[](http://www.calstatela.edu/centers/hipic) CIS5560 Term Project Tutorial

**Authors:**[**Ruchi Singh**](https://www.linkedin.com/in/ruchi-singh-68015945/)**; Kulwinder Kaur; Ovanes Mikaelian**

**Instructor:**[**Jongwook Woo**](https://www.linkedin.com/in/jongwook-woo-7081a85)

**Date: 05/16/2017**

**Lab Tutorial**

Ruchi Singh ([rsingh26@calstatela.edu)](mailto:rsingh26@calstatela.edu))

Kulwinder Kaur ([kkaur7@calstatela.edu)](mailto:kkaur7@calstatela.edu))

Ovanes Mikaelien ([omikael@calstatela.edu](mailto:omikael@calstatela.edu))

05/16/2017

**Applications of Machine Learning Models for Yelp Local Business Data**

**Objectives**

**List what your objectives are.** In this hands-on lab, you will learn how to implement the following machine learning algorithms:

* Match-box Recommender
* Compare Two Class Logistic Regression and Two Class Boosted regression
* K-means Clustering( 3 and 5 clusters)
* Text- Analysis using n-gram and uni-gram

**Platform Spec**

* Microsoft Azure Machine Learning Studio
* # of nodes: 1
* Total Memory Size: 10 GB
* # of modules per experiment: 100

**Four steps to create an experiment using ML studio:**

1. Create a model

* [Step 1: Upload the data](https://azure.microsoft.com/en-us/documentation/articles/machine-learning-create-experiment/#step-1-get-data)
* [Step 2: Preprocess and clean data](https://azure.microsoft.com/en-us/documentation/articles/machine-learning-create-experiment/#step-2-preprocess-data)

1. Train the model

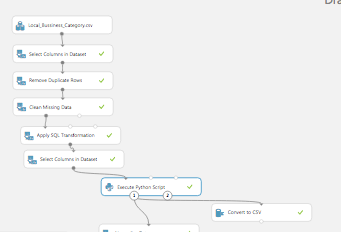
* [Step 3: Choose and apply a learning algorithm](https://azure.microsoft.com/en-us/documentation/articles/machine-learning-create-experiment/#step-4-choose-and-apply-a-learning-algorithm) from the set.

1. Score and test the model

* Step 4: Evaluate the model

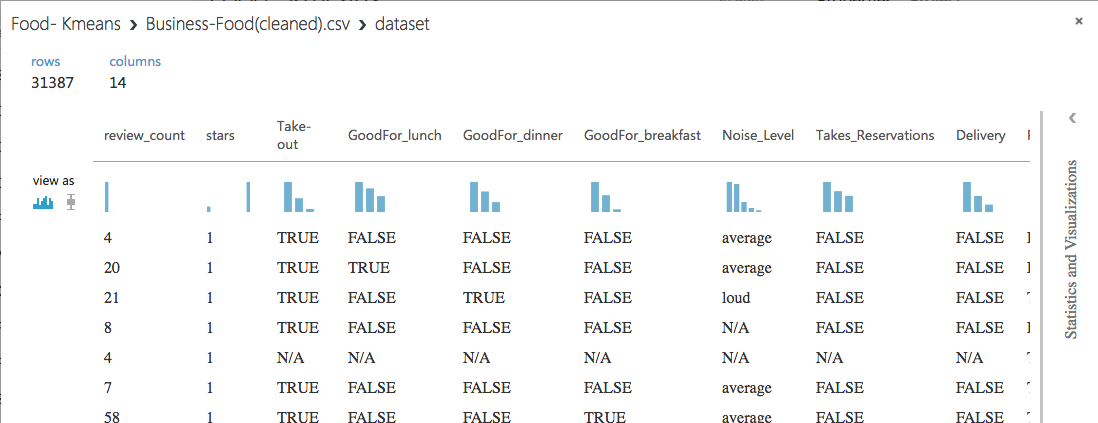
Step 1: Upload data

We saved the clean dataset as CSV from the Classification experiment to be used in the Clustering experiment for Food category.



The saved dataset in Azure is called Business-Food(Cleaned).csv.

