

Nam Ho Phan

Northeastern University

Assignment 5

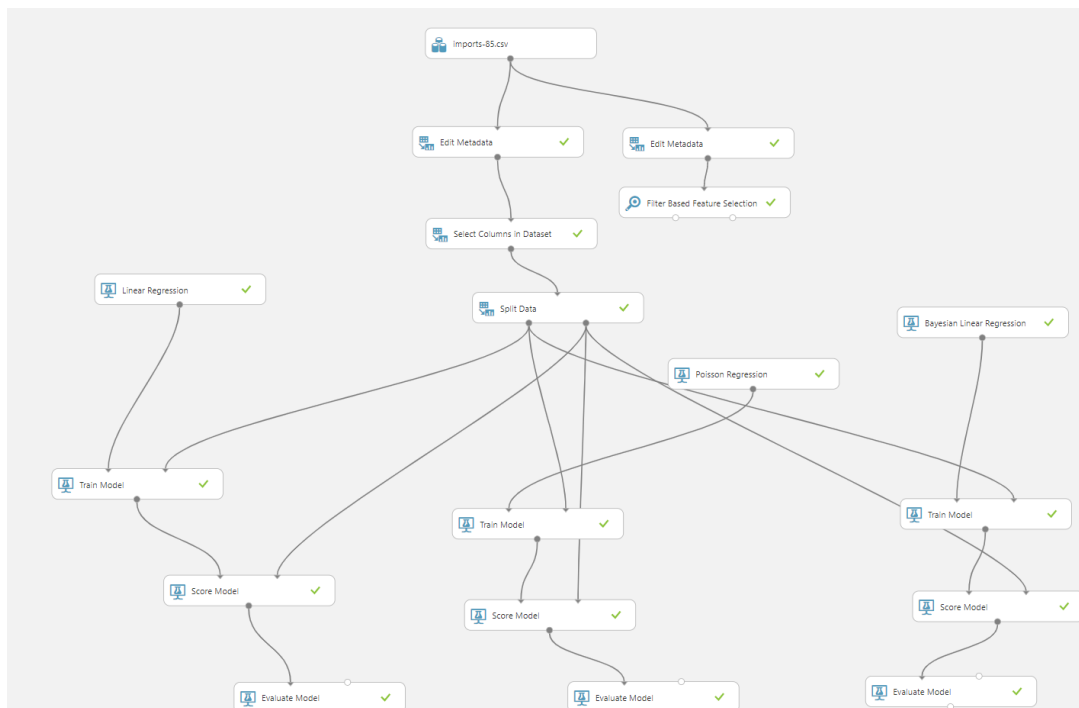
EAI6010

Applications of AI<https://northeastern.blackboard.com>

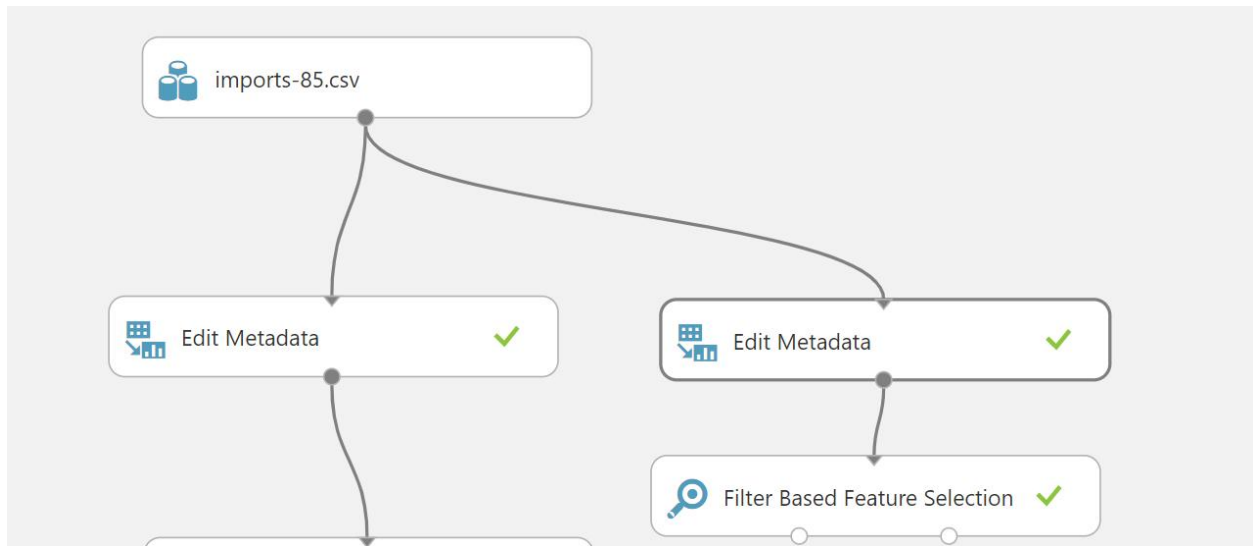
Problem 1:

