Spam Detection using Azure Automated ML

Spam refers to unsolicited or unwanted electronic messages, typically sent in bulk. These messages can come in various forms, such as emails, text messages, or social media posts, and often contain irrelevant or unwanted content. Spam is important to understand because it can have a negative impact on both individuals and businesses. For individuals, spam can be a nuisance, clogging up inboxes and wasting time. In some cases, spam can even be a security risk, as it may contain malicious links or attachments.

1. Azure Resources

1.1 Azure Subscription

In this project we are using **Microsoft Azure Machine Learning** you will need to sign into the Azure Portal at Microsoft Azure to create the resources. If you are using Azure portal for the first time you can get the subscription from Free Account for free.

1.2 Create Azure Resource Group

Azure resource group is a logical container in Microsoft Azure that is used to organize and manage related resources. A resource group is a way to group Azure resources such as virtual machines, storage accounts, and virtual networks our project will consist of the Spam data set, compute from training, then compute to run our web service for inferencing. There are two ways to create an Azure Resource Group.

1.2.1 Create Azure Resource Group - Use Azure Portal

To use the Azure web portal to create the Azure Resource Group follow following steps - create an Azure Resource Group using portal

1.2.2 Create Azure Resource Group - Use Azure Portal

If you would rather create the Resource Group using bash then you can use Azure Cloud Shell .Execute the subsequent commands using Azure Cloud Shell (bash): `Command`

```
resourceGroupName=spam$RANDOM-rg
location=SouthCentralUS

az group create \
   --name $resourceGroupName \
   --location $location
```

1.3 Create Azure Machine Learning Workspace

Azure Machine Learning is a cloud service for accelerating and managing the machine learning project lifecycle. we can create a model in Azure Machine Learning or use a model built from an open-source platform, such as Pytorch, TensorFlow, or scikit-learn. There are also two ways to create an Azure Resource Group.

1.3.1 Create Azure Machine Learning Workspace - Use Azure Portal

To create Azure Machine Leaning resource from the Azure portal – Create Azure Machine Learning Workspace

1.3.2 Create Azure Resource Group - Use Azure Portal

Execute the subsequent commands using [Azure Cloud Shell](bash) to create an Azure Machine Learning Workspace `Command`

```
workspace=spam-$RANDOM

az extension add -n azure-cli-ml

az ml workspace create -w $workspace -g $resourceGroupName  --sku enterprise
```

2. Spam Classification with Automated Machine Learning

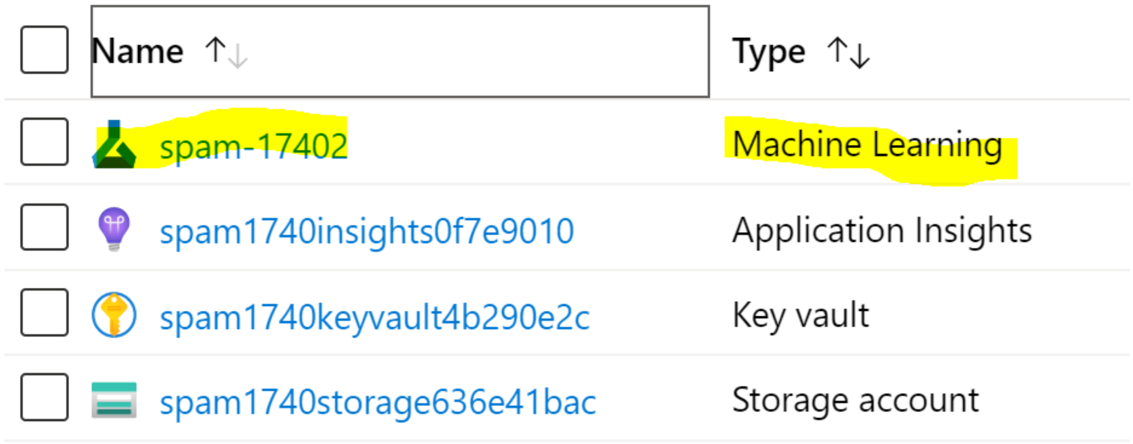
Using the Azure resources that we have collected now we can use Azure Automated Machine Learnig in the training of best classification model.

2.1 Downloading Spam Dataset

First of all we need to download the Spam Dataset - this data set came from the University of California Irvine ML Repository. The comma-delimited file that automated ml will use for training is contained in the zip file. Once you've downloaded the file, please unzip it.

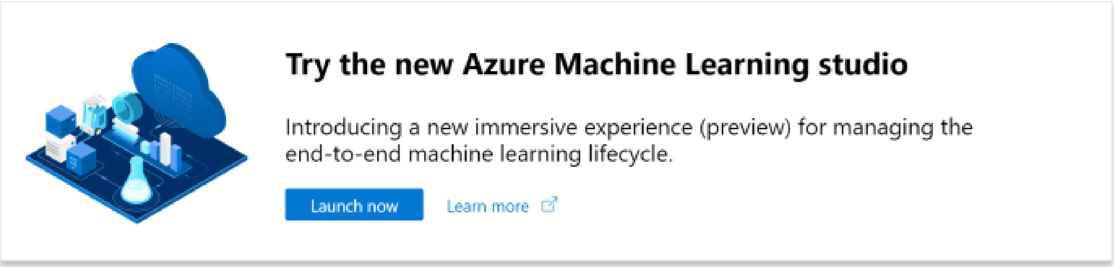
2.2 Creating an Automated ML Run

Please access your Azure Machine Learning Workspace by clicking on the Azure Resource Group you created in the Azure Portal. I created **spam-17402** using the bash script in this post; yours will have a different name.



□	Name ↑↓	Type ↑↓
□	spam-17402	Machine Learning
□	spam1740insights0f7e9010	Application Insights
□	spam1740keyvault4b290e2c	Key vault
□	spam1740storage636e41bac	Storage account

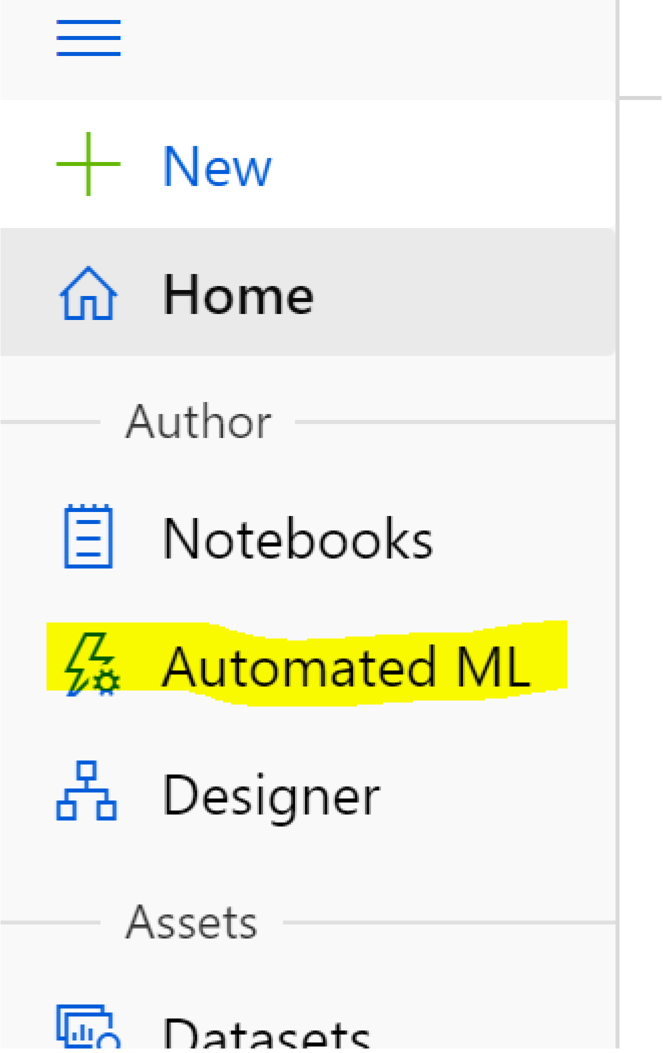
Once you have accessed the Machine Learning Resource then click **Launch now** where is says **Try the new Azure Machine Learning studio** towards the center of the screen as the remainder of the post will be using this new UX.

