IST 615 Final Report of Spring 2021

-Machine Learning On AWS

Zeyang Zhou

Contents

[1. Introduction 2](#_Toc71360016)

[1.1 What is SageMaker. 2](#_Toc71360017)

[1.2 Amazon SageMaker Features 3](#_Toc71360018)

[2. Objective 4](#_Toc71360019)

[3. Experiment 4](#_Toc71360020)

[3.1 Dataset 4](#_Toc71360021)

[3.2 Configuring Amazon SageMaker 5](#_Toc71360022)

[3.3 Tasks 8](#_Toc71360023)

[3.4 Result 9](#_Toc71360024)

[3.5 Conclusion 11](#_Toc71360025)

[References: 13](#_Toc71360026)

[Appendix 14](#_Toc71360027)

## 1. Introduction

### 1.1 What is SageMaker.

Amazon SageMaker is a fully managed machine learning service based on cloud that helps users create, design, train, tune, and deploy machine-learning models in a production-ready hosted environment. Traditionally, machine learning is considered as a very helpful way to classification or regression problems. However, high time consuming, too many computing resources and need for math or statistic background restrict its applications in business or industry and limit large scale use of it.

To solve these problems, Amazon released their first version SageMaker in November 2017 and it is the first IDE, Integrated development environment for machine learning. With the help of SageMaker, data scientists and developers can quickly prepare, build, train, deploy machine learning models and monitor their progresses.

I choose this topic, as I think cloud platform and data-oriented mode are the future trend of business. If we combine machine learning technologies and cloud services, it may be a very interesting product. Thus, I want to study further on it.



Figure 1: SageMaker

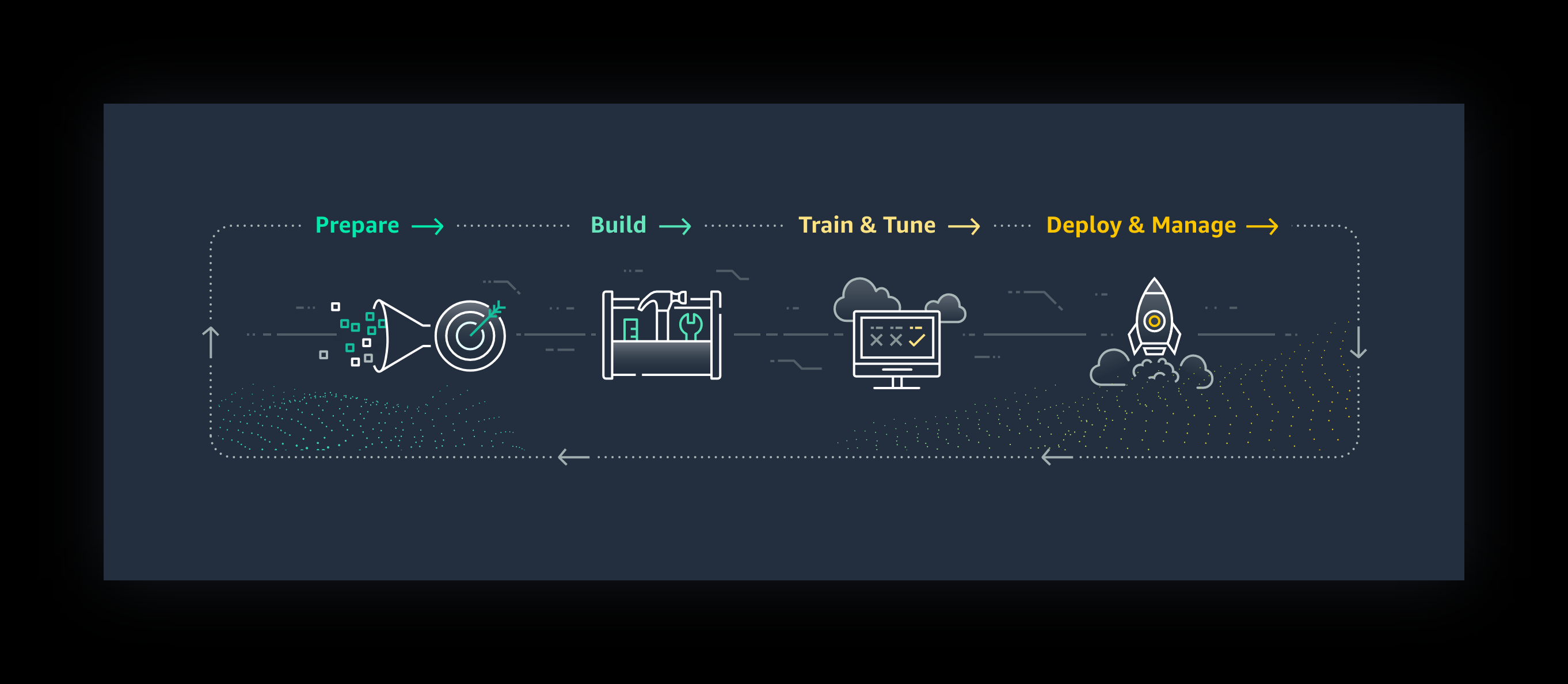


Figure 1: Machine learning life cycle

### 1.2 Amazon SageMaker Features

Amazon SageMaker includes the following features: SageMaker Studio, SageMaker Model Registry, SageMaker Projects, SageMaker Model Building Pipelines, SageMaker ML Lineage Tracking, SageMaker Data Wrangler, SageMaker Feature Store, SageMaker JumpStart, SageMaker Clarify, SageMaker Edge Manager, SageMaker Ground Truth, Amazon Augmented AI, SageMaker Studio Notebooks, SageMaker Experiments, SageMaker Debugger, SageMaker Autopilot, SageMaker Model Monitor, SageMaker Neo, SageMaker Elastic Inference, Reinforcement Learning and so on.

Since there are too many features of Amazon SageMaker (e.g., SageMaker Clarify, SageMaker Data Wrangler, SageMaker JumpStart and etc.), it just focuses on three key features: SageMaker Studio, SageMaker Autopilot and Notebook instance.

SageMaker Studio: An integrated machine learning environment where you can build, train, deploy, and analyze your models all in the same application.

SageMaker Autopilot: Users without machine learning knowledge can quickly build classification and regression models.

SageMaker Studio Notebooks: The next generation of SageMaker notebooks that include AWS Single Sign-On (AWS SSO) integration, fast start-up times, and single-click sharing.



Figure 3: How SageMaker Studio works

## 2. Objective

This final report will showcase the basic use of AWS machine learning services. A demo including configuration of SageMaker and machine learning life cycle will be shown later. In addition, one of the most important functions of SageMaker is to help data scientists do data analysis simply. Three goals will be achieved in this report:

* Learn how to create and run machine learning models on cloud.
* Demonstrate how to do data analysis with Amazon SageMaker Studio and explore the functions of SageMaker.
* A dataset from Kaggle will be used as a sample to test the performance of SageMaker Studio.

Cloud services which will be used in the experiment are SageMaker Studio, Notebook instance, SageMaker Autopilot, S3 bucket, AWS Identity and Access Management (IAM).

## 3. Experiment

### 3.1 Dataset

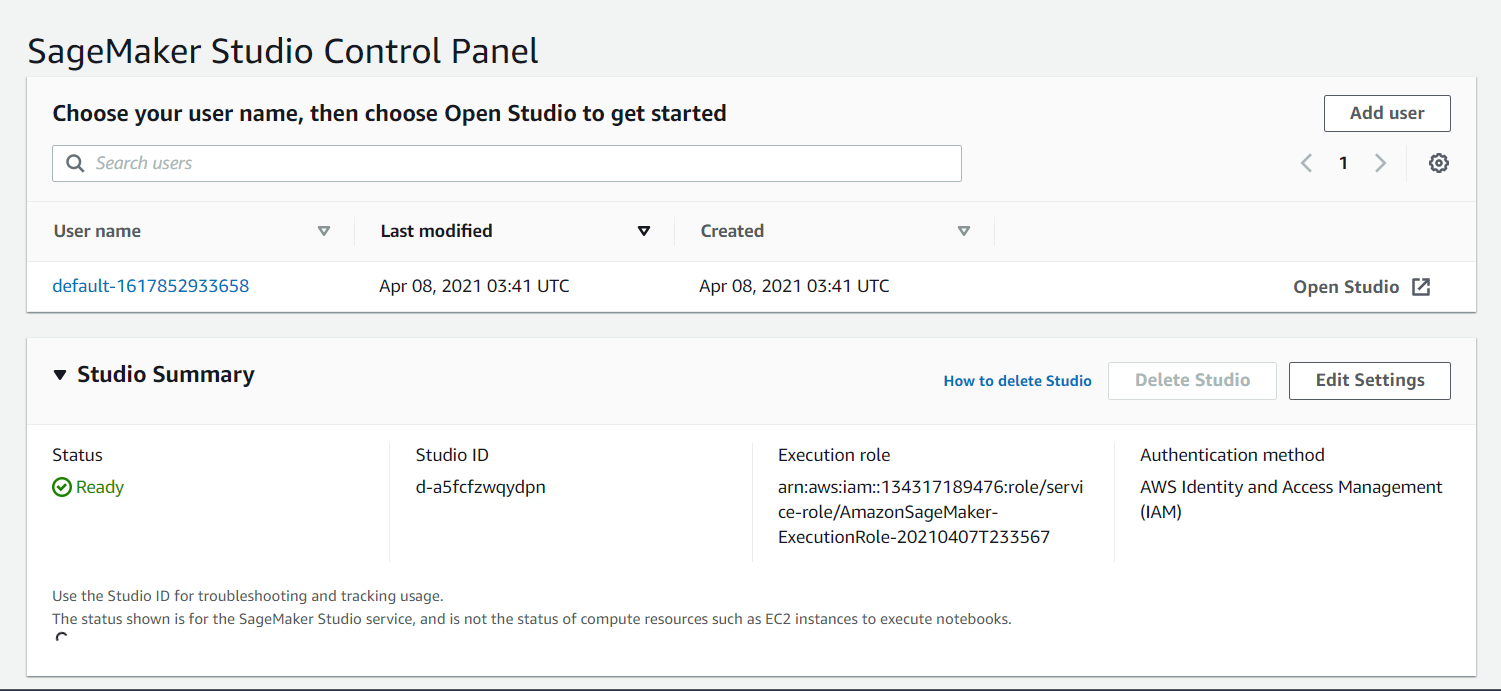
In this study, a small size dataset about video games sales will be used as the test demo which can show the performance and procedures of machine learning in SageMaker. The dataset used in this demo is found at Kaggle (Kaggle, 2016) which contains a list of video games with sales greater than 100,000 copies from 1980 to 2016. The data set is a Statistical data set, which first column contains names of games and the remaining columns consists of Platforms, such as PC, PS4, and genres like action, fighting and several critical game attributes. A thorough description will be provided in the appendix.



### 3.2 Configuring Amazon SageMaker

1. SageMaker Studio:

We cannot directly open SageMaker Studio until we choose or create an IAM user ID. Amazon SageMaker ask their users to use IAM to manage access to SageMaker from the purpose of data security. With IAM identity-based policies, users can specify allowed or denied actions and resources as well as the conditions under which actions are allowed or denied (e.g., fetch dataset from s3 bucket, run machine learning algorithm on EC2 instance).

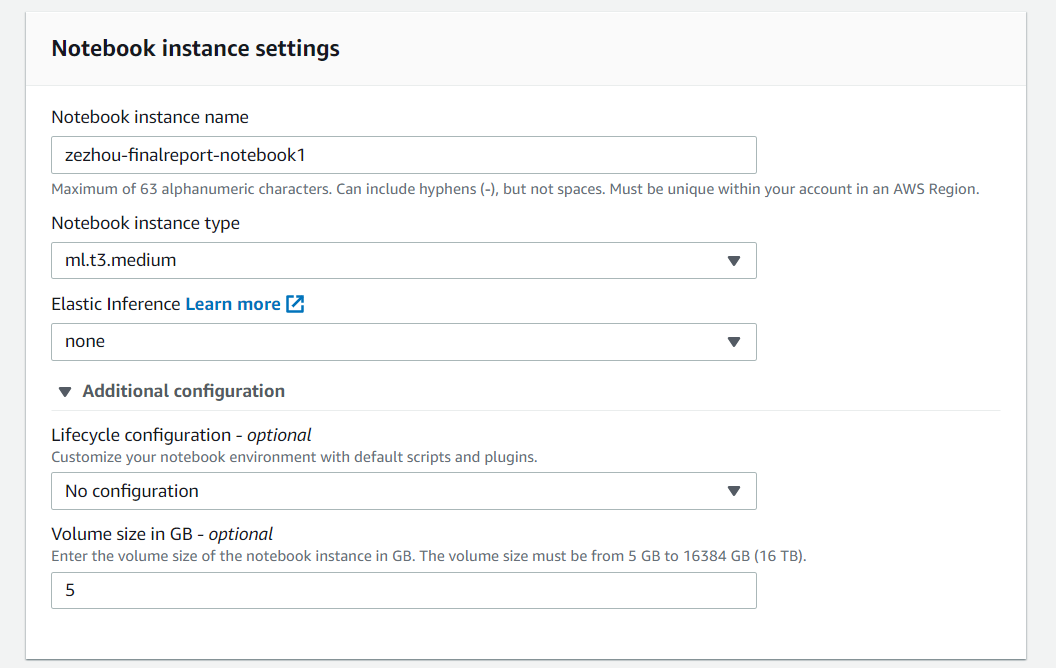


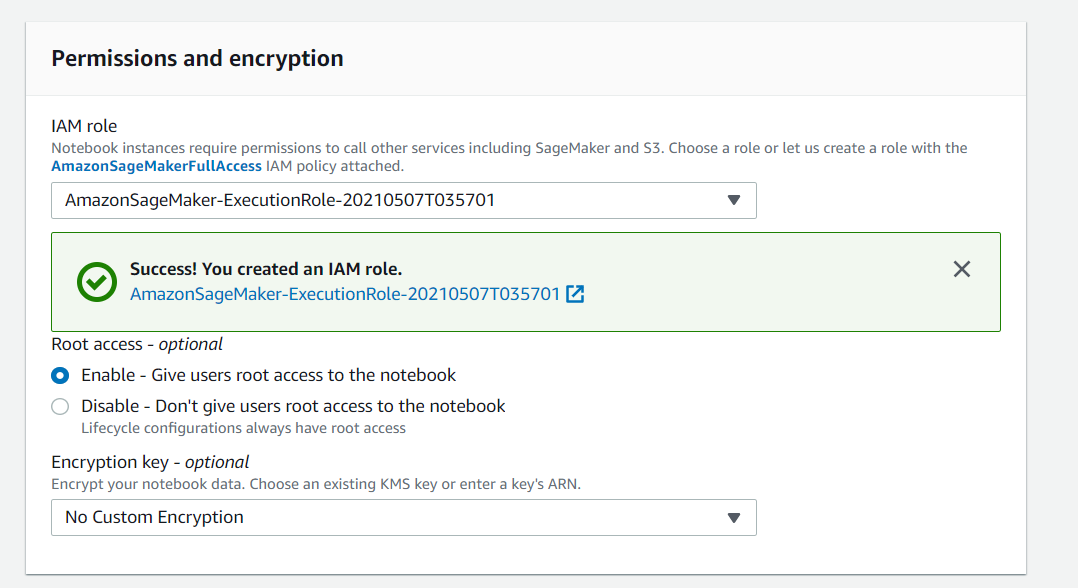
2. Notebook Instance:

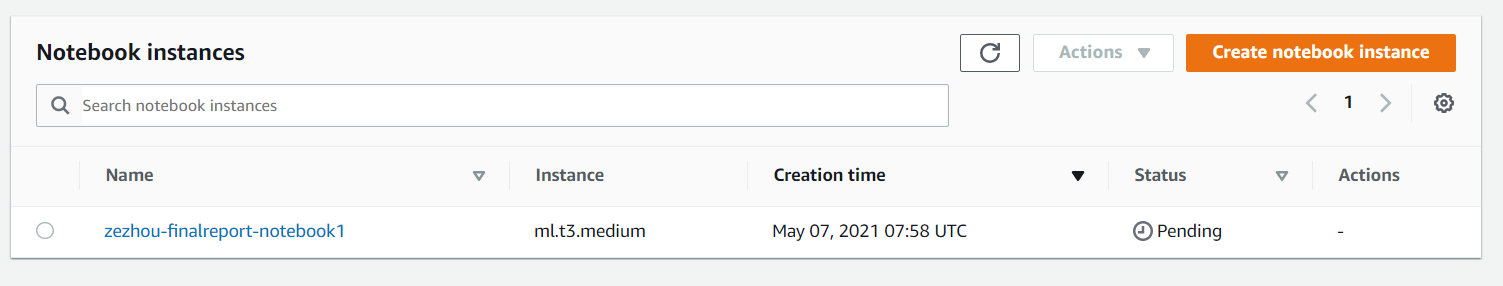
Before within notebook, it begins with EC2 instance. In SageMaker, this is a fully managed instance which means it will not show up under EC2 instance console and users cannot access to this instance by SSH.

First, we need to pick the appropriate shape, size, and family of EC2 instance. EC2 instances have four family, {t, m, c, p} from small to large. Take ml.t2.medium as an example, “ml” means machine learning, “t” means tiny family, “2” means version 2 (latest version: 3) and “medium” means that its size is medium. In this case, I pick a small one since the size of dataset I choose is 2 MB and there are just 50 credits. Therefore, I do not want to waste too many credits on storage and running algorithms. Second, I set the EBS (Elastic Block store) volume as 5GB, the default size which is enough for this case.

I skip security setting such as root volume access to the instance or internet connect because it’s just a toy example. At last, I choose to train the model by local inference. And then, it is time that the Jupyter notebook is ready.

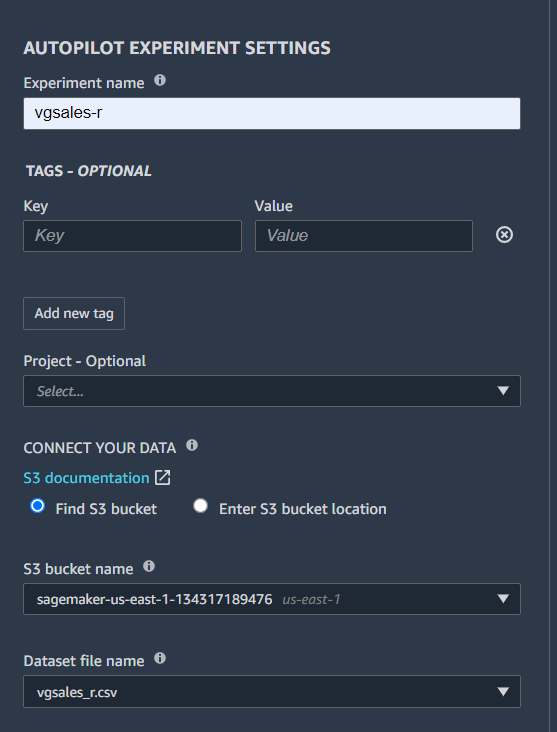
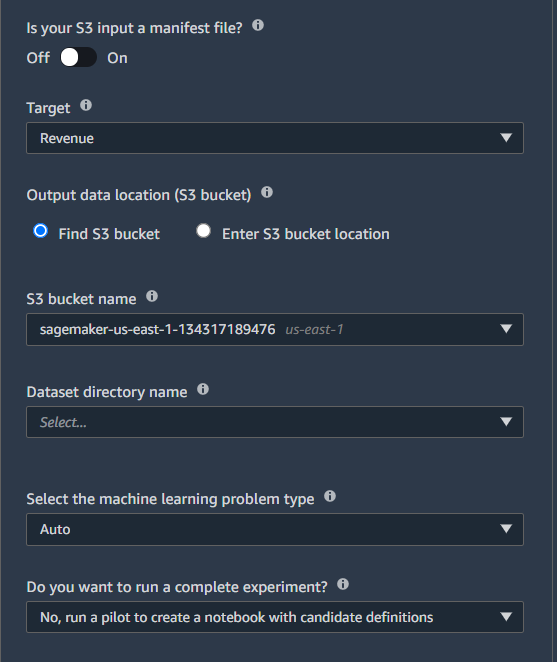


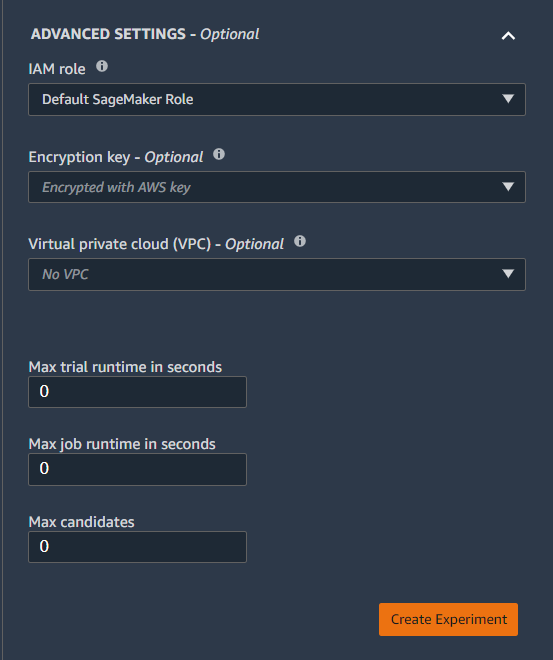




3. SageMaker Autopilot

SageMaker Autopilot will create 250 trials (models) in one experiment, and it cannot be changed. It help people move fast on basic machine learning problems like regression and classification which maximum size is 5 GB. The work procedures are three steps. First, uploading your data into s3 bucket. Second, picking your target column. Third, running your experiment on SageMaker. Autopilot will generate two notebooks for you, one is about data exploratory and the other is about candidatae generate notebook.



### 3.3 Tasks

1. Notebook Instance:

To test the performance of it, I prepare a Xgboost machine learning model. The code will be attached in the appendix which is an original Sagmaker example, and I take it as the template. In Notebook, what should be included are:

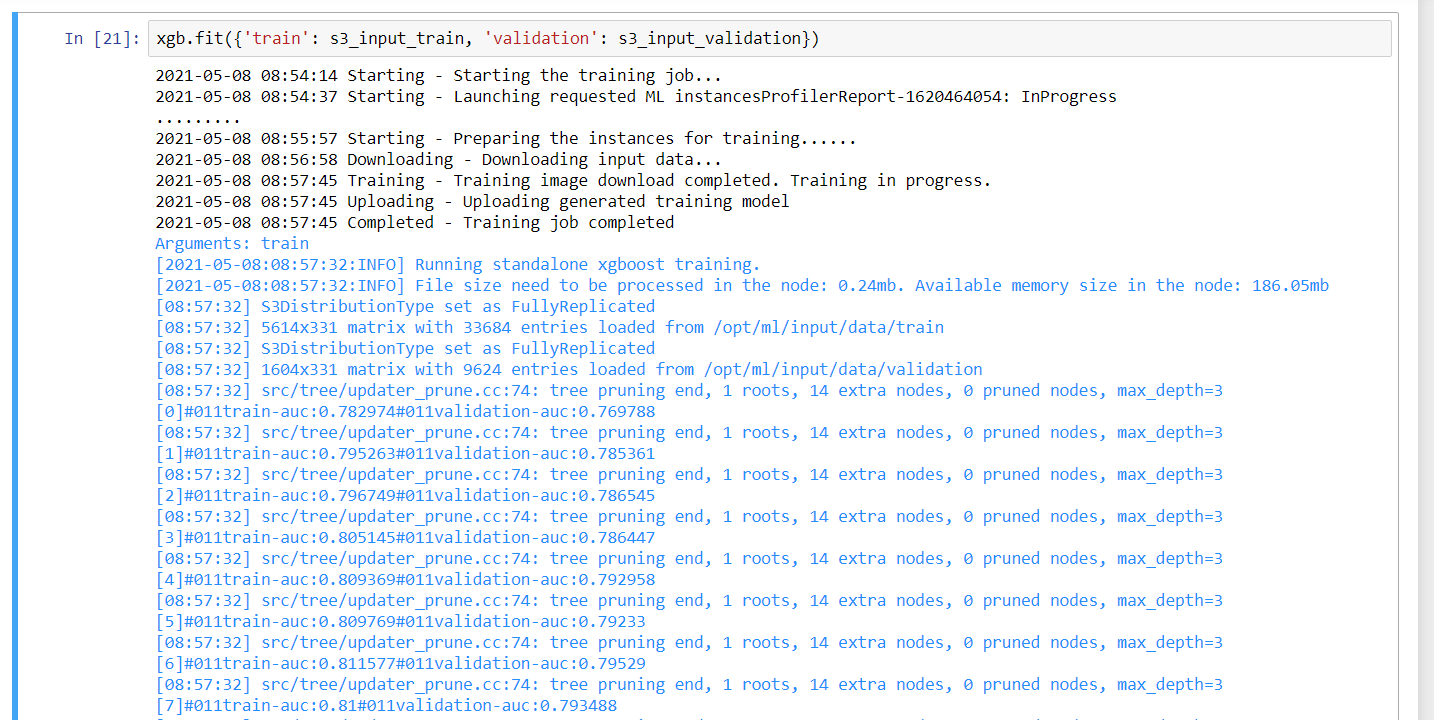
1. Create or choose s3 **bucket** and upload the dataset in it.
2. Fetch the dataset, vgsales from s3 bucket.
3. Do data preprocessing and data transformation for Xgboost.
4. Split data set into training dataset (70%), test dataset (20%) and validation dataset (10%).
5. Create SageMaker estimator with 1 training instance which size is m5.large and set the hyperparameters.
6. Train the machine learning model and deploy it.
7. At last, set the end point and do prediction.

2. SageMaker Autopilot:

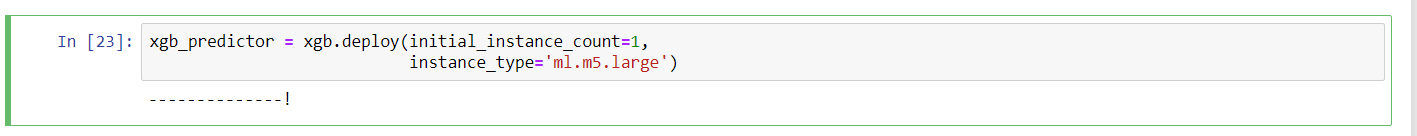
1. Upload the dataset to s3 bucket.
2. Create experiment in SageMaker studio.
3. Wait for the end of 5 autopilot steps: preprocessing, training, feature engineering, tuning parameters and deploy machine learning model.

## 3.4 Result

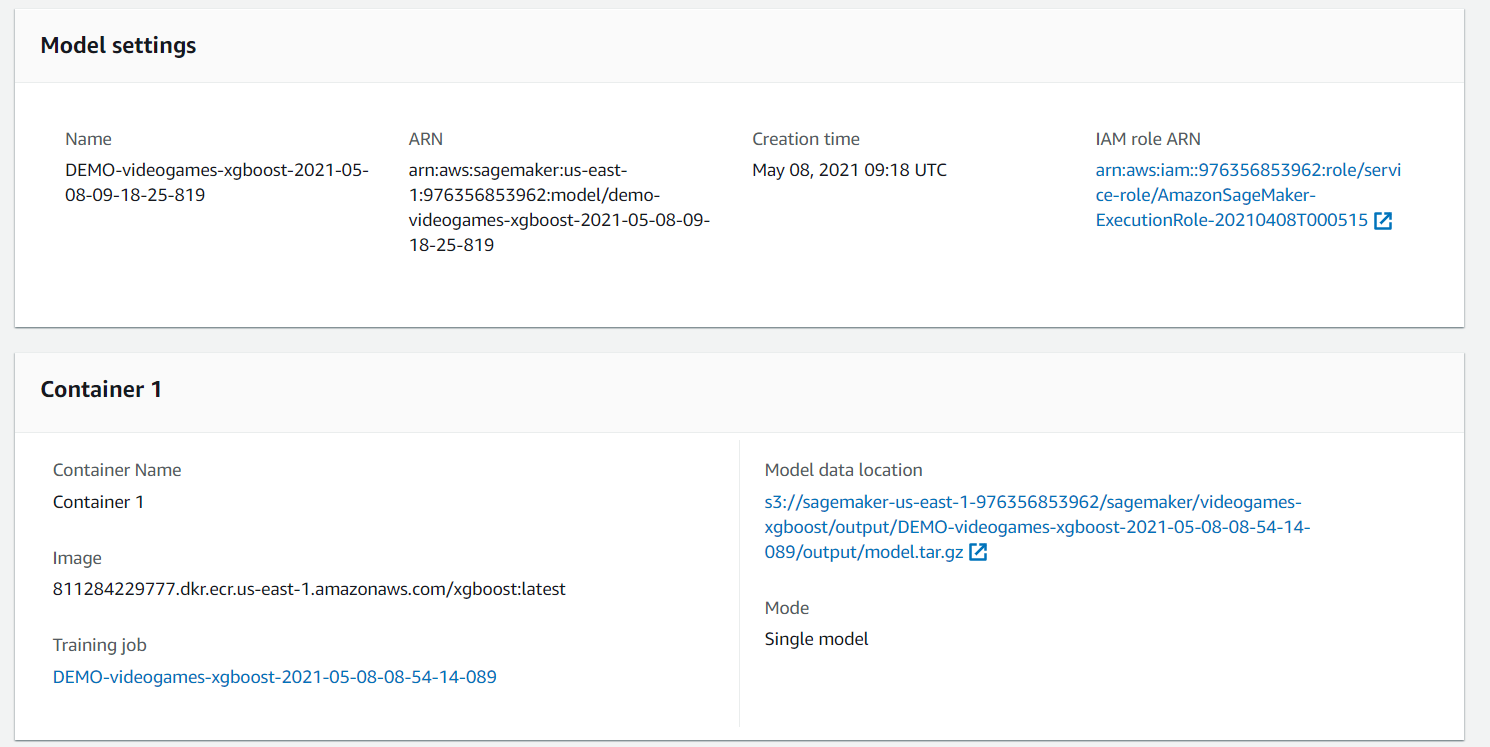
Notebook Instance



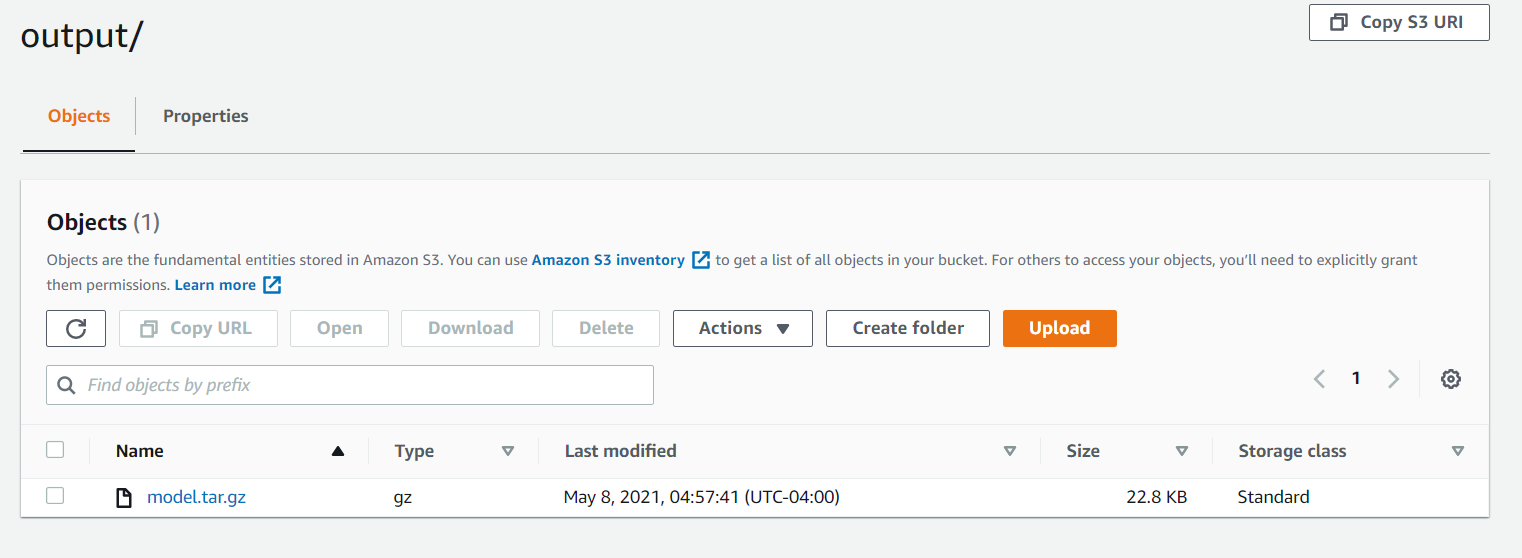
Training



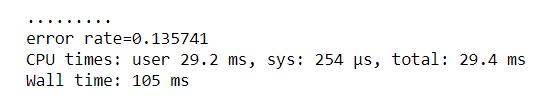
Deploying



Model



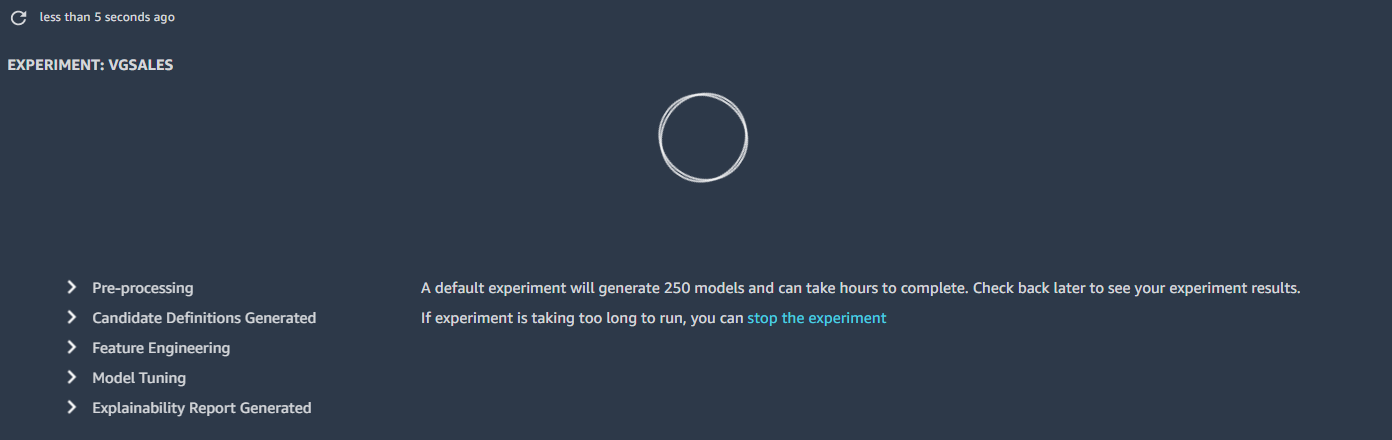
