CIS 5560 Term Project Tutorial

Nikita Shendkar; Ruchita Shinde; Sohong Chakraborty; Shahnawaz Khan

Instructor: **Jongwook Woo**

Date: 05/10/2018

Lab
Tutorial

Ruchita Shinde (rshinde@calstatela.edu) , Nikita Shendkar (nshendk@calstatela.edu) , Sohong Chakraborty (schakra2@calstatela.edu) , Shahnawaz Khan (skhan30@calstatela.edu)

Analysis of National Stock Exchange of India

Objective:

- To extract data from Kaggle.com for NSE stocks of India.
- Create Accounts in AzureML and Databricks.
- To Split, Evaluate and use Cross validator and Linear regression model on the dataset .
- Visualize using AzureML and Databricks by coding through python and Spark ML.
- https://gallery.cortanaintelligence.com/Experiment/NSE-Analysis-2

Platform Specification:

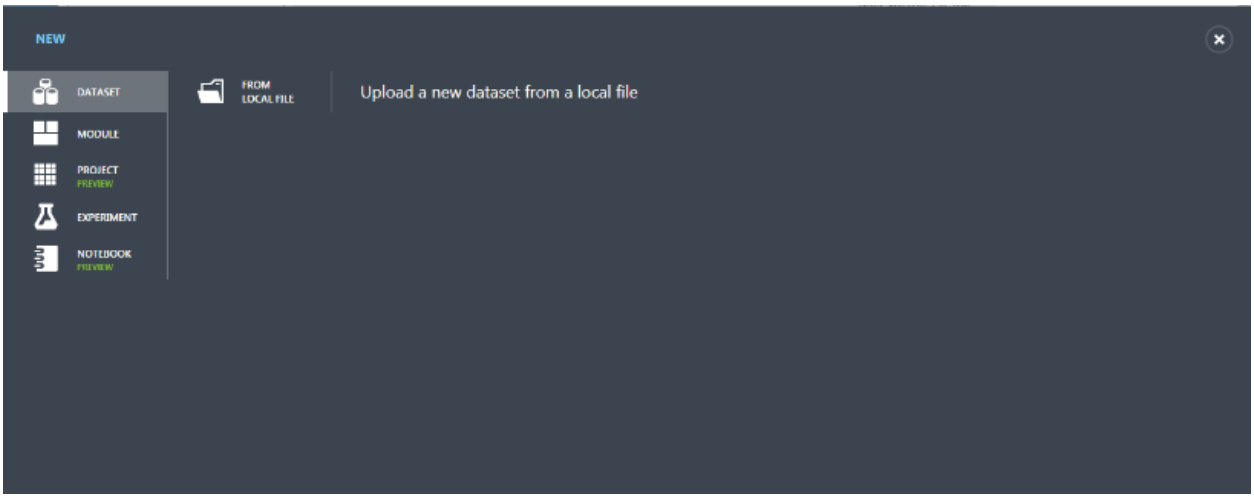
- Azure ML account
- Databricks Community Account (Apache Spark 2.3.0 and Scala 2.11)
- A web browser and Internet connection
- CPU Speed: ~3.4GHz
- # of nodes: 1
- Total Memory Size: 10GB

Sign into Azure ML Studio

1. Open a browser and browse to https://studio.azureml.net.
2. Click Sign In and sign in using the Microsoft account associated with your free Azure ML account

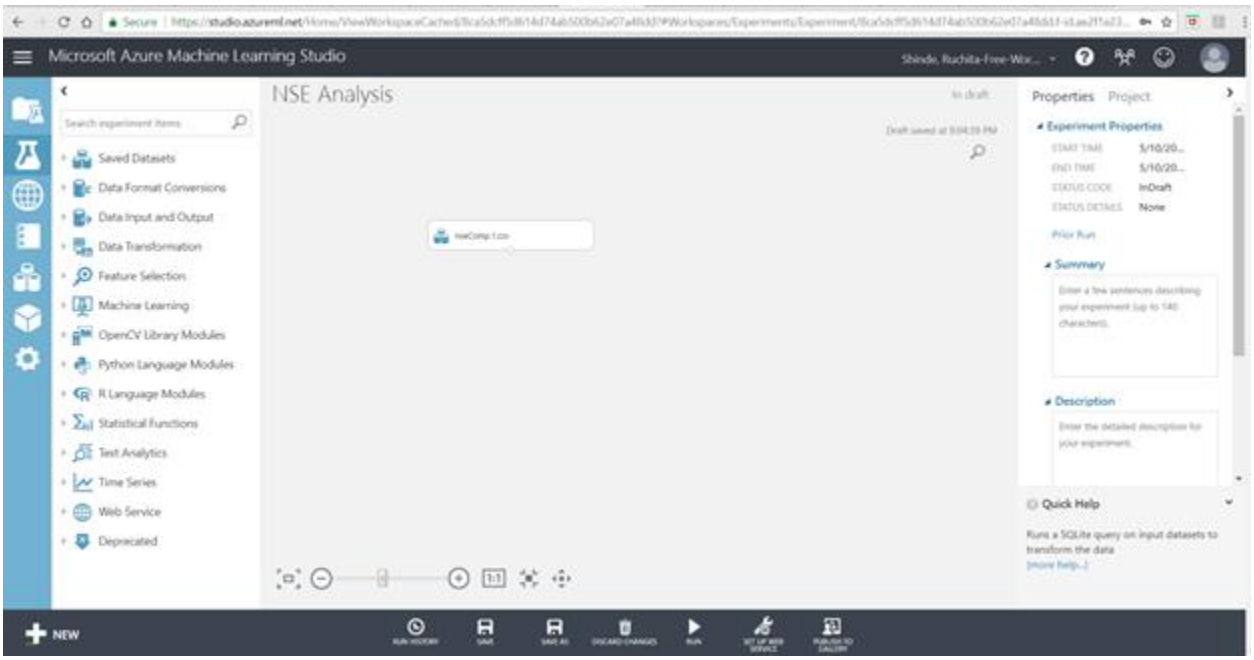
Step 1: Upload the Data Set from local file

This step is to upload datasets: nseComp.1.csv from local file system.