1. Start a new experiment by clicking +NEW at the bottom of the Machine Learning Studio window, select EXPERIMENT, and then select Blank Experiment. Select the default experiment name at the top of the canvas and rename it to Food- Kmeans .
2. To the left of the experiment canvas is a palette of datasets and modules. Search for the dataset Business-Food(Cleaned).csv to use for the experiment.
3. Drag the dataset to the experiment canvas.
4. To see what this data looks like, click the output port at the bottom of the Business-Food(Cleaned) dataset, and then select **Visualize**.

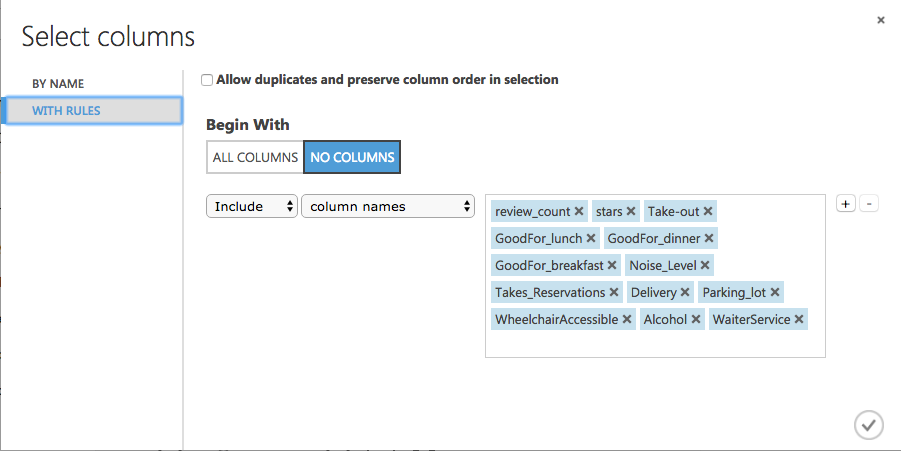


Step 2: Preprocess and Clean the Data

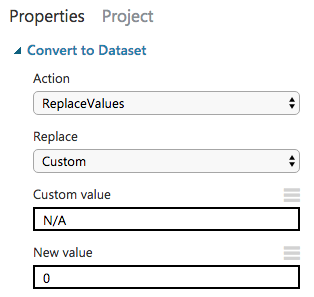
A dataset usually requires some preprocessing before it can be analyzed. You might have noticed the missing values present in the columns of various rows. These missing values need to be cleaned so the model can analyze the data correctly. In our case, we'll remove any rows that have missing values.

First, we'll remove the not required columns, and then we'll remove any row that has missing data.

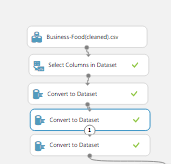
1. Search for project columns in the Search box at the top of the module palette to find the [Project Columns](https://msdn.microsoft.com/library/azure/1ec722fa-b623-4e26-a44e-a50c6d726223/) module, then drag it to the experiment canvas and connect it to the output port of the Yelp (Raw) dataset. This module allows us to select which columns of data we want to include or exclude in the model.
2. Select the [Project Columns](https://msdn.microsoft.com/library/azure/1ec722fa-b623-4e26-a44e-a50c6d726223/) module and click Launch column selector in the Properties pane.
3. Make sure all the relevant columns to Food Category are selected in the filter drop-down list, Begin With. This directs [Project Columns](https://msdn.microsoft.com/library/azure/1ec722fa-b623-4e26-a44e-a50c6d726223/) to pass through all the columns [review\_count,stars,Take-out,GoodFor\_lunch,GoodFor\_dinner,GoodFor\_breakfast,Noise\_Level,Takes\_Reservations,Delivery,Parking\_lot,WheelchairAccessible,Alcohol,WaiterService].



