

## HI6039 Predictive Analytics

# Tutorial 5: Regression Analysis

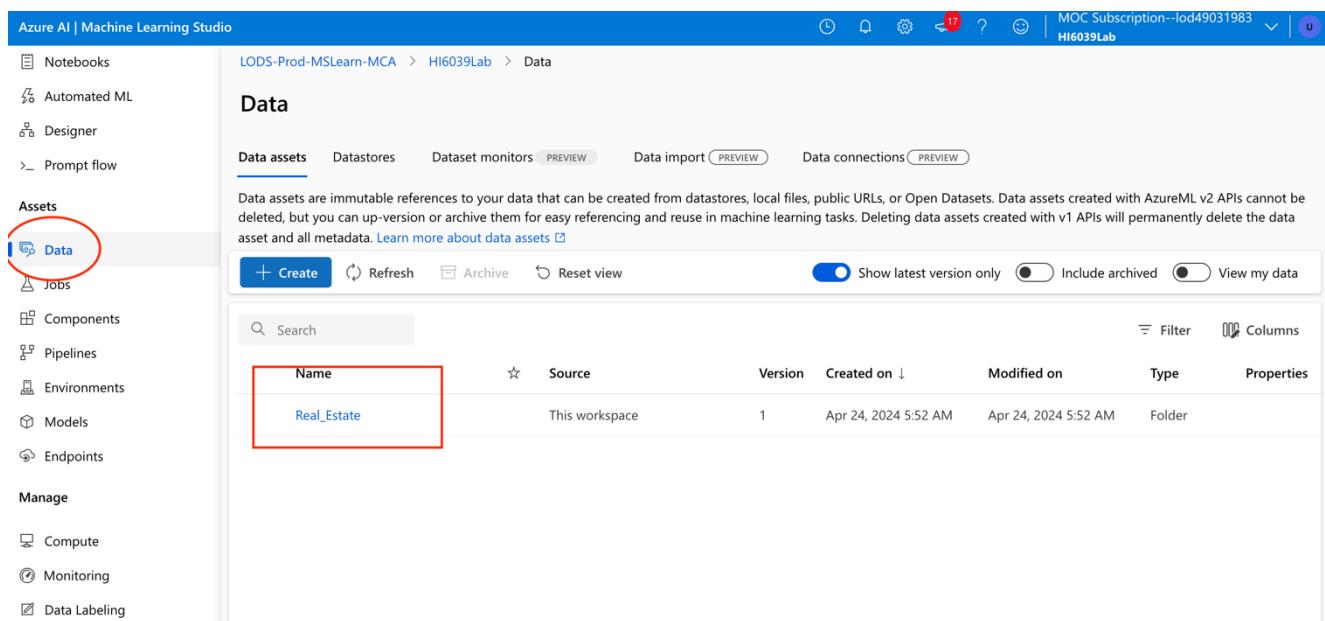
Before proceeding with the tasks ahead, make sure you have created **resource**, **workspace**, and **compute** in MS Azure Machine Learning. If you need guidance on this, please refer to Tutorial 4.

## Create dataset

You need to download ‘Real estate price prediction’ dataset (Real estate.csv) from the following URL

<https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction>

and create a dataset ‘Real\_Estate’



The screenshot shows the Azure AI | Machine Learning Studio interface. The left sidebar has sections for Notebooks, Automated ML, Designer, Prompt flow, Assets (with 'Data' highlighted and circled in red), Jobs, Components, Pipelines, Environments, Models, Endpoints, and Manage. The main area is titled 'Data' and shows 'Data assets'. It includes tabs for Data assets, Datastores, Dataset monitors, Data import (PREVIEW), and Data connections (PREVIEW). Below these are buttons for Create, Refresh, Archive, and Reset view, and filters for Show latest version only, Include archived, and View my data. A search bar and filter/columns buttons are also present. A table lists datasets, with one row for 'Real\_Estate' highlighted by a red box. The table columns are Name, Source, Version, Created on, Modified on, Type, and Properties.

Name	Source	Version	Created on	Modified on	Type	Properties
Real_Estate	This workspace	1	Apr 24, 2024 5:52 AM	Apr 24, 2024 5:52 AM	Folder	

# Regression Analysis

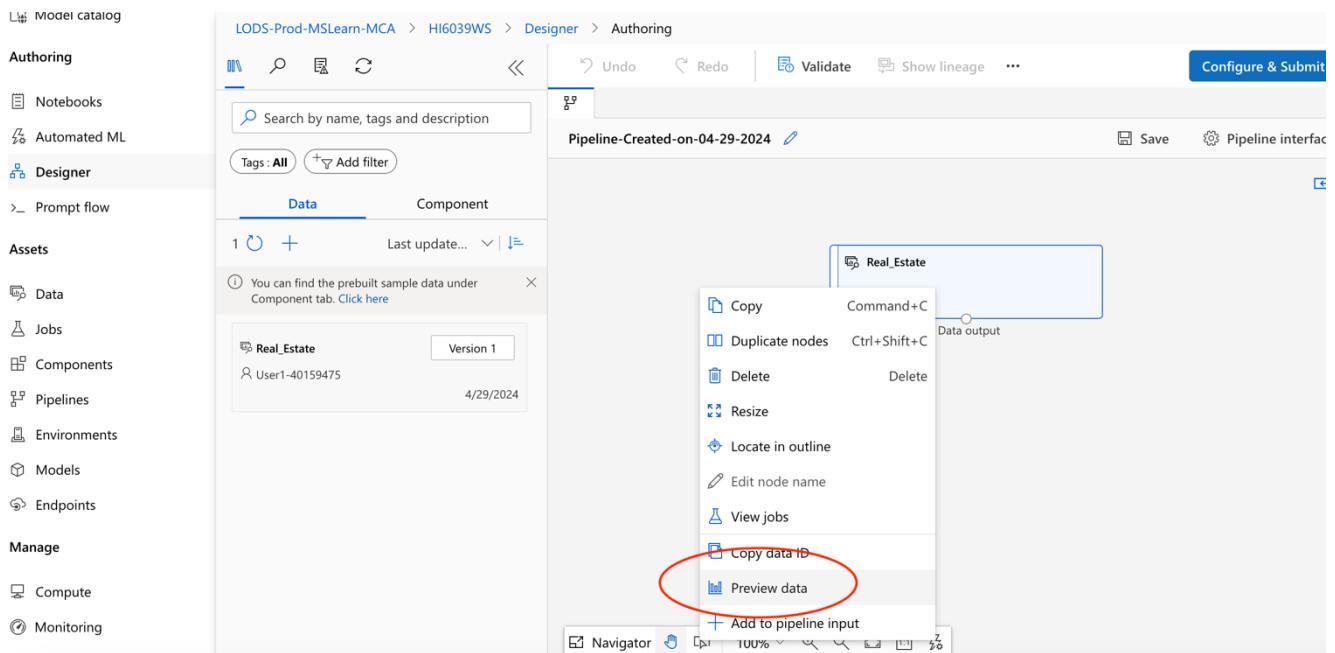
1. Select ‘Designer’ and click ‘Create a new pipeline using classic prebuilt components’.

The screenshot shows the Azure AI | Machine Learning Studio interface. On the left, the navigation bar includes 'All workspaces', 'Home', 'Model catalog', 'Authoring' (which is expanded to show 'Notebooks', 'Automated ML', 'Designer' (circled in red), and 'Prompt flow'), 'Assets' (with 'Data', 'Jobs', 'Components', 'Pipelines', 'Environments', 'Models', and 'Endpoints'), and a search bar. The main area is titled 'Designer' and 'New pipeline'. It features a 'Classic prebuilt' tab (selected) and a 'Custom' tab. A descriptive text states: 'This low-code option uses existing prebuilt components and earlier dataset types (tabular, file), and is best suited for data processing and traditional machine learning tasks like regression and classification. This option continues to be supported but will not have any new components added.' Below this are four sample pipeline icons: 'Create a new pipeline using classic prebuilt components' (circled in red), 'Image Classification using DenseNet', 'Binary Classification using Vowpal Wabbit Model - ...', and 'Wide & Deep based Recommendation - Rest...'. The 'Pipelines' section shows a table with columns: Name, Pipeline type, Updated on, and Created by. Buttons for Refresh, Delete, and Reset view are at the top of the table, along with a search bar and filter/columns options.