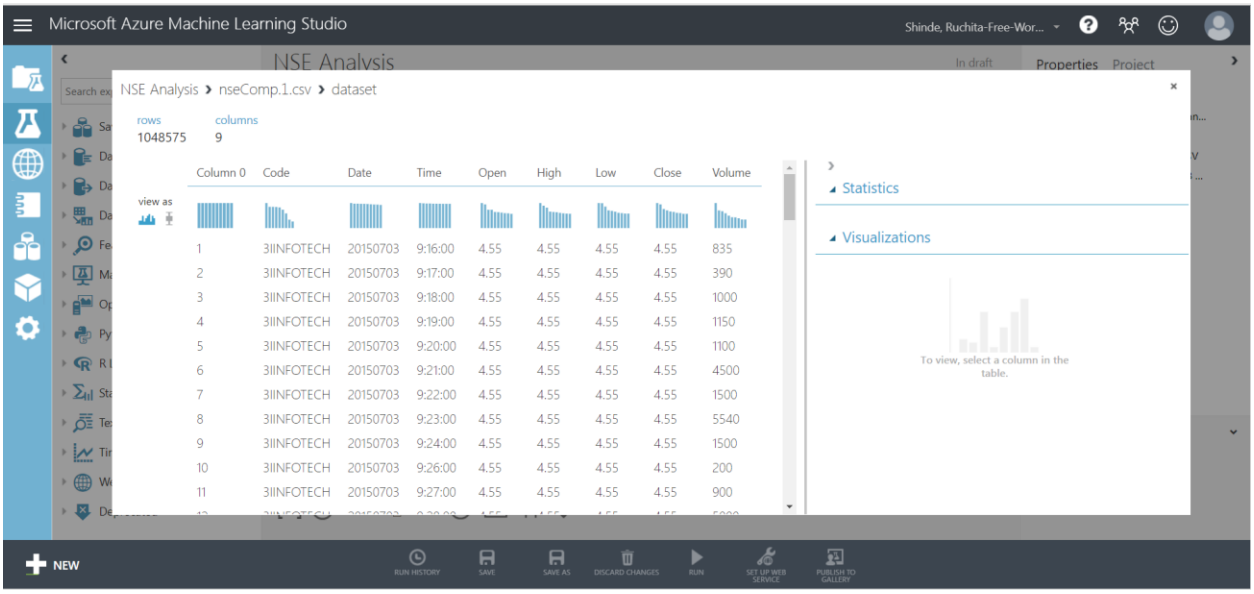
Steps to upload:

- Select +New option and then click on dataset => from local file => open the location where dataset is stored.
- Enter a name for the new dataset: nseComp.1.csv
- Select a type for the new dataset: Generic CSV file with a header(.csv)
- Provide an optional description: nseComp.1.csv



Step 2: Visualize the Dataset in Azure ML

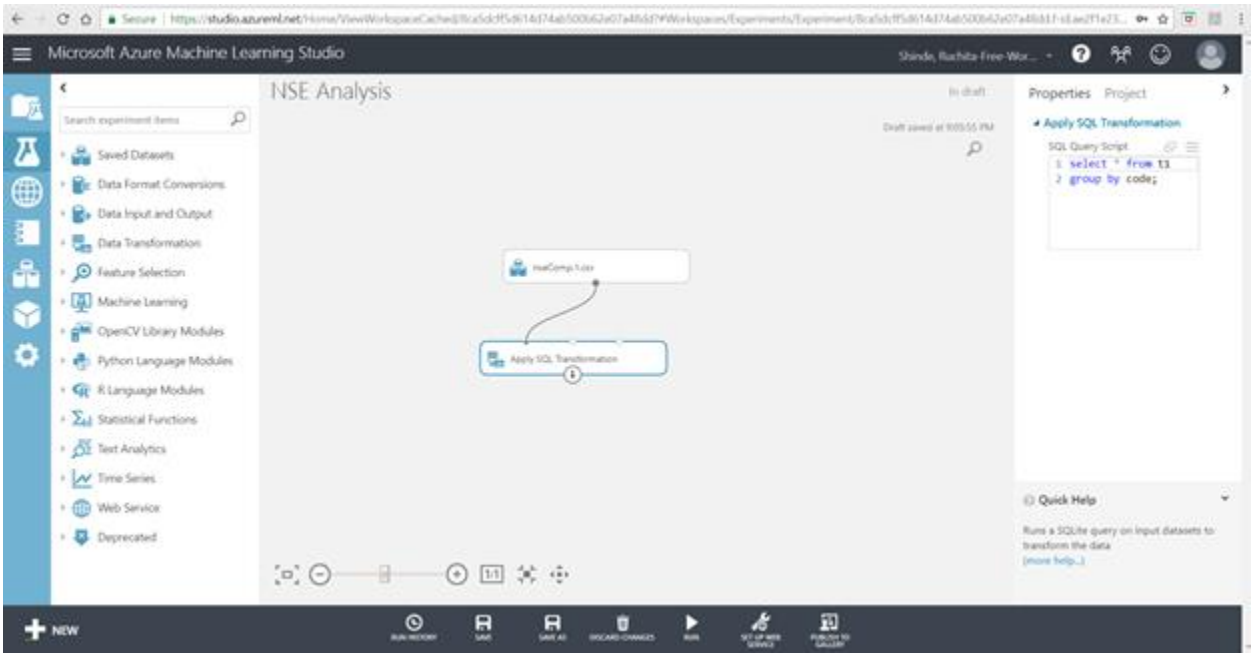
This step verifies that the data set is uploaded and contains all the data from the source file.



