[](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.

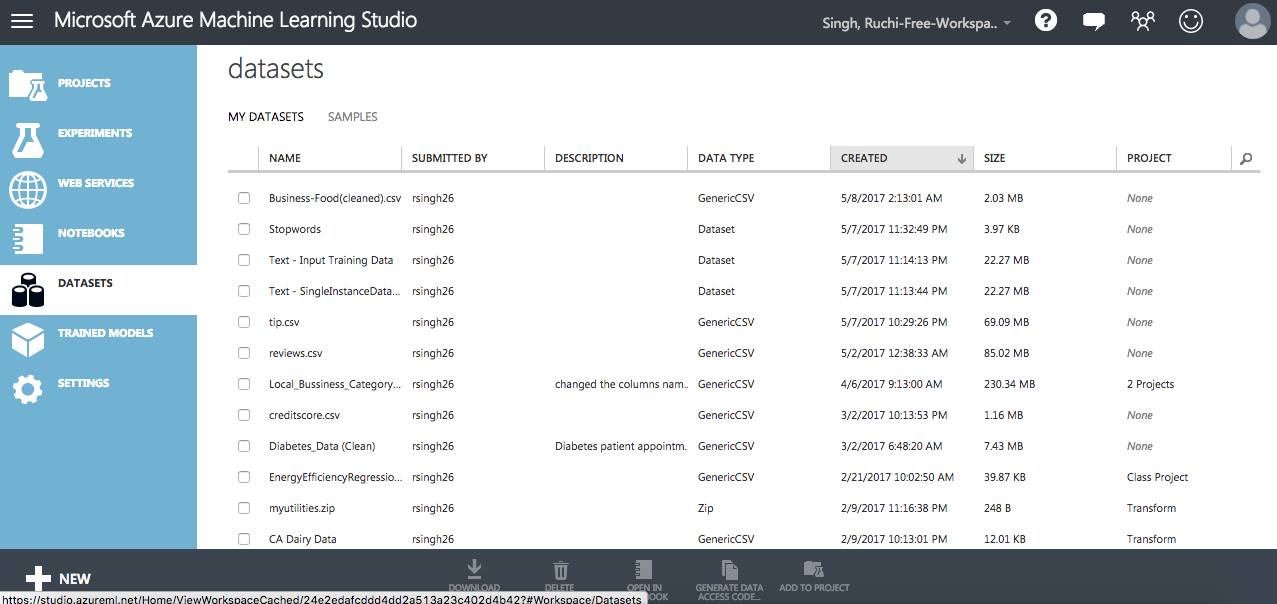
1. Score and test the model

* Step 4: Evaluate the model

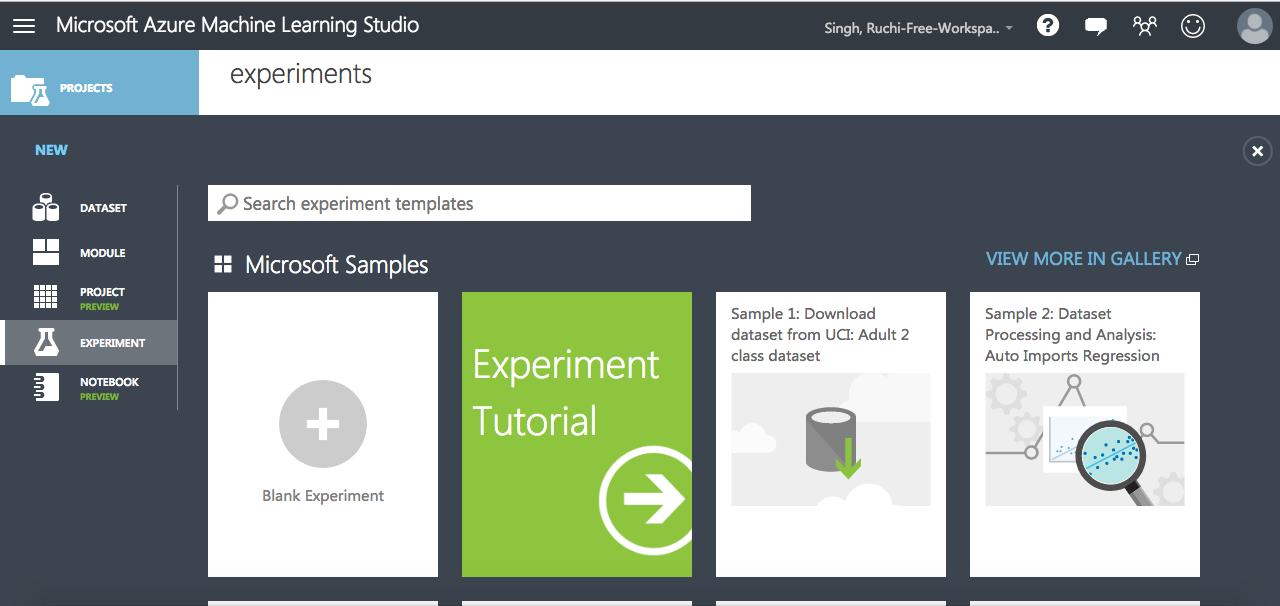
Step 1: Upload data

**Explain what this step is for.** This step is to get data manually….

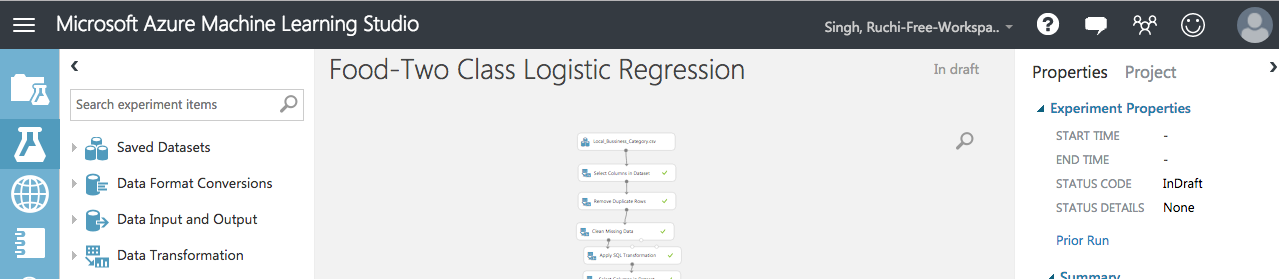
1. Create an account for Azure Machine Learning Studio.
2. Upload the Yelp dataset to Azure. This dataset includes entries like business ID, Name, Location, Ratings, Review Count, Category of the business etc.
3. Go to Dataset tab on left hand side, click on **+ New** sign and upload the dataset file. When you click on my datasets then the dataset you have uploaded will appear.



