Lab Tutorial

Shuvadeep Kundu, skundu@calstatela.edu; Ankita Paul, apaul11@calstatela.edu;
Kevin Xu, hxu9@calstatela.edu

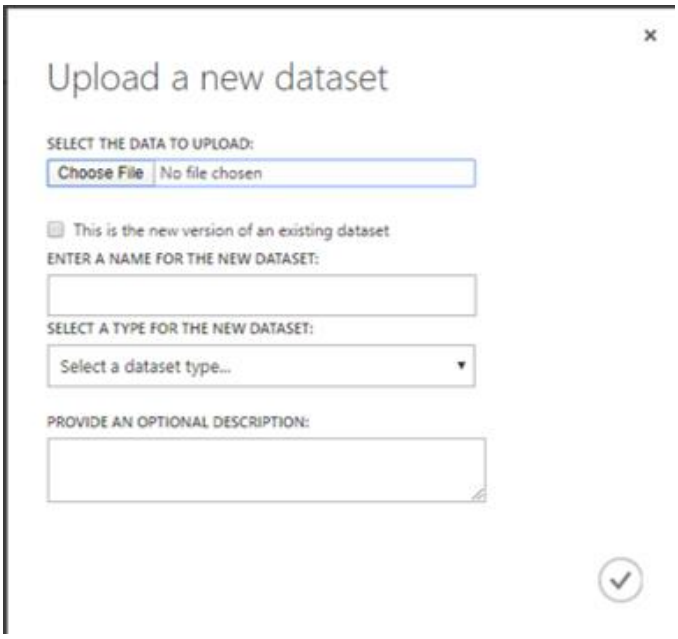
Iowa Liquor Sales Predictive Analytics

This tutorial will show you how to set up the following machine learning algorithms in Microsoft Azure ML to build models to predict liquor sales in the state of Iowa with Linear Regression, Boosted Decision Tree Regression, and Decision Forest Regression.

Data Preparation ➡ Splitting the data ➡ Training the model ➡ Evaluating the model

Creating a Sample dataset from the original big dataset

- Open a browser and go to https://studio.azureml.net and sign into your Azure ML account.
- In the Azure ML Studio, go to the **Experiments** page, click **New** at the bottom of the page. Then in the collection of Microsoft samples, select **Blank Experiment**. This creates a blank experiment. Give it the title **Sampling**.
- With the **Sampling** experiment open, at the bottom left, click **NEW**. Then in the **NEW** dialog box, click the **DATASET** tab.
- Go to **FROM LOCAL FILE**, choose file **iowa_Liquor_Sales.csv** and click **Ok** to upload a new dataset.



- Once the dataset is uploaded, in the **Sampling** experiment items pane on the left, expand **Saved Datasets**, expand **My Datasets**, and drag **iowa_Liquor_Sales.csv** to the experiment canvas in the middle of the page.
- Search for **Partition and Sample** module and drag it to the canvas below the dataset module. Connect the output of the **iowa_Liquor_Sales.csv** dataset to the **Dataset** input of the **Partition and Sample** module. In the **Properties** pane of the **Partition and Sample** module, select the fields as following:

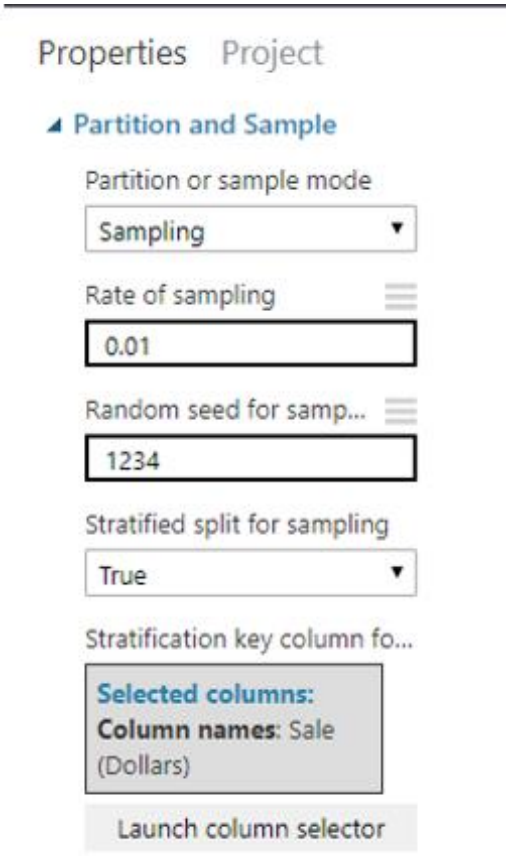
Partition or sample mode: Sampling

Rate of sampling: 0.01

Random seed for sampling: 1234

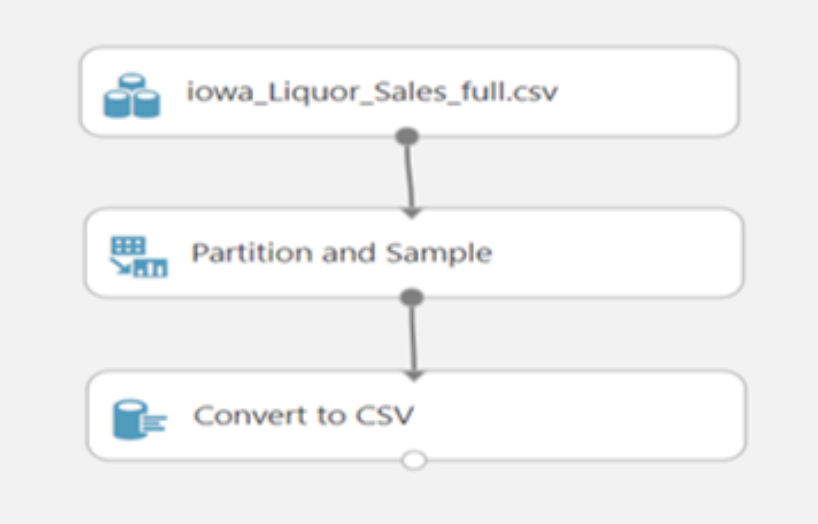
Stratified split for sampling: True

Launch Column Selector: Sale (Dollars)



- Search for **Convert to CSV** module and drag it to the canvas. Connect the output port (**Results Dataset**) of *Partition and Sample* module to the **Dataset** port of **Convert to CSV** module.

The experiment would look like this:

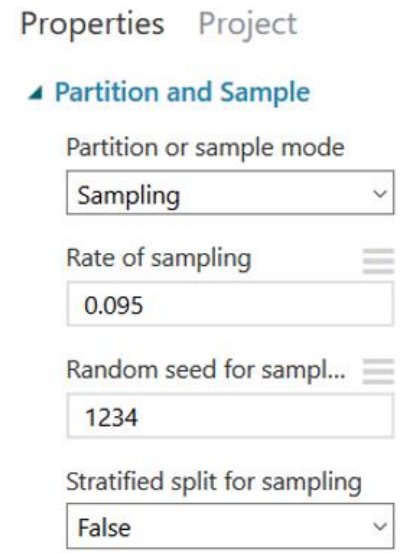


- **Save** the experiment and **Run** it. Once it has finished running, right click on the **Results dataset** (output) port of the **Convert to CSV** module and **Download** the sampled dataset. Name it as **sample.csv** and save it. Upload this new dataset **sample.csv** to the Datasets page of Azure ML Studio in the same way as the *iowa_Liquor_Sales.csv* dataset was uploaded. We would use this **sample.csv** dataset to build our Machine Learning models and run experiment on it.