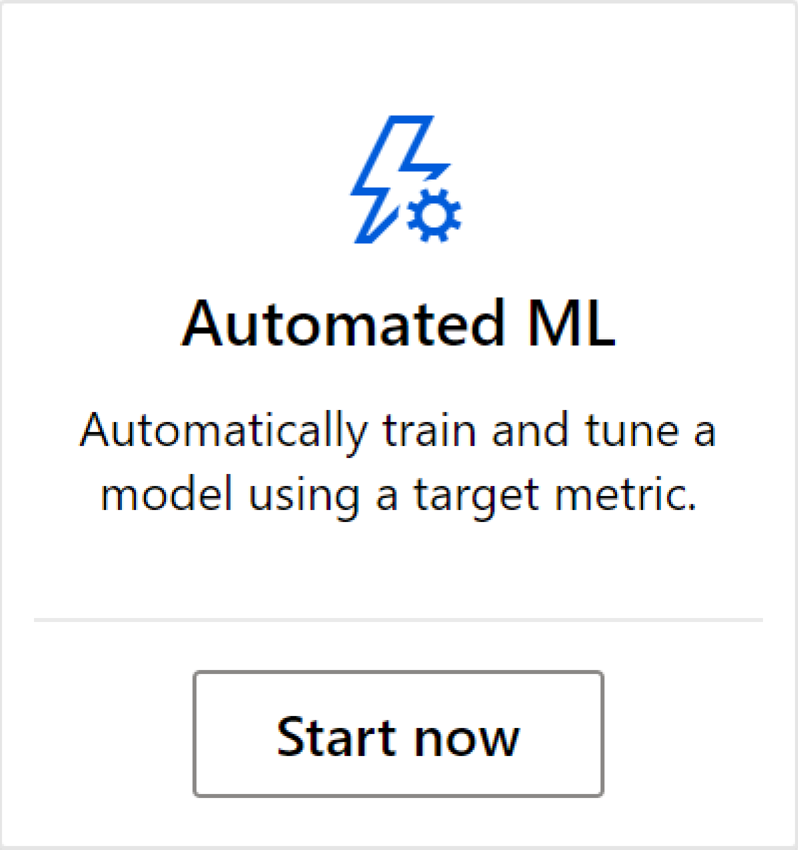


Try the new Azure Machine Learning studio

Introducing a new immersive experience (preview) for managing the end-to-end machine learning lifecycle.

Now you can select Automated ML from the left navigation or choose Automated ML **‘Start Now’**.

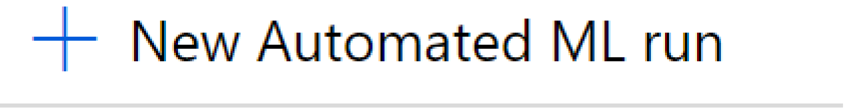




Automated ML

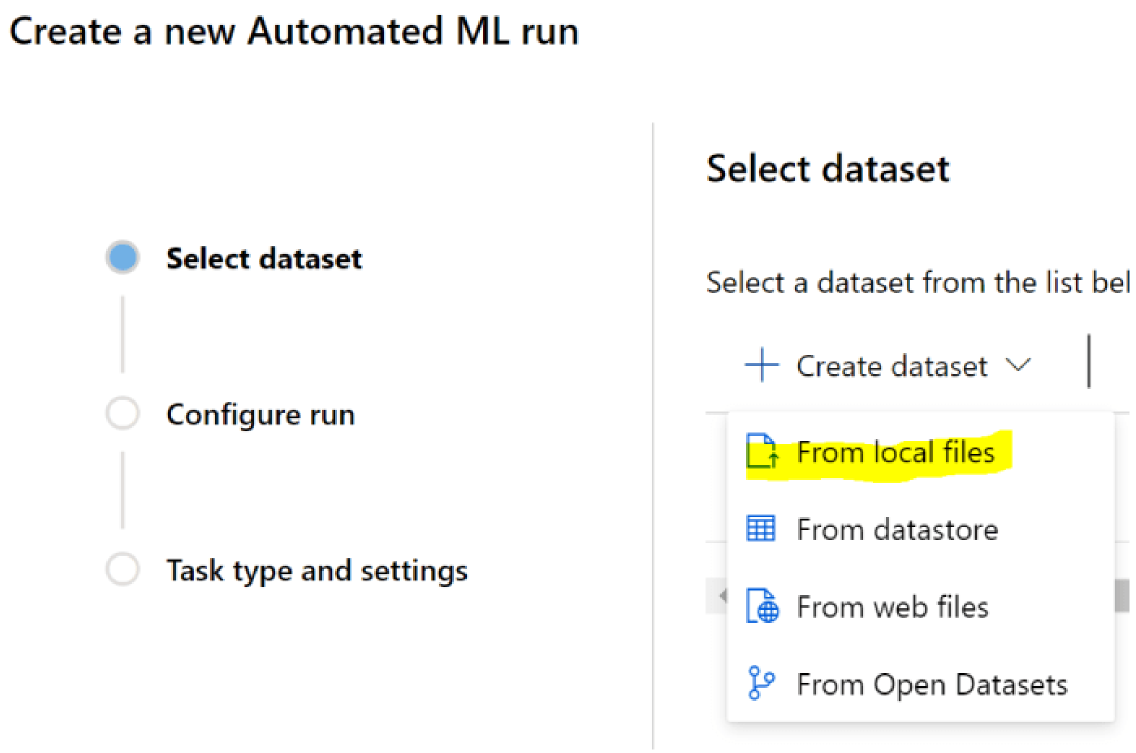
Automatically train and tune a model using a target metric.

Once on the Automated ML screen, select **New Automated ML Run**.