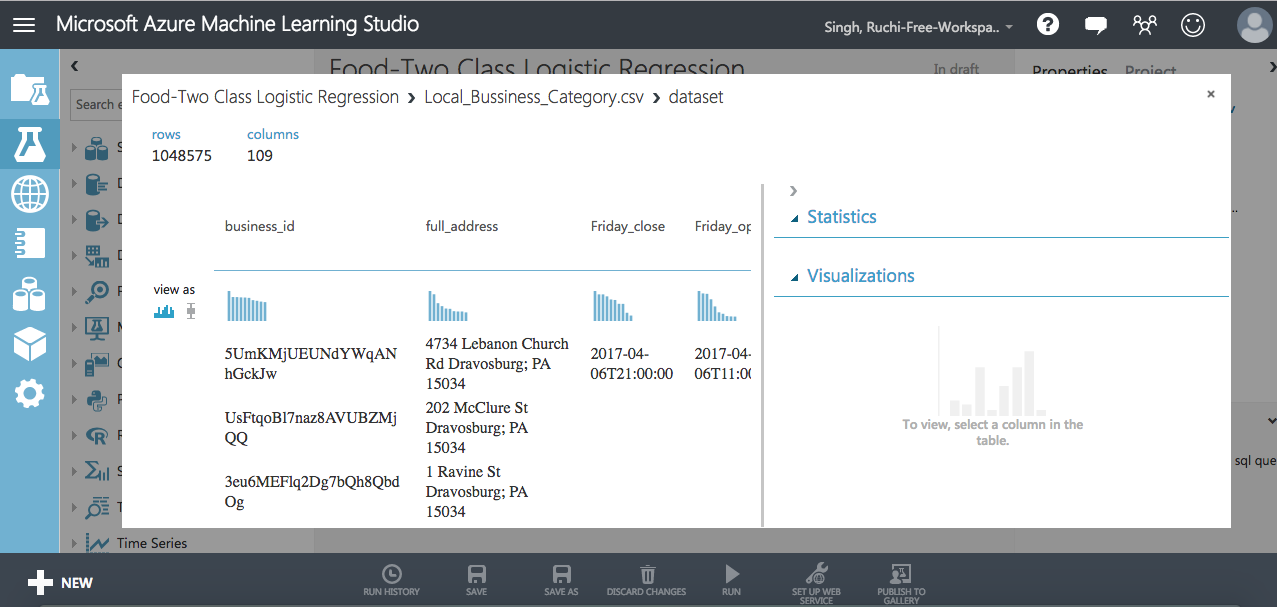
1. Start a new experiment by clicking +NEW at the bottom of the Machine Learning Studio window, select EXPERIMENT, and then select Blank Experiment. Select the default experiment name at the top of the canvas and rename it to something meaningful, for example, **Business Ratings prediction** or **Yelp dataset analysis** etc.



1. To the left of the experiment canvas is a palette of datasets and modules. Search for the dataset you want to use for the experiment.
2. Drag the dataset to the experiment canvas. In this case, upload Yelp dataset which we have uploaded earlier in my datasets.



1. To see what this data looks like, click the output port at the bottom of the Yelp dataset, and then select **Visualize**.

