**Regression 1: Build a Machine Learning Model**

**In this article**

You will set up automated machine learning training jobs using Azure Machine Learning Automated ML in [Azure Machine Learning studio](https://learn.microsoft.com/en-us/azure/machine-learning/overview-what-is-azure-machine-learning?view=azureml-api-2#studio). Automated ML is a process where Azure Machine Learning selects the best machine learning algorithm for your data.

**Prerequisites**

* An Azure subscription.
* An Azure Machine Learning workspace or compute instance.
* The data asset to use for the Automated ML training job. This tutorial describes how to select an existing data asset or create a data asset from a data source, such as a local file, web url, or datastore.

**Important**

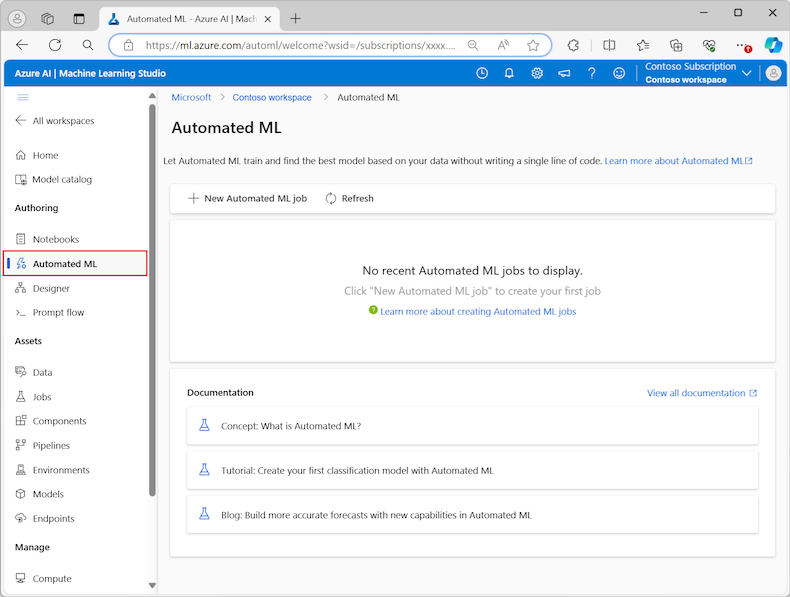
There are two requirements for the training data:

* + The data must be in tabular form (columns and rows).
  + The value to predict (the *target* column) must be present in the data.

**Create experiment**

Create and run an experiment by following these steps:

1. Sign in to Azure Machine Learning studio at ml.azure.com and select your workspace.
2. On the left menu, select **Automated ML**:

[](https://learn.microsoft.com/en-us/azure/machine-learning/media/how-to-use-automated-ml-for-ml-models/automated-ml-overview-large.png?view=azureml-api-2#lightbox)

1. Select **New automated ML job**

By default, the process selects the **Train automatically** option on the **Training method** tab and continues to the configuration settings.

1. On the **Basics settings** tab, enter the **Job** name and **Experiment** name. Enter like this:   
   **FirstName\_LastName\_Regression1\_Job**. For example, **John\_Smith\_ Regression1\_Job**.  
   **FirstName\_LastName\_ Regression1\_Experiment**. For example, **John\_Smith\_ Regression1\_Experiment**.
2. **Description**: Figure out the factors affecting customer rating.
3. Select **Next** to continue.

**Identify data asset**

On the **Task type & data** tab, you specify the data asset for the experiment and the machine learning model to use to train the data.

**Create a Dataset**

To create a new data asset, follow these steps:

1. To create a new data asset from a file on your local computer, select **Create**.
2. On the **Data type** page:
   1. Enter the **Data asset** name as “**Dataset\_Regression\_supermarket\_sales\_binary**”.
   2. For the **Type**, select **Tabular** from the dropdown list.
   3. Select **Next**.
3. On the **Data source** page, select **From local files**. The dataset “Dataset\_Regression\_Supermarket\_Sales\_binary” is on the course web site in Canvas in the respective module. Download the dataset to the location of your choice.

Machine Learning studio adds extra options to the left menu for you to configure the data source.

1. Select **Next** to continue to the **Destination storage type** page, where you specify the Azure Storage location to upload your data asset.  
   1. For the **Datastore type**, select **Azure Blob Storage**.
   2. In the list of datastores, select *workspaceblobstore*.
   3. Select **Next**.
2. On the **File and folder selection** page, use the **Upload files or folder** dropdown menu and select the **Upload files** or **Upload folder** option.  
   1. Browse to the location where you saved the data file and select **Open**.
   2. After the files upload, select **Next**.