Step 2: Query the table

Add Apply SQL Transformation module to the experiment and query the dataset .

*select * from t1 group by code;*

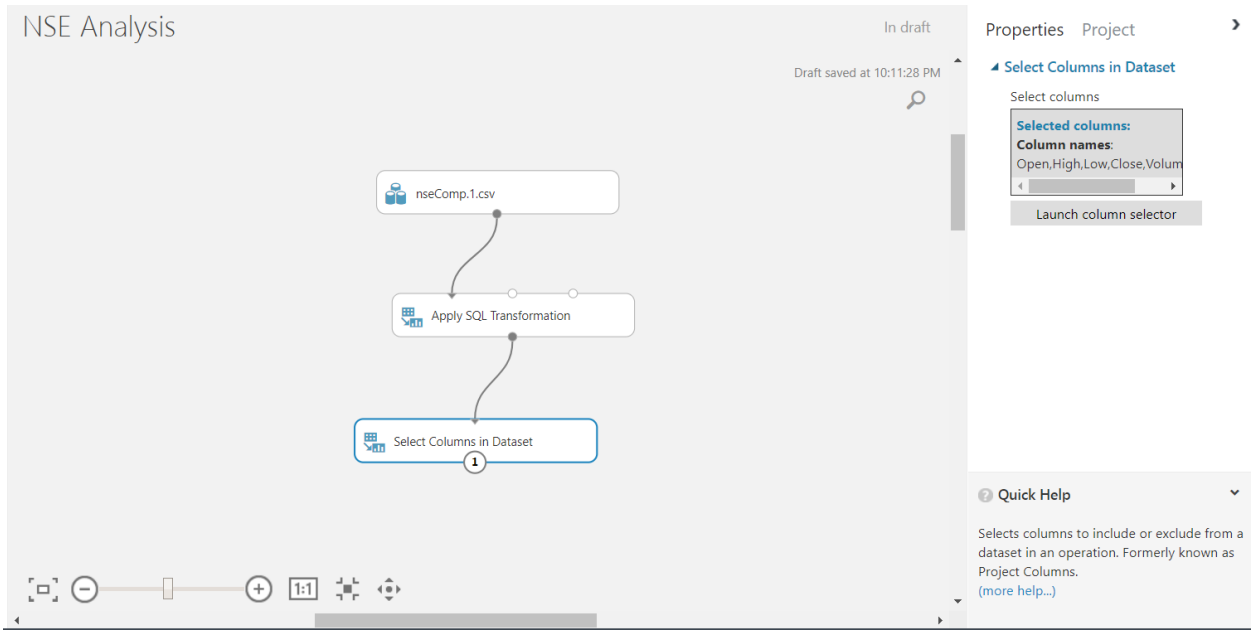


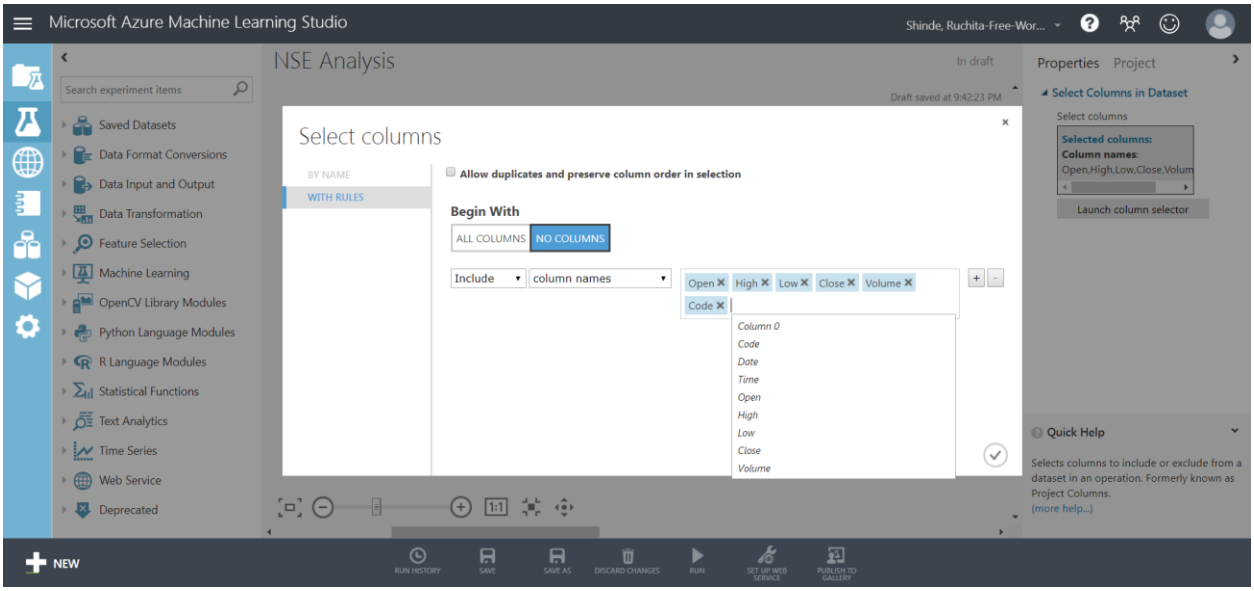
Create a New Model

Step 4: Select required Columns

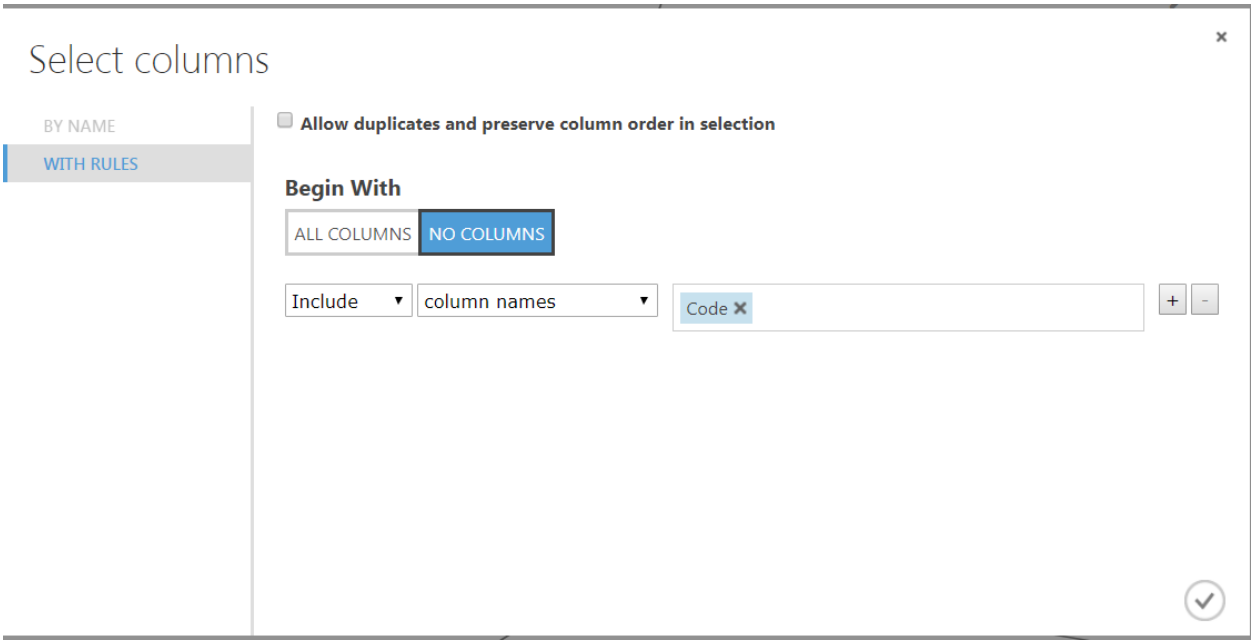
- Search for the Select Columns in Dataset (Project Columns) module and drag it onto your canvas. Connect the Results Dataset output of the Apply SQL Transformation to the dataset input of Select Columns in Dataset module.
- With the Select Columns in Dataset (Project Columns) module selected, in the properties pane, launch the column selector, and include the following columns:
    - Open
    - Close
    - High
    - Low
    - Volume
    - Code

The canvas should look as follows for the above steps.