Model in s3 bucket



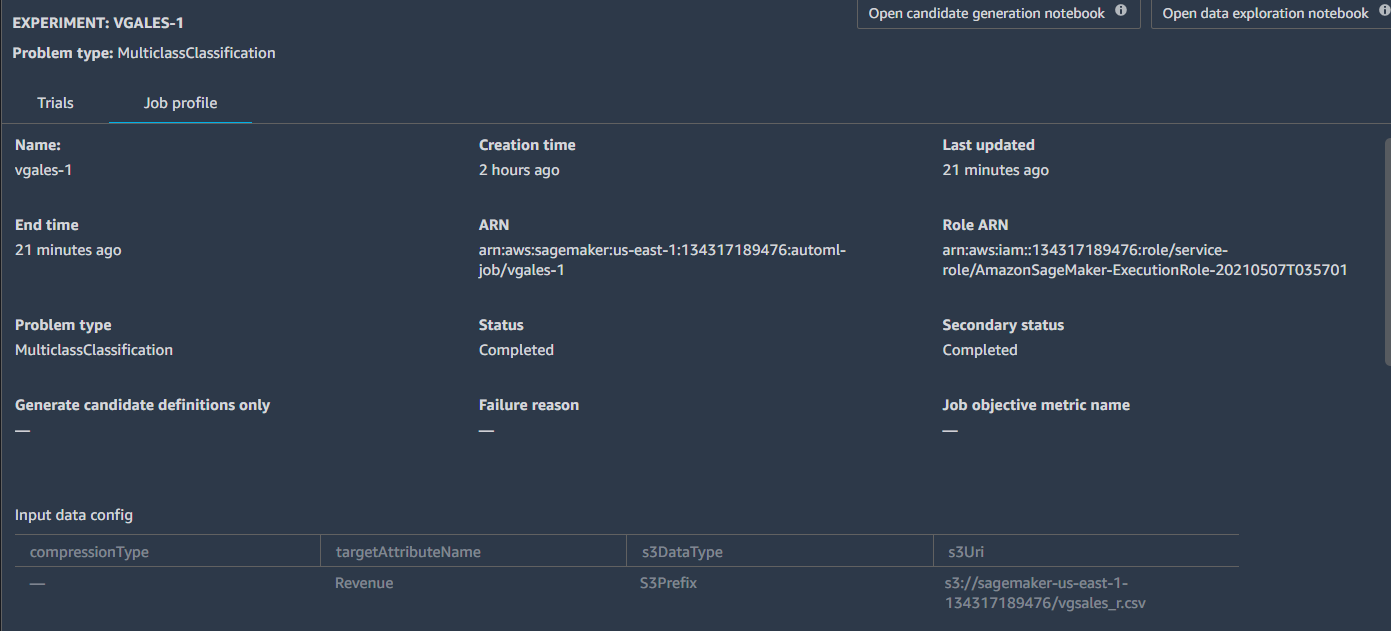
Accuracy

SageMaker Autopilot

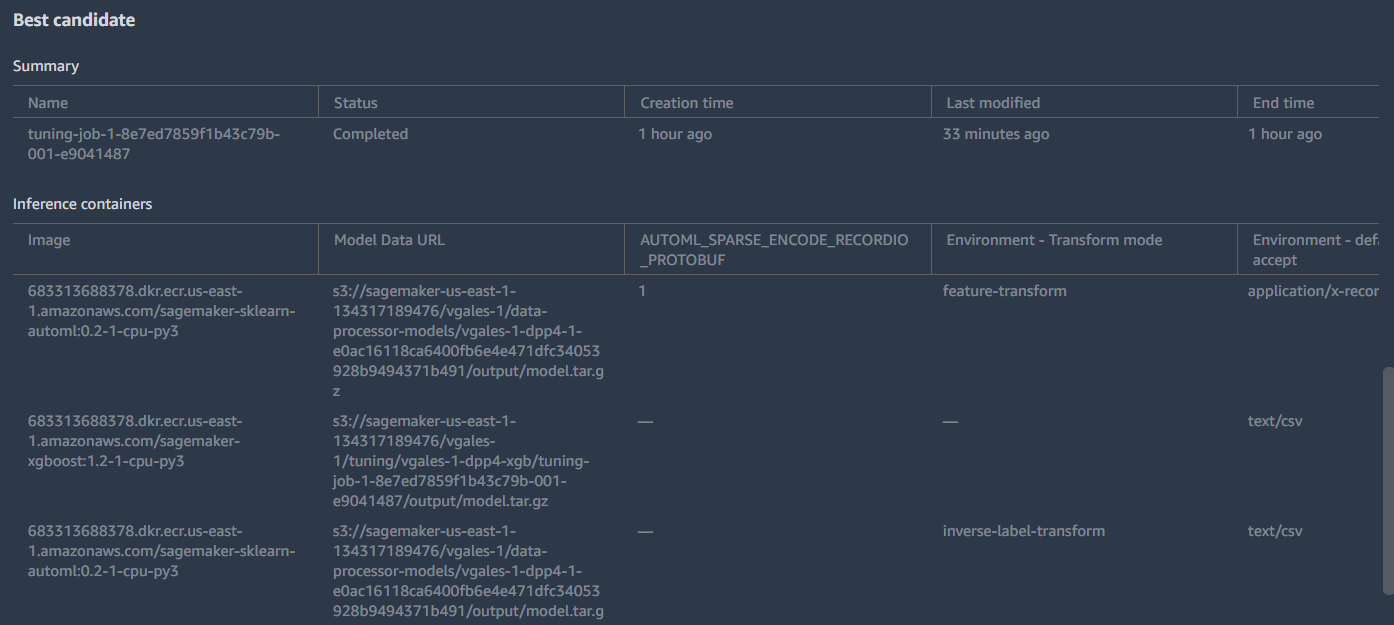
Two notebook reports generated by SageMaker Autopilot are attached in appendix.