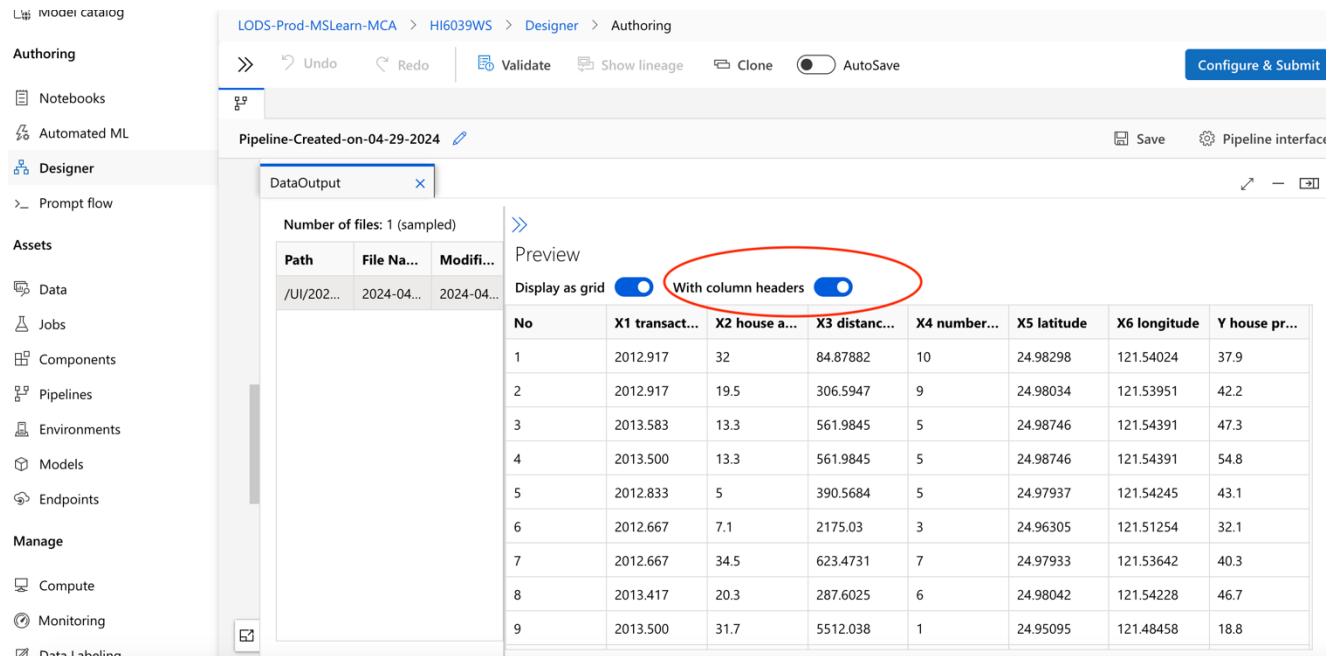
2. Drag the dataset ‘Real\_Estate’ and drop it to the pipeline.

The screenshot shows the Azure AI | Machine Learning Studio interface under the 'Authoring' section. The left sidebar is identical to the previous screenshot. The main area is titled 'Authoring' and shows a 'Pipeline-Created-on-04-29-2024' pipeline. The 'Data' tab is selected (circled in red). A dataset named 'Real\_Estate' (version 1, last updated 4/29/2024) is highlighted with a red circle. The pipeline interface on the right shows a single component labeled 'Real\_Estate' with a 'Data output' connection. Top navigation includes Undo, Redo, Validate, Show lineage, Clone, and a blue 'Configure & Submit' button.

3. Right-click ‘Real\_Estate’ dataset and select ‘Preview data’.



4. Turn on ‘With column headers’. Now, you can see the content of ‘Real\_Estate’ dataset that contains 8 attributes.



## 5. Close the 'DataOutput'

The screenshot shows the Azure Data Studio Designer interface. On the left, the sidebar lists 'Authoring' sections: Notebooks, Automated ML, Designer (which is selected), Prompt flow, Assets (Data, Jobs, Components, Pipelines, Environments, Models, Endpoints), Manage (Compute, Monitoring, Data Labeling), and Model catalog. In the center, a pipeline named 'Pipeline-Created-on-04-29-2024' is displayed. A 'DataOutput' component is selected, showing a preview of a CSV file with 28 rows and 8 columns. The first row of the preview table is:

No	X1 transact...	X2 house a...	X3 distanc...	X4 number...	X5 latitude	X6 longitude	Y house pr...
20	2012.667	1.5	23.38284	7	24.96772	121.54102	47.7
21	2013.417	4.5	2275.877	3	24.96314	121.51151	29.3
22	2013.417	10.5	279.1726	7	24.97528	121.54541	51.6
23	2012.917	14.7	1360.139	1	24.95204	121.54842	24.6
24	2013.083	10.1	279.1726	7	24.97528	121.54541	47.9
25	2013.000	39.6	480.6977	4	24.97353	121.53885	38.8
26	2013.083	29.3	1487.868	2	24.97542	121.51726	27
27	2012.667	3.1	383.8624	5	24.98085	121.54391	56.2
28	2013.250	10.4	276.449	5	24.95593	121.53913	33.6

## 6. Select the 'Component' option and then type 'Select columns' into the search box.

The screenshot shows the Azure Data Studio Model catalog interface. On the left, the sidebar lists 'Authoring' sections: Notebooks, Automated ML, Designer (selected), Prompt flow, Assets (Data, Jobs, Components, Pipelines, Environments, Models, Endpoints), Manage (Compute, Monitoring), and Model catalog. In the center, a search bar at the top contains the text 'select columns'. Below it, a navigation bar has tabs for 'Data' and 'Component' (which is highlighted with a red circle). A list of components is shown, with two items circled in red: 'Select Column Transform' and 'Select Column in Dataset'. Both items are described as transformations that select a subset of columns from a dataset. To the right, a preview window shows a dataset named 'Real\_Estate' with 1 row and 1 column. The bottom of the screen shows standard UI elements like Navigator, zoom controls, and a status bar.

7. Drag ‘Select Columns in Dataset’ component and drop it to the pipeline. Then, connect ‘Real\_Estate’ dataset to the ‘Selected Columns in Dataset’ component.

The screenshot shows the Azure Machine Learning Designer interface. On the left, the 'Designer' section is selected in the navigation bar. The main area displays a search bar with 'select columns' and a list of components under the 'Component' tab. One component, 'Select Columns in Dataset', is highlighted with a red circle. The pipeline canvas on the right shows a flow starting from a 'Real\_Estate' dataset, which connects to a 'Select Columns in Dataset' component named 'select\_columns\_in\_dataset'. The output of this component is labeled 'Results dataset'. A red circle also highlights this component in the pipeline.

8. Double-click the ‘Selected Columns in Dataset’ component and select ‘Edit column’.

The screenshot shows the Azure Machine Learning Designer interface with the 'Select Columns in Dataset' component selected. The properties pane on the right is open, showing various settings like 'Select columns' (which has a red circle around it) and 'Output settings'. At the bottom of the properties pane, there is a blue button labeled 'Edit column', which is also circled in red.

9. Select ‘With rules’, ‘Columns indices’, enter ‘2-8’, and click ‘Save’.

The screenshot shows the 'Select columns' dialog in the Azure AI | Machine Learning Studio Designer interface. The 'With rules' radio button is selected. The 'Column indices' dropdown shows '2-8'. The 'Save' button is highlighted with a red circle.

10. Click ‘Close’.

The screenshot shows the 'Pipeline-Created-on-04-29-2024' pipeline details in the Azure AI | Machine Learning Studio Designer interface. The 'Component' tab is selected. The 'Select Columns in Dataset' component is highlighted. The 'Edit column' button is highlighted with a red circle.