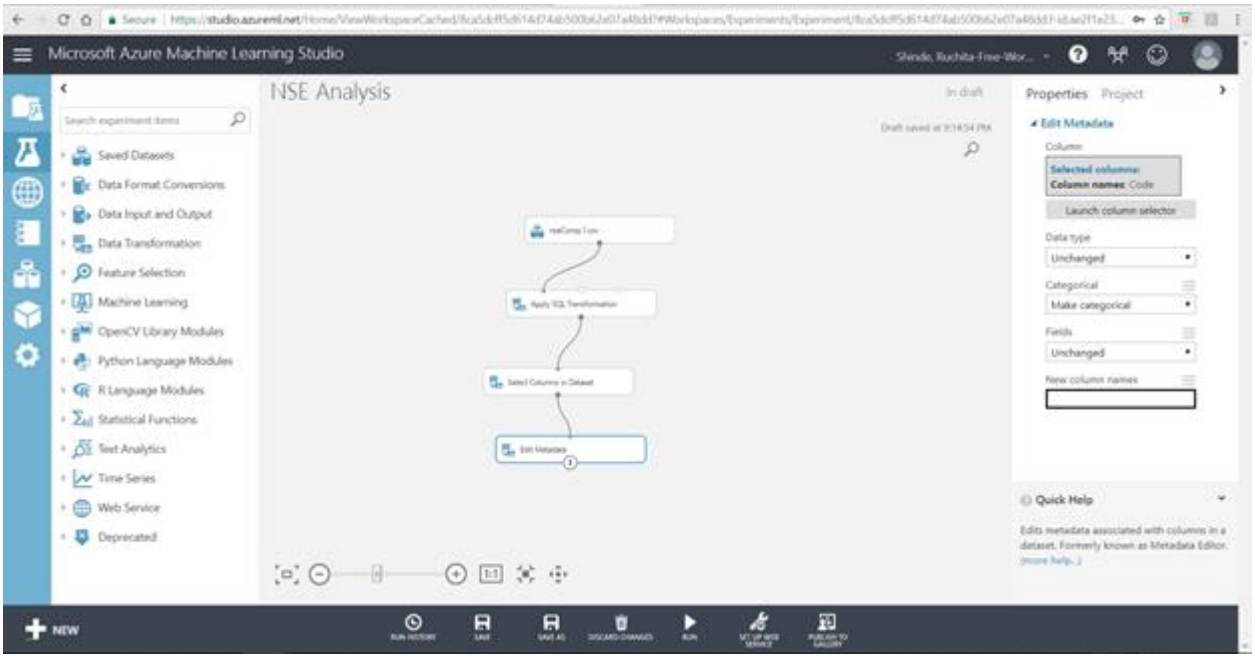




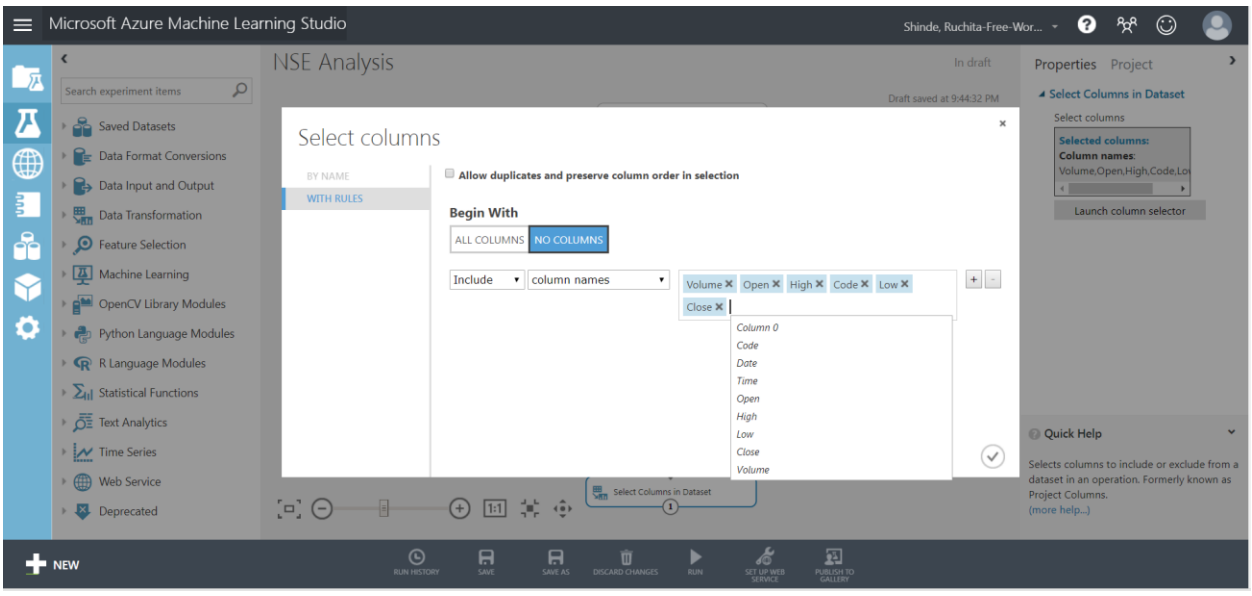
- Search for the Edit Metadata (Metadata Editor) and drag it onto the canvas. Connect the output of the Select Columns in Dataset (Project Columns) to the input of the Edit Metadata (Metadata Editor).
- Click the Edit Metadata (Metadata Editor) and in the properties pane, launch the Column Selector. Select as below:



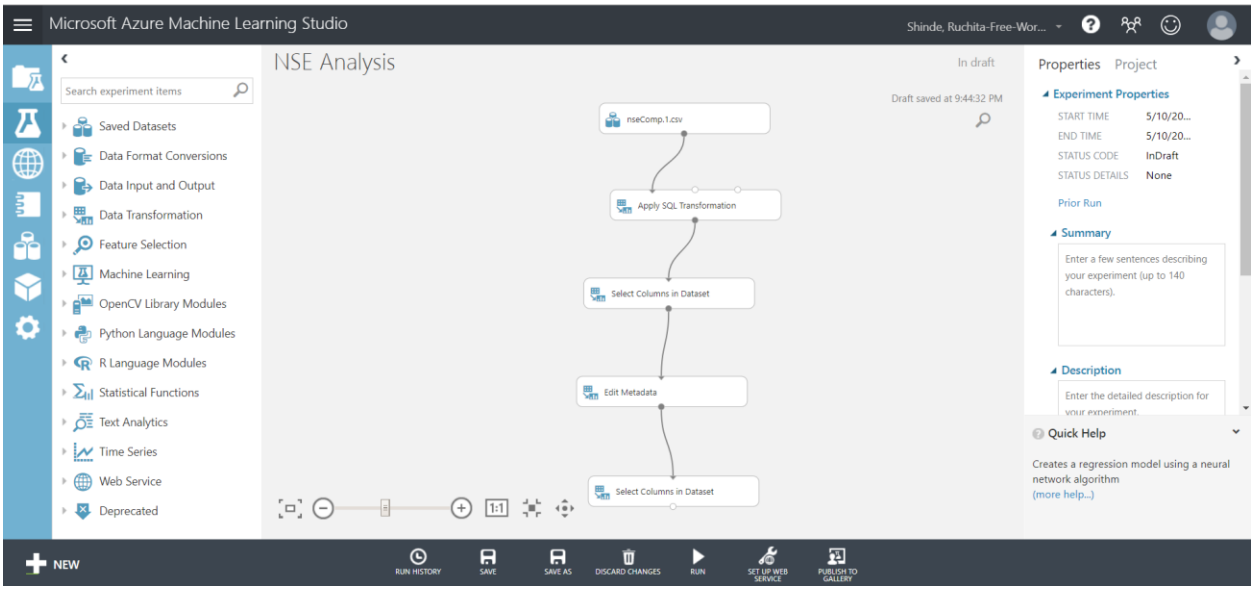
- In the Categorical drop down list, select Make Categorical.