Machine Learning studio validates and uploads your data.

1. Check your uploaded data on the **Settings** page for accuracy. The fields on the page are prepopulated based on the file type of your data.
2. Check the file settings below to verify everything is OK. You should not have to make any changes.

| **Field** | **Description** |
| --- | --- |
| **File format** | Defines the layout and type of data stored in a file. |
| **Delimiter** | Identifies one or more characters for specifying the boundary between separate, independent regions in plain text or other data streams. |
| **Encoding** | Identifies what bit to character schema table to use to read your dataset. |
| **Column headers** | Indicates how the headers of the dataset, if any, are treated. |
| **Skip rows** | Indicates how many, if any, rows are skipped in the dataset. |

1. Select **Next** to continue to the **Schema** page. This page is also prepopulated based on your **Settings** selections. You can configure the data type for each column, review the column names, and manage columns:  
   1. To change the data type for a column, use the **Type** dropdown menu to select an option. No need to take any action.
   2. To exclude a column from the data asset, toggle the **Include** option for the column. No need to take any action.
   3. **Click on the InvoiceID radio button to exclude it from the model**.
   4. **Click on the RatingDecimal radio button to exclude it from the model**.
   5. **Exclude the GrossMarginPercentage and the CostofGoodsSold**.
2. Click **Next** to continue to the **Review** page. Review the summary of your configuration settings for the job, and then select **Create**.

**Configure training model**

When the data asset is ready, Machine Learning studio returns to the **Task type & data** tab for the **Submit an Automated ML job** process. The new data asset is listed on the page.

Follow these steps to complete the job configuration:

1. Expand the **Select task type** dropdown menu and choose the training model to use for the experiment. The options include classification, regression, time series forecasting, natural language processing (NLP), or computer vision. **Select “Regression” to predict continuous values**.
2. Click on the radio button to the left of the left of the Dataset\_Regression\_supermarket\_sales\_binary.  
     
   A screenshot of a computer

   Description automatically generated
3. Select **Next** to continue to the **Task settings** tab.
4. In the **Target column** dropdown list, select the column to use for the model predictions, in this case **select RatingBinary**.
5. Click on **View additional configuration settings** to see the choices but leave the default values.
6. **Expand the Limits drop down and enter 30 for the Experiment timeout (minutes)**
7. **For Validate and Test enter the settings below**:
8. For **Validation Type** select **Automatic**.
9. For **Test Data** select **Train-Test Split.**
10. For **Percentage test of data** enter **10.**

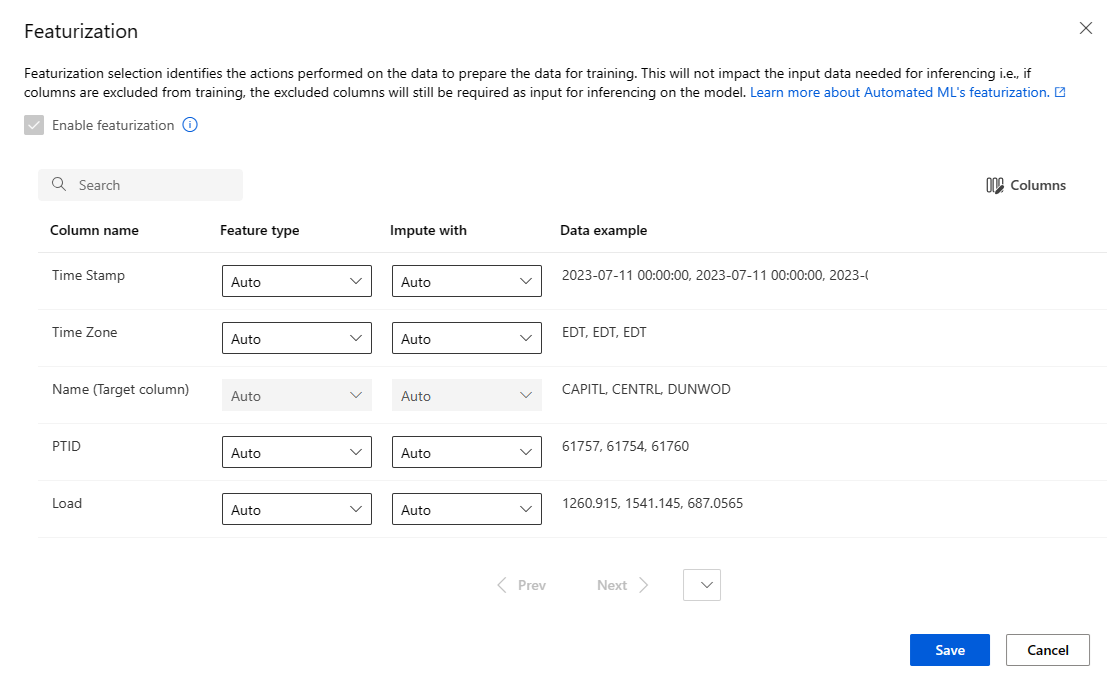
**Configure featurization settings – Optional – Read through will not adjust**

Click on **View featurization settings** option to see actions to perform on the data in preparation for training. **You do not need to make any changes for this lab**!

