CIS5560 Term Project Tutorial

Authors: Shanmathi Arul Murugan, Ashwin.P.Karthik, Kaushik Sridharan

Instructor: Jongwook Woo

Date: 05/19/2018

Lab Tutorial

Shanmathi Arul Murugan(sarulmu@calstatela.edu), Ashwin.P.Karthik(akarthi@calstatela.edu)

Kaushik Sridharan(ksridha@calstatela.edu)

05/19/2018

Predictive Analysis of salary For Different Job Titles

Objectives

The aim of this tutorial is to predict the Salary of different job titles based on the available features from the available features from the dataset by utilizing Machine learning Algorithms and build accurate models using AzureML:

- Get the data set from the links
- Create Azure ML models
- Use appropriate Machine Learning algorithms:

1. **Regression**: Boosted Decision Tree Regression, Linear Regression

- Use Python scripts
- Choose best model based on appropriate metrics- RMSE, Accuracy, Precision, Recall.

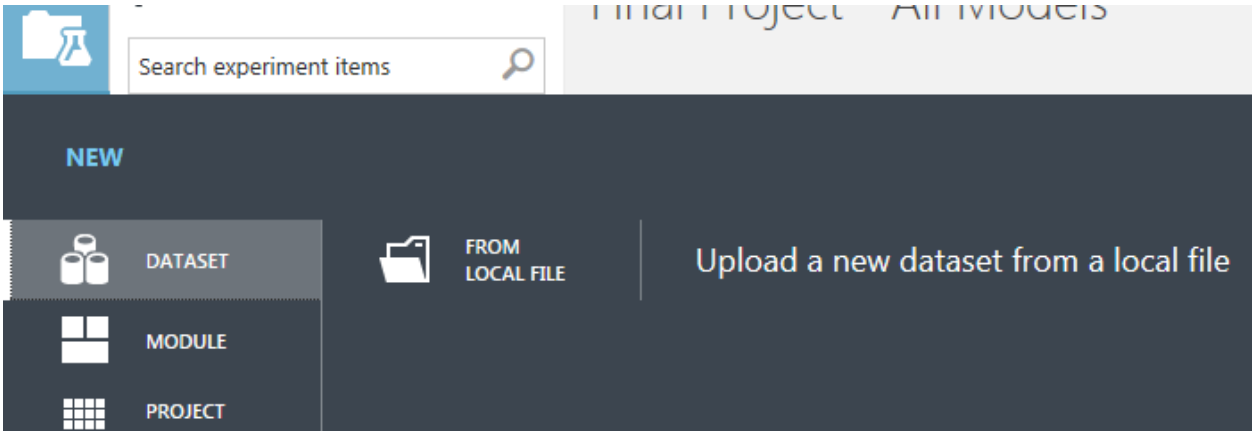
What You'll Need

To complete this lab, you will need the following:

- An Azure ML account
- A web browser and Internet connection
- The lab files for this lab
- Python Anaconda – our classroom should have a python editor
- The "NYC Jobs" dataset (see *Prepare the Data* steps below)

Task 1: Prepare and Upload Data

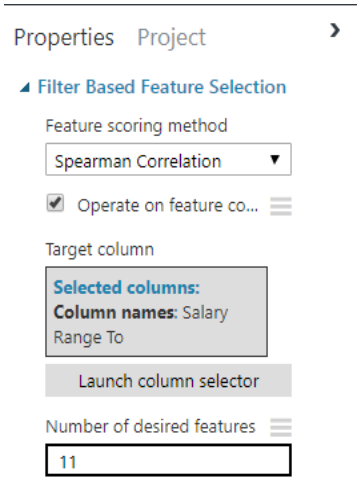
1. Open a browser and browse to https://studio.azureml.net .Then sign in using the Microsoft account associated with your Azure ML account.
2. Download the data from the Below Three Links:
   https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2
3. Save the data on your computer as CSV and name the file.
4. Once downloaded, Open the files in Excel and Select all the Columns containing Salary Figures with a "$" sign. To remove the $ symbol, right click→ Format cells→Change Format type to Number.
5. At Azure ML studio you need to upload the input csv files as follows:
   New→Datasets→From Local File



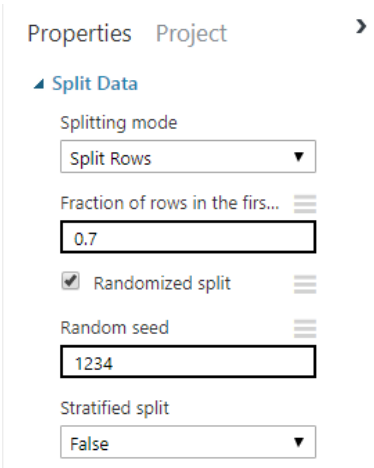
Task 2: Model Building for Prediction of Different Job Title's Salary