## 11. Click ‘Configure & Submit’.

The screenshot shows the Azure AI | Machine Learning Studio Designer interface. On the left, there's a sidebar with various options like Notebooks, Automated ML, Designer (which is selected), Prompt flow, Assets, Data, Jobs, Components, Pipelines, Environments, Models, Endpoints, Manage, Compute, Monitoring, Data Labeling, and Linked Services. The main area displays a pipeline diagram. At the top right, there are several icons and the text "MOC Subscription--lod49054667" and "HI6039-WS". A prominent blue button labeled "Configure & Submit" is located at the top right of the main workspace, with a red circle highlighting it to indicate the next step.

## 12. Select ‘Create new’, enter a new experiment name, and click ‘Next’

The screenshot shows the "Set up pipeline job" dialog box. The left side has a vertical navigation bar with steps: Basics, Inputs & outputs, Runtime settings, and Review + Submit. The Basics section is currently active. It contains fields for "Experiment name" (radio buttons for "Select existing" and "Create new", with "Create new" selected and a red circle around it), "New experiment name \*" (text input field containing "HI6039-Ex1", with a red circle around it), "Job display name" (text input field containing "Pipeline-Created-on-04-29-2024"), "Job description" (text input field containing "Pipeline created on 20240430"), and "Job tags" (a Name:Value pair input field). At the bottom, there are "Review + Submit", "Back", "Next", and "Close" buttons, with the "Next" button highlighted with a red circle.

13. Click 'Next'.

Azure AI | Machine Learning Studio

LODS-Prod-MSLearn-MCA > HI6039-WS > Designer > Authoring

Set up pipeline job

Basics

Inputs & outputs

Inputs

No inputs

Outputs

No outputs

Review + Submit Back Next Close

14. Select compute type 'Compute instance', select the compute you have already created, and select 'Next'.

Azure AI | Machine Learning Studio

LODS-Prod-MSLearn-MCA > HI6039-WS > Designer > Authoring

Set up pipeline job

Basics

Inputs & outputs

Runtime settings

Select compute type

Compute instance

Select Azure ML compute instance

HI6039-Compute1 - Running

Create Azure ML compute instance Refresh Compute

Default datastore

Select datastore \*

workspaceblobstore

Advanced settings

Continue on step failure

Review + Submit Back Next Close

## 15. Click 'Submit'.

The screenshot shows the 'Set up pipeline job' configuration screen. The left sidebar lists various ML components like Notebooks, Automated ML, Designer, etc. The main area shows a pipeline component 'select columns' with its details. On the right, the 'Review + Submit' section is open, showing 'Basics' (Job display name: Pipeline-Created-on-04-29-2024), 'Inputs & outputs' (None), and 'Runtime settings'. The 'Review + Submit' step is highlighted with a green checkmark. At the bottom right of the review panel, there is a 'Submit' button, which is circled in red.

## 16. Now, checking 'Notifications', you can see the pipeline (a job) has been submitted and it is running.

The screenshot shows the 'Notifications' panel after the pipeline has been submitted. It displays three notifications:

- Success: Pipeline job has been submitted. View Details
- Pipeline job "Pipeline-Created-on-04-29-2024" in experiment "HI6039-Ex1" Running Job details
- Submit pipelineRun succeeded.

Each notification is highlighted with a red circle. The 'Dismiss all' button is also circled in red.

## 17. Waiting until the job has completed.

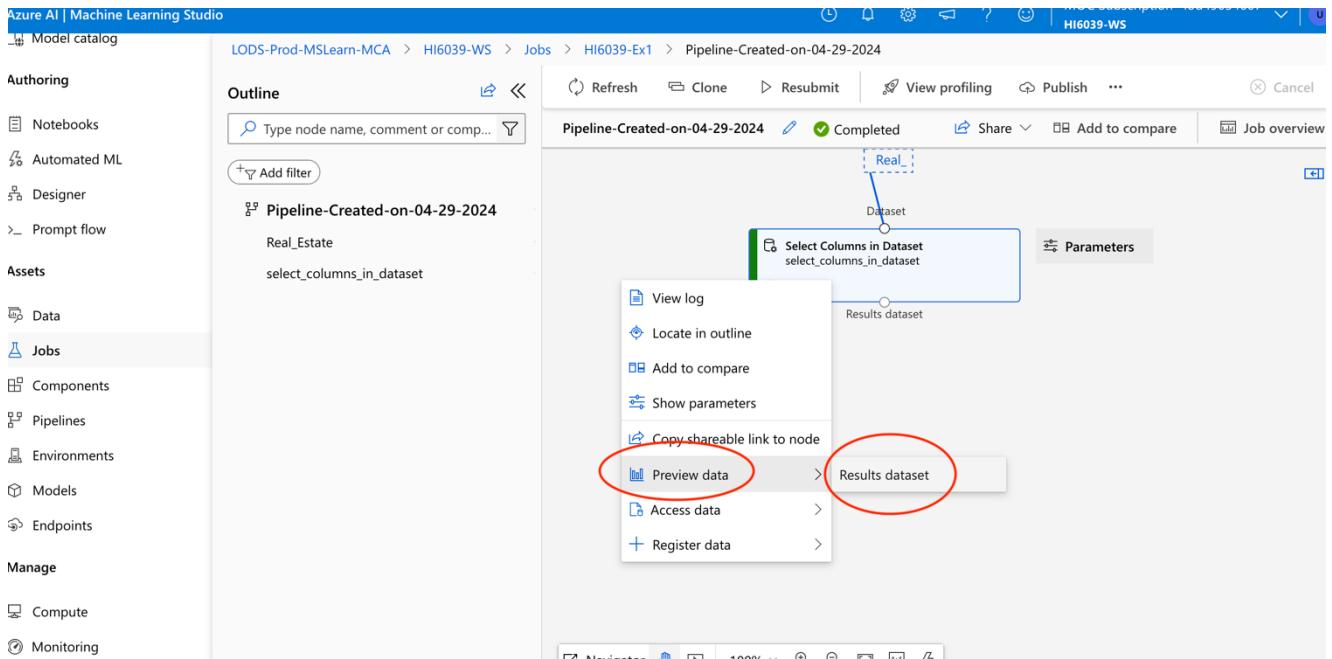
The screenshot shows the Azure AI | Machine Learning Studio interface. On the left, the navigation menu includes Notebooks, Automated ML, Designer (which is selected and highlighted with a red circle), Prompt flow, Assets, Data, Jobs, Components, Pipelines, Environments, Models, Endpoints, Manage, Compute, Monitoring, Data Labeling, and Linked Services (PREVIEW). The main workspace displays a pipeline titled "Pipeline-Created-on-04-29-2024". The pipeline consists of two components: "Real\_Estate" (Data) and "Select Columns in Dataset select\_columns\_in\_dataset" (Component). A "Data output" node connects the Real\_Estate dataset to the second component. The second component has a "Dataset" output node and a "Results dataset" output node. The pipeline status bar at the bottom indicates "100%". To the right, a "Notifications" panel shows three successful messages: "Pipeline job 'Pipeline-Created-on-04-29-2024' in experiment 'HI6039-Ex1' Completed" (Job details, April 30, 2024 4:17 PM), "Submit pipelineRun succeeded." (April 30, 2024 4:14 PM), and "Compute 'HI6039-Compute1' provisioning succeeded" (Compute details, April 30, 2024 4:02 PM). A "Dismiss all" button is also present in the notifications panel.