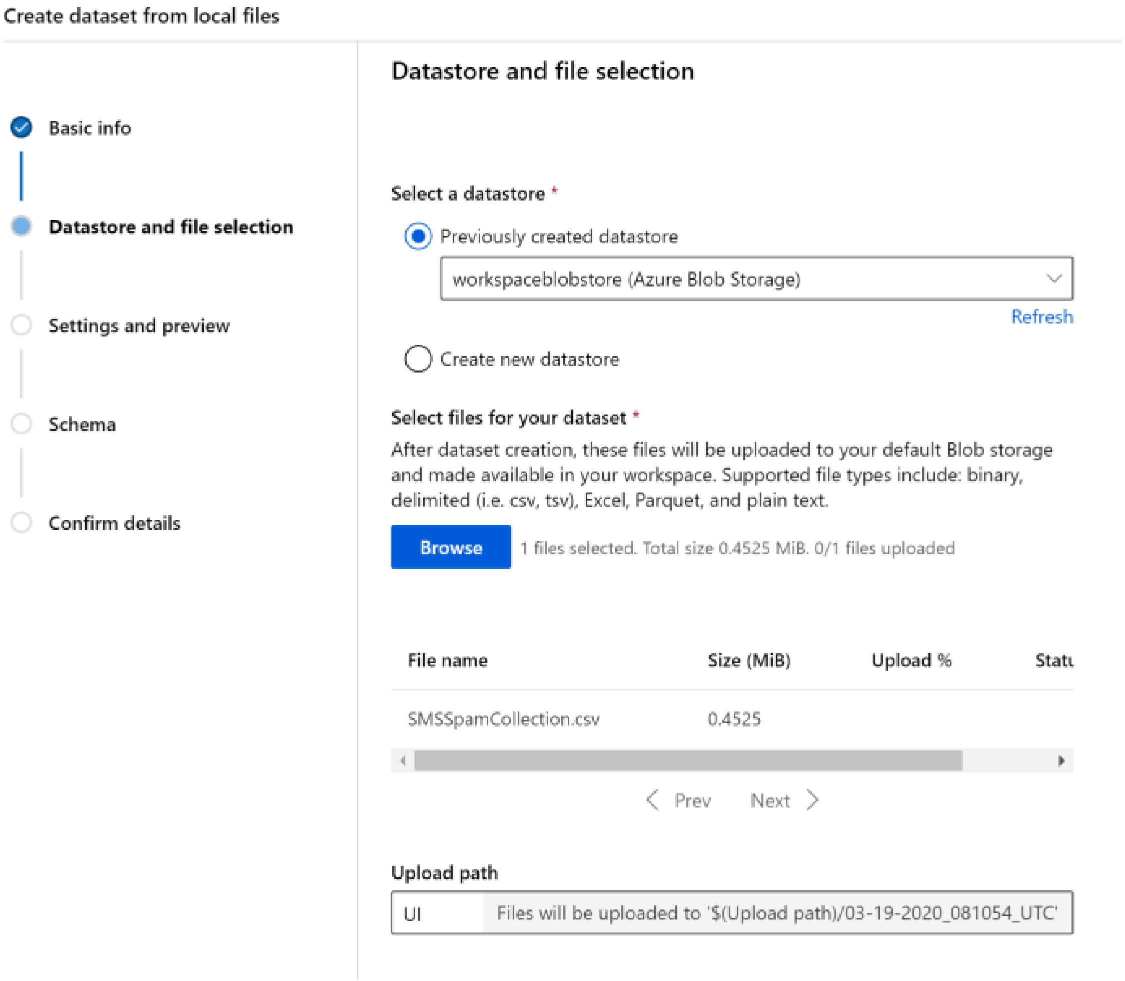


2.2.1 Automated ML Dataset

Now, we will create dataset in our Automated ML run using the database that we have downloaded. Select **Create dataset**, then select **From local files**.



Give name and description as this dataset can be used in other models and accessed via the SDK. Select **Browse** and find the downloaded and unzipped spam csv file.



Create dataset from local files

Datastore and file selection

Select a datastore *

● Previously created datastore

workspaceblobstore (Azure Blob Storage)

○ Create new datastore

Select files for your dataset *

After dataset creation, these files will be uploaded to your default Blob storage and made available in your workspace. Supported file types include: binary, delimited (i.e. csv, tsv), Excel, Parquet, and plain text.

1 files selected. Total size 0.4525 MiB. 0/1 files uploaded

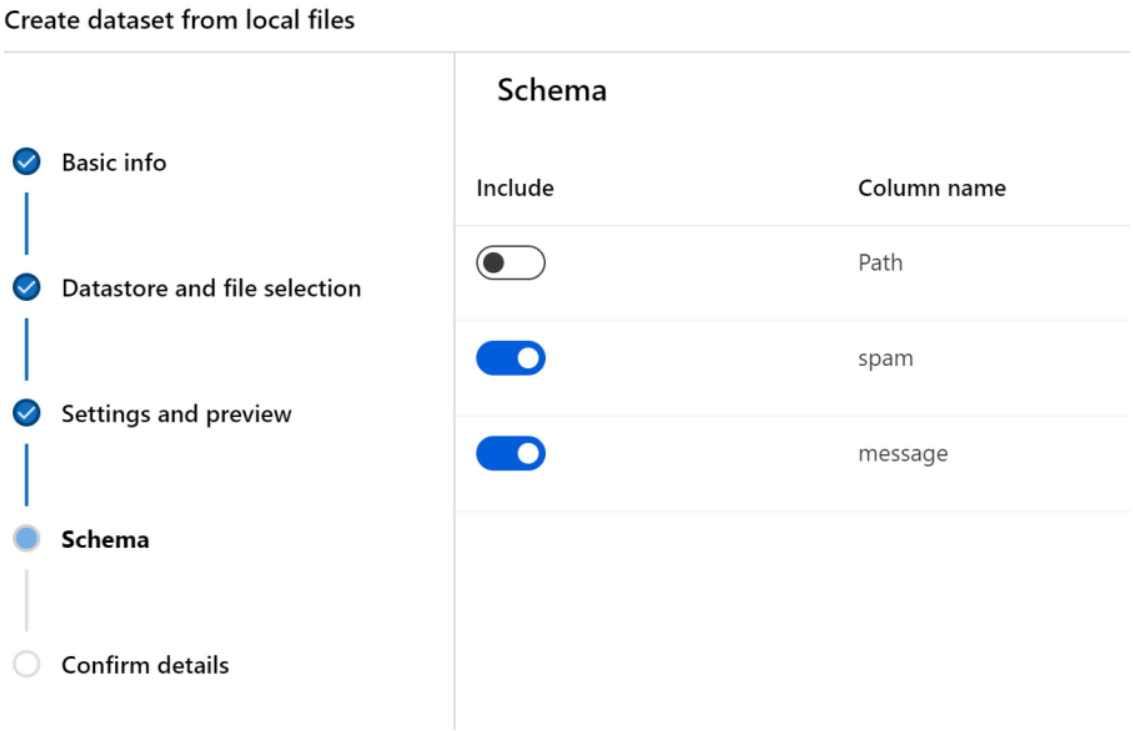
File name	Size (MiB)	Upload %	Statu
SMSSpamCollection.csv	0.4525		

Upload path

UI Files will be uploaded to '$(Upload path)/03-19-2020_081054_UTC'

Moving to the next step after selecting the csv file. Leave the default settings and schema selection:

Setting	Value
File Format	Delimited
Delimeter	Comma
Encoding	UTF-8
Column Headers	Use headers from the first file
Skip rows	None



Create dataset from local files

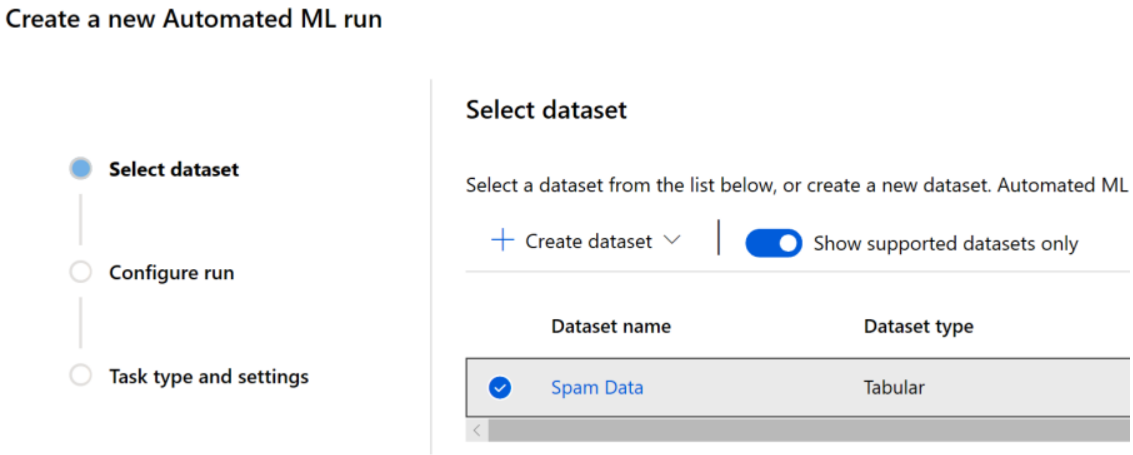
- ✓ Basic info
- ✓ Datastore and file selection
- ✓ Settings and preview
- ● **Schema**
- ○ Confirm details

Schema

Include	Column name
○	Path
●	spam
●	message

2.2.2 Automated ML Compute

As we have added the dataset now we are ready to configure our run. Please make sure to select the dataset you created before configuring your run. Then select **Next**.