- Search again for the Select Columns in Dataset (Project Columns) module and drag it onto your canvas. Connect the Results Dataset output of Edit Metadata module to the dataset input port of Select Columns in Dataset module.
- With the Select Columns in Dataset (Project Columns) module selected, in the properties pane, launch the column selector, and include the following columns:
    - Open
    - Close
    - High
    - Low
    - Volume
    - Code



The canvas should look as follows for the above steps.

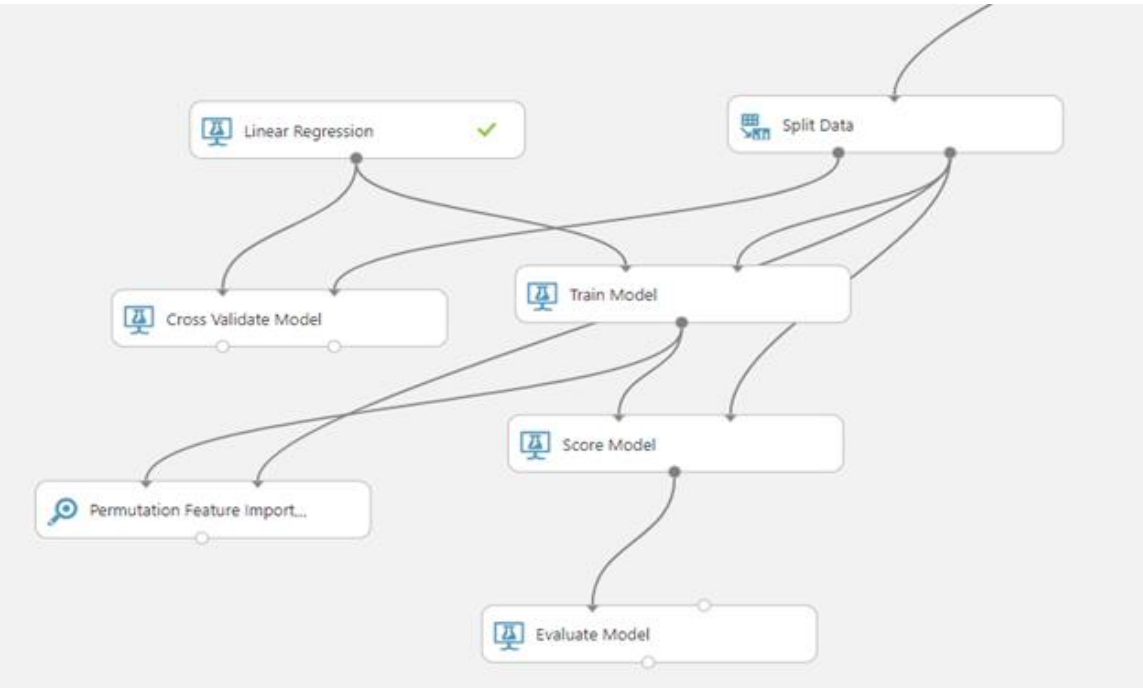


Module 1: Linear Regression

- Search for the Split Data (Split) module. Drag this module onto your experiment canvas. Connect the Results dataset output port of the Select Columns in Dataset (Project Columns) module to the Dataset input port of the Split Data (Split) module. Set the Properties of the Split Data (Split) module as follows:
    - Splitting mode: Split Rows
    - Fraction of rows in the first output: 0.5
    - Randomized Split: Unchecked
    - Random seed: 3456
    - Stratified Split: False

- Search for the Linear Regression module. Make sure you have selected the regression model version of this algorithm. Drag this module onto the canvas. Set the Properties if this module as follows:
    - Solution method : Online Gradient Descent
    - Create trainer mode : Single parameter
    - Learning rate : 0.1
    - Number of training epochs : 10
    - L2 regularization weight : 0.001
    - Normalize features , Average final hypothesis is average, Decrease learning rate as iterations progress : Checked
    - Random number seed : 3456

- Added the Train Model module, Score module and Cross validation module to the canvas.
- In cross validator module selected column Volume in the column selector with Random seed as 3456.
- Added Permutation Feature Importance module on the model with Selected columns Volume and Metric for measuring performance : Regression - Mean absolute error.
- The columns are connected as shown in the screenshot below.