Linear Regression

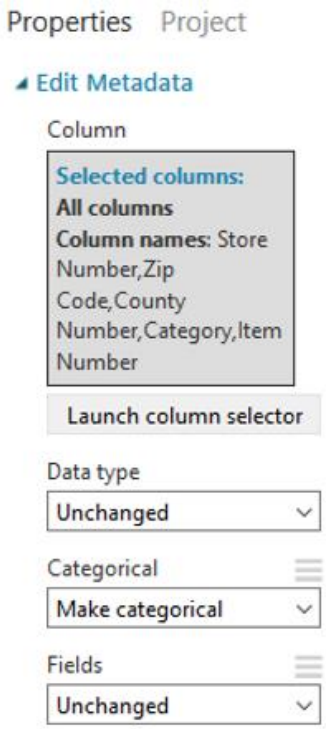
Data Preparation

- Open a browser and browse to https://studio.azureml.net. Then sign in using the Microsoft account associated with your Azure ML account.
- Create a new blank experiment and give it the title **IOWA – Linear Regression – Cross validation + Tune Model Hyperparameters**.
- In the experiment items pane on the left, expand **Saved Datasets**, expand **My Datasets**, and drag **sample.csv** to the experiment canvas.
- Search for **Partition and Sample** module and connect the output port of **sample.csv** module to the **Dataset** port of **Partition and Sample** module. In the **Properties** pane of Partition **and Sample** module, set the configurations as follows:

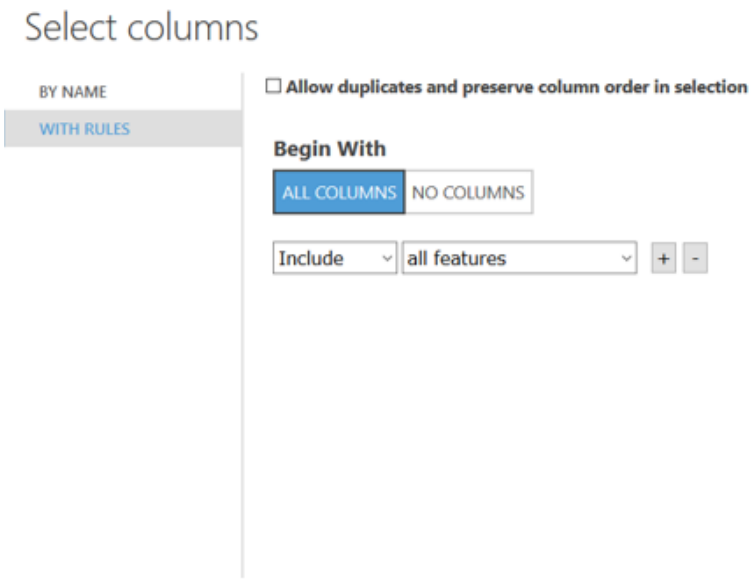


- Search for **Edit Metadata** module and drag it to the canvas. Connect the **Results Dataset** (output) port of *Partition and Sample* module to the **Dataset** port of *Edit Metadata* module. In the Properties pane of *Edit Metadata* module, **Launch Column Selector. With Rules > Begin With > All Columns** and select **Include column names.** Select the columns *Store Number, Zip Code, County Number, Category, Item Number* as follows :

Select **Ok**.
Set the rest of the configurations as follows:

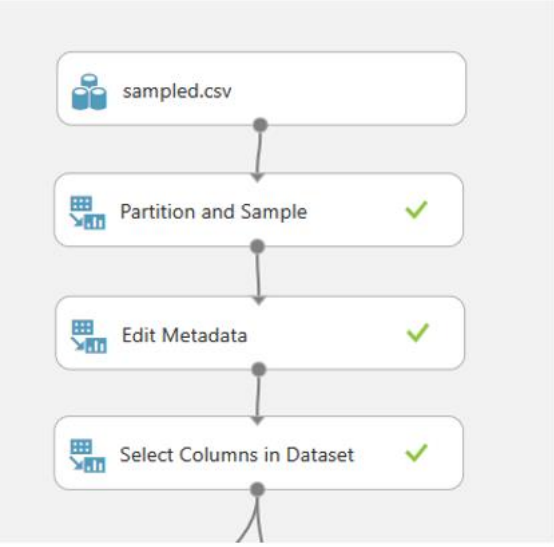


- Search for **Select Columns in Dataset** module. Connect the **Results Dataset** port of *Edit Metadata* module to the **Dataset** port of **Select Columns in Dataset** module. **Launch column selector** and select **All columns, All features** as shown below:



Save the experiment and run it.

The experiment would appear like this till now:

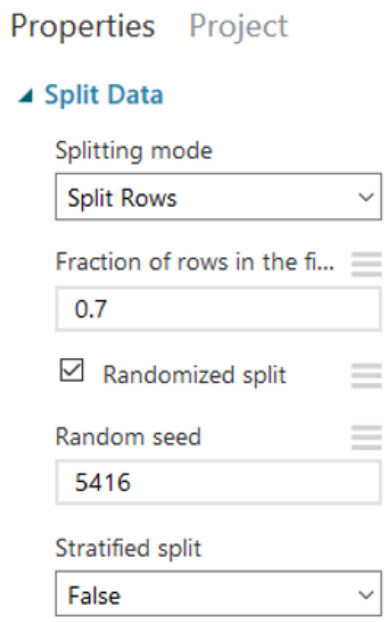


Splitting the data

Now that the data is prepared, we would split the data into Training dataset and Testing dataset.

- Search for the **Split Data** module and drag it onto the Canvas.
- Connect the **Results dataset** output of the **Select Columns in Dataset** module to the input of the **Split Data** module.

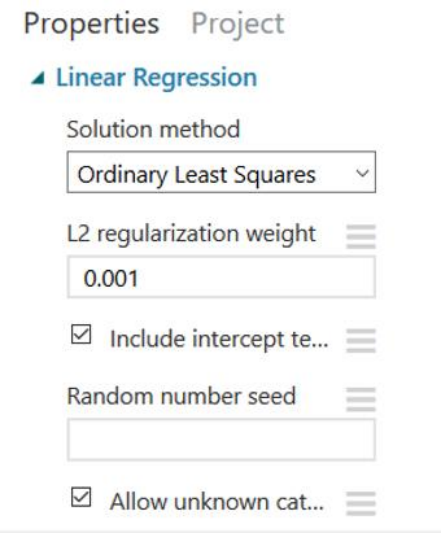
On the properties pane of the **Split Data** module, configure the properties as shown below:



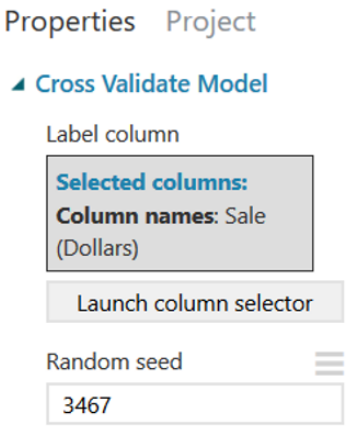
Training the model

Now that the data has been split, we can introduce the algorithm and then train the model.

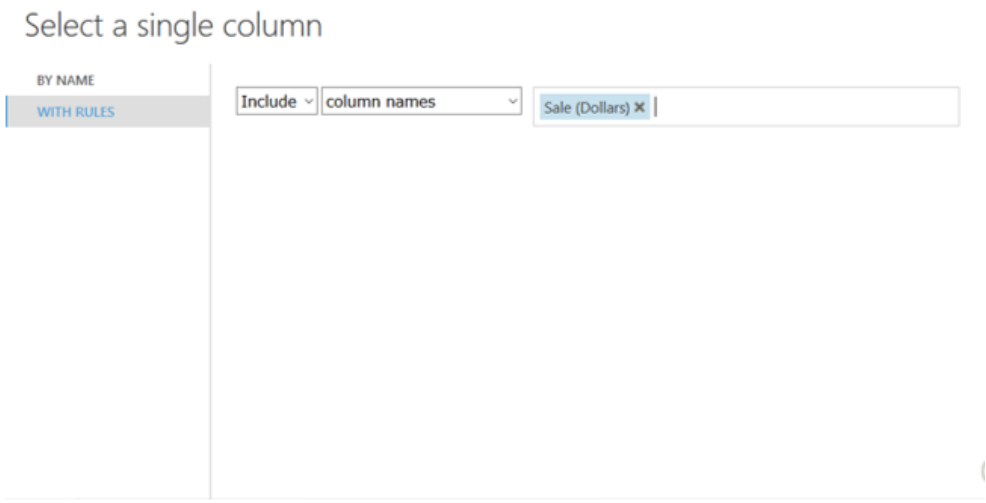
- Search for **Linear Regression** module and drag it to the canvas. Set the property as shown below:



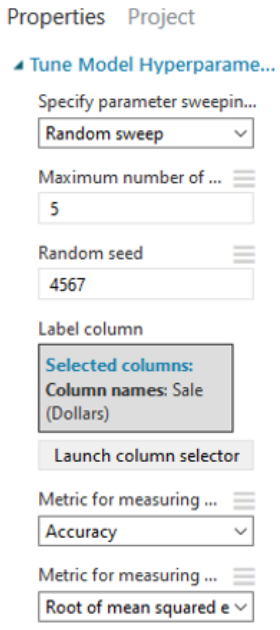
- Search for the **Cross Validate Model** module and drag it onto the canvas under the **Linear Regression** module. Connect the **Untrained model** output port of the **Linear** Regression module to the **Untrained model** input port (Left input port) of the **Cross Validate Model** module**.** Connect the **Results dataset** output of the **Select Columns in Dataset** module to the **Dataset** (right input port) port of the **Cross Validate Model** module. Set the property as shown below:



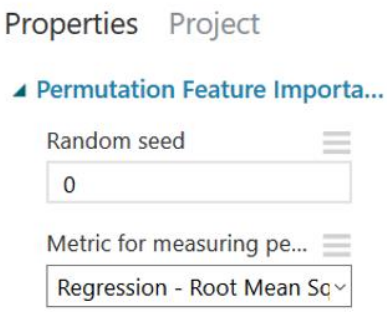
- Search for the **Train Model** module and drag it onto the canvas under the **Cross Validate Model** module. Connect the **Untrained model** output port of the **Linear Regression** module to the **Untrained model** input port (Left input port) of the **Train Model** module. Connect the **Results dataset1 (left output port) port** of the **Split Model** module to the **Dataset port (right input port)** of the **Train Model** module. In the properties pane, **Launch column selector** and select the column **Sale (Dollars)** as below:



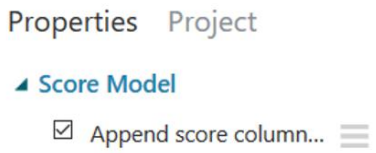
- Search for the **Tune Model Hyperparameters** module and drag it onto the canvas below the **Split Data** module. Connect the **Untrained model** output port of the **Linear Regression** module to the **Untrained model** input port (**Left input port**) of the **Tune Model Hyperparameters** module. Connect the **Results dataset1 (left output port) port** of the **Split Model** module to the **Training dataset port (middle port)** of the **Tune Model Hyperparameters** module. Set the configurations of the **Tune Model Hyperparameters** module as:
  - **Specify parameter sweeping mode :** Random Sweep
  - **Maximum number of runs on random sweep :** 5
  - **Random seed :** 4567
  - **Label column :** Sale (Dollars)
  - **Metric for measuring performance of classification :** Accuracy
  - **Metric for measuring performance of regression :** Root of mean squared error



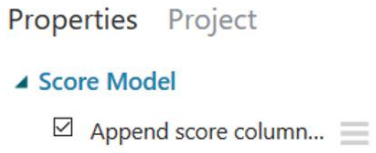
- Search for the **Permutation Features Importance** module and drag it onto the canvas. Connect the **Trained best model port (right output port)** of the **Tune Model Hyperparameters** module to the **Trained model input port (left input port)** of the **Permutation Features Importance** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Test data port (right input port)** of the **Permutation Features Importance** module. Set the property of the **Permutation Features Importance** module as shown below:



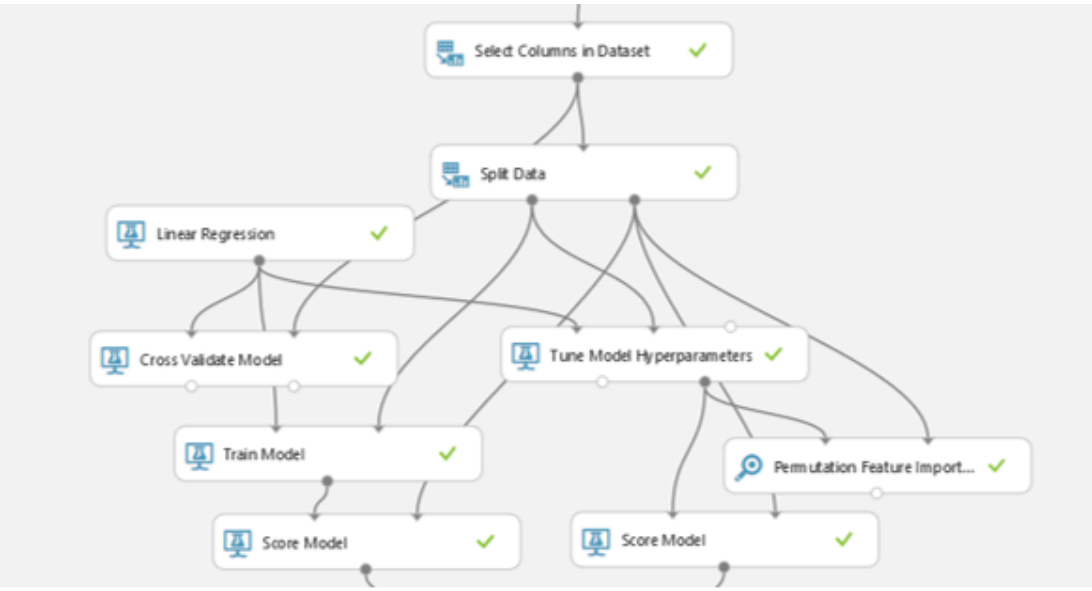
- Search for **Score Model** module and drag it onto the canvas under the **Train Model** module. Connect the **output port** of the **Train Model** module to the **Trained model input port (left input port)** of the **Score Model** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Dataset input port (right input port)** of the **Score Model** module. Set the properties as follows:



- Search for **Score Model** module again and drag this second **Score Model** module onto the canvas under the **Tune Model Hyperparameters** module. Connect the **Trained best model port (right output port)** of the **Tune Model Hyperparameters** module to the **Trained model input port (left input port)** of this second **Score Model** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Dataset input port (right input port)** of this **Score Model** module. Set the properties same as the first **Score Model** module, i.e,



The experiment would look like this till now:



Evaluating the model

- Search for the **Evaluate** module and drag it onto the canvas.
- Connect the **output port** of the **Score Model** module on the left side to the **left input port** of the **Evaluate** module. Connect the **output port** of the second **Score Model** module to the **right input port** of the **Evaluate** module. It would appear as below:

- Save and run the experiment.
- When the experiment has finished, **Visualize** the output from the **Evaluate** module. You would see that both the **Cross Validate Model** and **Tune Model Hyperparameters** have performed equally.

IOWA - Linear Regression - Cross Validation +Tuning... › Evaluate Model › Evaluation results

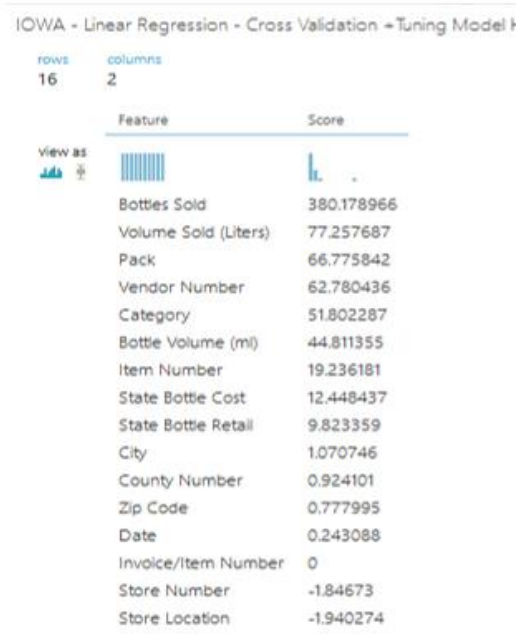
Metrics

Mean Absolute Error	91.888803
Root Mean Squared Error	137.145669
Relative Absolute Error	0.215055
Relative Squared Error	0.074045
Coefficient of Determination	0.925955

Metrics

Mean Absolute Error	91.888803
Root Mean Squared Error	137.145669
Relative Absolute Error	0.215055
Relative Squared Error	0.074045
Coefficient of Determination	0.925955

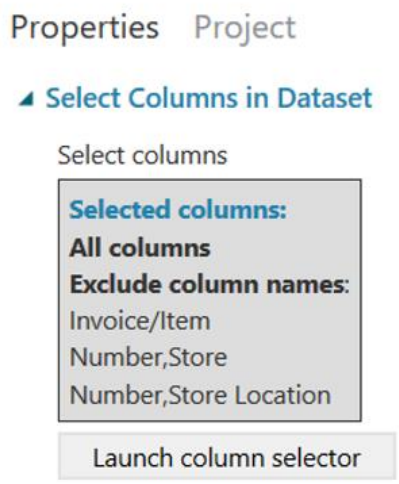
- You can also visualize the output from the **Permutation Feature Importance** module to see features with their scores of importance and can prune features with less importance. This might improve the model.