Create a new Automated ML run

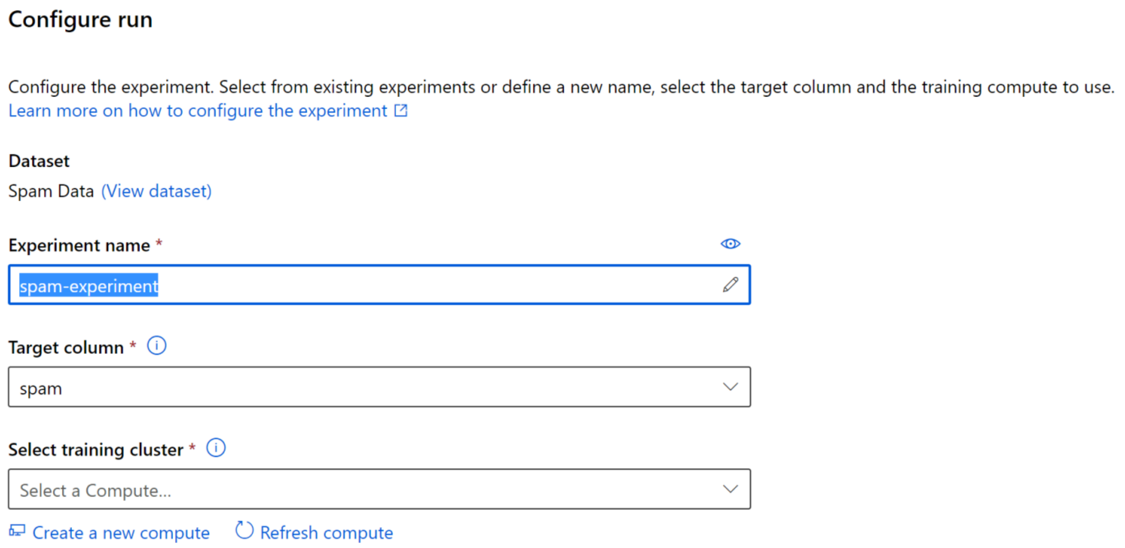
- ● **Select dataset**
- ○ Configure run
- ○ Task type and settings

Select dataset

Select a dataset from the list below, or create a new dataset. Automated ML

	Dataset name	Dataset type
✓	Spam Data	Tabular

Give your experiment a descriptive name, such as **spam-experiment**. The dataset's spam column is located in the **Target column** (label). You must build a new compute if this is your first training run with Azure Machine Learning; otherwise, you will have compute options under **Select training cluster**.



Configure run

Configure the experiment. Select from existing experiments or define a new name, select the target column and the training compute to use.
Learn more on how to configure the experiment

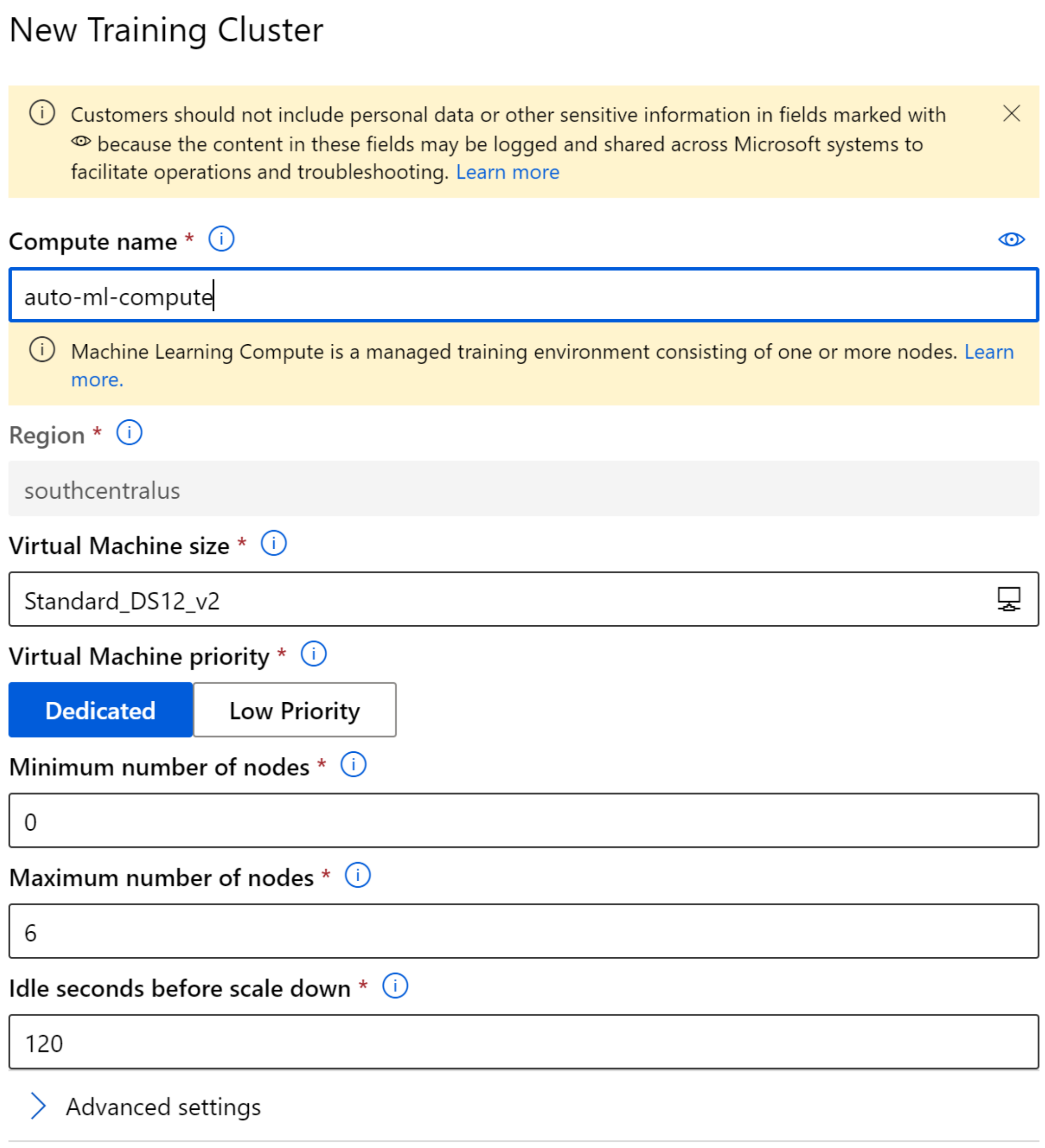
Dataset
Spam Data (View dataset)

Experiment name *
spam-experiment

Target column *
spam

Select training cluster *
Select a Compute...

Select **Create a new compute** to proceed with a new computer cluster. The name need not be specific to automl because compute can be used for other training runs. You can use the default settings for this post because provisioning the cluster's virtual machines will only take a short while.



New Training Cluster

ⓘ Customers should not include personal data or other sensitive information in fields marked with 👁 because the content in these fields may be logged and shared across Microsoft systems to facilitate operations and troubleshooting. Learn more

Compute name * ⓘ

auto-ml-compute

ⓘ Machine Learning Compute is a managed training environment consisting of one or more nodes. Learn more.