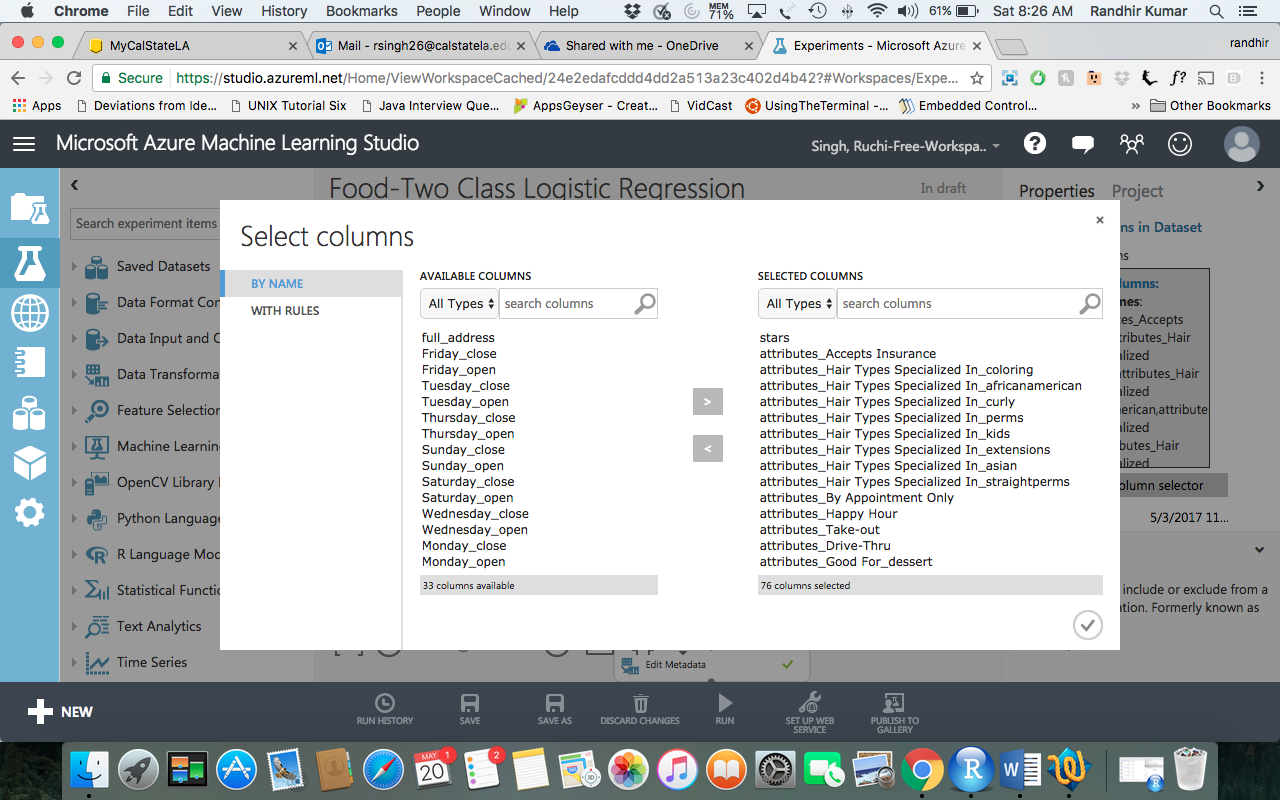


Step 2: Preprocess and Clean the Data

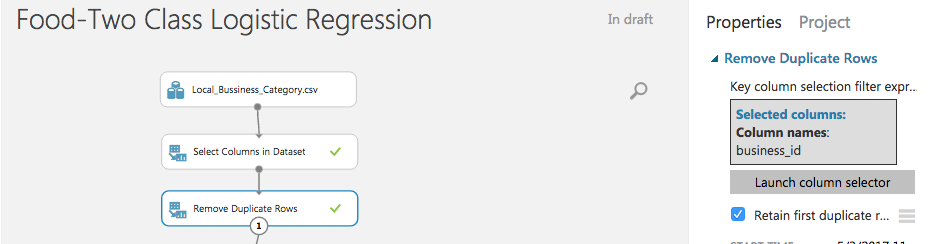
A dataset usually requires some preprocessing before it can be analyzed. You might have noticed the missing values present in the columns of various rows. These missing values need to be cleaned so the model can analyze the data correctly. In our case, we'll remove any rows that have missing values.

First we'll remove the normalized-losses column, and then we'll remove any row that has missing data.

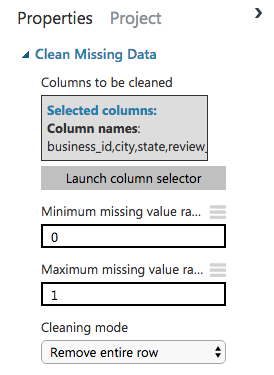
1. Search for project columns in the Search box at the top of the module palette to find the [Project Columns](https://msdn.microsoft.com/library/azure/1ec722fa-b623-4e26-a44e-a50c6d726223/) module, then drag it to the experiment canvas and connect it to the output port of the Yelp (Raw) dataset. This 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 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 (except those we're about to exclude).[ stars,attributes\_Accepts Insurance,attributes\_Hair Types Specialized In\_coloring,attributes\_Hair Types Specialized In\_africanamerican,attributes\_Hair Types Specialized In\_curly,attributes\_Hair Types Specialized In\_perms,attributes\_Hair Types Specialized In\_kids,attributes\_Hair Types Specialized In\_extensions,attributes\_Hair Types Specialized In\_asian,attributes\_Hair Types Specialized In\_straightperms,attributes\_By Appointment Only,attributes\_Happy Hour,attributes\_Take-out,attributes\_Drive-Thru,attributes\_Good For\_dessert,attributes\_Good For\_latenight,attributes\_Good For\_lunch,attributes\_Good For\_dinner,attributes\_Good For\_brunch,attributes\_Good For\_breakfast,attributes\_Caters,attributes\_BYOB,attributes\_Corkage,attributes\_Noise Level,attributes\_Takes Reservations,attributes\_DietaryRestrictions\_dairy-free,attributes\_DietaryRestrictions\_glutenfree,attributes\_DietaryRestrictions\_vegan,attributes\_DietaryRestrictions\_kosher, attributes\_DietaryRestrictions\_halal, attributes\_DietaryRestrictions\_soyfree, attributes\_DietaryRestrictions\_vegetarian,attributes\_Delivery,attributes\_DogsAllowed,attributes\_CoatCheck,attributes\_Smoking,attributes\_Ambience\_romantic,attributes\_Ambience\_intimate, attributes\_Ambience\_classy,attributes\_Ambience\_hipster,attributes\_Ambience\_divey,attributes\_Ambience\_touristy,attributes\_Ambience\_trendy,attributes\_Ambience\_upscale,attributes\_Ambience\_casual,attributes\_Parking\_garage,attributes\_Parking\_street,attributes\_Parking\_validated,attributes\_Parking\_lot,attributes\_Parking\_valet,attributes\_AgesAllowed,attributes\_WheelchairAccessible,attributes\_Open24Hours,attributes\_Music\_dj,attributes\_Music\_background\_music,attributes\_Music\_jukebox,attributes\_Music\_live,attributes\_Music\_video,attributes\_Music\_karaoke,attributes\_HasTV,attributes\_OutdoorSeating,attributes\_Attire,attributes\_Alcohol,attributes\_WaiterService,attributes\_Wi-Fi,attributes\_OrderatCounter,attributes\_AcceptsCreditCards,attributes\_BYOB/Corkage,attributes\_GoodforKids,attributes\_GoodForGroups,attributes\_GoodForDancing,business\_id,Category,city,state,review\_count]



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 module we need to add is Remove Duplicate rows. This will remove all the duplicate rows based on the Business id.

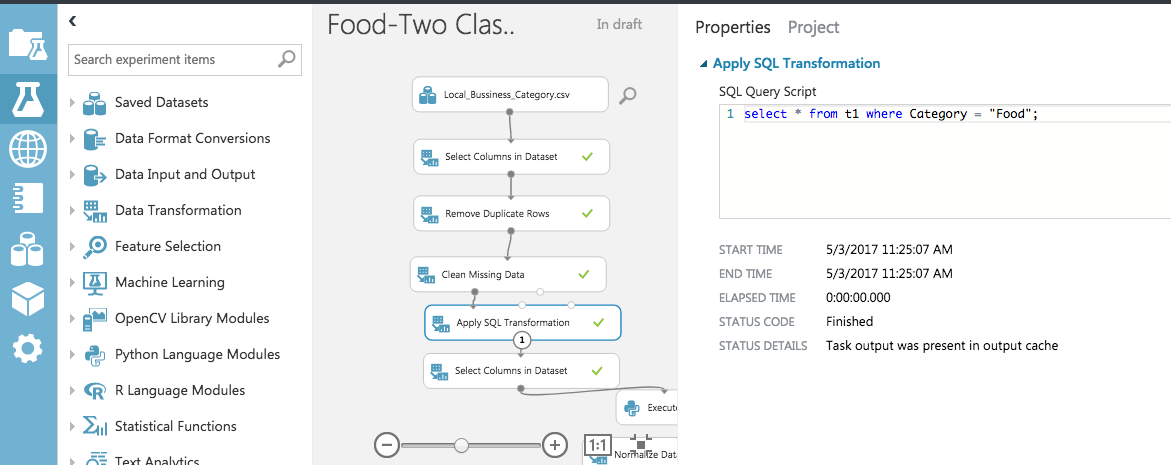


1. Now search for the module named clean missing data. Drag that to the experiment. Set the Properties pane. Fill minimum missing value as 0 and maximum missing value as 1 and select the columns business\_is, city state,review\_count, stars and category



Now that the data is clean our next steps will be to process the data.