## 18. Select 'Job' and then select the pipeline (the latest job).

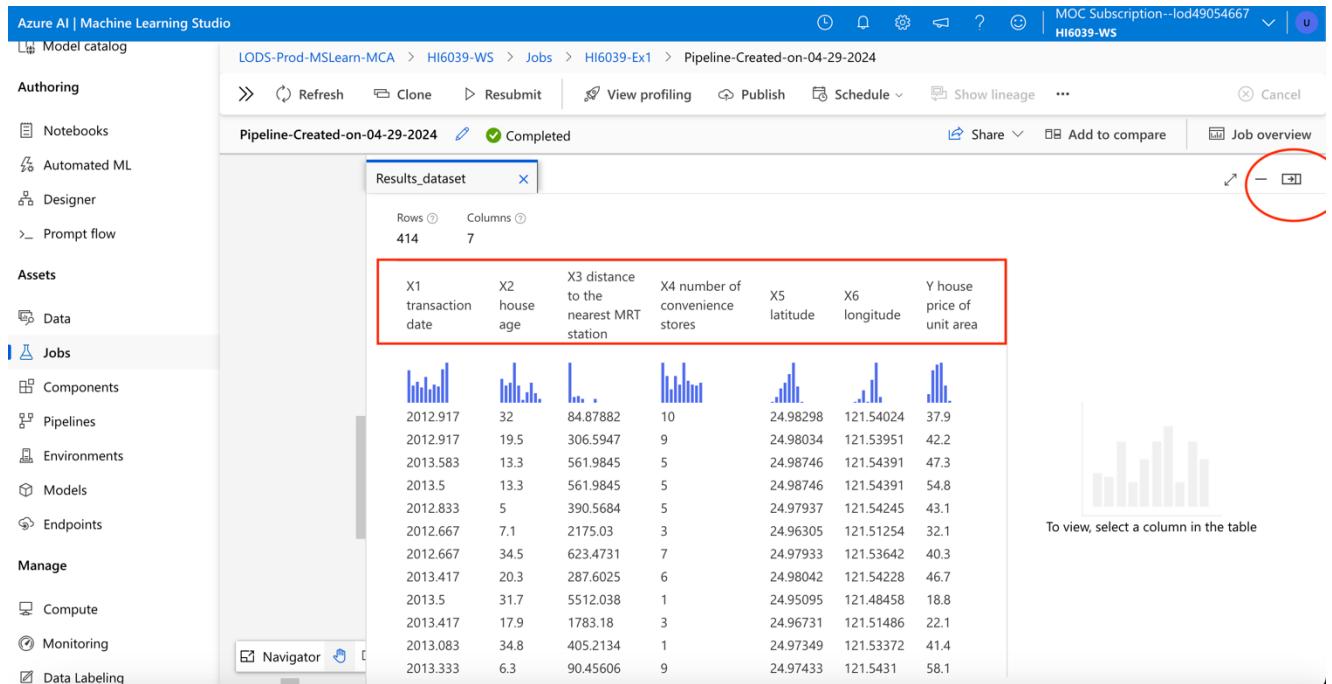
The screenshot shows the Azure AI | Machine Learning Studio interface. The navigation menu on the left is identical to the previous screenshot. The main workspace shows the "Jobs" list under the "All experiments" tab. The table has columns: Experiment, Latest job, Last submitted, and Created. One row is visible for the experiment "HI6039-Ex1", which has the latest job "Pipeline-Created-on-04-29-2024". This row is highlighted with a red circle. The "Last submitted" column shows "Apr 30, 2024 4:14 PM" and the "Created" column shows "Apr 30, 2024". The "Jobs" icon in the navigation menu is also circled in red.

Experiment	Latest job	Last submitted	Created
HI6039-Ex1	Pipeline-Created-on-04-29-2024	Apr 30, 2024 4:14 PM	Apr 30, 2024

19. Right-click ‘Select Columns in Dataset’ component, select ‘Review data’, and select ‘Results dataset’



20. Now, you can see 7 selected attributes. Click ‘close’ to continue.



## 21. Click 'Designer' and click the pipeline.

The screenshot shows the 'Designer' section of the Azure Machine Learning Studio. On the left, a sidebar lists various categories like 'Authoring', 'Assets', and 'Manage'. Under 'Authoring', 'Designer' is selected and highlighted with a red oval. The main area displays four prebuilt pipeline templates: 'Create a new pipeline using classic prebuilt components', 'Image Classification using DenseNet', 'Binary Classification using Vowpal Wabbit Model', and 'Wide & Deep based Recommendation - Rest...'. Below these, a table titled 'Pipelines' shows a single entry: 'Pipeline-Created-on-04-29-2024' (highlighted with a red oval), which is a 'Training' pipeline last updated on April 30, 2024, at 4:14 PM.

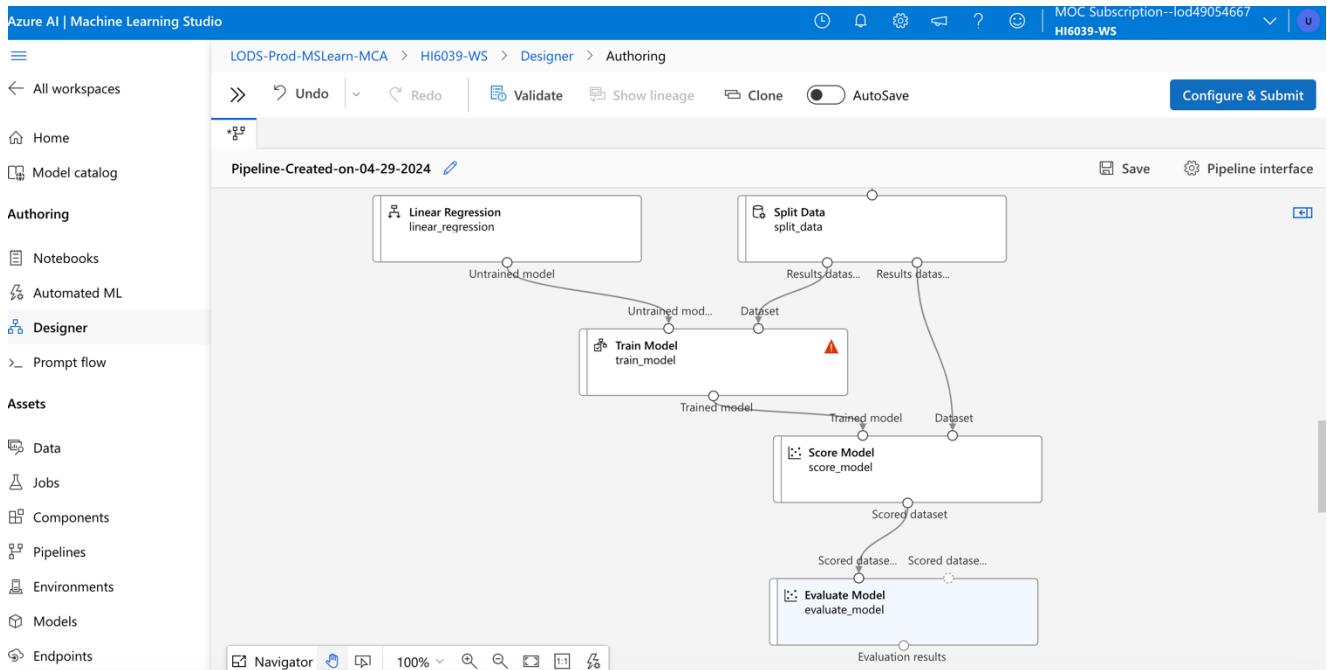
## 22. Search and add 'Split Data' component to the pipeline. Then, connect the 'Selected Columns in Dataset' to the 'Split Data' component.

The screenshot shows the 'Designer' interface in Azure Machine Learning Studio. The left sidebar shows 'Designer' selected. In the center, a search bar contains 'split data', and the results list shows the 'Split Data' component (highlighted with a red oval). The pipeline canvas on the right displays a flow starting from a 'Real\_Estate' dataset, which connects to a 'Select Columns in Dataset' component. The output of this component then connects to a 'Split Data' component (highlighted with a large red oval). The 'Parameters' pane on the right shows 'Select columns'.