Region * ⓘ

southcentralus

Virtual Machine size * ⓘ

Standard_DS12_v2

Virtual Machine priority * ⓘ

Minimum number of nodes * ⓘ

0

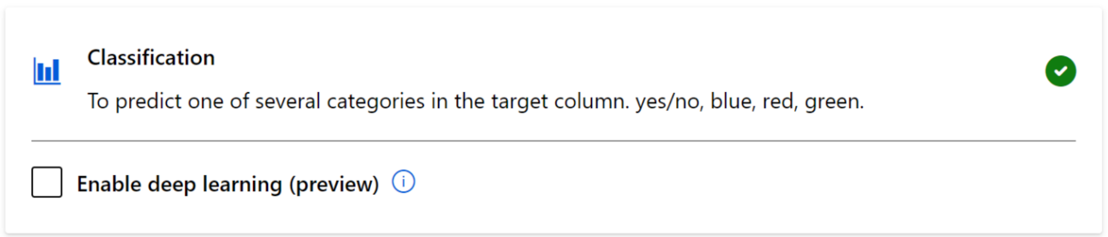
Maximum number of nodes * ⓘ

6

Idle seconds before scale down * ⓘ

120

```
The next step, Tack Type and Settings, requires you to choose Classification as the
task type after provisioning your compute and adding it to the run configuration.
```



Classification ✓

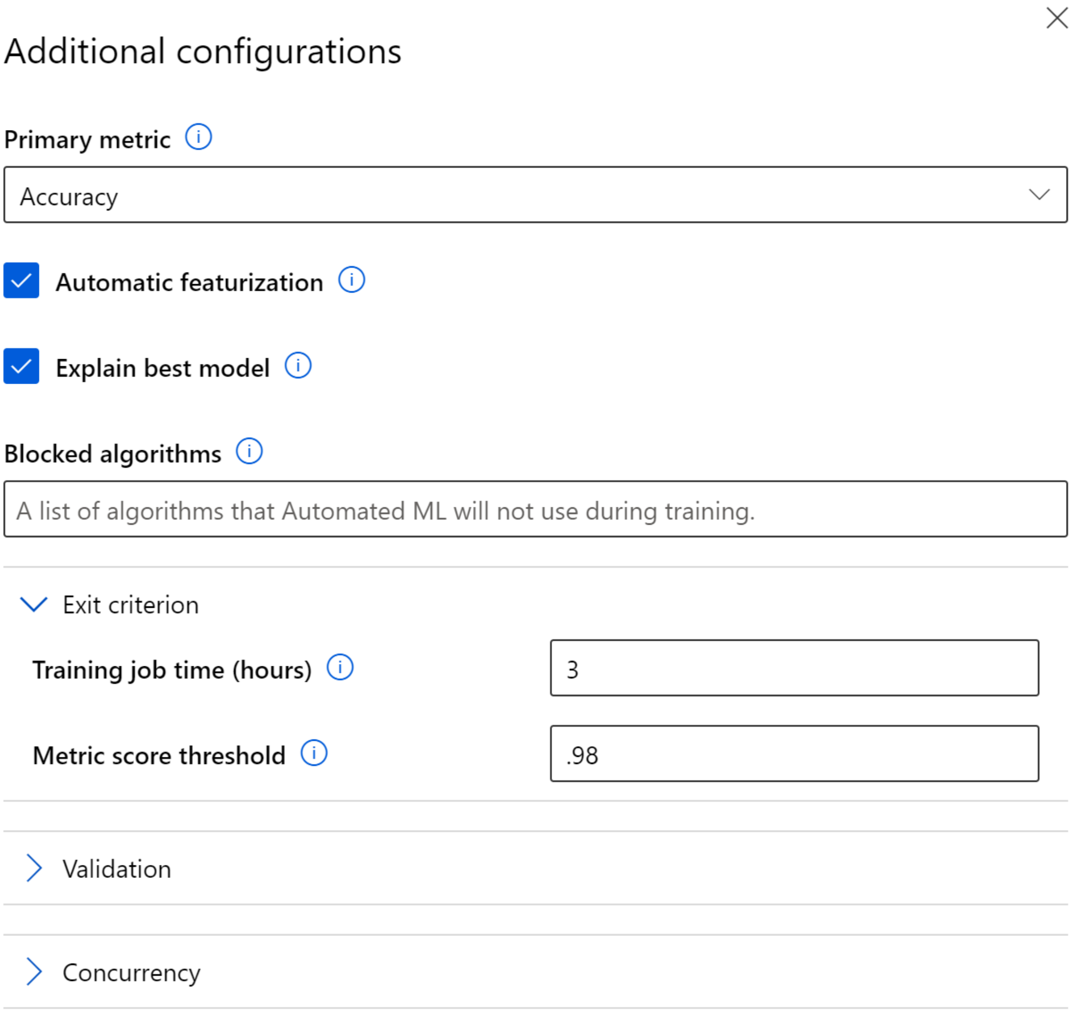
To predict one of several categories in the target column. yes/no, blue, red, green.

□ Enable deep learning (preview) ⓘ

You can modify the run setting by clicking **View additional configuration settings**.

View additional configuration settings

In this proect, I set an exit score of **0.98** for my run to early terminate when my accuracy reaches 0.98.



Additional configurations

Primary metric

Accuracy

☑ Automatic featurization

☑ Explain best model

Blocked algorithms

A list of algorithms that Automated ML will not use during training.

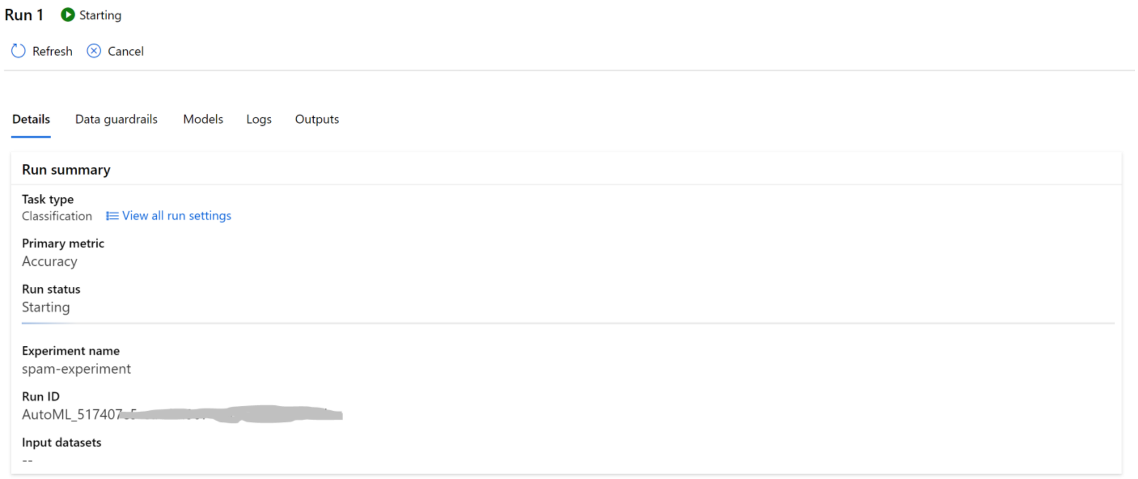
Exit criterion

Training job time (hours)	3
Metric score threshold	.98

Validation

Concurrency

Cogratulation - you have successfully completed first **Automated ML Run**.



Run 1 Starting

Refresh Cancel

Details | Data guardrails | Models | Logs | Outputs

Run summary

Task type
Classification View all run settings

Primary metric
Accuracy

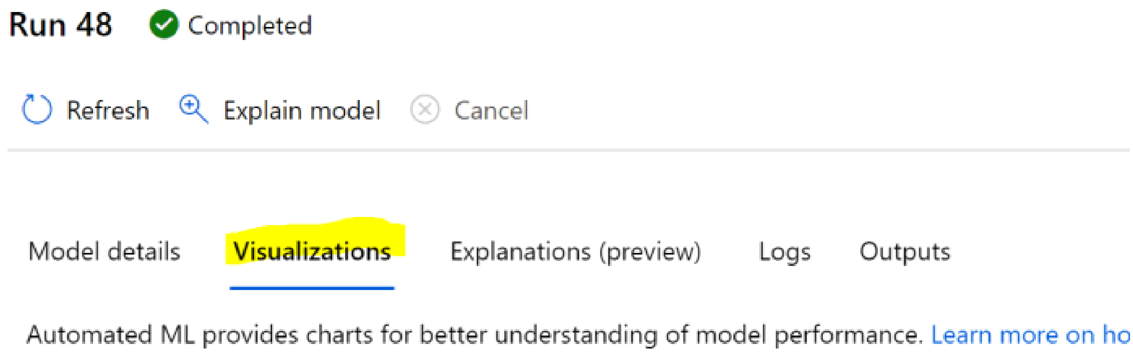
Run status
Starting

Experiment name
spam-experiment

Run ID
AutoML_517407-[redacted]

Input datasets
--

once your Run is complete, you canview the Visulizations section for the best model by selecting **Visualizations** or **Matrix**.