IOWA - Linear Regression - Cross Validation - Tuning Model H

rows 16 columns 2

Feature	Score
Bottles Sold	380.178966
Volume Sold (Liters)	77.257687
Pack	66.775842
Vendor Number	62.780436
Category	51.802287
Bottle Volume (ml)	44.811355
Item Number	19.236181
State Bottle Cost	12.448437
State Bottle Retail	9.823359
City	1.070746
County Number	0.924101
Zip Code	0.777995
Date	0.243088
Invoice/Item Number	0
Store Number	-1.84673
Store Location	-1.940274

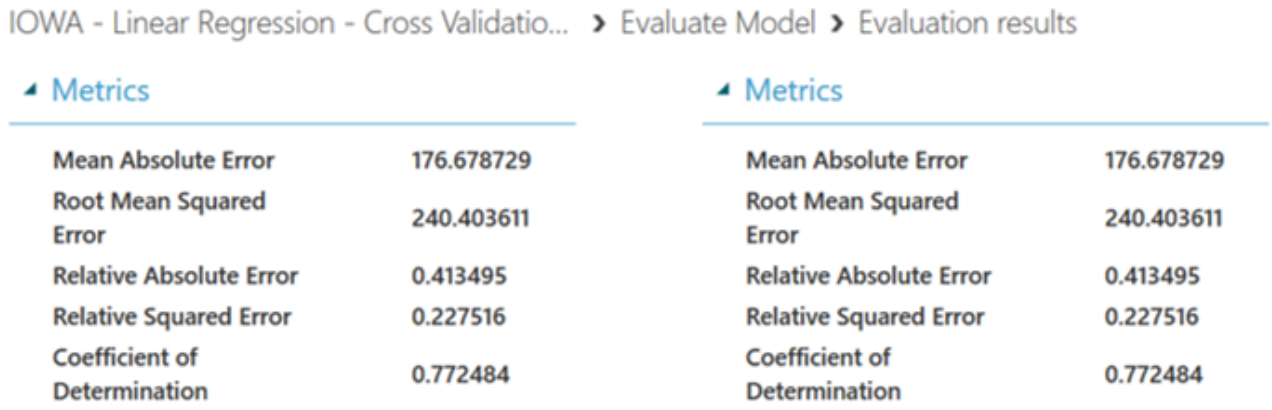
- Save a copy of this experiment by clicking **Save as** in the bottom of the page and name the experiment copy as **IOWA – Linear Regression – Cross validation + Tune Model Hyperparameters – New.**
- In the experiment **IOWA – Linear Regression – Cross validation + Tune Model Hyperparameters – New**, on the properties pane of **Select Columns in Dataset** module, **exclude** the columns *Invoice/Item Number, Store Number* and *Store Location* which had 0 and negative scores of feature importance.



It can be done as follows:

Launch column selector > WITH RULES > Begin With > ALL COLUMNS > Exclude > column names > *Invoice/Item Number, Store Number, Store Location*

- Save and run the experiment.
- When the experiment has finished, **Visualize** the output form the **Evaluate** module to see if any improvement is there. It turned out that the **RMSE** value has actually increased and the **Coefficient of Determination** value decreased.

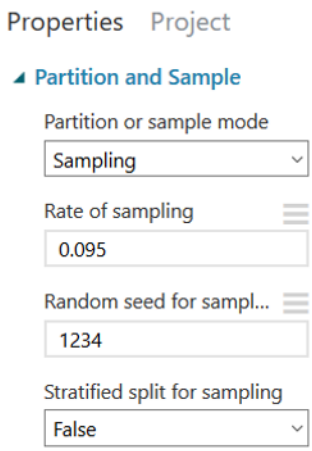


So, the model performed better before pruning the features.

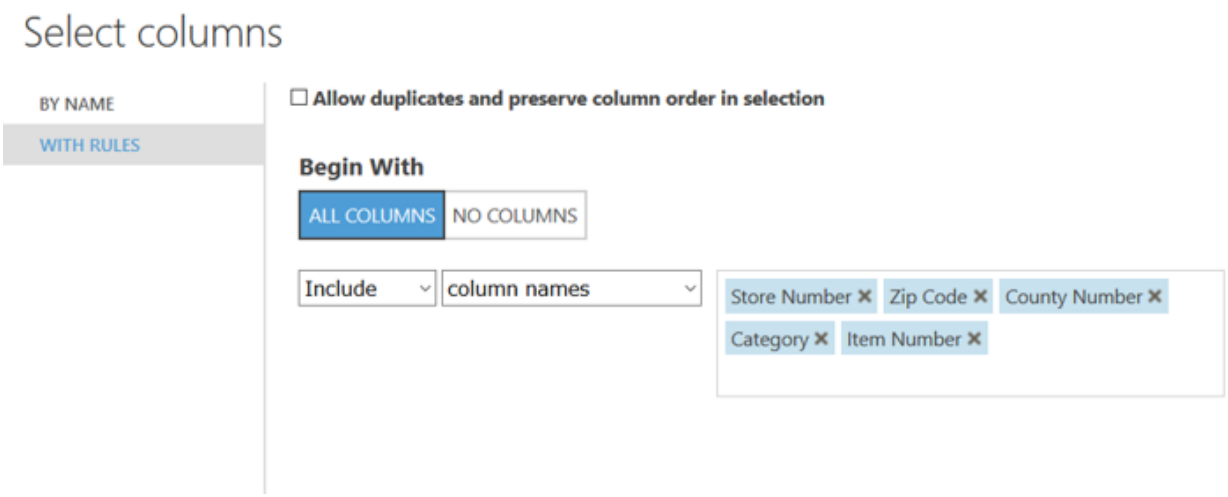
Boosted Decision Tree Regression

Data Preparation

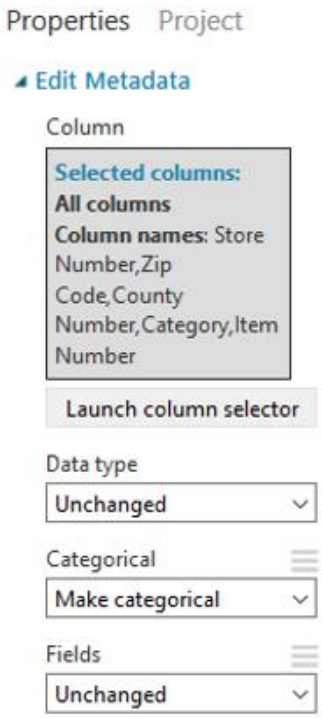
- Open a browser and browse to https://studio.azureml.net. Then sign in using the Microsoft account associated with your Azure ML account.
- Create a new blank experiment and give it the title **IOWA - Boosted Decision Tree Regression**.
- In the experiment items pane on the left, expand **Saved Datasets**, expand **My Datasets**, and drag **sample.csv** to the experiment canvas.
- Search for **Partition and Sample** module and connect the output port of **sample.csv** module to the **Dataset** port of **Partition and Sample** module. In the **Properties** pane of **Partition and Sample** module, set the configurations as follows:



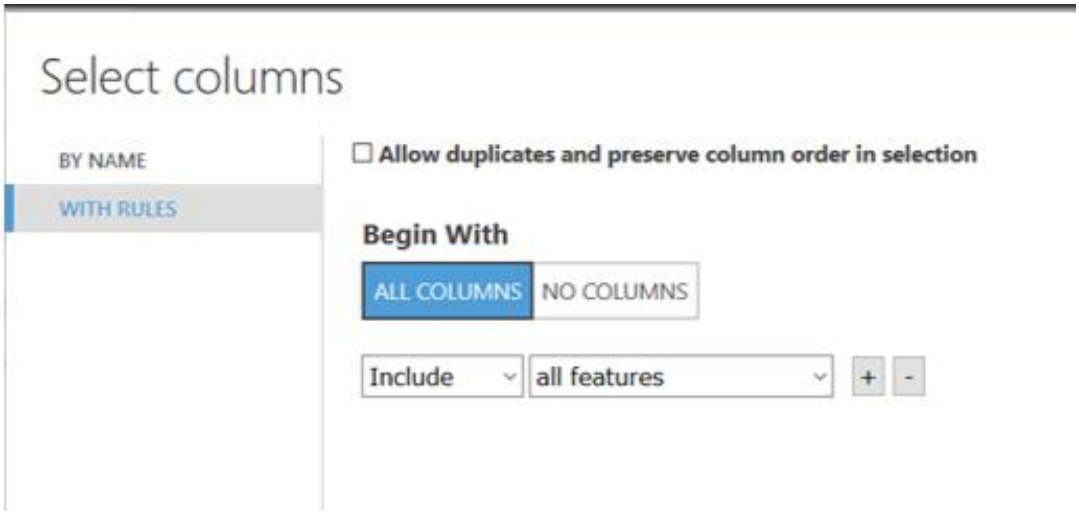
- Search for **Edit Metadata** module and drag it to the canvas. Connect the **Results Dataset** (output) port of *Partition and Sample* module to the **Dataset** port of *Edit Metadata* module. In the Properties pane of *Edit Metadata* module, **Launch Column Selector. With Rules > Begin With > All Columns** and select **Include column names.** Select the columns *Store Number, Zip Code, County Number, Category, Item Number* as follows :