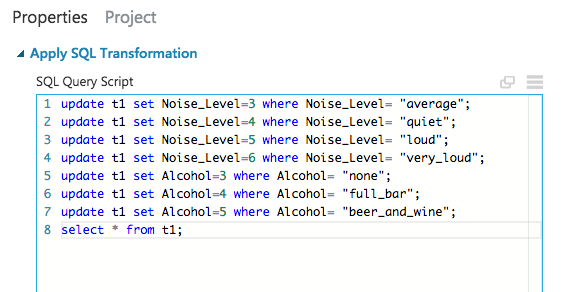
1. Click the check mark (OK) button to close the column selector. The properties pane for Project Columns shows that it will pass through all the selected columns from the dataset.
2. Next select the Convert to Dataset module to replace the N/A values to 0.



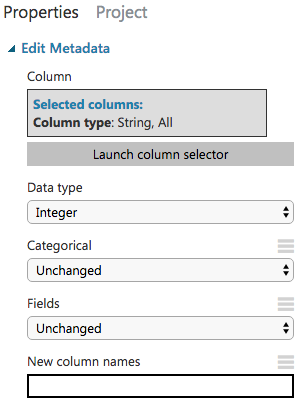
1. Add two more such modules to convert False to 1 and True to 2. After that the Experiment should look like below.



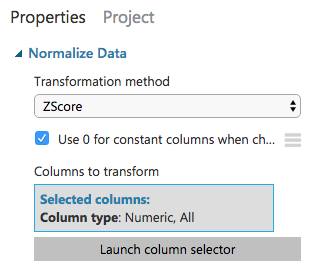
1. Next, we use an SQL Transformation module for writing an SQL query script. This query should update the various values for Noise level and Alcohol to numeric values.



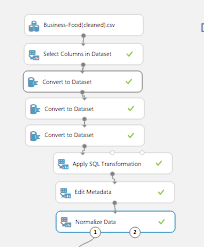
1. Add edit metadata to change all String type columns to Category.



1. Now we need to normalize all the numeric columns using the Normalize Data module.



1. After the completion of all the data cleaning and processing steps the experiment should look like below.



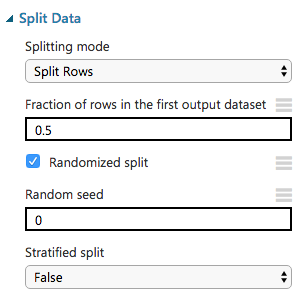
Step 3: Choose and apply machine learning algorithms:

Now that the data is ready, constructing two clustering model based on the intuitive grouping of the data. I am using data to train the model and then evaluate the model to understand which is better and why.

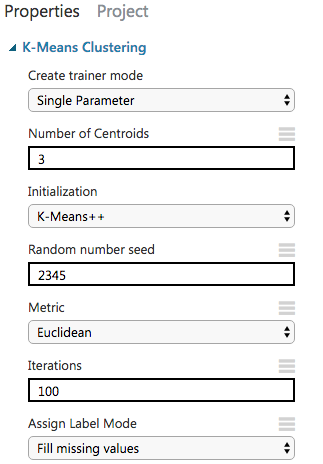
***Clustering*** is an un-supervised machine learning technique. Using the k-means clustering model we shall be able to classify the food business based on their attributes. It can help the business owners understand why a food business has low or high review count. It can also suggest the investor or new business owners, the combination of food business attributes that could lead to increase in review count.

Every column in our clean dataset is categories, thus there is no need for cleaning the outliers from the dataset. We shall use 3 cluster and 5 cluster K-means algorithms to compare which is better algorithm for our dataset.