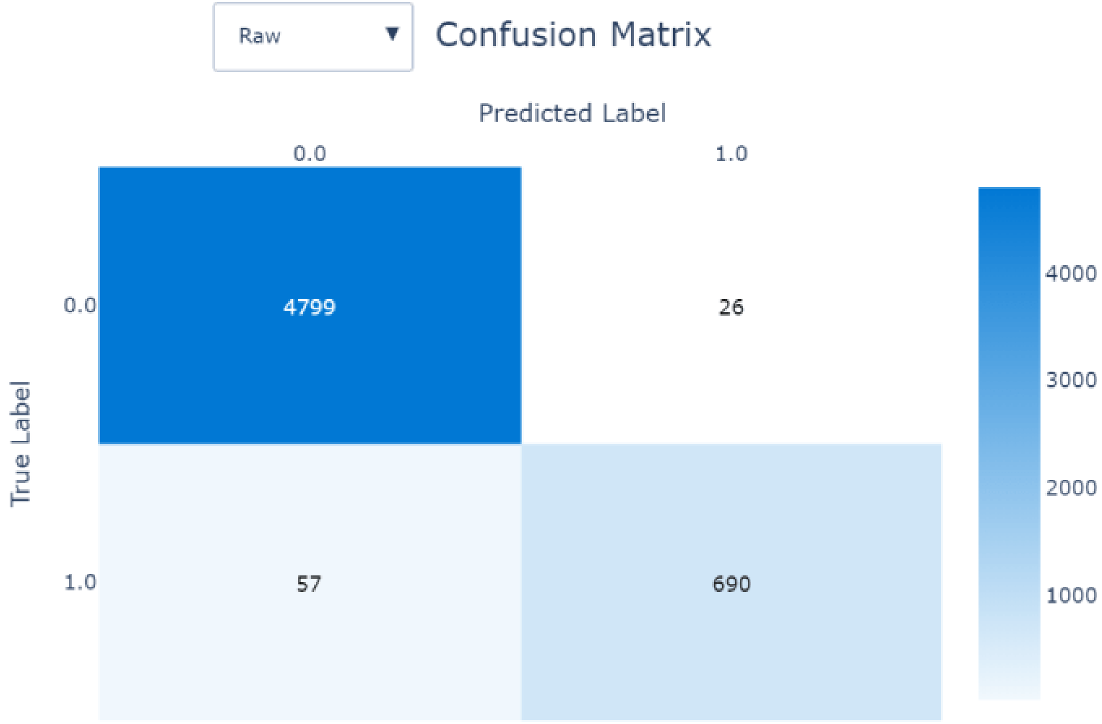
Run 48 Completed

Refresh Explain model Cancel

Model details | Visualizations | Explanations (preview) | Logs | Outputs

Automated ML provides charts for better understanding of model performance. Learn more on ho

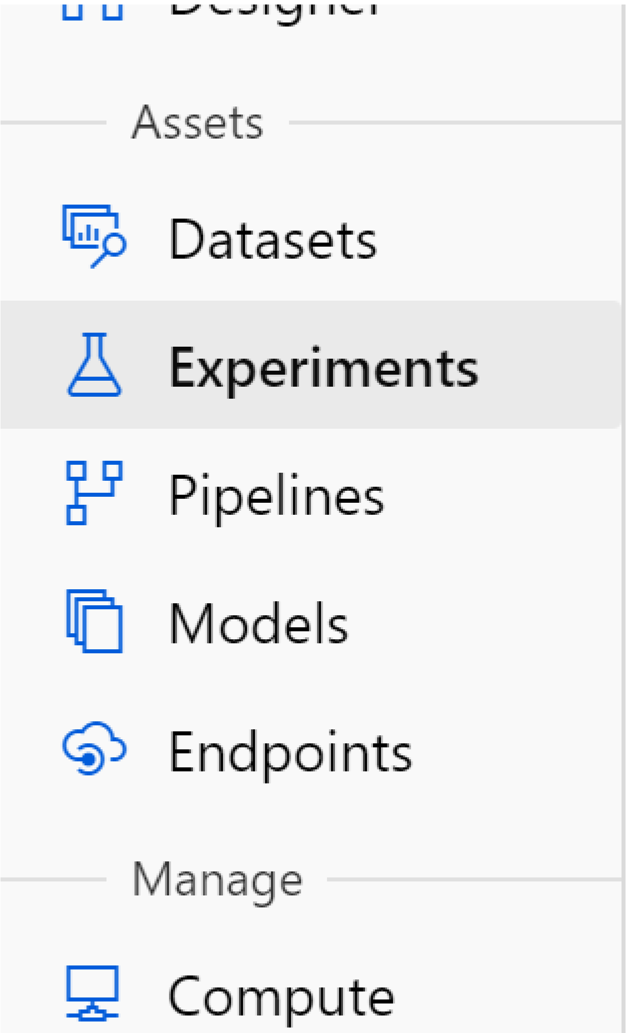
The **Confusion Matrix** for the **Best model** is shown below; It is possible that your matrix looks different.

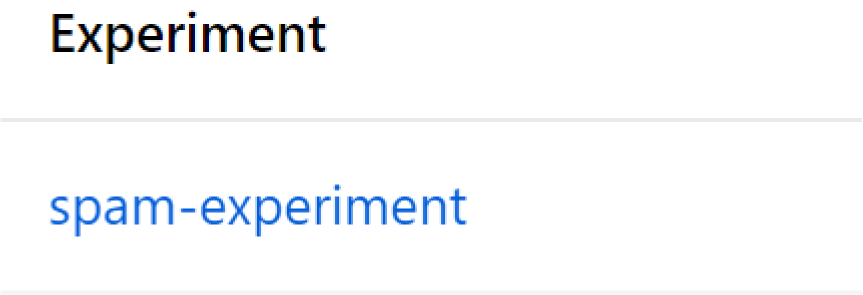


3. Deploying the ‘Best’ Automated ML Spam Model

3.1 Deploy the Best Model

To make downloading and deploying the model easier, automated machine learning will keep track of the best model from all the training runs. With the new Azure Machine Learning studio UX that we introduced in part 2.2 of this project, this is conveniently accessible. Go to your experiment by clicking Experiments on the left navigation, then select your experiment. I labelled mine **"spam-experiment"**, as can be seen below, to deploy your best model.