Select **Ok**.
Set the rest of the configurations as follows:

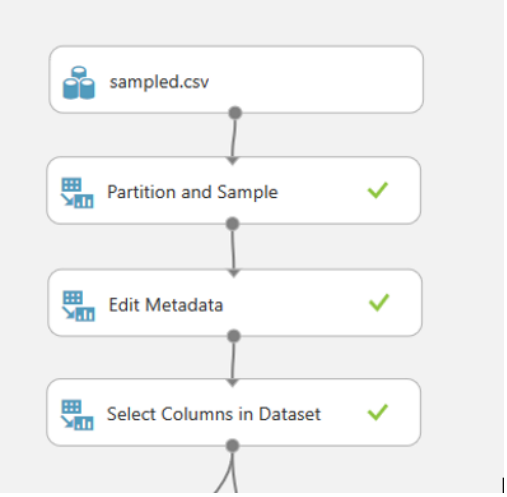


- Search for **Select Columns in Dataset** module. Connect the **Results Dataset** port of *Edit Metadata* module to the **Dataset** port of **Select Columns in Dataset** module. **Launch column selector** and select **All columns, All features** as shown below:



Save the experiment and run it.

The experiment would appear like this till now:

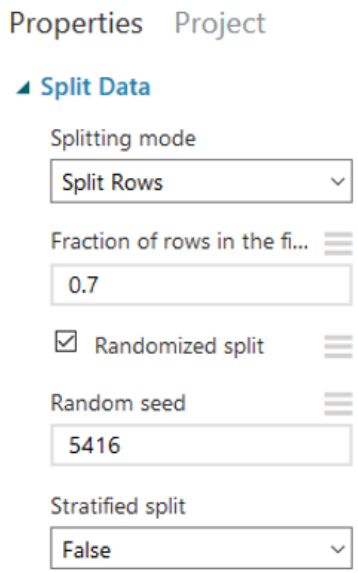


Splitting the data

Now that the data is prepared, we would split the data into Training dataset and Testing dataset.

- Search for the **Split Data** module and drag it onto the Canvas.
- Connect the **Results dataset** output of the **Select Columns in Dataset** module to the input of the **Split Data** module.

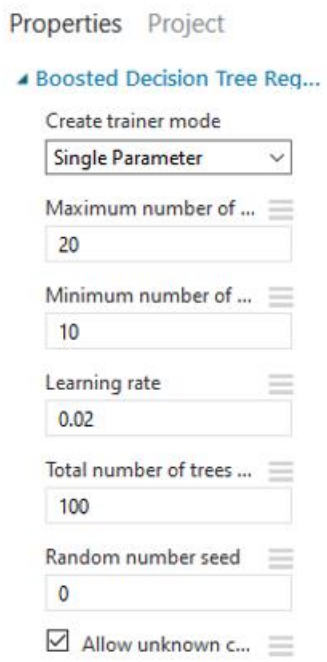
On the properties pane of the **Split Data** module, configure the properties as shown below



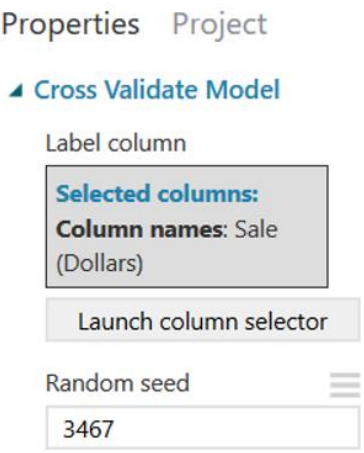
Training the model

Now that the data has been split, we can introduce the algorithm and then train the model.

- Search for the **Boosted Decision Tree Regression** module and drag it onto the canvas. Set the property as shown below:

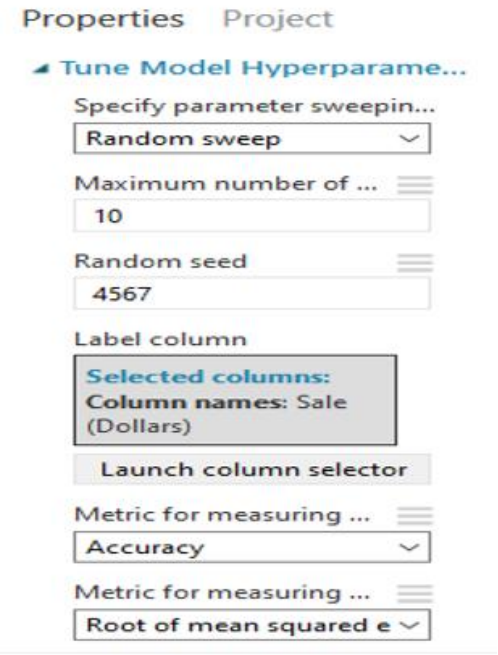


- Search for the **Cross Validate Model** module and drag it onto the canvas under the **Boosted Decision Tree Regression** module. Connect the **Untrained model** output port of the **Boosted Decision Tree Regression** module to the **Untrained model input port (Left input port)** of the **Cross Validate Model** module**.** Connect the **Results dataset** output of the **Select Columns in Dataset** module to the **Dataset (right input port) port** of the **Cross Validate Model** module. Set the property as shown below:

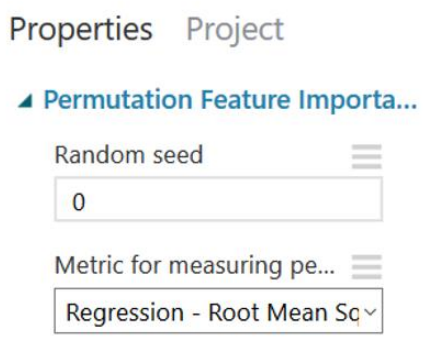


- Search for the **Train Model** module and drag it onto the canvas under the **Cross Validate Model** module. Connect the **Untrained model** output port of **Boosted Decision Tree Regression** module to the **Untrained model input port (Left input port)** of the **Train Model** module. Connect the **Results dataset1 (left output port) port** of the **Split Model** module to the **Dataset port (right input port)** of the **Train Model** module. In the properties pane, **Launch column selector** and select the column **Sale (Dollars)**.

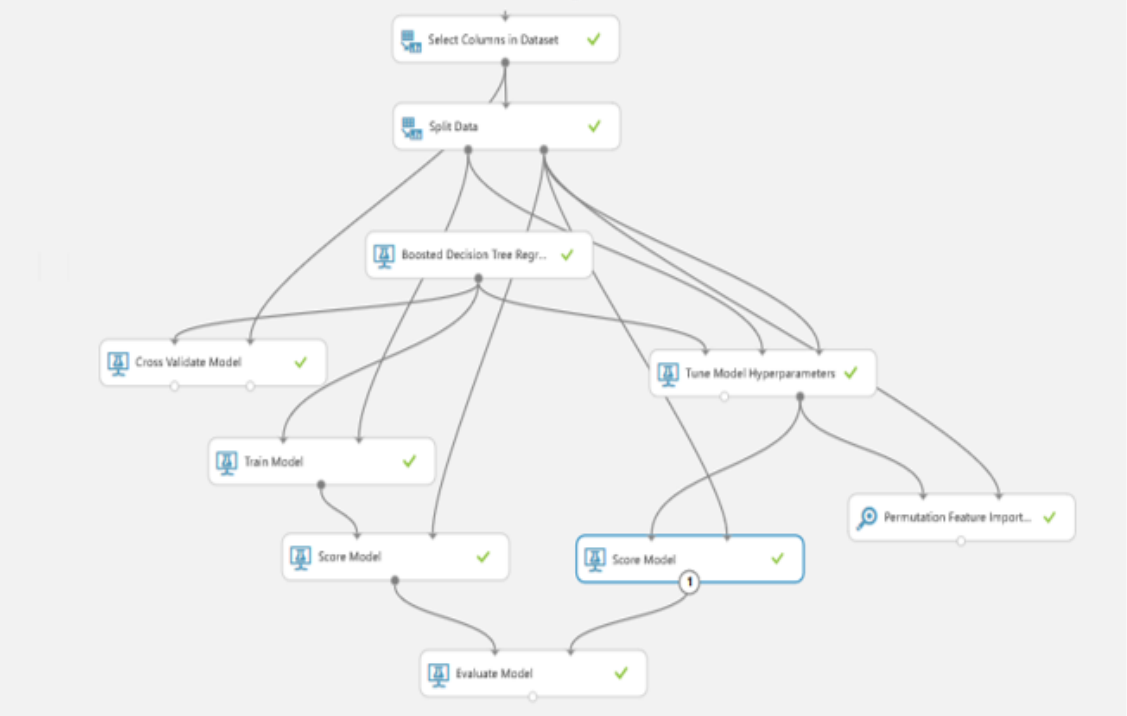
- Search for the **Tune Model Hyperparameters** module and drag it onto the canvas below the **Split Data** module. Connect the **Untrained model** output port of the **Boosted Decision Tree Regression** module to the **Untrained model** input port (**Left input port**) of the **Tune Model Hyperparameters** module. Connect the **Results dataset1 (left output port) port** of the **Split Model** module to the **Training dataset port (middle input port)** of the **Tune Model Hyperparameters** module. Connect the **Results dataset2 (right output port) port** of the **Split Model** module to the **Optional validation dataset port (right input port)** of the **Tune Model Hyperparameters** module. Set the configurations of the **Tune Model Hyperparameters** module as below :