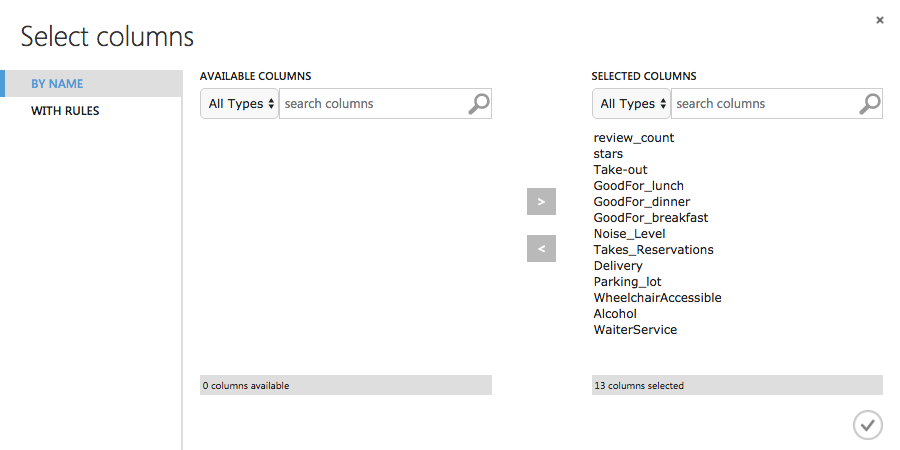
1. Run the experiment. This allows the Clean and processed data and then [Split Data](https://msdn.microsoft.com/library/azure/70530644-c97a-4ab6-85f7-88bf30a8be5f/) modules to pass column definitions to the modules we'll be adding next. We split the data to 0.5 to that we can used 50% of the data for 3 cluster K-means and 50% for 5 cluster K means.



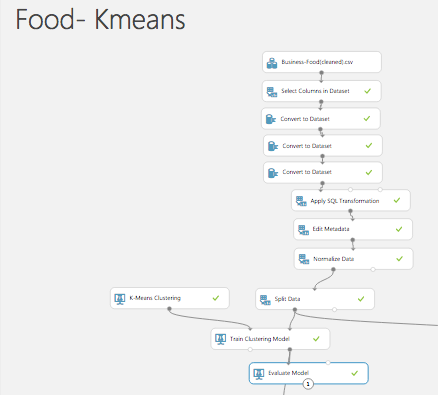
1. To select the learning algorithm, expand the Machine Learning category in the module palette to the left of the canvas, and then expand Initialize Model. This displays several categories of modules that can be used to initialize machine learning algorithms. For this experiment, select the K-Means Clustering module under the Clustering category and drag it to the experiment canvas. Configure it as below.



1. Find and drag the Train Clustering Model module to the experiment canvas. Connect the left input port to the output of the K-Means Clustering module. Connect the right input port to the training data output (left port) of the [Split Data](https://msdn.microsoft.com/library/azure/70530644-c97a-4ab6-85f7-88bf30a8be5f/) module.
2. Select the Train Clustering Model module, click Launch column selector in the Properties pane, and then select all the feature column. Based on these columns the model will cluster all the food business.

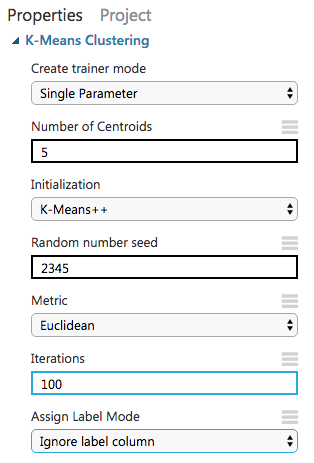


1. Add Evaluate Module to see the evaluation of 3 cluster K-Means Clustering algorithm.
2. Make sure the experiment looks like below and then run the experiment.

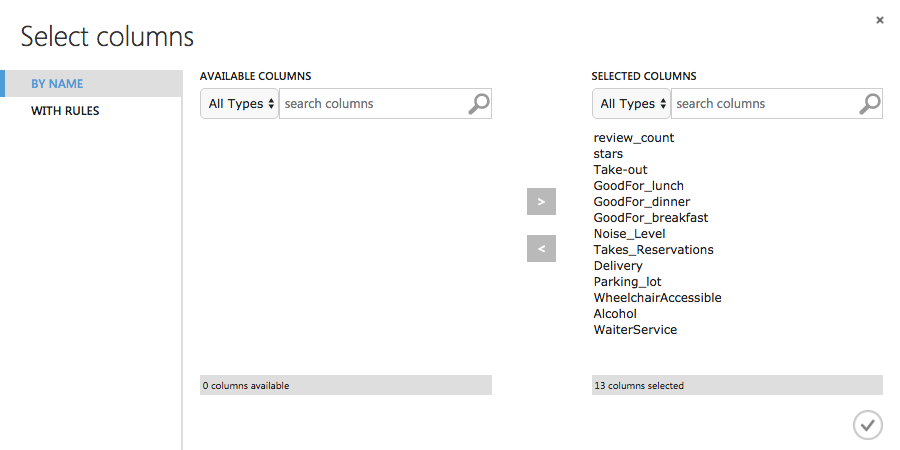


We can try another K-Means Clustering module with 5 cluster and evaluate it to see which model gives a better score. To do so we will need to follow the bellow steps.