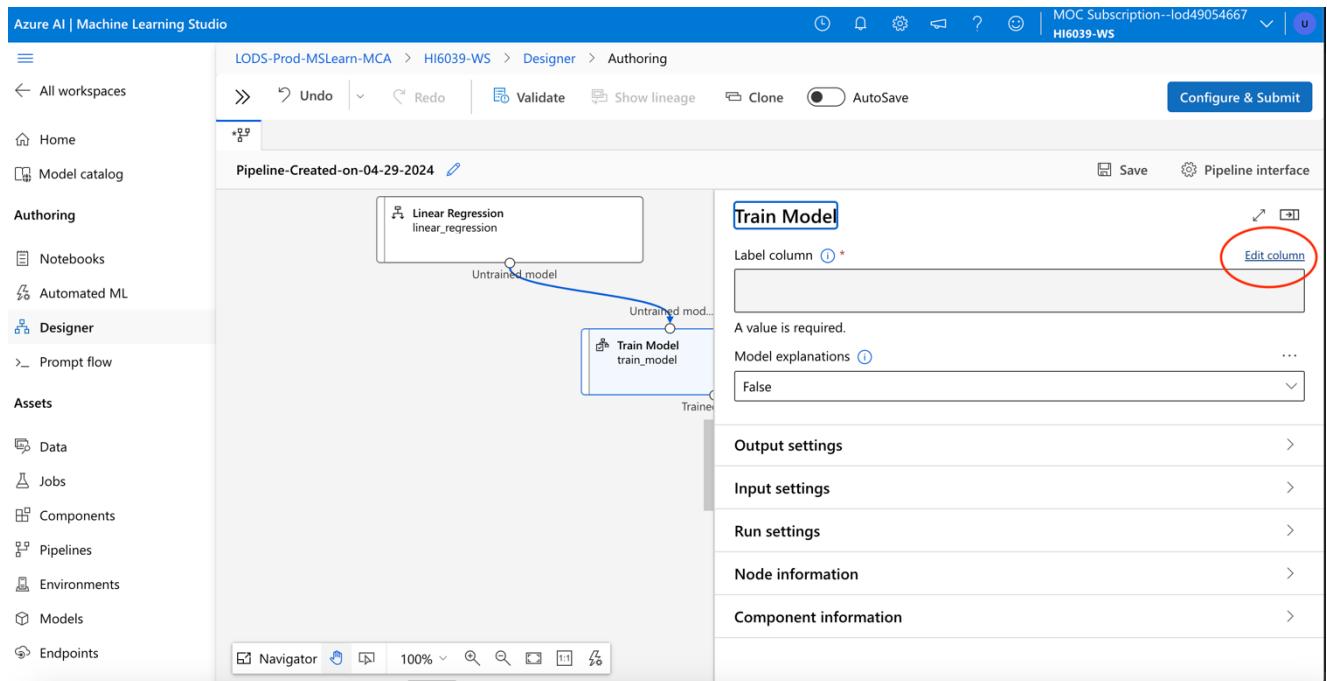
23. Double-click the ‘Split Data’ component and enter ‘0.7’ for the ‘Fraction of rows in the first output dataset’.

The screenshot shows the Azure Machine Learning Studio Designer interface. On the left, there's a sidebar with categories like Authoring, Assets, Manage, and Designer (which is selected). In the center, a search bar at the top has 'split data' typed into it. Below the search bar, there are tabs for 'Data' and 'Component'. Under the 'Component' tab, a list of components is shown, with 'Split Data' selected. The main area is titled 'Pipeline-Created-on-04-29-2024'. The 'Split Data' component configuration pane is open, showing settings for 'Split Rows' mode. The 'Fraction of rows in the first output dataset' field is highlighted with a red circle and contains the value '0.7'. Other settings include 'Randomized split' set to 'True' and 'Random seed' set to '0'. There are also sections for 'Output settings' and 'Input settings'.

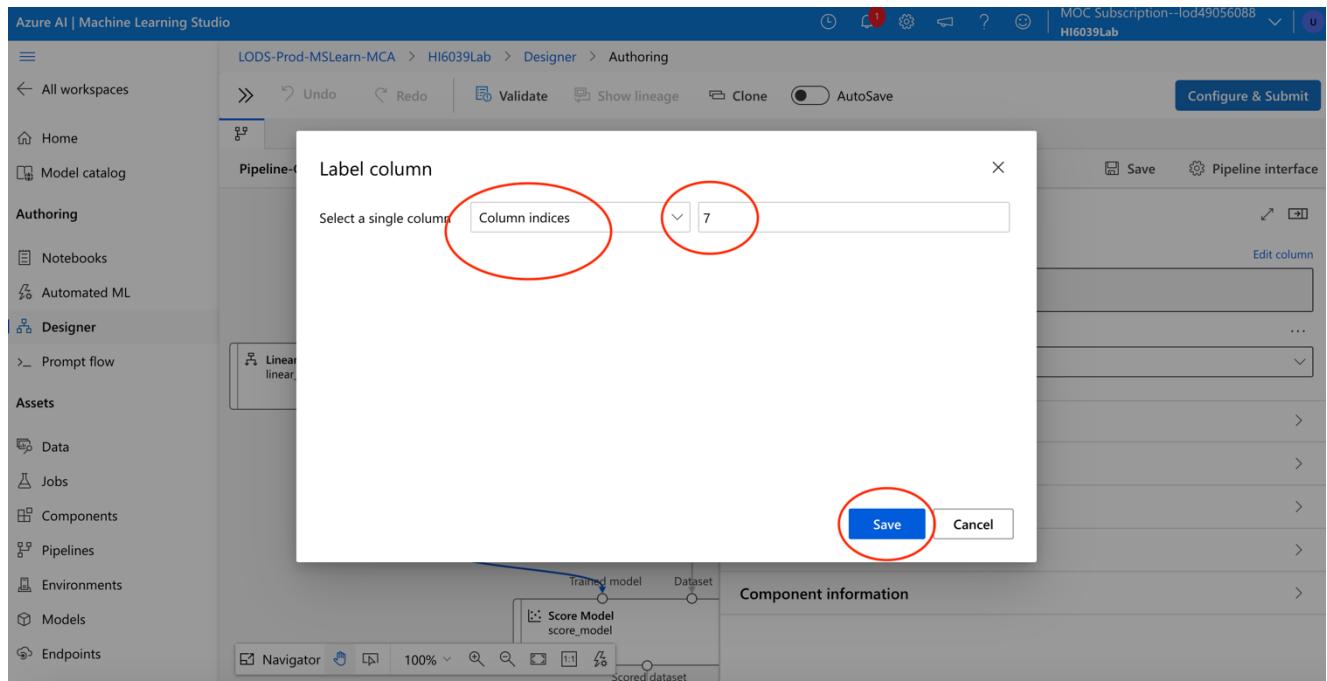
24. Search and add the following components to the pipeline: ‘Linear Regression’, ‘Train Model’, ‘Score Model’, and ‘Evaluate Model’. Then, connect them as shown in the screenshot below.



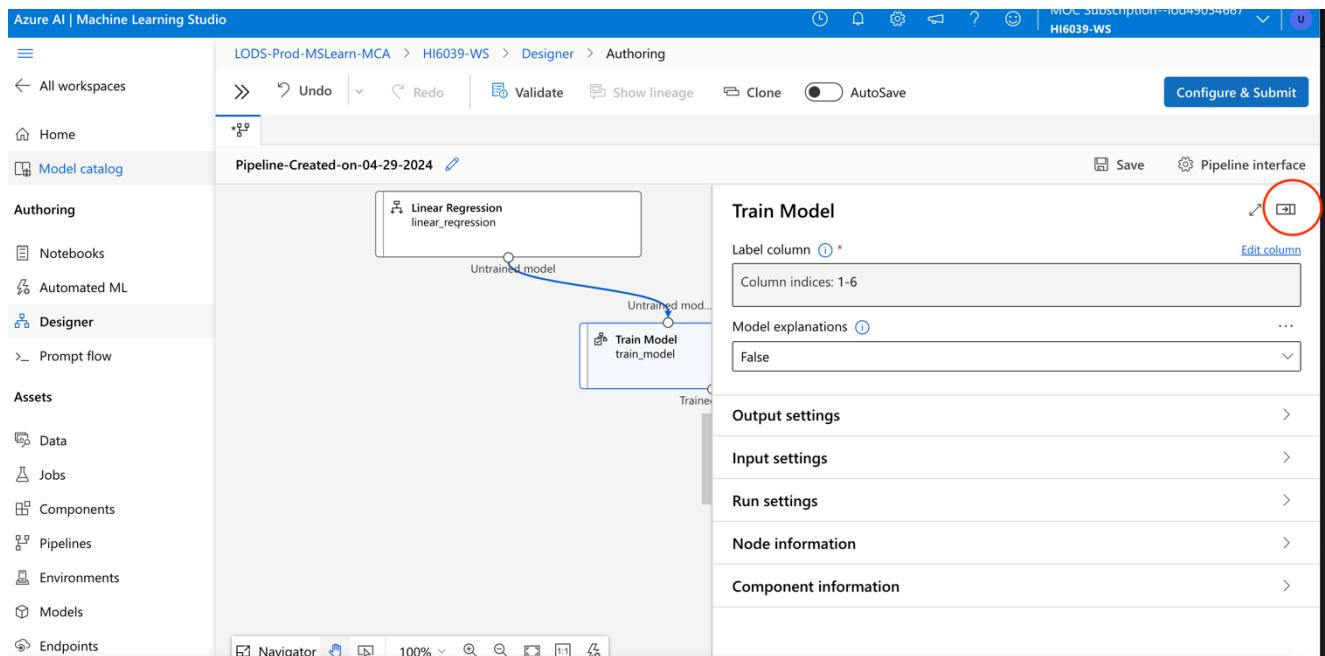
25. Double-click ‘Train Model’ component, and select ‘Edit column’.