- Search for the **Permutation Features Importance** module and drag it onto the canvas. Connect the **Trained best model port (right output port)** of the **Tune Model Hyperparameters** module to the **Trained model input port (left input port)** of the **Permutation Features Importance** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Test data port (right input port)** of the **Permutation Features Importance** module. Set the property of the **Permutation Features Importance** module as shown below:



- Search for **Score Model** module and drag it onto the canvas under the **Train Model** module. Connect the **output port** of the **Train Model** module to the **Trained model input port (left input port)** of the **Score Model** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Dataset input port (right input port)** of the **Score Model** module.
- Search for **Score Model** module again and drag this second **Score Model** module onto the canvas under the **Tune Model Hyperparameters** module. Connect the **Trained best model port (right output port)** of the **Tune Model Hyperparameters** module to the **Trained model input port (left input port)** of this second **Score Model** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Dataset input port (right input port)** of this **Score Model** module. The experiment figure would look like this till now:



Evaluating the model

- Search for the **Evaluate** module and drag it onto the canvas.
- Connect the **output port** of the **Score Model** module on the left side to the **left input port** of the **Evaluate** module. Connect the **output port** of the second **Score Model** module to the **right input port** of the **Evaluate** module. It would appear as below:

- Save and run the experiment.
- When the experiment has finished, **Visualize** the output from the **Evaluate** module. We would see that the **Tune Model Hyperparameter** has performed better, with lower **RMSE** value and higher **Coefficient of Determination** value.

IOWA - Boosted Decision Tree Regression › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	206.890108
Root Mean Squared Error	267.974043
Relative Absolute Error	0.484202
Relative Squared Error	0.282693
Coefficient of Determination	0.717307

Metrics

Mean Absolute Error	119.92497
Root Mean Squared Error	173.140216
Relative Absolute Error	0.28067
Relative Squared Error	0.118012
Coefficient of Determination	0.881988

- You can also visualize the output from the **Permutation Feature Importance** module to see features with their scores of importance and can prune features with less importance. This might improve the model.

IOWA - Boosted Decision Tree Regression › Permutation F

rows 16 columns 2

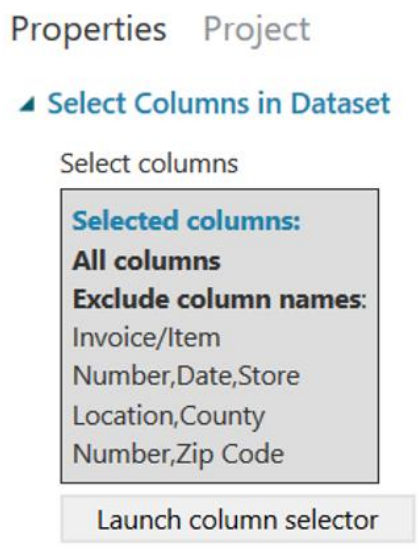
view as

Feature	Score
Bottles Sold	454.818329
Pack	81.371853
Vendor Number	75.608935
Category	54.132026
Bottle Volume (ml)	39.98075
Volume Sold (Liters)	6.885084
State Bottle Cost	2.324521
Store Number	0.948541
City	0.295601
Item Number	0.219255
Zip Code	0.167606
State Bottle Retail	0.015421
Invoice/Item Number	0
Date	0
Store Location	0
County Number	-0.122389

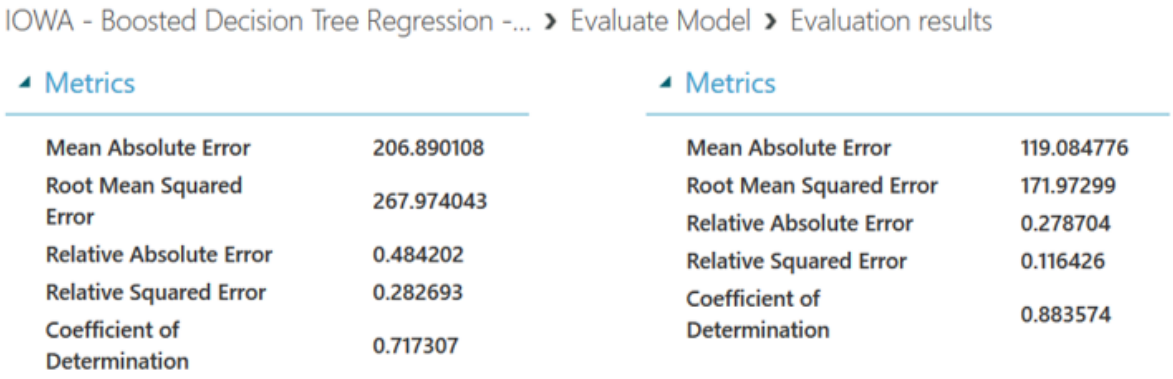
- Save a copy of this experiment by clicking **Save as** in the bottom of the page and name the experiment copy as **IOWA - Boosted Decision Tree Regression – New.**
- In the experiment **IOWA - Boosted Decision Tree Regression – New**, on the properties pane of **Select Columns in Dataset** module, **exclude** the columns *Invoice/Item Number, Date, Store Location, County Number, Zip Code*.

It can be done as follows:

Launch column selector > WITH RULES > Begin With > ALL COLUMNS > Exclude > column names > *Invoice/Item Number, Date, Store Location, County Number, Zip Code*



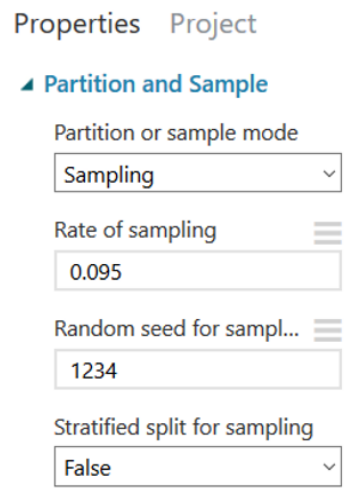
- Save and run the experiment.
- When the experiment has finished, **Visualize** the output form the **Evaluate** module to see if any improvement is there. It turned out that the **RMSE** value has decreased by very negligible amount and the **Coeffiecient of Determination** has increased by very negligible amount.