Result of 1st model is as below:

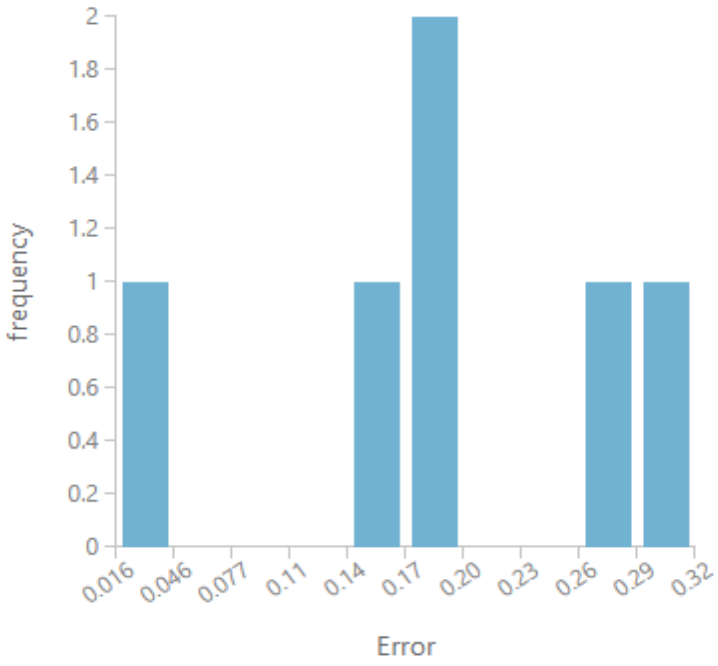
RMSE value is 0.207358.
The Coefficient of definition for Model 1 is 0.967752.

NSE Analysis › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	0.184494
Root Mean Squared Error	0.207358
Relative Absolute Error	0.184494
Relative Squared Error	0.032248
Coefficient of Determination	0.967752