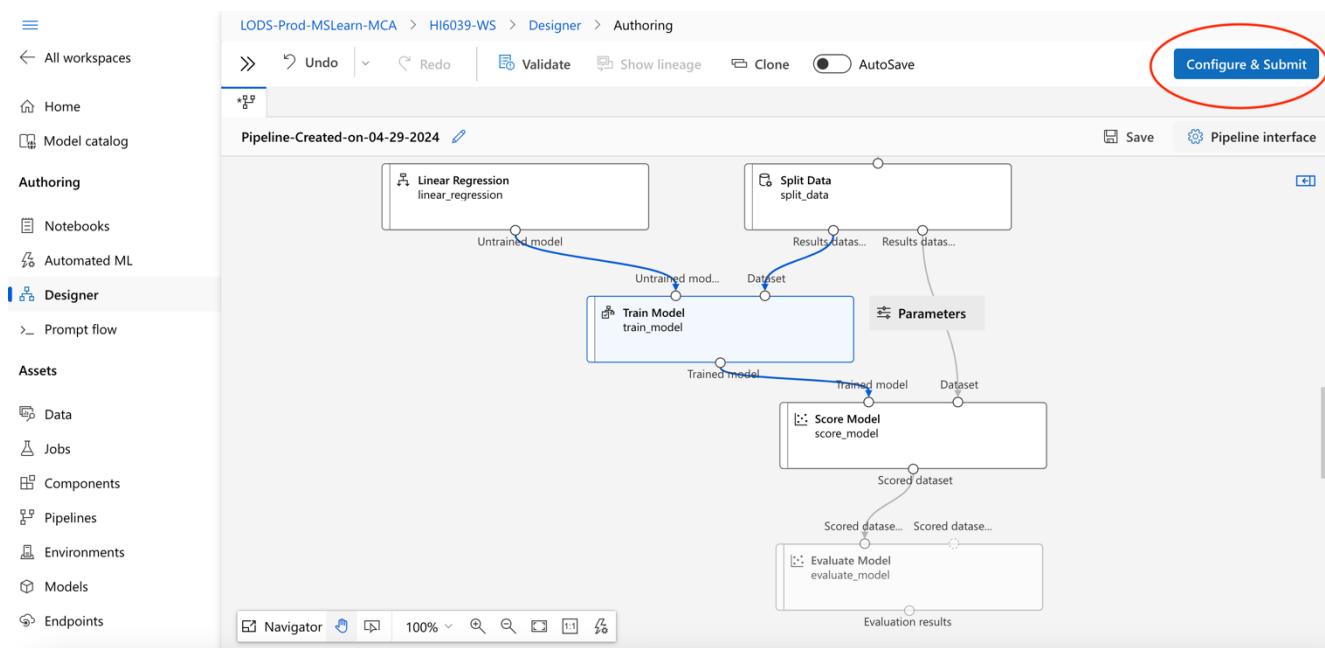
26. Select ‘Column indices’, enter ‘7’, and click ‘Save’.



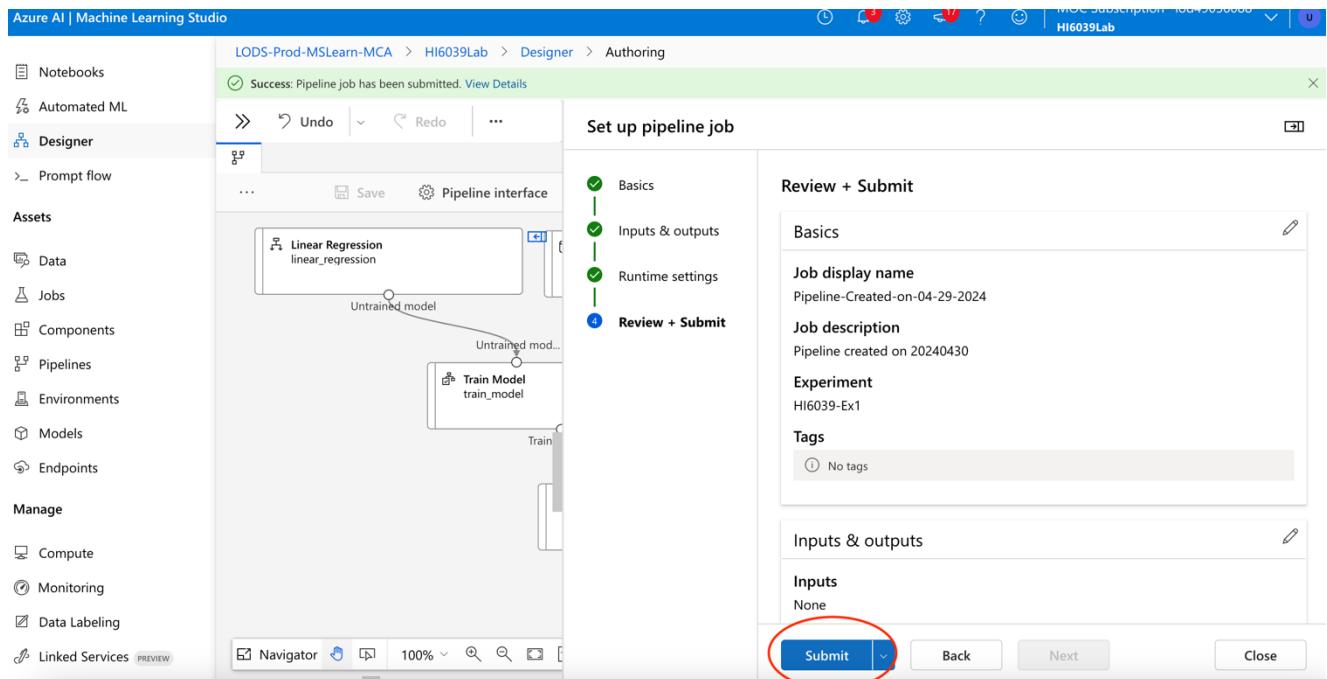
27. Click ‘Close’.