Azure Machine Learning Studio is very useful for data scientists because it can integrate the cloud and model building for real-time analysis. This platform will help beginners easily understand the pipeline of building predictive model, and it saves time for experts as they do not have to spend time coding.

For this data set, I will decide to use Linear Regression, Poisson Regression and Bayesian Linear Regression because it is suitable for continuous target. Before I go into details of model, I show the overview of my pipeline as below:



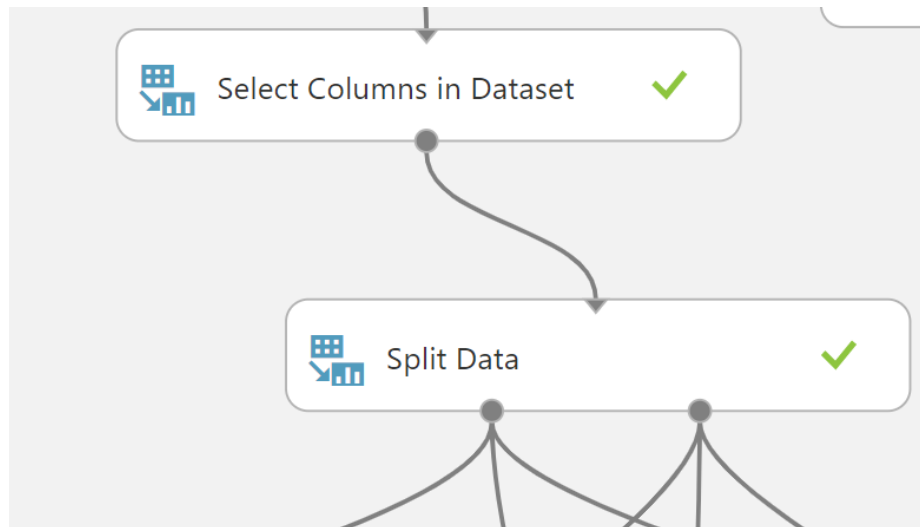
Firstly, we must convert the string data into categorical data type using ‘Edit Metadata’ because we will need this type of data for preprocessing. Then, we will use ‘Filter Based Feature Selection’ to find the high correlation between label and feature.



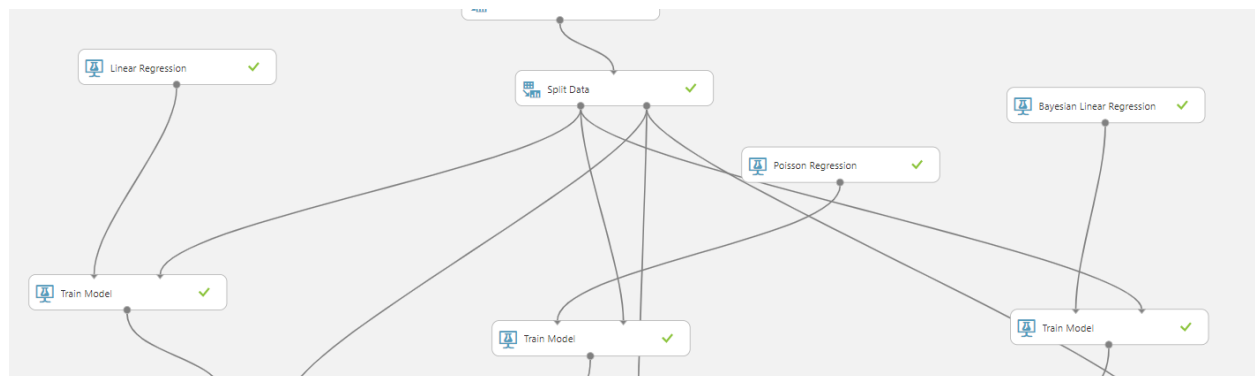
As you can see, the highest correlation with price is make, engine-size, curb-weight, horsepower.