1. Create an Experiment and name it CIS5560Final. Go to My Datasets from the Saved Datasets and search for the "NYC Jobs" dataset you uploaded. Drag it on the Canvas.
2. Next, drag the Select Columns in Dataset module to the canvas and click Launch column selector and select the following columns
   - **Job ID**
   - **Business Title**
   - **Full-Time Part-Time Indicator**
   - **Salary Range To**
   - **Hours/Shift**
   - **Posting Type**
   - **Level**
   - **Job Category**
   - **Work Location**
   - **Agency**
   - **Title Code No**
   - **Salary Frequency**
   - **Division/Work Unit**

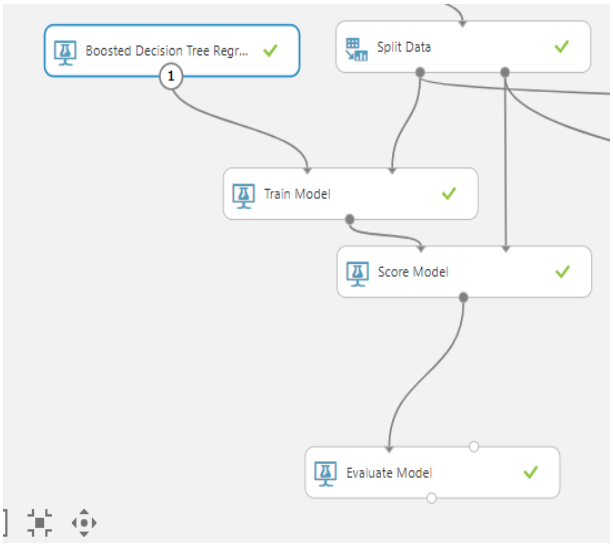
3. Thereafter, drag the **Filter Based Feature Selection** onto the canvas and set the following parameters for the same.

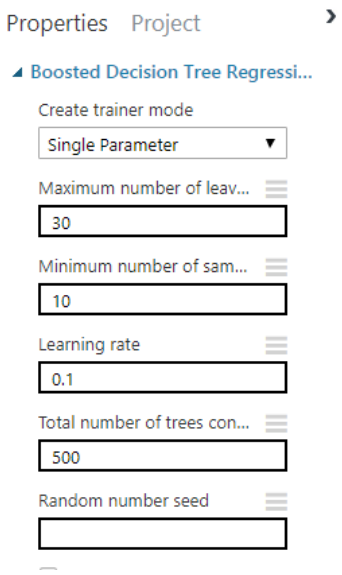


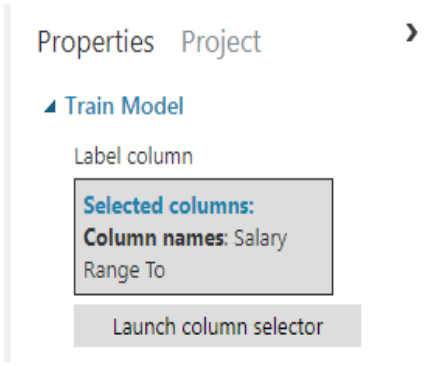
4. Choose the Split Data module and drag it to the canvas and set the following Parameters:



5. Now, choose the **Boosted Decision Tree Regression** Algorithm, Train Model and Test Model onto the canvas and make all the parameters and connections as shown below:

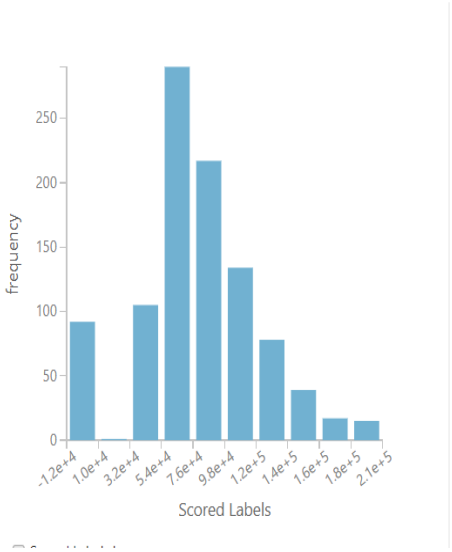






6. Right click and Visualize the Scored dataset output port. The Scored Labels are the predicted values of our test. You should see a similar result as below:

Title Code No	Level	Job Category	Full-Time/Part-Time indicator	Salary Frequency	Work Location	Scored Labels
10050	M1	Technology, Data & Innovation	Full-Time	Annual	1 Centre St., N.Y.	118843.390625
10050	M1	Information Technology & Telecommunications	Full-Time	Annual	59 Maiden Lane	120903.632813
30080	2	Legal Affairs Social Services	Full-Time	Annual	4 World Trade Center	84418.882813
10095	M5	Finance, Accounting, &	Full-Time	Annual	123 William Street, NY,	143694.765625

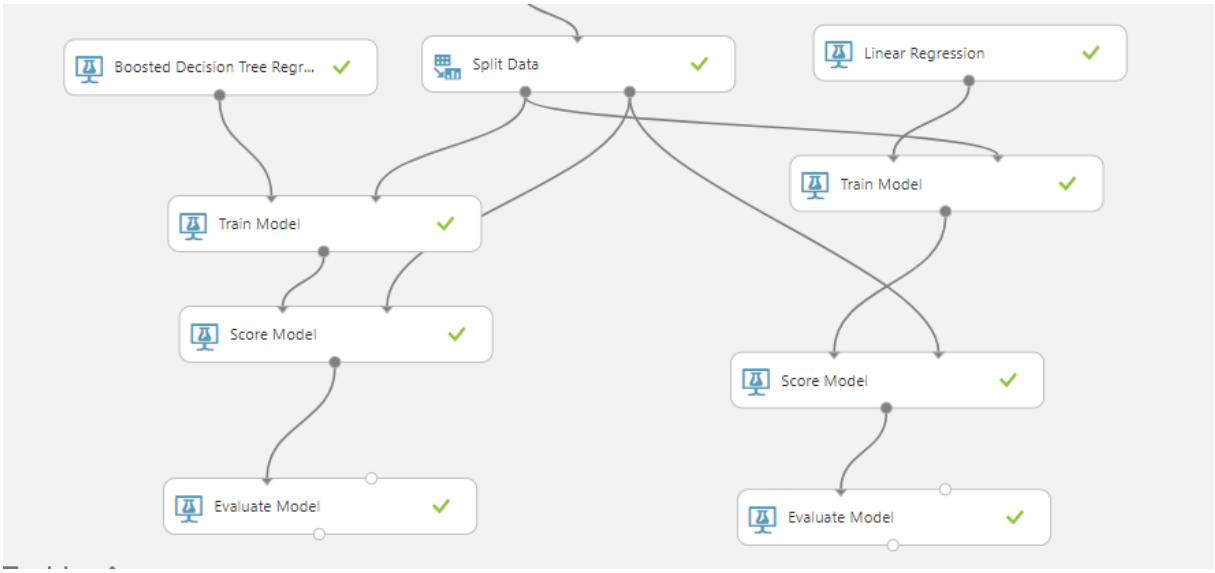


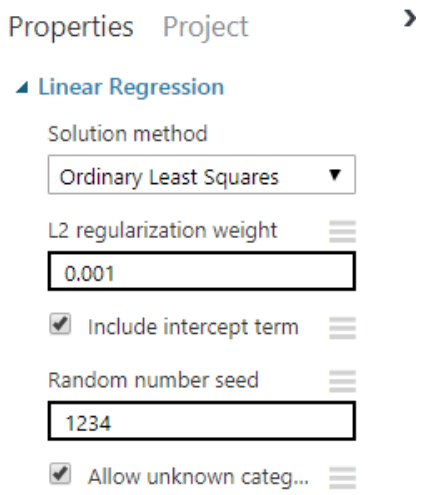
7. At the end, Save and Run the Experiment. Visualize the output of the bottom most Evaluate model and you can see that the **RMSE** for the **Boosted Decision Tree Regression** model and the **Coefficient of Determination**.

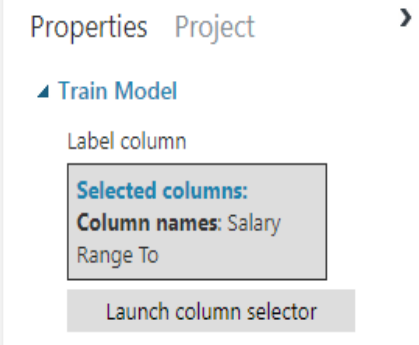
Metrics

Mean Absolute Error	7675.41378
Root Mean Squared Error	12436.852182
Relative Absolute Error	0.243189
Relative Squared Error	0.085812
Coefficient of Determination	0.914188

8. Copy the **Train Model, Score Model, Evaluate Model** and paste it. Now connect the results dataset1 of the split data to the dataset of the train model, then trained model of the train model to the trained model of the score model. Then connect the result dataset2 to the data set of the scored model and then finally connect scored model to evaluate model. Then select **Linear Regression Algorithm** and connect it to train model. Which should look like the following image:

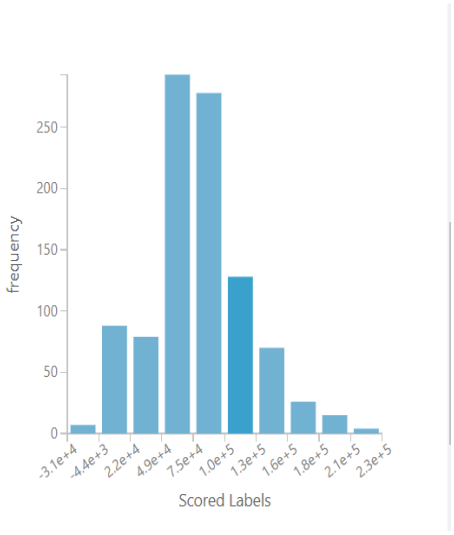






9. Right click and Visualize the Scored dataset output port. The Scored Labels are the predicted values of our test. You should see a similar result as below:

Title Code No	Level	Job Category	Full-Time/Part-Time indicator	Salary Frequency	Work Location	Scored Labels
10050	M1	Technology, Data & Innovation	Full-Time	Annual	1 Centre St., N.Y.	120065.199796
10050	M1	Information Technology & Telecommunications	Full-Time	Annual	59 Maiden Lane	125069.056275
30080	2	Legal Affairs Social Services	Full-Time	Annual	4 World Trade Center	58873.265578
10095	M5	Finance, Accounting, &	Full-Time	Annual	123 William Street, NY,	143972.58467



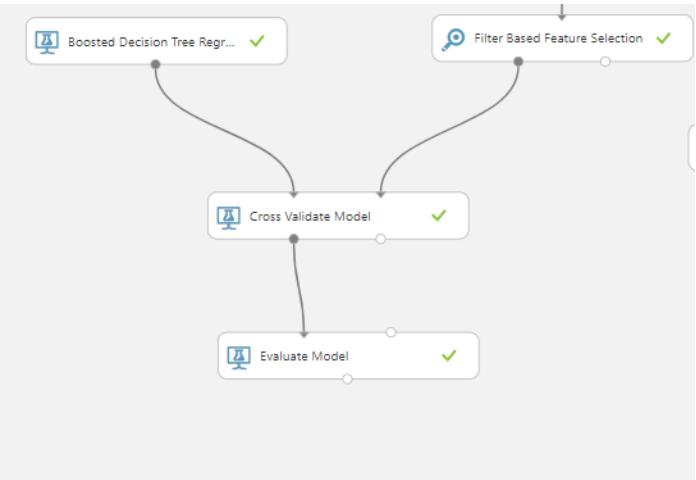
10. Now, Save and Run the Experiment. Visualize the output of the bottom most Evaluate model and you can see that the **RMSE** for the **Linear Regression** model and the **Coefficient of Determination**.

Metrics

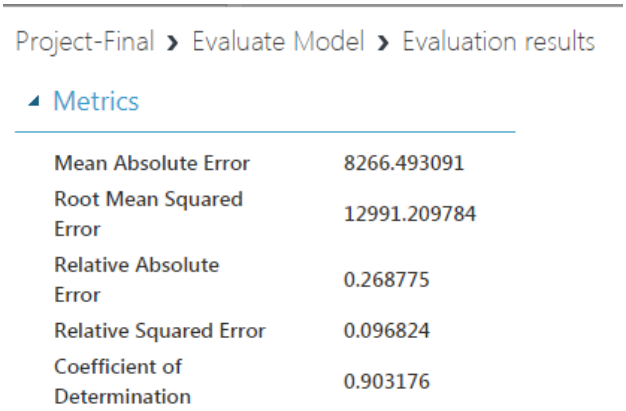
Mean Absolute Error	4394.189922
Root Mean Squared Error	9961.10068
Relative Absolute Error	0.139226
Relative Squared Error	0.055048
Coefficient of Determination	0.944952

Tune Model Parameters using Cross Validator (Optional)

1. Do the same steps from 1 to 3 in Task 2.
2. Then drag the cross validate model and connect the filtered data set output of filter-based feature selection to the cross validate model and then drag boosted decision tree regression algorithm to the cross validate model.
3. Now drag the evaluate model and connect the cross validate model output to it.
4. If you have done all the steps, then it will appear something like this on the image given below.



5. Now click on the Evaluate model and visualize the result. Which should provide you with results such as the one given below



References

1. **URL's of Data Sources:**
   https://data.cityofnewyork.us/Housing-Development/DOB-Job-Application-Filings/ic3t-wcy2

2. **GitHub URL:**

3. **Other Reference:** https://studio.azureml.net