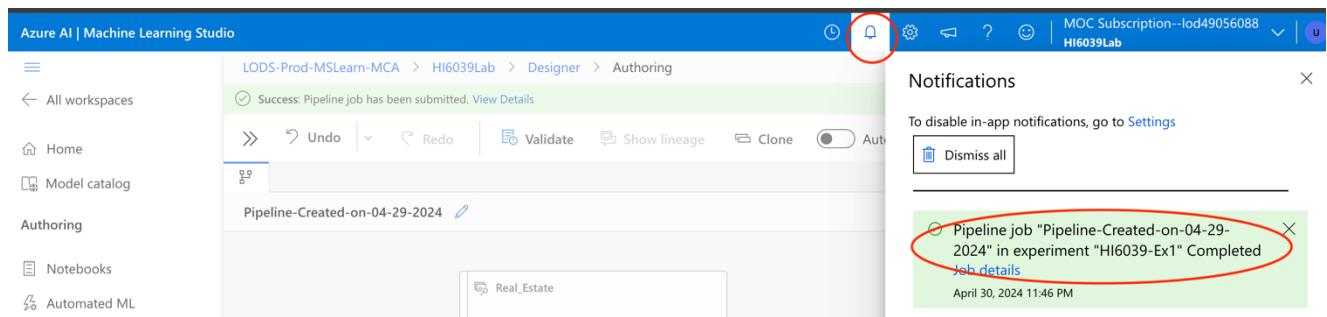
28. Click ‘Configure & Submit’.



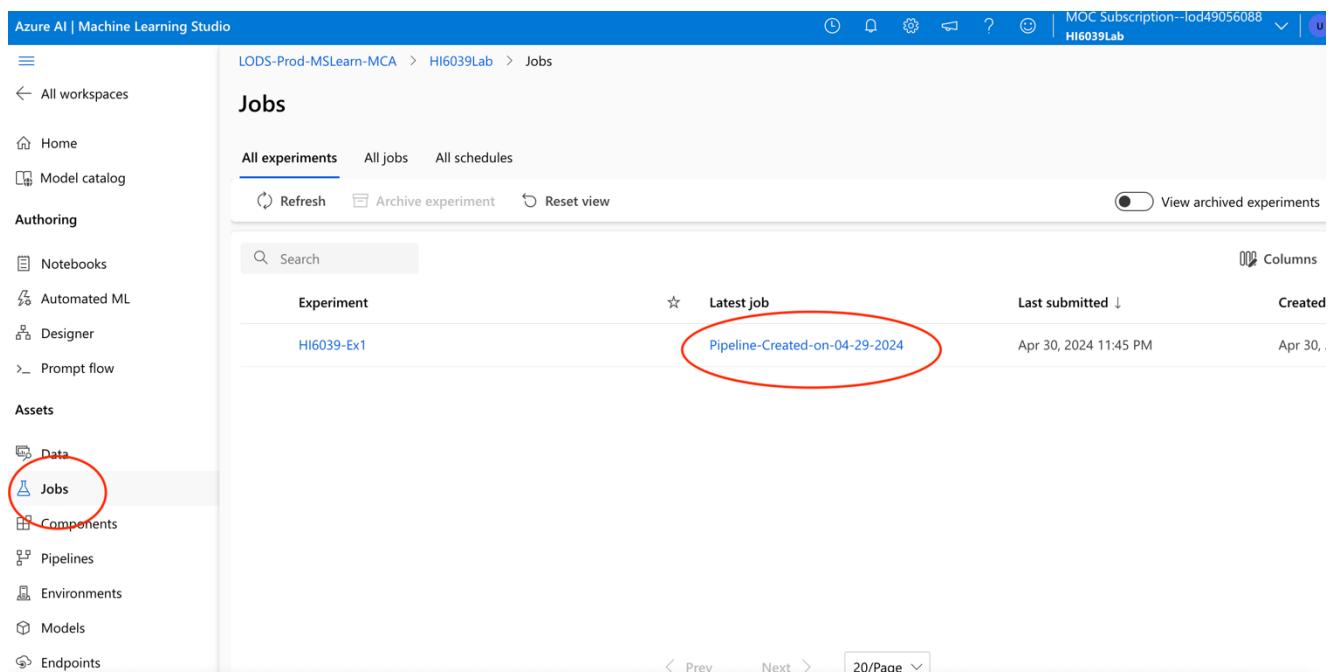
## 29. Click 'Submit'.



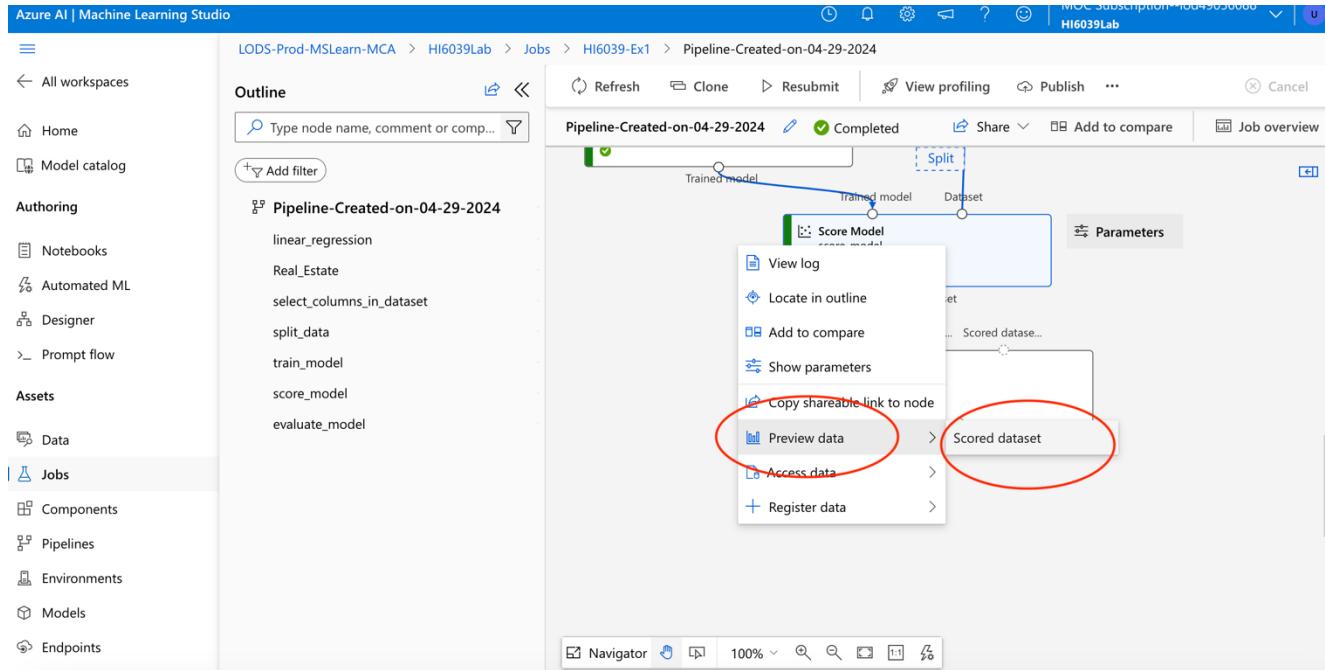
## 30. Waiting until the job has completed.