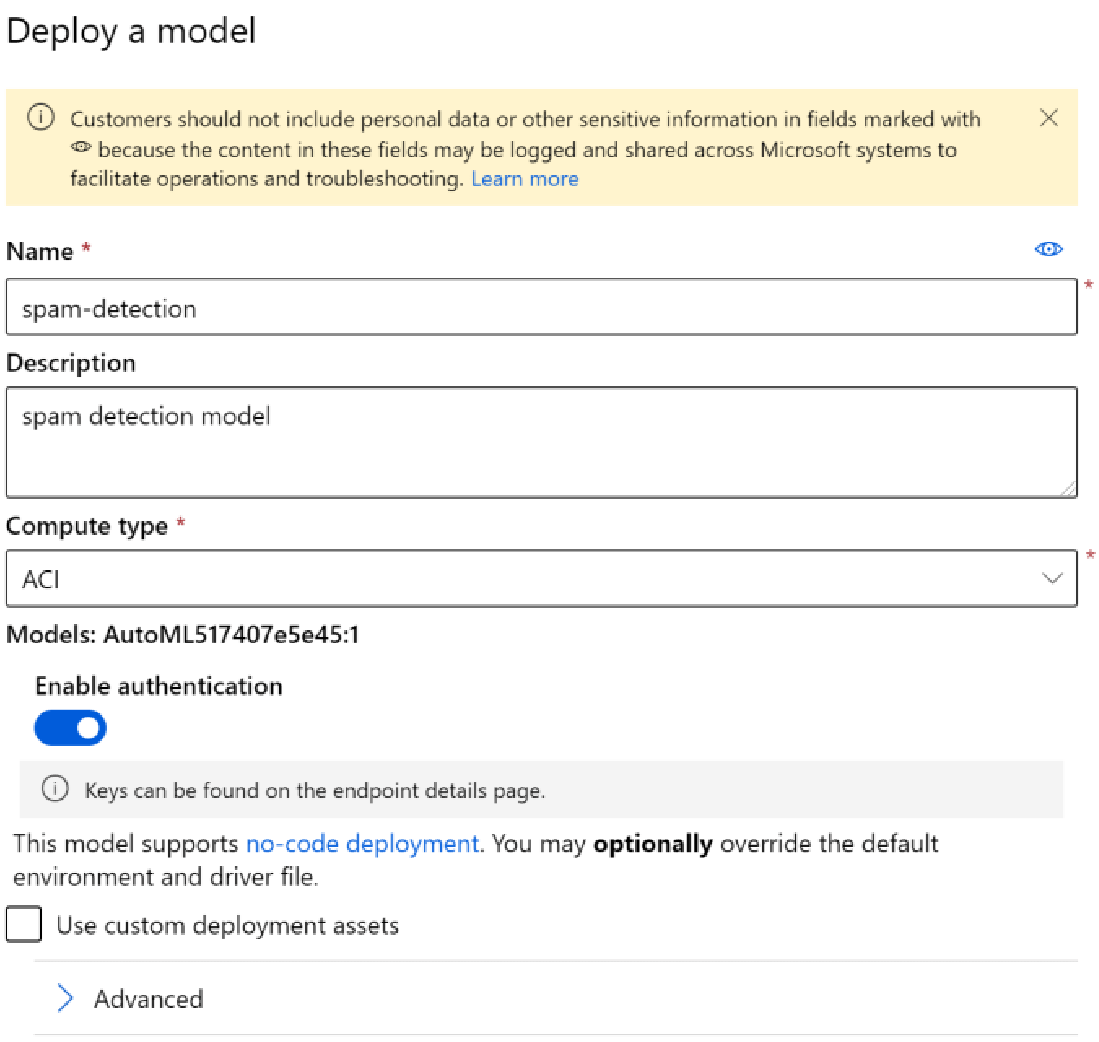


Experiment
spam-experiment

3.2 Deployment Configuration

As this is the quickest and easiest way to operate a container in Azure without having to manage any VMs, we will deploy to Azure Container Instances (ACI) for this project.

Anyone can call your webservice if you leave Enable authentication off. It is important to **Enable Authentication** now.



Deploy a model

ⓘ Customers should not include personal data or other sensitive information in fields marked with 👁 because the content in these fields may be logged and shared across Microsoft systems to facilitate operations and troubleshooting. Learn more

Name * : spam-detection

Description: spam detection model

Compute type * : ACI

Models: AutoML517407e5e45:1

Enable authentication: ●

ⓘ Keys can be found on the endpoint details page.

This model supports no-code deployment. You may **optionally** override the default environment and driver file.

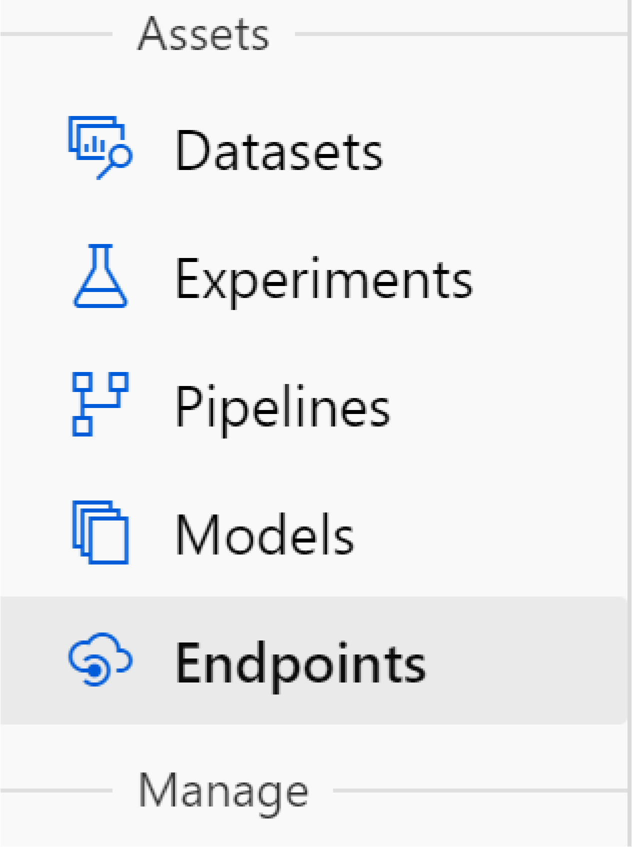
□ Use custom deployment assets

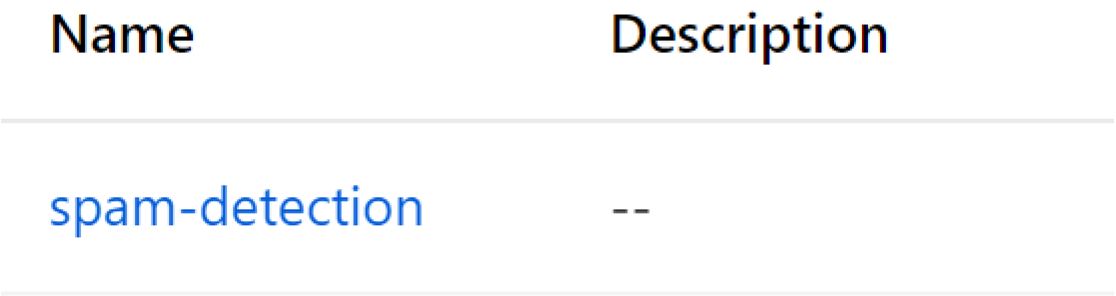
It will take few minutes to deploy ACI with the docker image and authentication/routing sidecar containers after registering the model and building the docker image. When our endpoint is in a **healthy deployment condition**, as described in the following, the deployment will be finished and ready.

