# Predictive Modeling with Regression Algorithms

# Data: Housing Data

## Get the Data

* Go to the URL to download the dataset: <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>

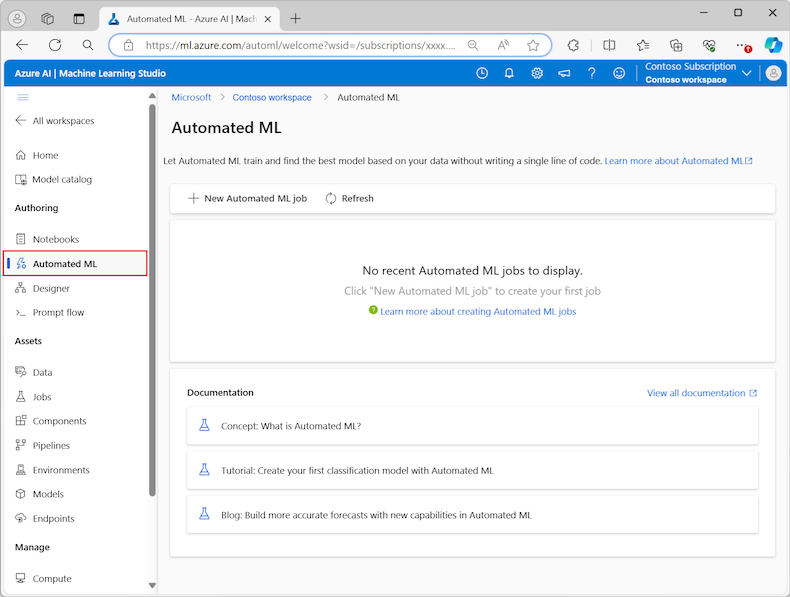
or download the dataset from the class web site on Canvas:

* At the upper left corner of the window, click the black download button.
* Open the zip folder and copy the file **Dataset\_Regression\_kc\_house\_data.csv** to a location of your preference.

## Create the Experiment

Create the experiment:

1. Go to ml.azure.com to open your working environment.
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**
2. In the left pane for the **Training Method**, Azure selects the **Train automatically** option on the **Training method** tab and continues to **Basic settings**.
3. On the **Basics settings** tab, enter the **Job** name and **Experiment** name. Enter like this:   
   **FirstName\_LastName\_LAB2\_Job**. For example, John\_Smith\_LAB2\_Job.  
   **FirstName\_LastName\_LAB2\_Experiment**. For example, John\_Smith\_LAB2\_Experiment.
4. **Description**: Enter "Figure out the factors affecting House Prices."
5. Select **Next** to continue.

**Task Type and Data**

1. In the **Select Task Type** drop down menu, select **Regression**.
2. For **Select Data** click **Create** to create a new data asset from the file on your local computer..
3. On the page **Set the name and type for your data asset**:  
   1. Enter the **Data asset** name as “kc\_house\_data\_”.
   2. For the **Type**, select **Tabular** from the dropdown list.
   3. Select **Next**.
4. On the page **Choose a source for your data asset**
   1. click **From local files**
   2. click **Next** to continue
5. On the page **Select a datastore**, we specify the Azure Storage location to upload our data.  
   1. For the **Datastore type**, select **Azure Blob Storage**.
   2. In the list of datastores, select *workspaceblobstore*.
   3. Select **Next**.
6. On the page **Choose a file or folder**
   1. click on the **Upload files or folder** dropdown menu and select the **Upload files** option.
   2. Browse to the location where you saved the “kc\_house\_data” file and select **Open**.
   3. After the files upload, select **Next**.

1. On the **Settings** page browse your data to check the values and see if anything is unusual or out of order.
2. Check the file settings below to have an idea of how your data is formatted and 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 id radio button to exclude it from the model**.
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 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. In the **Select task type** dropdown menu the choice **Regression** should already be there. If not, select it. The options include classification, regression, time series forecasting, natural language processing (NLP), or computer vision.
2. Click on the radio button to the left of the left of the file **kc\_house\_data**.
3. Select **Next** to continue to the **Task settings**.
4. In the **Target column** dropdown list, select the column to use for the model predictions, in this case **select price**.
5. Click on **View additional configuration settings** to see the choices but leave the default values.
6. Click on **View Featurization settings – Read through will not adjust**

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. When your data contains non-numeric columns, featurization is always enabled.  
You could configure each available column, as desired. The following table summarizes the customizations currently available via the studio.

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

1. **Expand the Limits drop down and enter 30 for the Experiment timeout (minutes)**
2. **For Validate and Test enter the settings below**:  
   1. For **Validation Type** select **Automatic**.
   2. For **Test Data** select **Train-Test Split.**
   3. For **Percentage test of data** enter **10.**
   4. **Click Next**.

**Limits for the job – Read through, do 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** 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, your experiment is running.