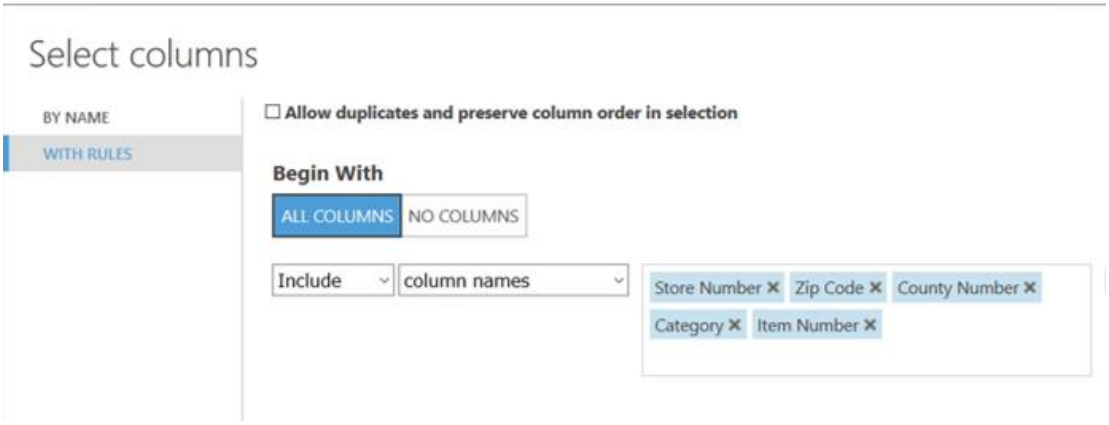
Decision Forest Regression

Data Preparation

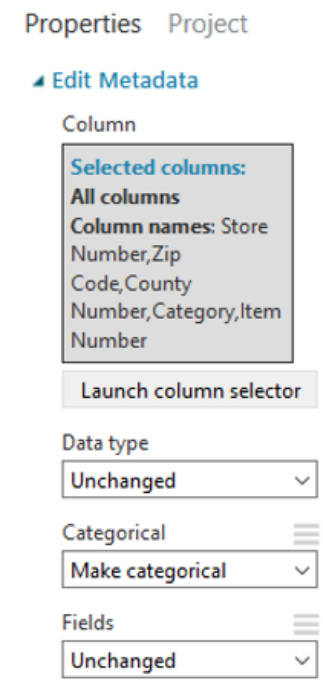
- Open a browser and browse to https://studio.azureml.net. Then sign in using the Microsoft account associated with your Azure ML account.
- Create a new blank experiment and give it the title **IOWA - Decision Forest Regression**.
- In the experiment items pane on the left, expand **Saved Datasets**, expand **My Datasets**, and drag **sample.csv** to the experiment canvas.
- Search for **Partition and Sample** module and connect the output port of **sample.csv** module to the **Dataset** port of **Partition and Sample** module. In the **Properties** pane of **Partition and Sample** module, set the configurations as follows:



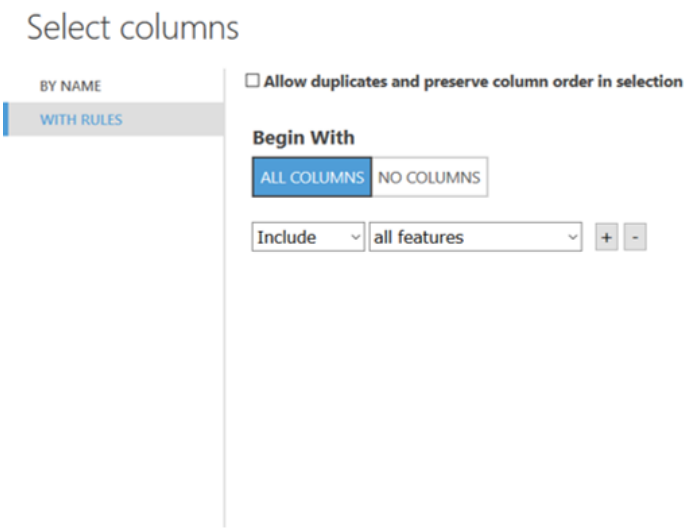
- Search for **Edit Metadata** module and drag it to the canvas. Connect the **Results Dataset** (output) port of *Partition and Sample* module to the **Dataset** port of *Edit Metadata* module. In the Properties pane of *Edit Metadata* module, **Launch Column Selector. With Rules > Begin With > All Columns** and select **Include column names.** Select the columns *Store Number, Zip Code, County Number, Category, Item Number* as follows :



Select **Ok**.
Set the rest of the configurations as follows:

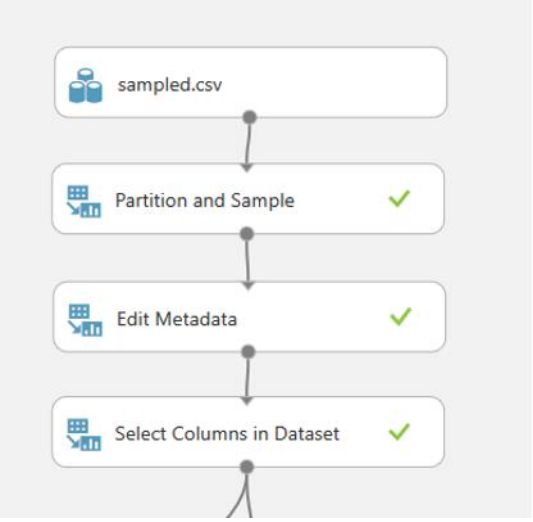


- Search for **Select Columns in Dataset** module. Connect the **Results Dataset** port of *Edit Metadata* module to the **Dataset** port of **Select Columns in Dataset** module. **Launch column selector** and select **All columns, All features** as shown below:



Save the experiment and run it.

The experiment would appear like this till now:

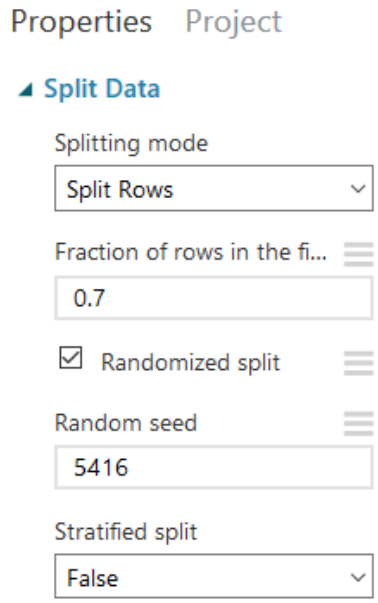


Splitting the data

Now that the data is prepared, we would split the data into Training dataset and Testing dataset.

- Search for the **Split Data** module and drag it onto the Canvas.
- Connect the **Results dataset** output of the **Select Columns in Dataset** module to the input of the **Split Data** module.

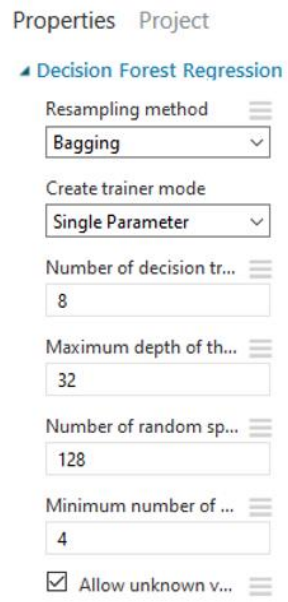
On the properties pane of the **Split Data** module, configure the properties as shown below



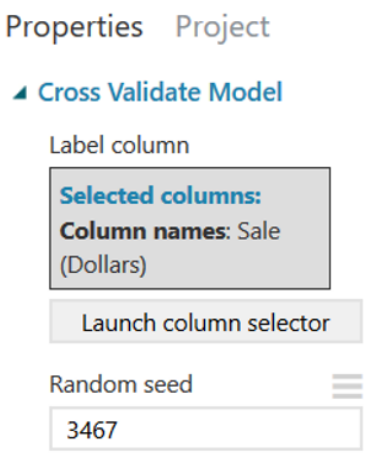
Training the model

Now that the data has been split, we can introduce the algorithm and then train the model.

- Search for the **Decision Forest Regression** module and drag it onto the canvas. Set the property as shown below

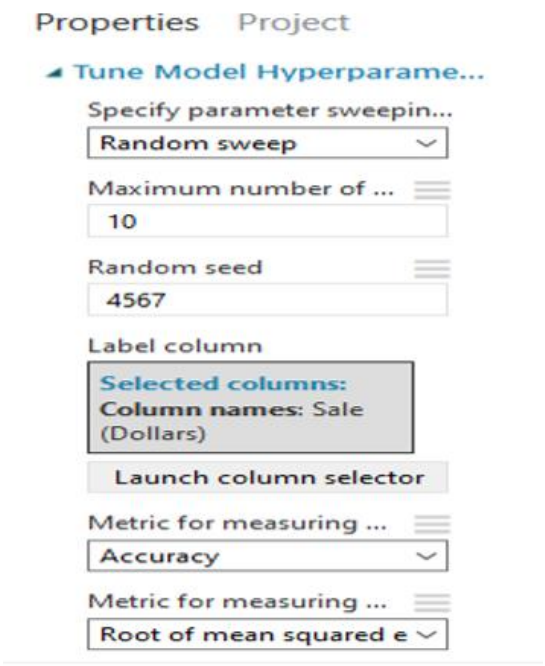


- Search for the **Cross Validate Model** module and drag it onto the canvas under the **Decision Forest Regression** module. Connect the **Untrained model** output port of the **Decision Forest Regression** module to the **Untrained model input port (Left input port)** of the **Cross Validate Model** module**.** Connect the **Results dataset** output of the **Select Columns in Dataset** module to the **Dataset (right input port) port** of the **Cross Validate Model** module. Set the property as shown below:

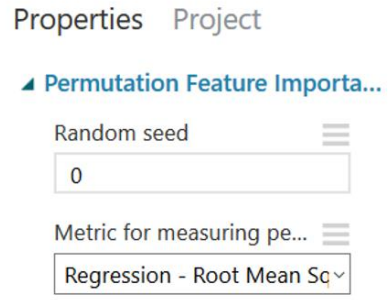


- Search for the **Train Model** module and drag it onto the canvas under the **Cross Validate Model** module. Connect the **Untrained model** output port of **Decision Forest Regression** module to the **Untrained model input port (Left input port)** of the **Train Model** module. Connect the **Results dataset1 (left output port) port** of the **Split Model** module to the **Dataset port (right input port)** of the **Train Model** module. In the properties pane, **Launch column selector** and select the column **Sale (Dollars)**.
- Search for the **Tune Model Hyperparameters** module and drag it onto the canvas below the **Split Data** module. Connect the **Untrained model** output port of the **Decision Forest Regression** module to the **Untrained model** input port (**Left input port**) of the **Tune Model Hyperparameters** module. Connect the **Results dataset1 (left output port) port** of the **Split Model** module to the **Training**

dataset port (middle input port) of the **Tune Model Hyperparameters** module. Connect the **Results dataset2 (right output port) port** of the **Split Model** module to the **Optional validation dataset port (right input port)** of the **Tune Model Hyperparameters** module. Set the configurations of the **Tune Model Hyperparameters** module as below :

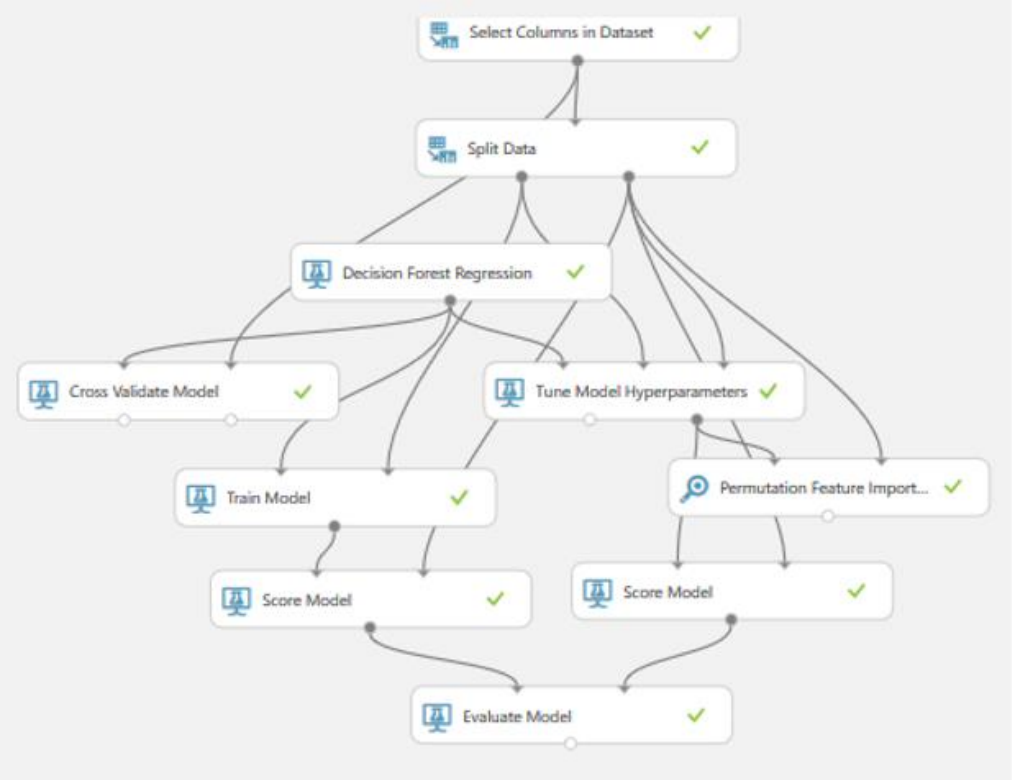


- Search for the **Permutation Features Importance** module and drag it onto the canvas. Connect the **Trained best model port (right output port)** of the **Tune Model Hyperparameters** module to the **Trained model input port (left input port)** of the **Permutation Features Importance** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Test data port (right input port)** of the **Permutation Features Importance** module. Set the property of the **Permutation Features Importance** module as shown below:



- Search for **Score Model** module and drag it onto the canvas under the **Train Model** module. Connect the **output port** of the **Train Model** module to the **Trained model input port (left input port)** of the **Score Model** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Dataset input port (right input port)** of the **Score Model** module.
- Search for **Score Model** module again and drag this second **Score Model** module onto the canvas under the **Tune Model Hyperparameters** module. Connect the **Trained best model port (right**

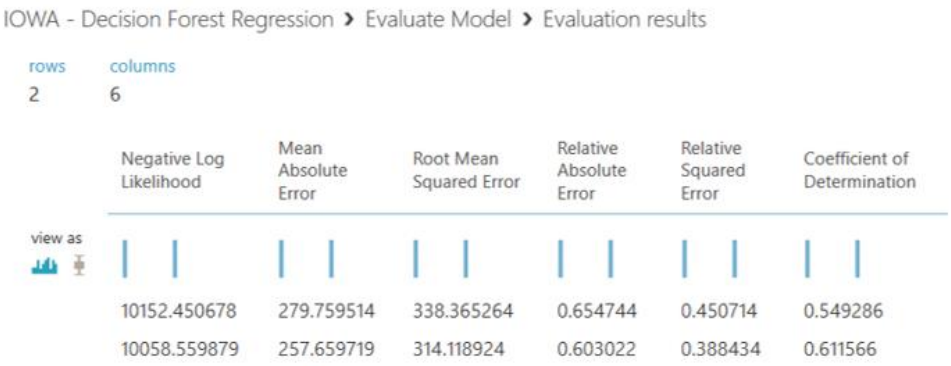
output port) of the **Tune Model Hyperparameters** module to the **Trained model input port (left input port)** of this second **Score Model** module. Connect the **Results dataset2 port (right output port)** of the **Split Data** module to the **Dataset input port (right input port)** of this **Score Model** module. The experiment figure would look like this till now:



Evaluating the model

- Search for the **Evaluate** module and drag it onto the canvas.
- Connect the **output port** of the **Score Model** module on the left side to the **left input port** of the **Evaluate** module. Connect the **output port** of the second **Score Model** module to the **right input port** of the **Evaluate** module. It would appear as below:

- Save and run the experiment.
- When the experiment has finished, **Visualize** the output from the **Evaluate** module. We would see that the **Tune Model Hyperparameter** has performed better, with lower **RMSE** value and higher **Coefficient of Determination** value.

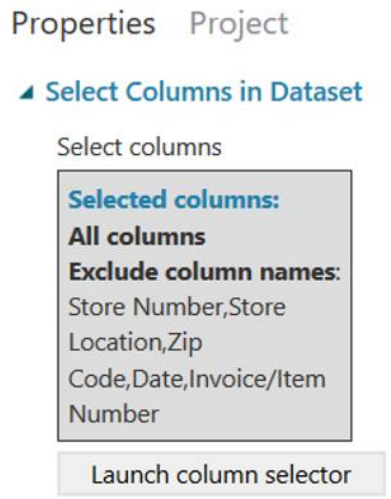


- You can also visualize the output from the **Permutation Feature Importance** module to see features with their scores of importance and can prune features with less importance. This might improve the model.

- Save a copy of this experiment by clicking **Save as** in the bottom of the page and name the experiment copy as **IOWA - Decision Forest Regression - New.**
- In the experiment **IOWA - Decision Forest Regression - New**, on the properties pane of **Select Columns in Dataset** module, **exclude** the column names *Store Number, Store Location, Zip Code, Date, Invoice/Item Number.*

It can be done as follows:

Launch column selector > WITH RULES > Begin With > ALL COLUMNS > Exclude > column names > *Store Number, Store Location, Zip Code, Date, Invoice/Item Number*



- Save and run the experiment.
- When the experiment has finished, **Visualize** the output form the **Evaluate** module to see if any improvement is there. It turned out that the **RMSE** value has decreased by considerable amount and the **Coeffiecient of Determination** has increased. However, the **Cross Validate Model** performed better this after pruning the features.