	price	make	engine-size	curb-weight	horsepower
view as					
	1	0.882438	0.872335	0.834415	0.810533

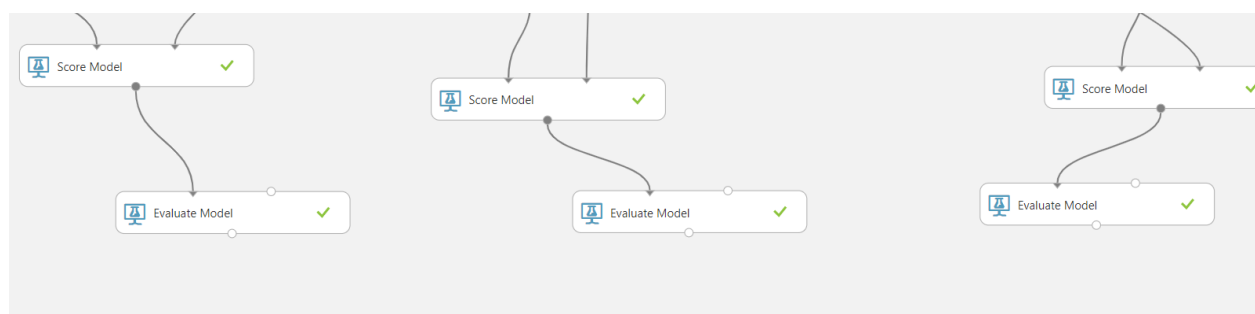
After that, we can use “Select Columns in Dataset” to select top 4 correlated variables as requirement. To split data into train and test set, we can use ‘Split Data’ with 70/30 as it will validate the accuracy of our result.



After splitting data, I create Train Model to connect the model and the train data.



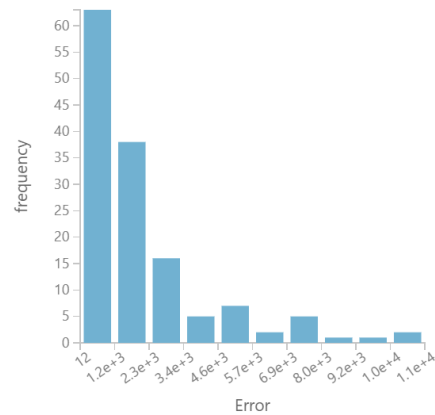
Finally, the 'Evaluate Model' will help me to evaluate the result of these 3 models, so I can compare it to make the final decision.



- **Linear Regression:**

Mean Absolute Error	2041.623918
Root Mean Squared Error	3019.82511
Relative Absolute Error	0.327722
Relative Squared Error	0.139189
Coefficient of Determination	0.860811

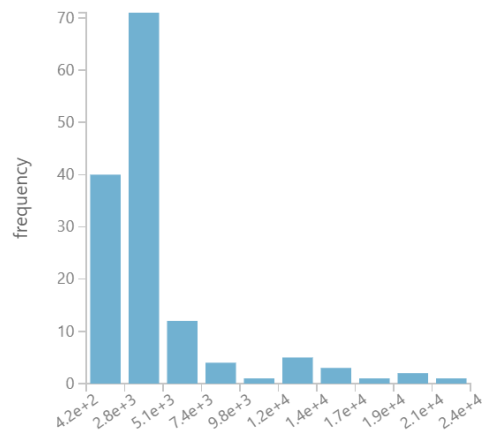
Error Histogram











- Poisson Regression:**

Mean Absolute Error	4604.059242
Root Mean Squared Error	6180.362756
Relative Absolute Error	0.739045
Relative Squared Error	0.583
Coefficient of Determination	0.417

Error Histogram



- Bayesian Linear Regression:**

	Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
view as  						
	1311.116423	2148.467394	2896.227614	0.344872	0.128028	0.871972

- **Result:**

The Mean Absolute Error represents to the difference between the actual values and predicted values. Therefore, the higher the mean absolute error, the more the errors our model will possibly have. With around 2041 score of Mean absolute error, the Linear Regression Model would be expected to apply into system because it shows the high accuracy of model when it is applied on new instances.