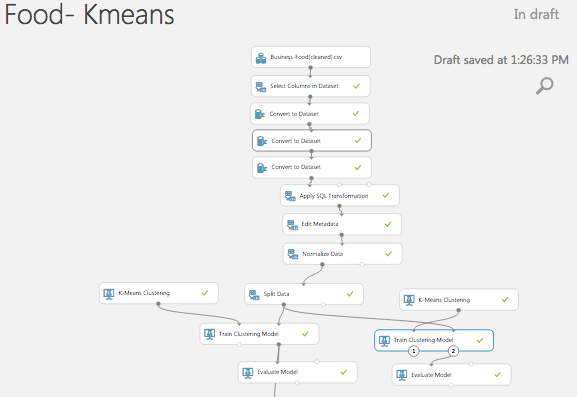
1. Select another the K-Means Clustering module under the Clustering category and drag it to the experiment canvas. Configure it as below.



1. Find and drag the Train Clustering Model module to the experiment canvas. Connect the left input port to the output of the K-Means Clustering module. Connect the right input port to the training data output (left port) of the [Split Data](https://msdn.microsoft.com/library/azure/70530644-c97a-4ab6-85f7-88bf30a8be5f/) module.
2. Select the Train Clustering Model module, click Launch column selector in the Properties pane, and then select all the feature column. Based on these columns the model will cluster all the food business.



1. Add Evaluate Module to see the evaluation of 3 cluster K-Means Clustering algorithm.
2. Make sure the experiment looks like below and then run the experiment.

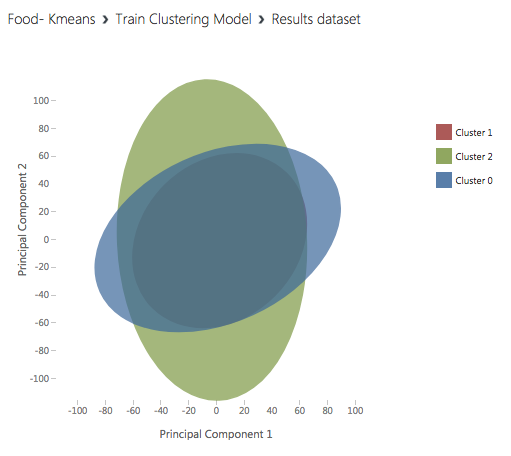


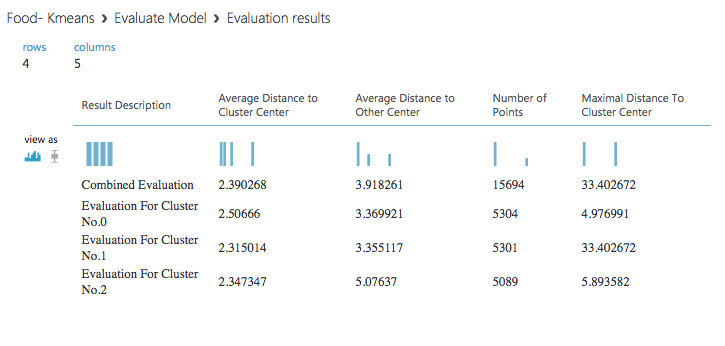
Step 4: Visualization:

To compare the two models, we need to visualize the output of Training module and Evaluate module first for 3 Cluster K-Means Clustering and then for 5 Cluster K-Means Clustering.