3.3 Consuming the Web Service

3.3.1 Web Service Endpoint

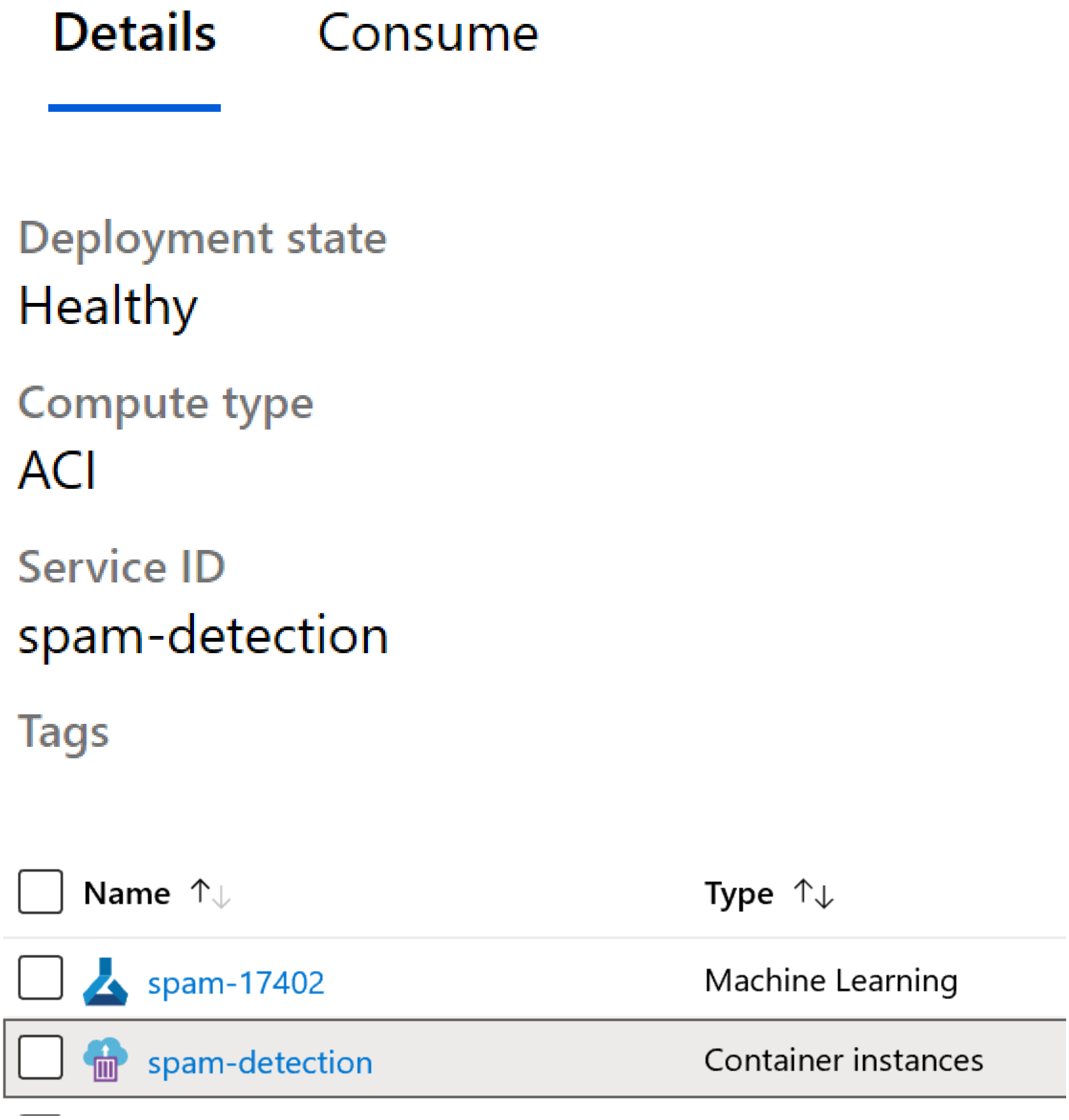
Navigate to left navigation menu, to view your deployed enpoint go to **Endpoints**. Select your Endpoint, in my case **spam-detection**.





Name	Description
spam-detection	--

Once the **Deployment state** is **Healthy**, the endpoint is prepared for inferencing. As mentioned in this project's section 3.2, this could take a while. The Azure Container Instance resource will also be visible in the newly established Resource Group. (same resource group as Azure Machine learning).



Details Consume

Deployment state
Healthy

Compute type
ACI

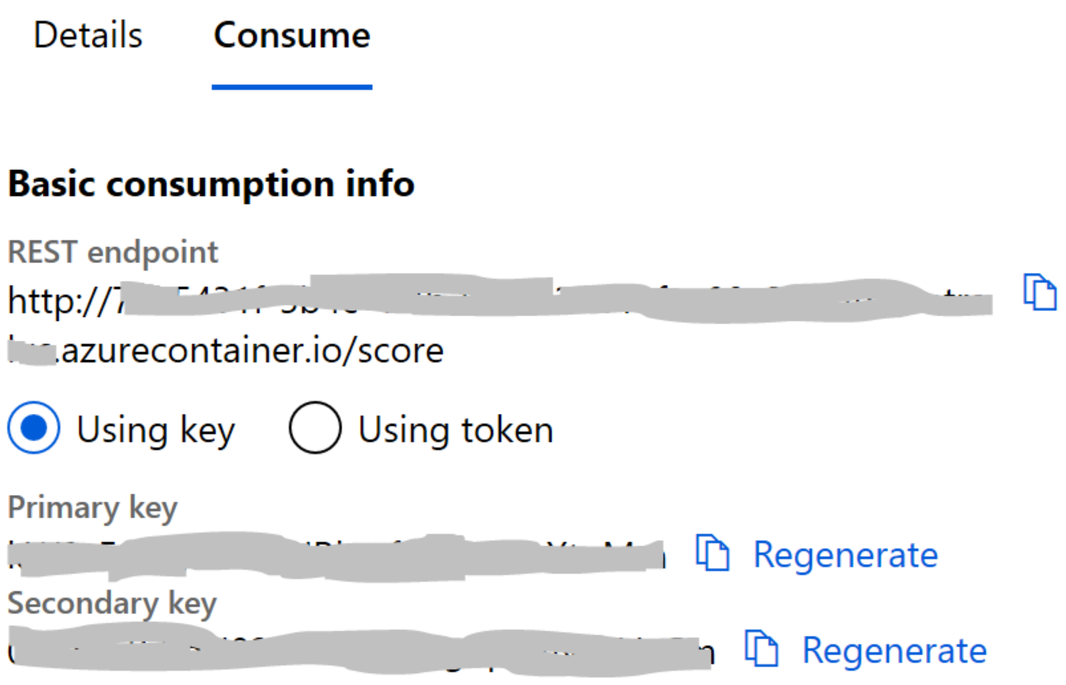
Service ID
spam-detection

Tags

□ Name ↑↓	Type ↑↓
□ spam-17402	Machine Learning
□ spam-detection	Container instances

3.3.2 Consumer Web Service in Python

Here is a straightforward Python example that uses the installed web service to score messages. The Consume area of your spam detection endpoint is where you can find the your **REST endpoint URL** and **Primary** or **Secondary key**.



Basic consumption info

REST endpoint

http://[redacted]
[redacted].azurecontainer.io/score

● Using key ○ Using token

Primary key

[redacted] Regenerate

Secondary key

[redacted] Regenerate

Python Code

Paste your endpoint url and key into the following code to score the example messages.

```python
import pandas as pd
import json
import requests

url = 'http://<your url>.azurecontainer.io/score'
key = '<your key>'

spam = {'spam':[1,0,1]}
data = {"data":['PRIVATE! Your 2003 Account Statement for 078','Send me the new number at my work','Free e-book']}

input_data = json.dumps(data)

headers = {'Content-Type':'application/json'}

#for AKS deployment you'd need to the service key in the header as well
headers = {'Content-Type':'application/json',  'Authorization':('Bearer '+ key)}

resp = requests.post(uri, input_data, headers=headers)

print("POST to url", uri)
print("input data:", input_data)
print("label:", spam['spam'])
print( resp.text)
```

Output from PyCharm

```
input data: {"data": ["PRIVATE! Your 2003 Account Statement for 078", "Send me the new number at my work", "Free e-book"]}
label: [1, 0, 1]
 "{\"result\": [1.0, 0.0, 1.0]}"
```

Importance of Spam Detection

Spam Detection refers to the process of identifying and filtering out unwanted or unsolicited messages, particularly in email communication. Spam messages can include a variety of content, such as advertisements, phishing attempts, malware, and fraudulent schemes.

- Protection against malicious attacks: Spam emails can contain malware, viruses, and phishing attempts that can harm individuals or organizations. Spam detection helps to identify and filter out these malicious emails, reducing the risk of a successful attack.

- Saving time and resources: Spam emails can flood inboxes, wasting time and resources by forcing individuals to sift through irrelevant messages. Spam detection can automatically filter out these emails, allowing individuals to focus on important tasks.

- Avoiding scams: Spam emails often contain scams that can lead to financial loss or identity theft. By detecting and filtering out these emails, individuals and organizations can