1. Next we use an SQL Transformation module for writing an SQL query script. In this experiment I wish to predict the popularity of Food business. Thus lets write a query to select rows with category Food.



1. Now we need to select columns relevant to Food category. Use a Column Selector again and select the columns, review\_count,stars,attributes\_Good For\_lunch,attributes\_Good For\_dinner,attributes\_Good For\_breakfast,attributes\_Take-out,attributes\_Takes Reservations,attributes\_Parking\_lot, attributes\_Delivery, attributes\_WheelchairAccessible,attributes\_Alcohol,attributes\_WaiterService,attributes\_Wi-Fi,attributes\_Noise Level
2. Next we write a Python script int Execute Python Script module to categorize stars < 3 as 0 and stars > 3 as 1.

def set\_readmit\_class(x):

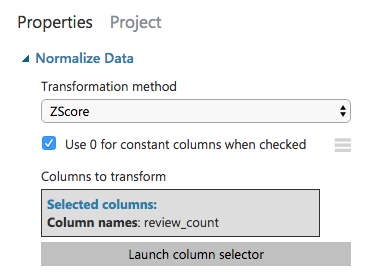
return [0 if (y < 3) else 1 for y in x]

def azureml\_main(df):

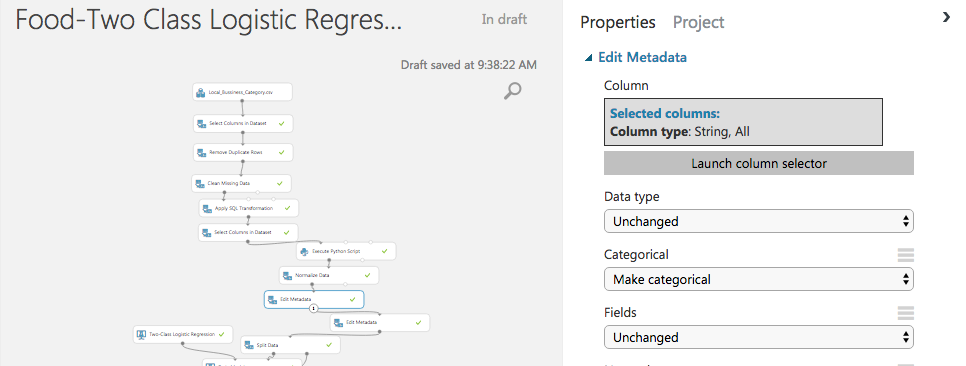
df['stars'] = set\_readmit\_class(df['stars'])

return df

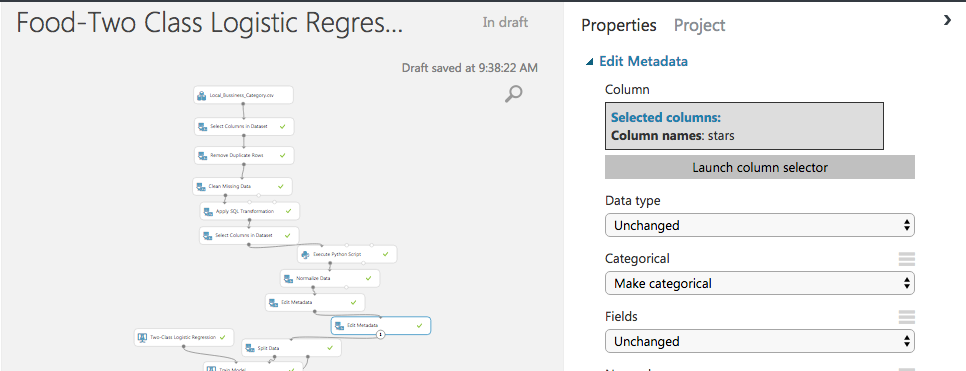
1. Now we need to normalize the review\_count using the Normalize Data module.



1. Change all String type columns to Category using Edit Metadata.



1. We also need to convert stars to Category using another Edit Metadata.



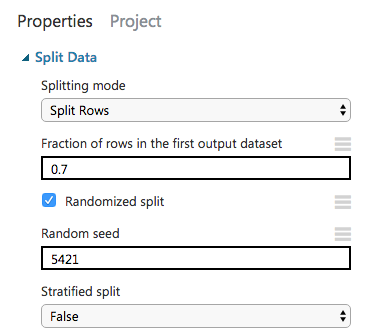
Step 3: Choose and apply machine learning algorithms:

Now that the data is ready, constructing a predictive model consists of training and testing. I am using data to train the model and then test the model to see how close it's able to predict ratings.

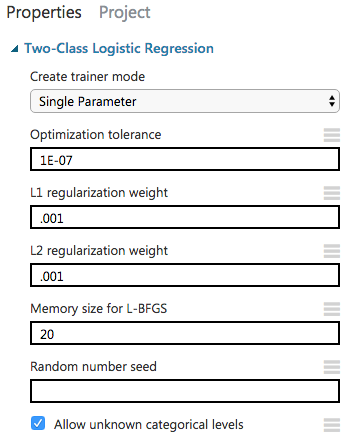
***Classification*** and ***Regression*** are 2 types of supervised machine learning techniques. We want to predict the popularity of the business, which can be value 0 meaning unpopular (stars < 3) and 1 meaning popular (stars > 3), so we'll use a Classification model.

Our data for both training and testing by splitting it into separate training and testing sets. Select and drag the [Split Data](https://msdn.microsoft.com/library/azure/70530644-c97a-4ab6-85f7-88bf30a8be5f/) module to the experiment canvas and connect it to the output of the last Edit Metadata module. Set **Fraction of rows in the first output dataset** to 0.7 and Random Seed of 5421. This way, we'll use 70 percent of the data to train the model, and hold back 30 percent for testing.