3 Cluster K-Means Clustering

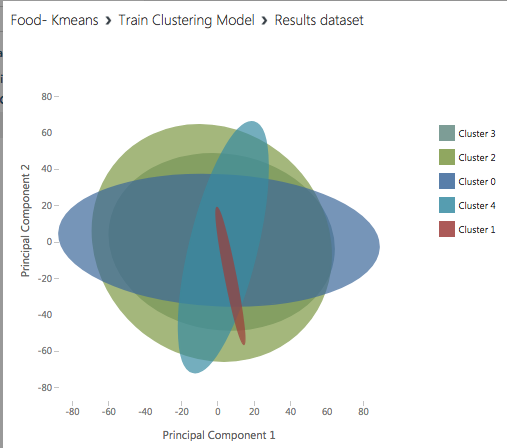
1. The Visualize the right output (result dataset) of the train clustering module. It shows the 3 clusters created by the algorithm. There are two distinct ellipses shown on this projection of the first two principle components. The major axes (long dimension) of each ellipse are in a distinct, nearly orthogonal, indicating the two clusters have good separation. The last cluster is circular in shape and is completely overlapped.

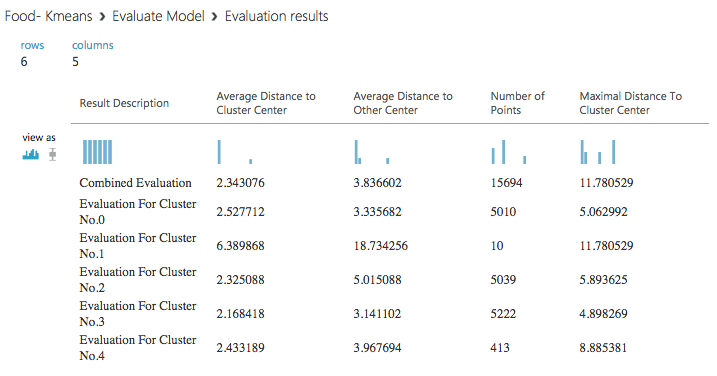


1. Visualize the Evaluate model to of see the final result of our experiment. This shows that there are 3 cluster centers and the Maximum Distance to Cluster Center for cluster. Of all the 3 clusters 33.4 is the maximum distance.

5 Cluster K-Means Clustering

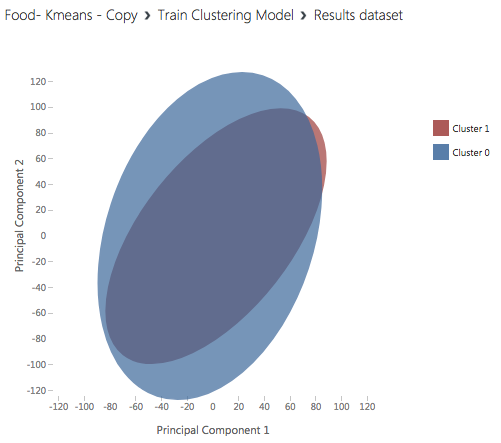
1. The Visualize the right output (result dataset) of the train clustering module. It shows the 3 clusters created by the algorithm. The ellipse for three clusters (Cluster 0, Cluster 1 and Cluster 4) looks about the same. However, the ellipses for the other two clusters (Cluster 2 and Cluster 3) are nearly circular. Further the direction of the major axes of the ellipses for Cluster 2 and Cluster 3 are close to that of Cluster 4. The separation of the clusters in this projection is poor and it appears that five clusters are too many for this dataset.



1. Visualize the Evaluate model to of see the final result of our experiment. This shows that there are 5 cluster centers and the Maximum Distance to Cluster Center for cluster. Of all the 5 clusters 11.7 is the maximum distance.

Step 5: Comparison of models for clustering of Food business:

From the Train Clustering module, we find out that 5 cluster has less distinct ovals compared to the 3 cluster, making 5 cluster less favorable for our dataset. From the Evaluation module, we find out that the maximum distance of the points plotted from the cluster center is grater in 3 cluster as compared to the 5 cluster. This means that the in 5 cluster the points are better clustered together. To conclude a better cluster, we tested a 2-cluster medal and following was the result on training the model.



The major axes of the ellipses for Cluster 0 and Cluster 1 are close and the separation of the clusters in this projection is poor. This helps us conclude that cluster 3 and cluster 5 have better results. Reducing the number of clusters will not improve the results further.