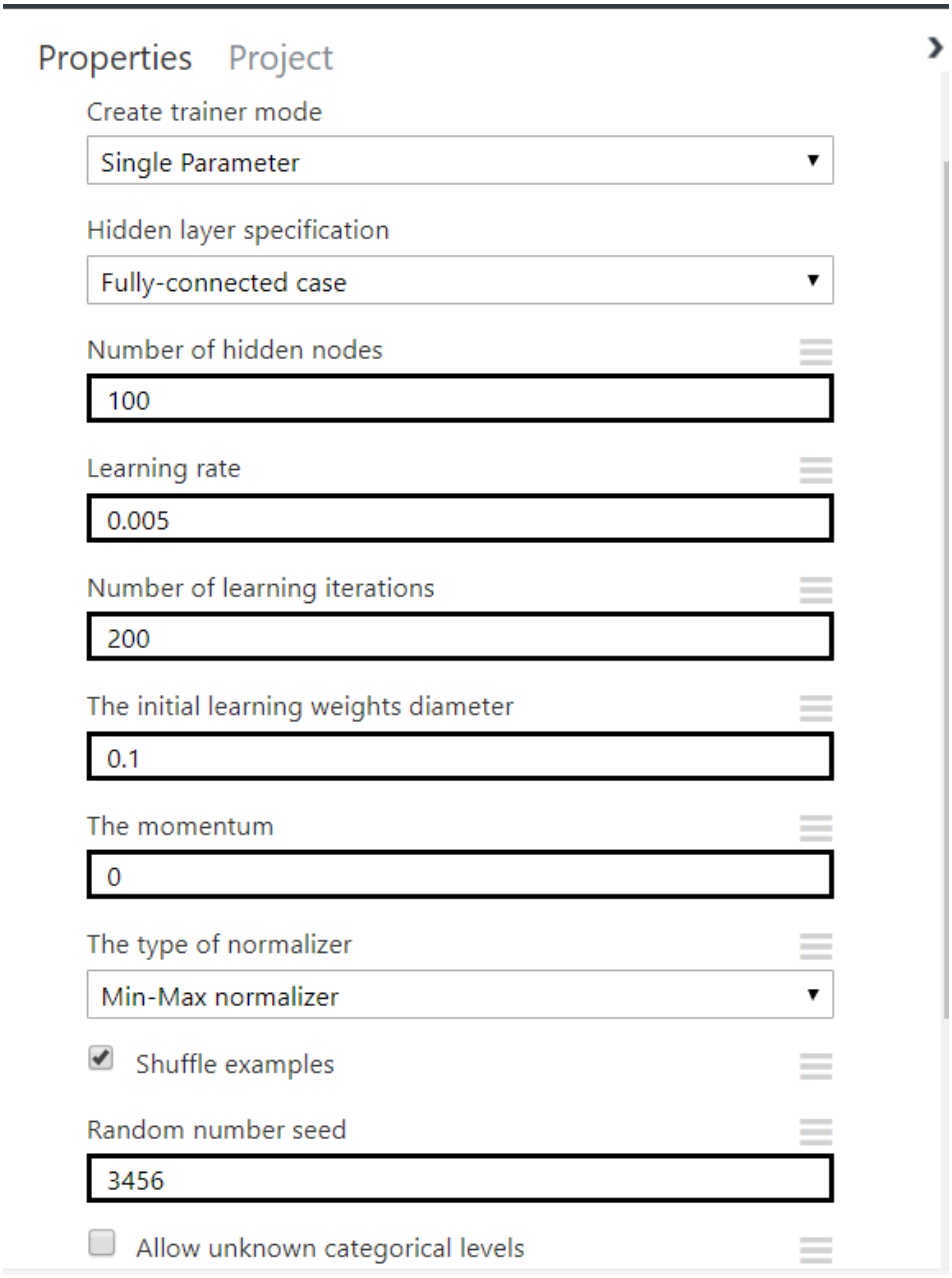
Error Histogram



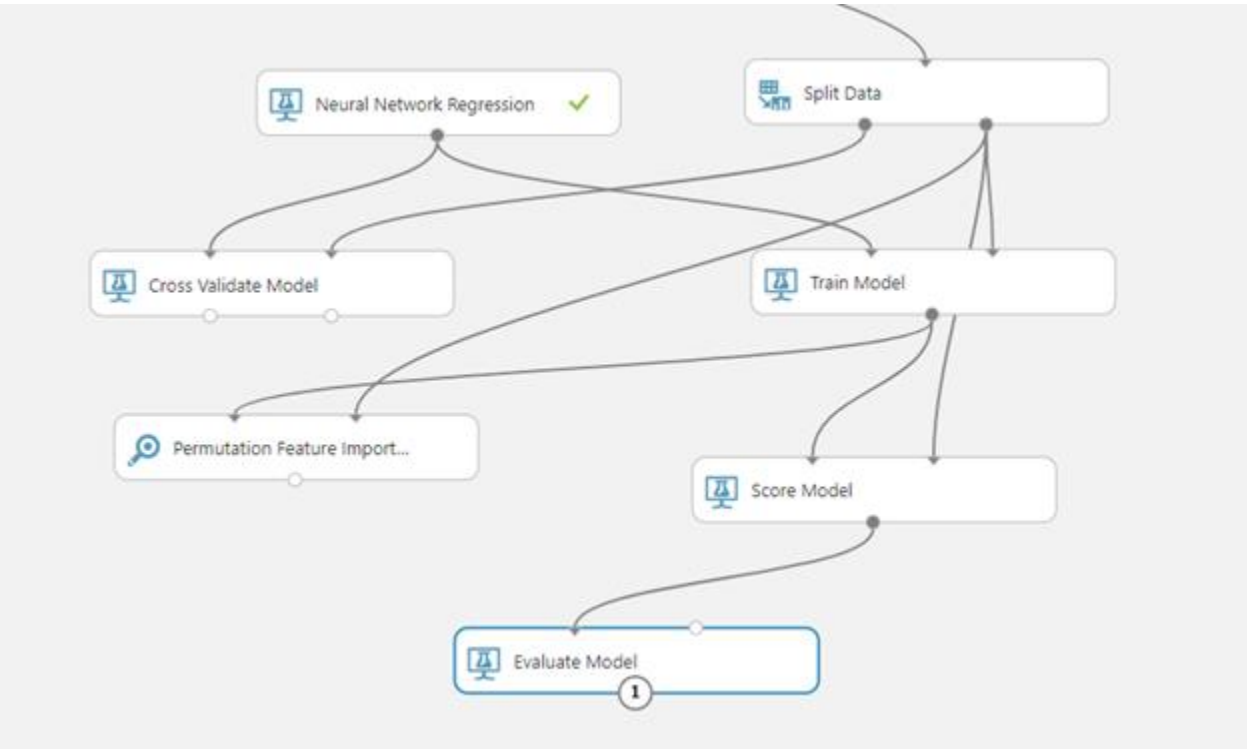
Module 2: Neural Network

Neural Network regression is used in the second model instead of Linear regression model to find the COD and rmse value.

Below properties are changed for Neural Network Regression model as per our dataset to get a good RMSE value.



- The columns are connected as shown in the screenshot below.



Result of 2nd Model is as follows:

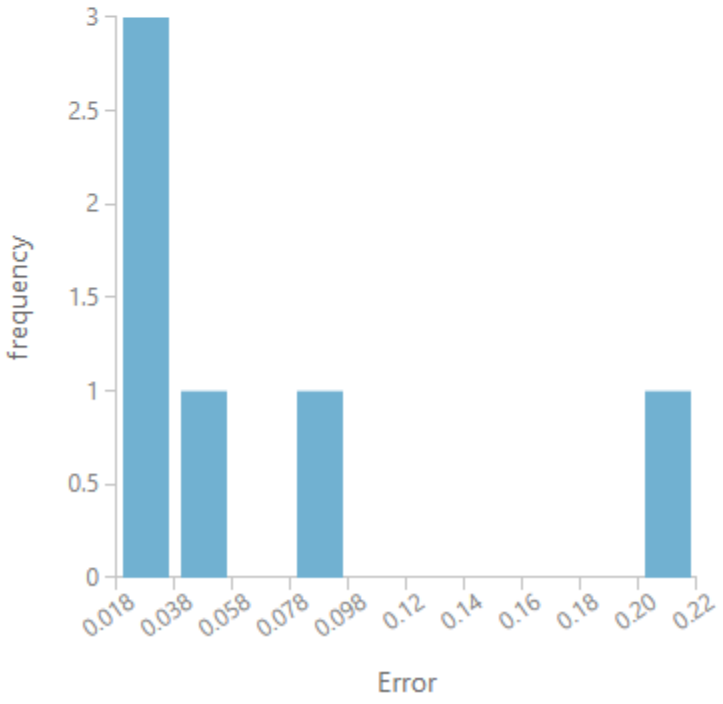
RMSE Value: 0.098181
Coefficient of determination: 0.99277

NSE Analysis › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	0.06938
Root Mean Squared Error	0.098181
Relative Absolute Error	0.06938
Relative Squared Error	0.00723
Coefficient of Determination	0.99277

Error Histogram



References:

- https://gallery.cortanaintelligence.com/Experiment/NSE-Analysis-2
- https://www.kaggle.com/ramamet4/nse-company-stocks/data