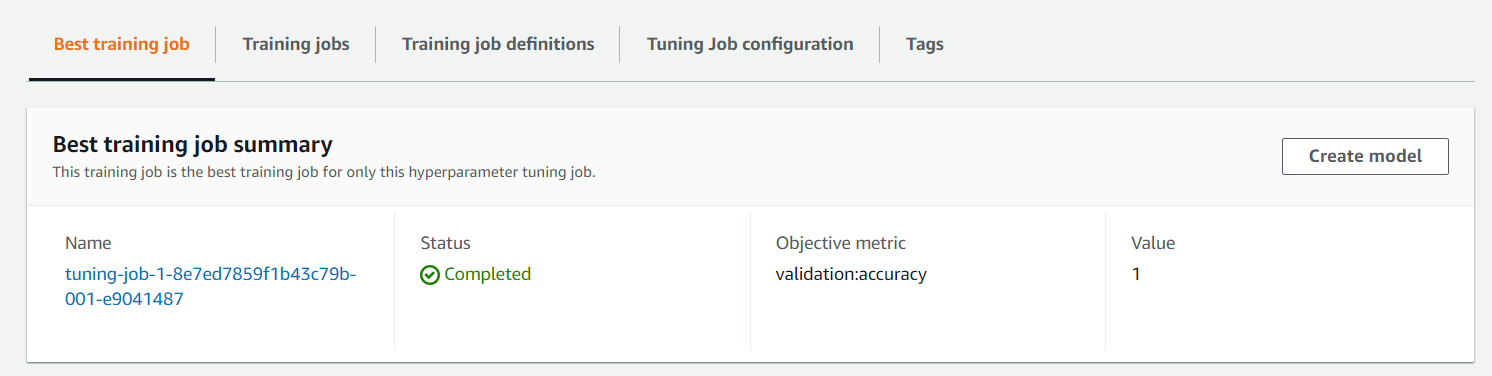
First step of Autopilot



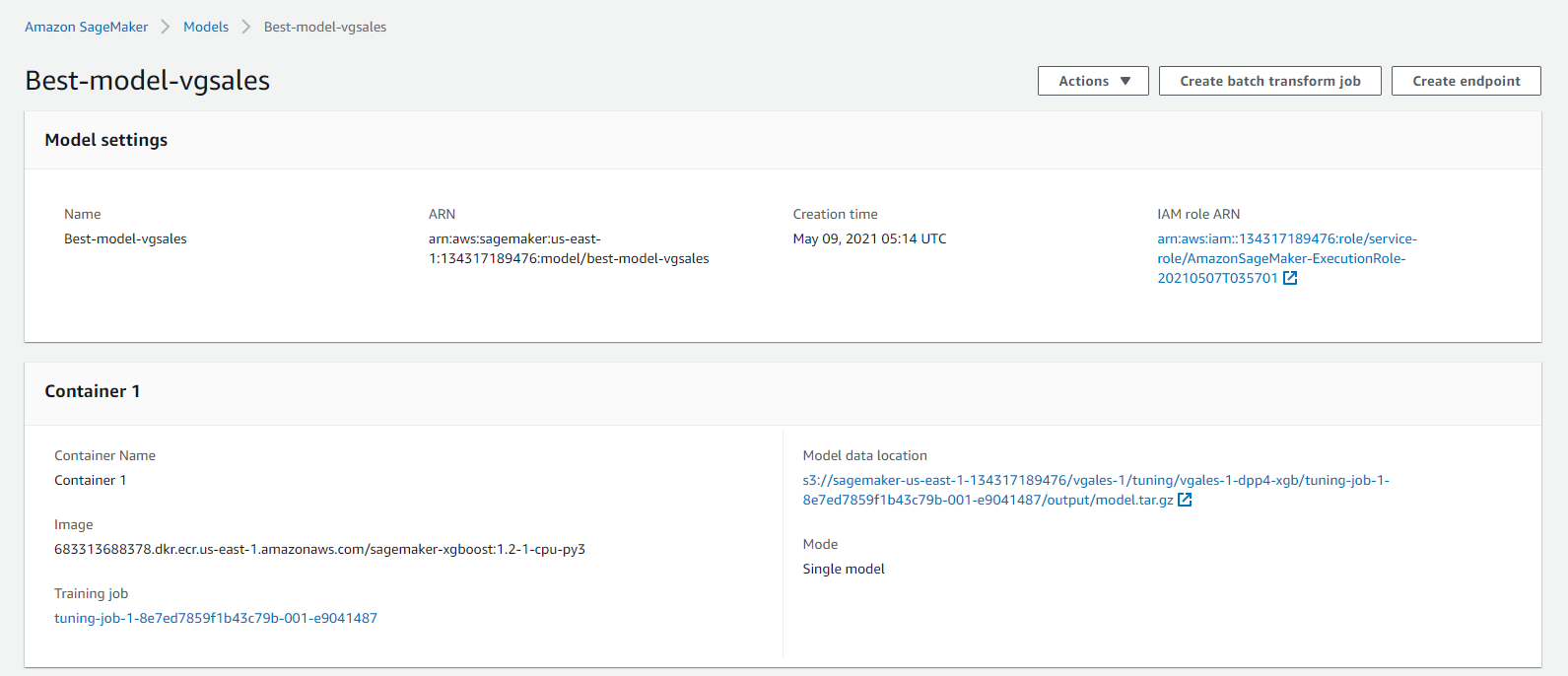
Completed trials



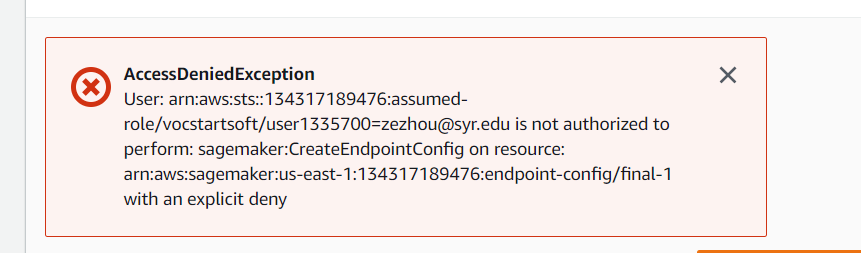
Best Candidate and Three containers



Best Hyperparameters



Model



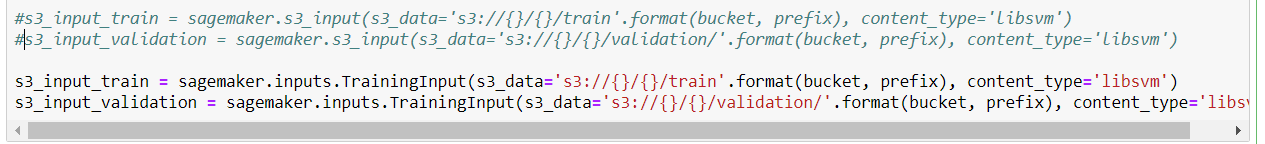
The best hyperparameters generated by SageMaker Autopilot can be used to create the machine learning model for predicting popular video games in the future.