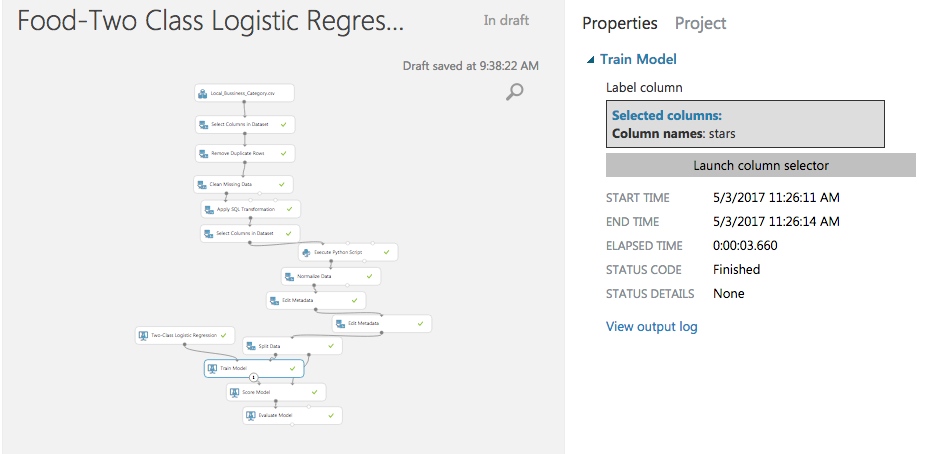
1. Run the experiment. This allows the Clean and processed data and then [Split Data](https://msdn.microsoft.com/library/azure/70530644-c97a-4ab6-85f7-88bf30a8be5f/) modules to pass column definitions to the modules we'll be adding next.



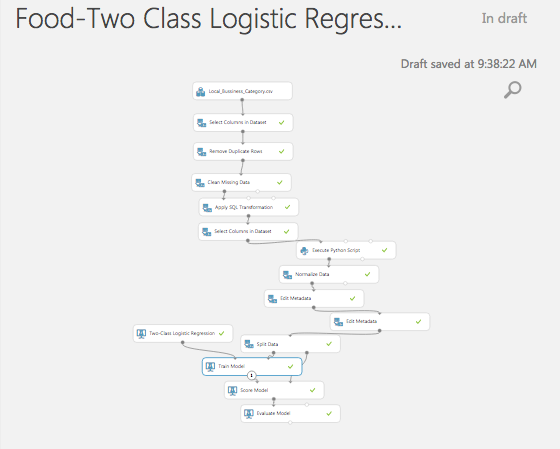
1. To select the learning algorithm, expand the Machine Learning category in the module palette to the left of the canvas, and then expand Initialize Model. This displays several categories of modules that can be used to initialize machine learning algorithms. For this experiment, select the Two Class Logistic Regression module under the Classification category and drag it to the experiment canvas.



1. Find and drag the [Train Model](https://msdn.microsoft.com/library/azure/5cc7053e-aa30-450d-96c0-dae4be720977/) module to the experiment canvas. Connect the left input port to the output of the Two class Logistic Regression 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.

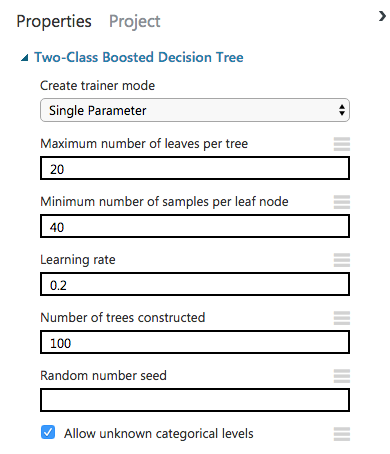


1. Select the [Train Model](https://msdn.microsoft.com/library/azure/5cc7053e-aa30-450d-96c0-dae4be720977/) module, click Launch column selector in the Properties pane, and then select the Stars column. This is the value that our model is going to predict.
2. Add Score Module and an Evaluate Module to see the evaluation of Two Class logistic regression.
3. Run the experiment.

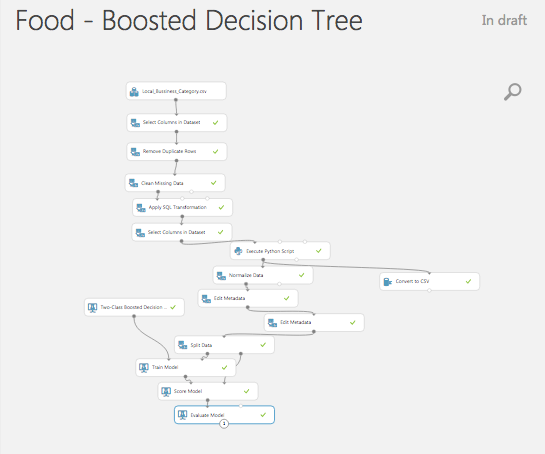


We can try another classification machine learning algorithm and evaluate it to see which model gives a better score. To do so we will need to follow the bellow steps after removing Two class Logistic regression module from the experiment.

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 Two Class Boosted Decision Tree module under the Classification category and drag it to the experiment canvas.



1. Find and drag the [Train Model](https://msdn.microsoft.com/library/azure/5cc7053e-aa30-450d-96c0-dae4be720977/) module to the experiment canvas. Connect the left input port to the output of the Two class Logistic Regression 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 Model](https://msdn.microsoft.com/library/azure/5cc7053e-aa30-450d-96c0-dae4be720977/) module, click Launch column selector in the Properties pane, and then select the Stars column. This is the value that our model is going to predict.
3. Add Score Module and an Evaluate Module to see the evaluation of Two Class Boosted Decision Tree.
4. Run the experiment.

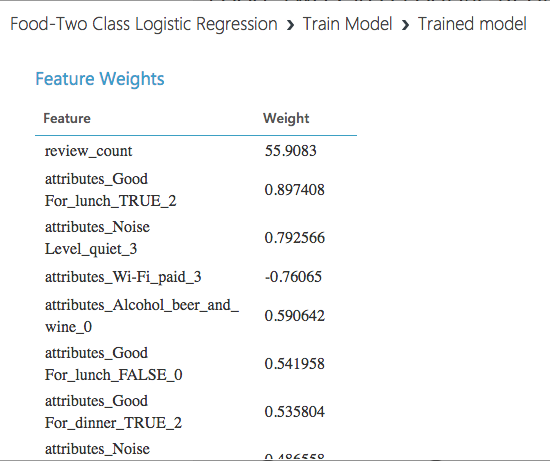


Step 4: Visualization:

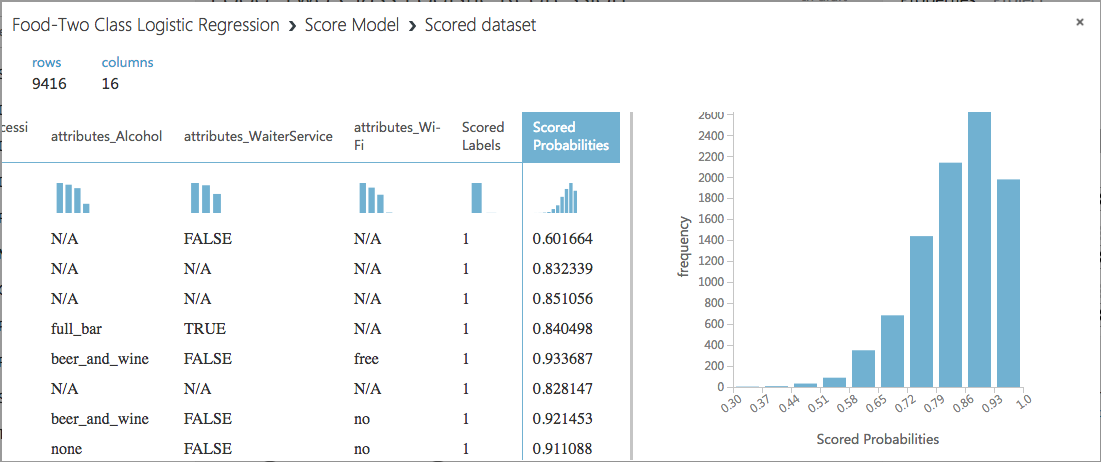
To compare the two models we need to visualize the output of Training module, Score module and Evaluate module first for Two class Logistic regression and then for Two class Boosted decision tree.

Two Class Logistic Regression

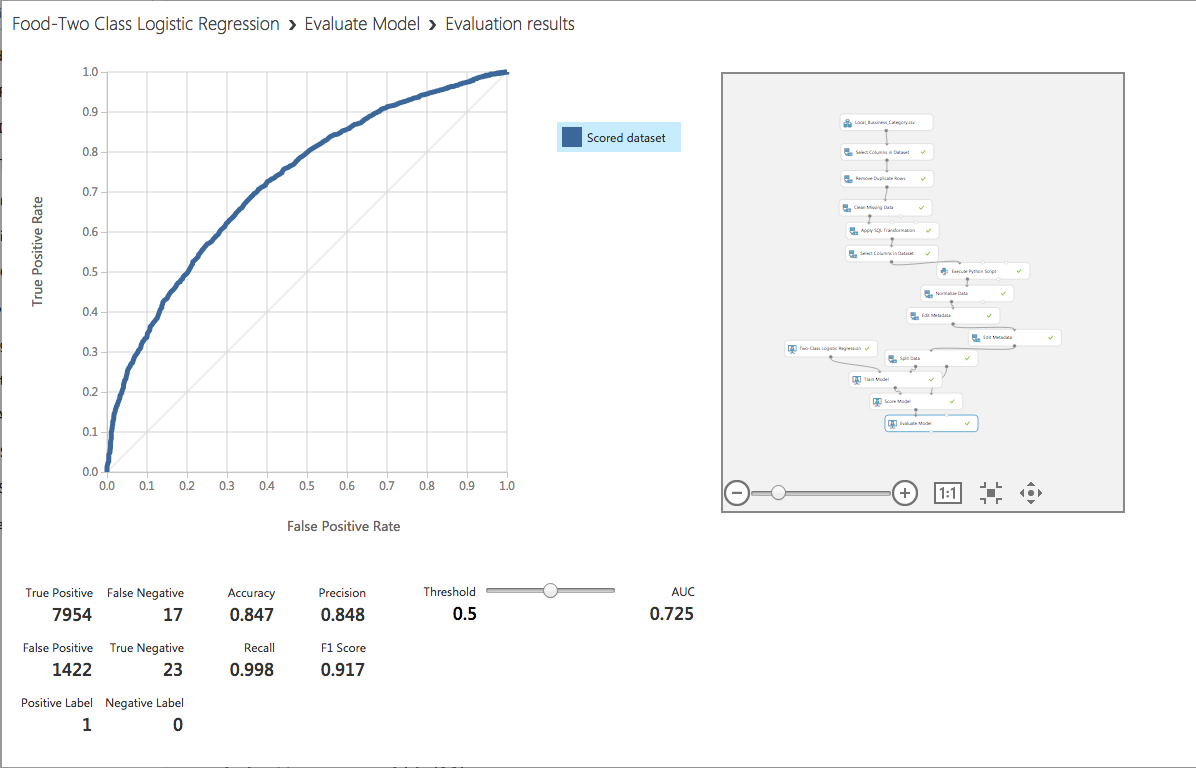
1. The Visualize the output of the train module. It shows the weight of the features for the prediction.



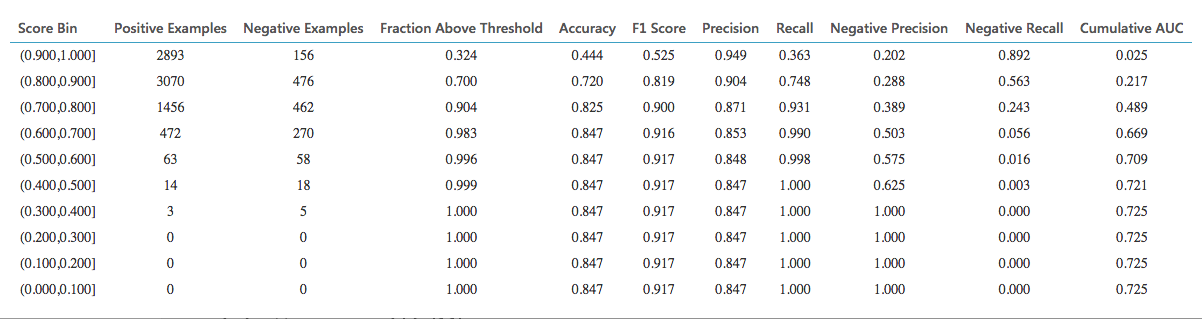
1. Visualize the Score model to see the Scored label and the Scored probabilities. Select the Scored probability column in the table to see the probability distribution of the prediction. It can be seen thet the frequency of the probability 0.7-1 is very high, which indicates that the probability of our perdic to be correct is very high.



1. Visualize the Evaluate model to of see the final result of our experiment. As expected it shows a very good accuracy of 0.8, recall of 0.9 and AUC of 0.7

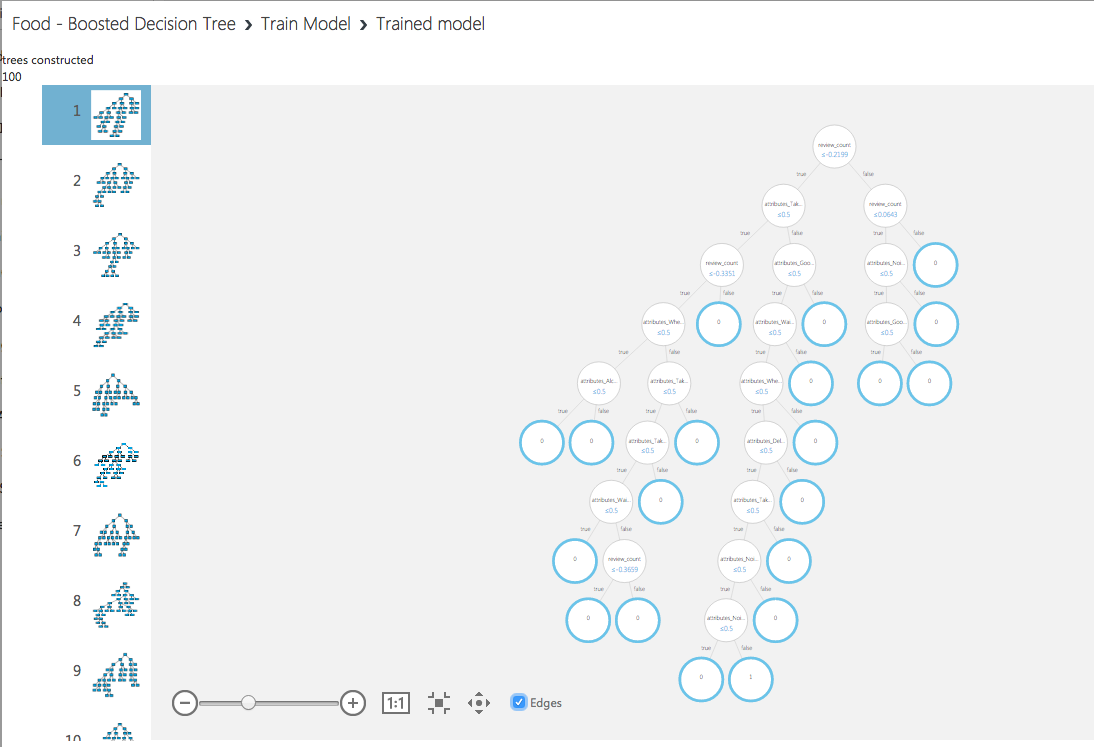


This the detail of every score bin and its prediction.

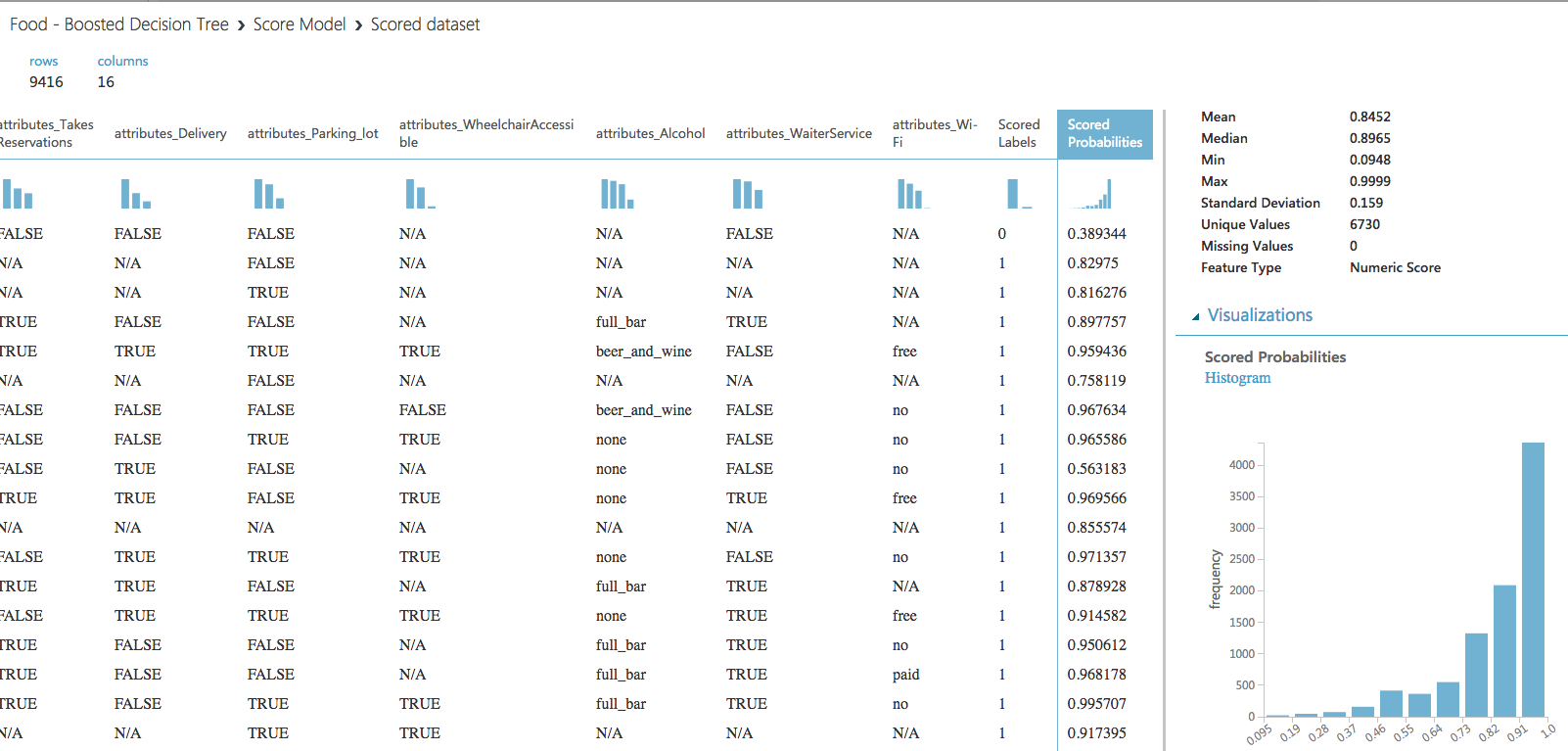


Two Class Boosted Decision Tree

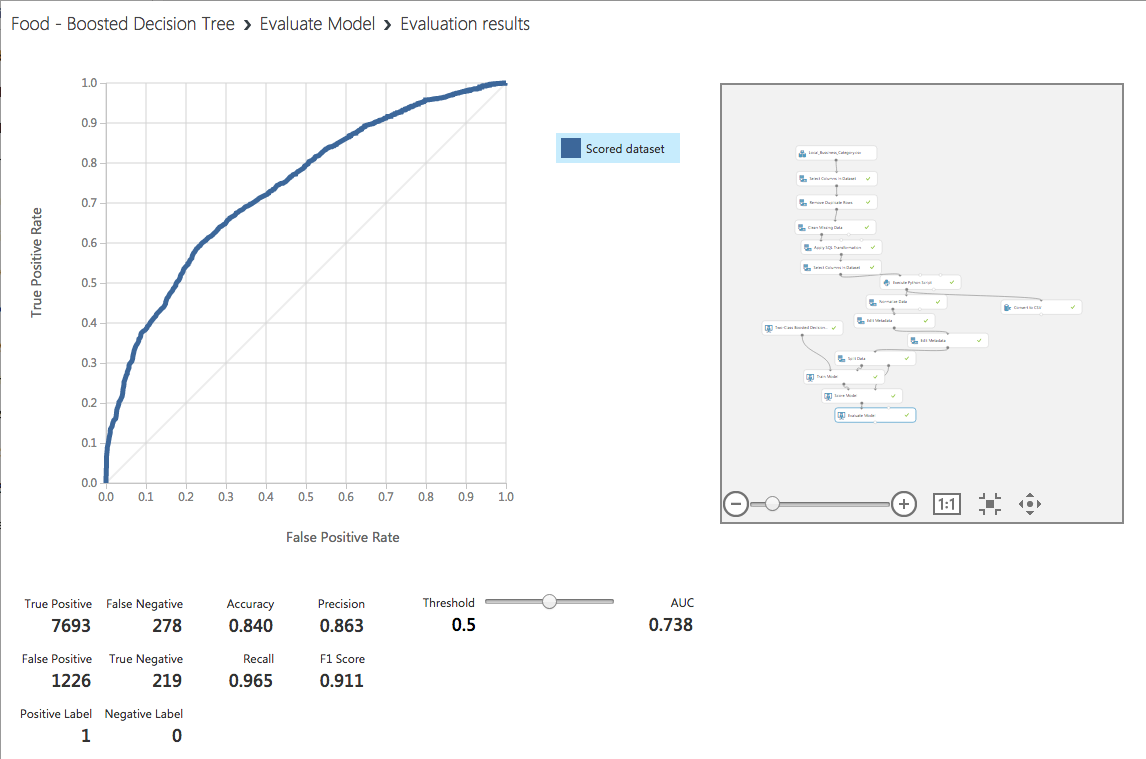
1. The Visualize the output of the train module. It shows the weight of the features for the prediction.



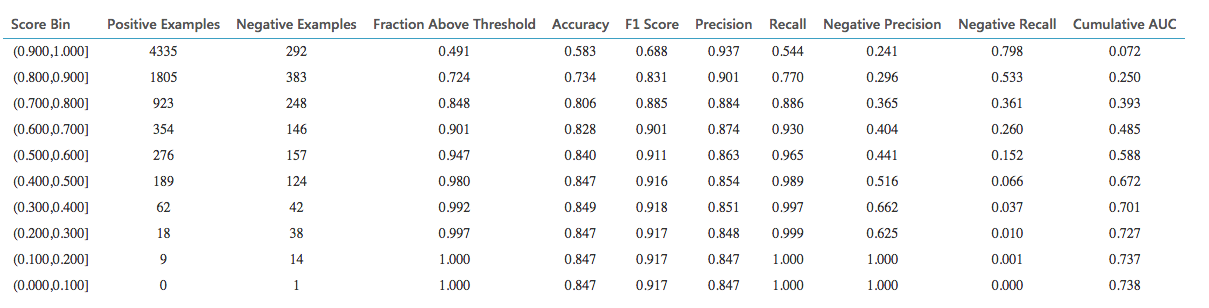
1. Visualize the Score model to see the Scored label and the Scored probabilities. Select the Scored probability column in the table to see the probability distribution of the prediction. It can be seen thet the frequency of the probability 0.7-1 is very high, which indicates that the probability of our perdic to be correct is very high.



1. Visualize the Evaluate model to of see the final result of our experiment. As expected it shows a very good accuracy of 0.8, recall of 0.9 and AUC of 0.7



This is the detail of every score bin and its prediction.



Step 5: Comparison of models to Predict popularity of business:

The two Class Logistic Regression and Two class Boosted Decision tree both give a good score of AUC = 0.7 which indicates that either of the two models can be used to predict the popularity of the business. The Two class Boosted Decision Tree has slightly better score than the logistic regression. These classification models are apt to predict whether a business will get less than 3 stars and will be unpopular or will get more that 3 stars and will be popular , depending upon the attributes of the business like, good for breakfast, good for lunch, good for dinner, parking lot, takeout, delivery, alcohol, waiter service, wi-fi and noise level. Therefore, either of the two models in our case can be used to have a good prediction accuracy.

**References**:

1. Dataset URL: ww[w.yelp.com/dataset\_challenge/dataset](https://www.yelp.com/dataset_challenge/dataset)
2. Github: <https://github.com/rsingh26/DataScience/tree/master/MachineLearning>
3. URL of Refereces :   <https://gallery.cortanaintelligence.com/Experiment/Text-Classification-Step-1-of-5-data-preparation-3>