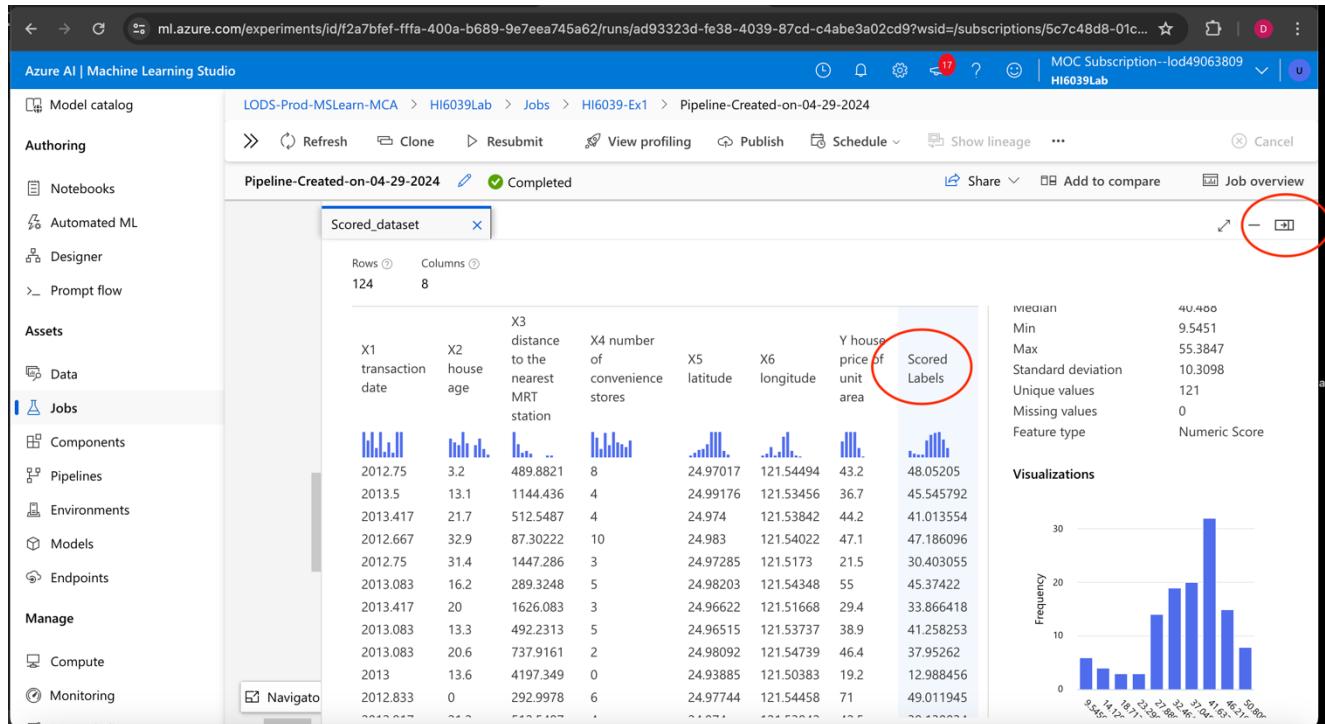
## 31. Click 'Job', then click the pipeline.



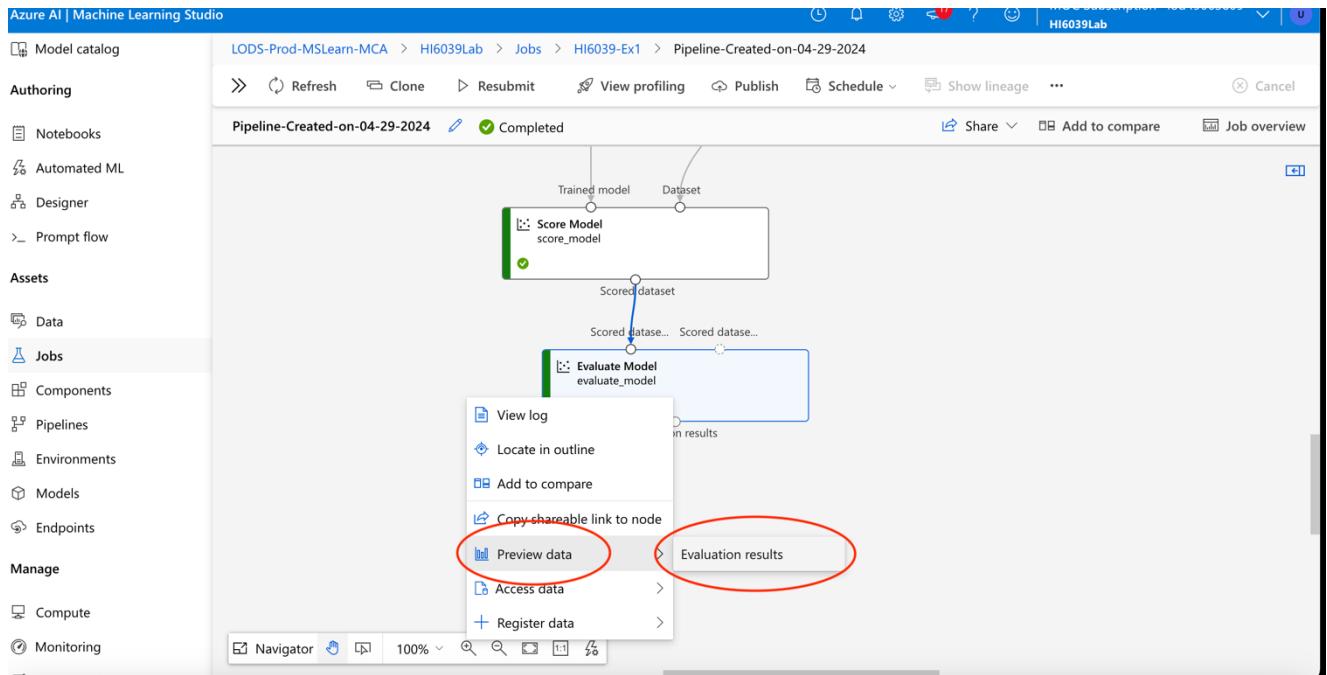
32. Right-click ‘Score model’ component, select ‘Preview data’, and select ‘Scored dataset’.



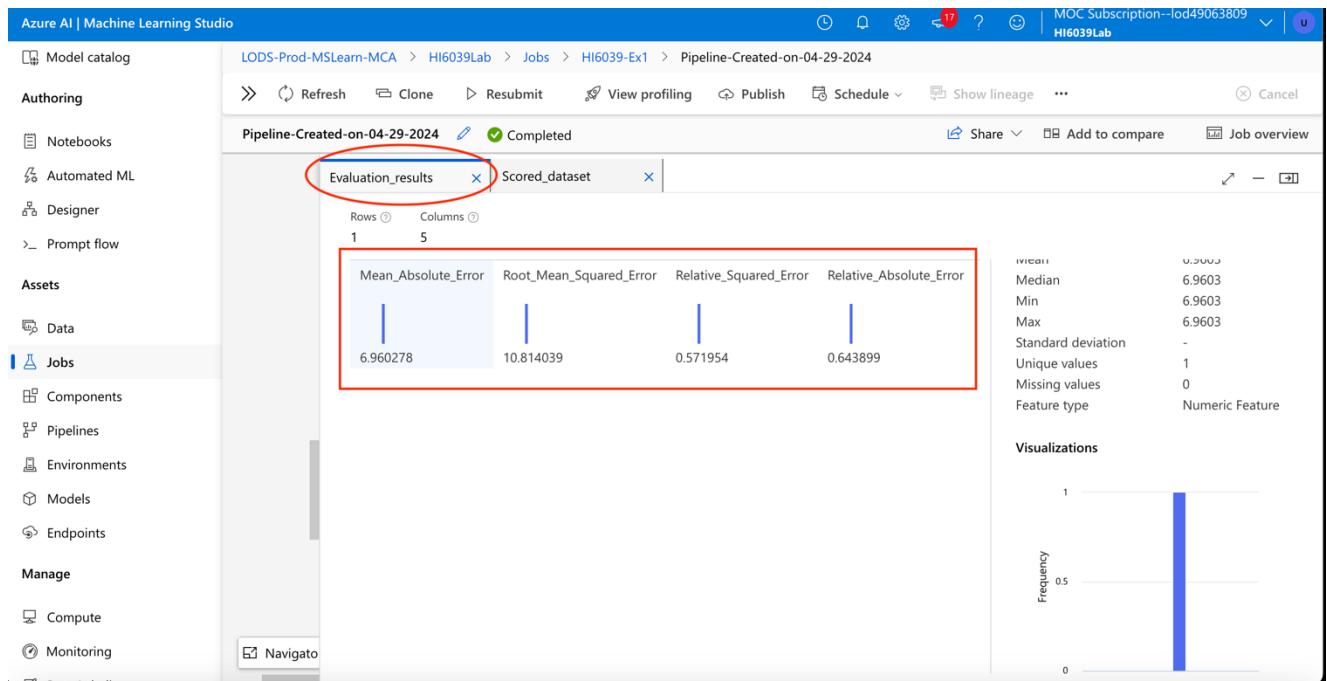
33. Now, you can see the predicted score for label/target attributes in the testing dataset. Click ‘Close’ to continue.



34. Right-click the ‘Evaluate Model’ component, select ‘Preview data’, and choose ‘Evaluation results’.



35. Now, you can see the performance of the predictive model.



---The end---