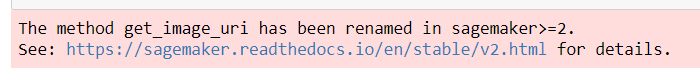
For the restriction of AWS Educate Account, I cannot create endpoint to do prediction work. But, it has been done in the SageMaker notebook instance in starter educate account (ml.m5.large instance).

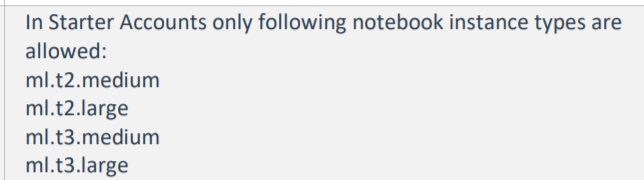
## 3.5 Conclusion

To conclude, Amazon SageMaker is really friendly to people who do not have machine learning or data science background. As what mentioned in its website, SageMaker really help developers and data scientists save so much time in configuration and training works. Also, with SageMaker, they can share their models, hyperparameters or training results in one second by one click which is beneficial for teamwork.

However, there are several problems in using Amazon SageMaker. First, some documentation and examples in SageMaker notebook instance are outdated. What they use is the framework in SageMaker version 1.x, but it is SageMaker version 2.x now. And in some examples, the kernel is still python 2 and what sagemaker support is kernel, python 3. For example, function sagemaker.s3.input has been replaced by sagemaker.inputs.TrainingInput. Since SageMaker is not a very popular cloud service, I cannot easily find the solutions or answers for them in internet or communication which is the biggest problem in this final project.





Second is the restriction on AWS SageMaker in the educate account, as shown below, instances can be used in educate account is just t family instances. And if you want to manage or host your machine learning models in AWS, the minimum instance size is m5.large. it wasted me a lot of time and, finally, I have to create a starter account which has the permission to use m5.large instance. In addition, the 50 credits are not enough to do machine learning tasks because one autopilot experiment may cost 10-20 credits and I just can run autopilot one time to two times in the whole project. And there is also a problem that Autopilot cannot handle dataset with many missing values. If they want to process a dataset with many NAs, they need to preprocess it before uploading it to s3 bucket. However, as what I mentioned before, SageMaker Autopilot is created for people without data science or machine learning background, and they cannot solve it. To information researchers or data scientists, there is no need to use SageMaker Autopilot. So, AWS SageMaker should improve their preprocess capability and optimize their target customers’ experience. 

Educate Account

Finally, so many terminologies in SageMaker make me confused. Pipeline, tensflow, edge manager, augmented ai and so on are difficult to people who are really freshmen in machine learning or data science. To be honest, I still cannot understand some specific SageMaker algorithms requirements without the latest documentation. So, I think some much more detailed and lucid instructions should be provided by AWS SageMaker except the existing instructions like “5 minutes to deploy your machine learning model” or 10 minutes to become data scientist.

## References:

*Amazon SageMaker – Machine Learning – Amazon Web Services*. (n.d.). Amazon Web Services, Inc. Retrieved May 6, 2021, from <https://aws.amazon.com/sagemaker/>

*Amazon SageMaker Studio—Amazon SageMaker*. (n.d.). Retrieved May 6, 2021, from <https://docs.aws.amazon.com/sagemaker/latest/dg/studio.html>

*An Introduction To AWS SageMaker*. (n.d.). Retrieved May 6, 2021, from <https://www.simplilearn.com/tutorials/aws-tutorial/aws-sagemaker>

*Deploy Thousands of Models on SageMaker Real-Time Endpoints with Automatic Retraining Pipelines | by Emily Webber | May, 2021 | Towards Data Science*. (n.d.). Retrieved May 6, 2021, from <https://towardsdatascience.com/deploy-thousands-of-models-on-sagemaker-real-time-endpoints-with-automatic-retraining-pipelines-4eef7521d5a3>

*Video Game Sales*. (n.d.). Retrieved May 6, 2021, from <https://kaggle.com/gregorut/videogamesales>

*An Introduction to AWS SageMaker*. (n.d.). Retrieved May 6, 2021, from <https://www.simplilearn.com/tutorials/aws-tutorial/aws-sagemaker>

*Video-game-sales-xgboost*. (n.d.). Retrieved May 8, 2021, from <https://vgsales.notebook.us-east-1.sagemaker.aws/examples/preview?example_id=%2Fhome%2Fec2-user%2Fsample-notebooks%2Fintroduction_to_applying_machine_learning%2Fvideo_game_sales%2Fvideo-game-sales-xgboost.ipynb>

## Appendix

1. Data set:

This dataset contains a list of video games with sales greater than 100,000 copies. It was generated by a scrape of [vgchartz.com](http://www.vgchartz.com/).

Data dictionary:

* Rank - Ranking of overall sales
* Name - The games name
* Platform - Platform of the games release (i.e. PC,PS4, etc.)
* Year - Year of the game's release
* Genre - Genre of the game
* Publisher - Publisher of the game
* NA\_Sales - Sales in North America (in millions)
* EU\_Sales - Sales in Europe (in millions)
* JP\_Sales - Sales in Japan (in millions)
* Other\_Sales - Sales in the rest of the world (in millions)
* Global\_Sales - Total worldwide sales.
* Critic\_score - Aggregate score compiled by Metacritic staff
* Critic*count - The number of critics used in coming up with the Critic*score
* User\_score - Score by Metacritic's subscribers
* User*count - Number of users who gave the user*score
* Developer - Party responsible for creating the game
* Rating - The [ESRB](https://www.esrb.org/) ratings

1. Notebook: it is an old SageMaker Example (SageMaker version 1) which is outdated with huge number of renamed functions and different syntax. I modified it to fit version 2.0 of SageMaker.
2. SageMaker Autopilot reports
   1. SageMakerAutopilotCandidateDefinitionNotebook\_1
   2. SageMakerAutopilotDataExplorationNotebook
3. Link to demo videos:
4. Demo-1: <https://video.syr.edu/media/t/1_scw6yz6t>
5. Demo-2: <https://video.syr.edu/media/t/1_bsqq8ll7>

Link for the first demo: <https://video.syr.edu/media/t/1_scw6yz6t>

Link for the second demo: <https://video.syr.edu/media/t/1_bsqq8ll7>