The **Featurization** page shows default featurization techniques for your data columns. You can enable/disable automatic featurization and customize the automatic featurization settings for your experiment.

1. Select the **Enable featurization** option to allow configuration. When your data contains non-numeric columns, featurization is always enabled.
2. Configure each available column, as desired. The following table summarizes the customizations currently available via the studio.

| **Column** | **Customization** |
| --- | --- |
| **Feature type** | Change the value type for the selected column. |
| **Impute with** | Select what value to impute missing values with in your data. |

[](https://learn.microsoft.com/en-us/azure/machine-learning/media/how-to-use-automated-ml-for-ml-models/updated-featurization.png?view=azureml-api-2#lightbox)

**Configure limits for the job – Optional – Read through will not adjust**

The **Limits** section provides configuration options for the following settings:

|  |  |  |
| --- | --- | --- |
| **Setting** | **Description** | **Value** |
| **Max trials** | Specify the maximum number of trials to try during the Automated ML job, where each trial has a different combination of algorithm and hyperparameters. | Integer between 1 and 1,000 |
| **Max concurrent trials** | Specify the maximum number of trial jobs that can be executed in parallel. | Integer between 1 and 1,000 |
| **Max nodes** | Specify the maximum number of nodes this job can use from the selected compute target. | 1 or more, depending on the compute configuration |
| **Metric score threshold** | Enter the iteration metric threshold value. When the iteration reaches the threshold, the training job terminates. Keep in mind that meaningful models have a correlation greater than zero. Otherwise, the result is the same as guessing. | Average metric threshold, between bounds [0, 10] |
| **Experiment timeout (minutes)** | Specify the maximum time the entire experiment can run. After the experiment reaches the limit, the system cancels the Automated ML job, including all its trials (children jobs). | Number of minutes |
| **Iteration timeout (minutes)** | Specify the maximum time each trial job can run. After the trial job reaches this limit, the system cancels the trial. | Number of minutes |
| **Enable early termination** | Use this option to end the job when the score isn't improving in the short term. | Select the option to enable early end of job |

**Configure the compute**

1. Select **Next** to continue to the **Compute** tab.
2. For **compute type** you have multiple options such as [compute cluster](https://learn.microsoft.com/en-us/azure/machine-learning/concept-compute-target?view=azureml-api-2#azure-machine-learning-compute-managed), [compute instance](https://learn.microsoft.com/en-us/azure/machine-learning/concept-compute-instance?view=azureml-api-2) or [serverless](https://learn.microsoft.com/en-us/azure/machine-learning/how-to-use-serverless-compute?view=azureml-api-2). Select **serveless** for this lab which is the easiest way to apply the computations and the Studio handles the tasks automatically.
3. Click **Next** to continue. Creation of a new compute can take a few minutes.
4. Click **Submit Training Job**.
5. You might see an error that the page does not exist or it has been deleted. It is OK.

**View results**

Select **Finish** to run your experiment. The experiment preparation process can take up to 10 minutes. Training jobs can take an additional 2-3 minutes more for each pipeline to finish running.

**Note**

The algorithms Automated ML employs have inherent randomness that can cause slight variation in a recommended model's final metrics score, like accuracy. Automated ML also performs operations on data such as train-test split, train-validation split or cross-validation, as necessary. If you run an experiment with the same configuration settings and primary metric multiple times, you likely see variation in each experiment's final metrics score due to these factors.

**View experiment details**

The **Job Detail** screen opens to the **Details** tab. This screen shows you a summary of the experiment job including a status bar at the top next to the job number.

The **Models** tab contains a list of the models created ordered by the metric score. By default, the model that scores the highest based on the chosen metric is at the top of the list. As the training job tries more models, the exercised models are added to the list. Use this approach to get a quick comparison of the metrics for the models produced so far.

**View training job details**

Drill down on any of the completed models for the training job details. You can see performance metric charts for specific models on the **Metrics** tab. On this page, you can also find details on all the properties of the model along with associated code, child jobs, and images.

.