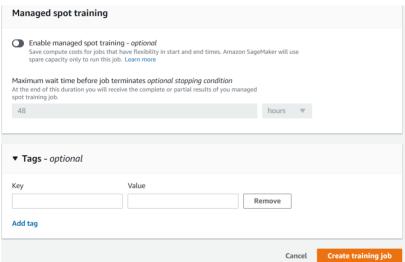
Maximum wait time before job terminates optional stopping condition  
At the end of this duration you will receive the complete or partial results of your managed spot training job.

48 hours

▼ Tags - optional

Key Value Remove Add tag

Cancel Create training job



48

Use this space to take notes:

## Slide 47

### ▶ Model Settings

Model settings

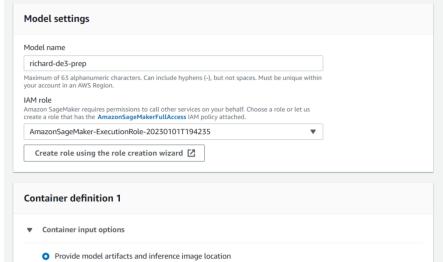
Model name richard-dct-prep  
Maximum of 64 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

IAM role Amazon SageMaker requires permission to call other services on your behalf. Choose a role or let us manage one that has the [AmazonSageMakerFullAccess](#) IAM policy attached.  
AmazonSageMaker-ExecutionRole-20230101T194235  
Create role using the role creation wizard

Container definition 1

Container input options

Provide model artifacts and inference image location



47

Use this space to take notes:

## Slide 48

### ► Model Settings

richard-de-prep

Job settings

Algorithm

Algorithm ARN

Training image  
683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.3-1

Additional volume size (GB)  
1

Maximum runtime (s)  
1800

Maximum wait time for managed spot training(s)  
-

Volume encryption key  
-

Input mode  
File

Clone Create model package Stop Create model

48

Use this space to take notes:

## Slide 49

### ► Model Settings

▼ Provide model artifacts and inference image options

Use a single model  
Use this to host a single model in this container.

Use multiple models  
Use this to host multiple models in this container.

Location of inference code image  
Type the repository name if the inference code image is stored in Amazon ECR.  
683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.3-1 → from training job

Location of model artifacts - optional  
Type the URL where model artifacts are stored in S3.  
s3://richard-de-prep/sagemaker-autosout/output/richard-de-prep/output/model → from s3 (model targt uri)

Container host name - optional  
Type the DNS host name for the container.

▼ Environment variables - optional

49

Use this space to take notes:

## Slide 50

### ► Model Settings

The screenshot shows the 'Model Settings' page in the Amazon SageMaker console. The 'Network' section contains an option to 'Enable network isolation' (unchecked) which is described as required for AWS Marketplace products. Below it is a 'VPC - optional' section with a dropdown menu set to 'No VPC'. Underneath is a 'Tags - optional' section with a table for adding key-value pairs. At the bottom right are 'Cancel' and 'Create model' buttons.



50

Use this space to take notes:

## Slide 51

### ► Endpoint Config Settings

The screenshot shows the 'Create endpoint configuration' page in the Amazon SageMaker console. It starts with a brief description of what endpoint configurations are used for. The 'Endpoint configuration' section has a field for 'Endpoint configuration name' containing 'richard-de3-prep'. Below it, 'Type of endpoint' is set to 'Provisioned'. Under 'Encryption key - optional', the 'No Custom Encryption' option is selected. At the bottom right are 'Cancel' and 'Create' buttons.



51

Use this space to take notes:

## Slide 52

### ▶ Endpoint Config Settings

The screenshot shows the 'Production' endpoint configuration page. At the top, there is a table with columns: Model name, Training job, Variant name, Instance type, Elastic inference, Initial instance count, Initial weight, and Actions. Below the table, a message says 'There are currently no resources.' A red box highlights the 'Create production variant' button at the bottom left of the table area. A modal window titled 'Add model' is open, showing a list of models with their creation times. One model, 'richard-de3-prep', is selected.

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

Create production variant

Add model

Name	Creation time
richard-de3-prep	Jan 21, 2023 19:41 UTC
richard-endpoint-de3	Dec 30, 2022 17:01 UTC
segmenter-endpoint-2022-12-30-11-21-29-341	Dec 30, 2022 11:37 UTC
segmenter-endpoint-2022-12-22-20-01-18-395	Dec 22, 2022 20:00 UTC
segmenter-endpoint-2022-12-11-11-44-26-214	Dec 11, 2022 11:44 UTC

Cancel Save

Use this space to take notes:

## Slide 53

### ▶ Endpoint Config Settings

The screenshot shows the 'Production' endpoint configuration page. The table now contains one row: 'richard-de3-prep' with 'richard-de3-prep' as the variant name, 'ml.m4.xlarge' as the instance type, 'none' as the elastic inference, '1' as the initial instance count, and '1' as the initial weight. The 'Actions' column shows 'Edit | Remove'. A red box highlights the 'Create endpoint configuration' button at the bottom right of the table area.

Model name	Training job	Variant name	Instance type	Elastic inference	Initial instance count	Initial weight	Actions
richard-de3-prep	richard-de3-prep	variant-name-1	ml.m4.xlarge	none	1	1	Edit   Remove

Create production variant

Create endpoint configuration

Use this space to take notes:

## Slide 54

### ▶ Endpoint Settings

Endpoint

Endpoint name  
Your application uses this name to access this endpoint.  
richard-de3-prep

Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

Attach endpoint configuration

Use an existing endpoint configuration  
Use an existing endpoint configuration or clone an endpoint configuration.

Create a new endpoint configuration  
Add models and configure the instance and initial weight for each model.



54

Use this space to take notes:

## Slide 55

### ▶ Endpoint Settings

Endpoint configuration

Name	ARN	Creation time
richard-de3-prep	arn:aws:sagemaker:us-east-1:046402772087:endpoint-config/richard-de3-prep	Jan 01, 2023 18:44 UTC
richard-autoscout-de03	arn:aws:sagemaker:us-east-1:046402772087:endpoint-config/richard-autoscout-de03	Dec 30, 2022 17:25 UTC
sagemaker-xgboost-2022-12-30-11-37-341	arn:aws:sagemaker:us-east-1:046402772087:endpoint-config/sagemaker-xgboost-2022-12-30-11-37-29-341	Dec 30, 2022 11:37 UTC

Select endpoint configuration

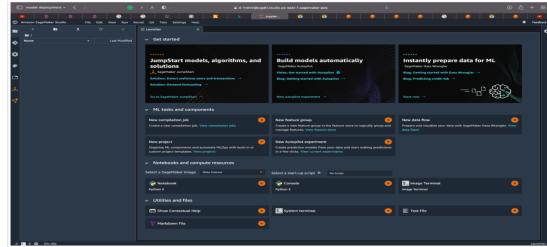


55

Use this space to take notes:

## Slide 56

### ► SageMaker Studio



58

Use this space to take notes: