**Advances in Data sciences**

**Report**

**Assignment3**

**Team 2:** Prajakta Pawar, Dhanshri More and Rutuja Magar

**1 Introduction**

We are using the publicly available dataset of Walmart Stores. This dataset contains the historical sales data of data for 45 Walmart stores located in different regions. Each store belongs to a particular type and contains a number of departments.

**2 Data Description**

**stores.csv**

This file contains anonymized information about the 45 stores, indicating the type and size of store.

**train.csv**

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

Store - the store number

Dept - the department number

Date - the week

Weekly\_Sales -  sales for the given department in the given store

IsHoliday - whether the week is a special holiday week

**test.csv**

This file is identical to train.csv, except we have withheld the weekly sales. You must predict the sales for each triplet of store, department, and date in this file.

**features.csv**

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

Store - the store number

Date - the week

Temperature - average temperature in the region

Fuel\_Price - cost of fuel in the region

MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.

CPI - the consumer price index

Unemployment - the unemployment rate

IsHoliday - whether the week is a special holiday week

**3 Data Pre-processing and cleaning**

* Read and merge the datasets from stores.csv, train.csv and features.csv.
* Convert the dates to character datatype.
* Consider a baseline date as '2010-02-05' and compute the number of days for each new date in our data.
* Split the given date into news columns Year, Month and Day.
* Sort the dataset on the basis of Store, Department and Date.
* We define four holidays viz: Labor Day, Christmas, Super Bowl and Thanksgiving.

**4 Linear Model**

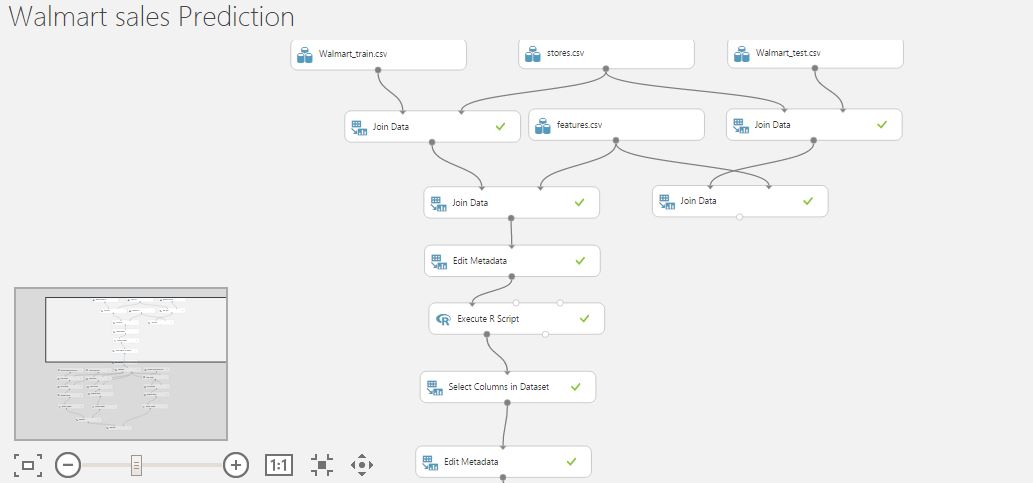
Predict the weekly sales depending on the parameters provided.

This helps in yearly or quarterly planning of the financials to achieve the set goals for the period. The revenue for this company is then estimated. We can include the increase in revenue for holiday season which compensates for the lag in revenue collection during regular days.

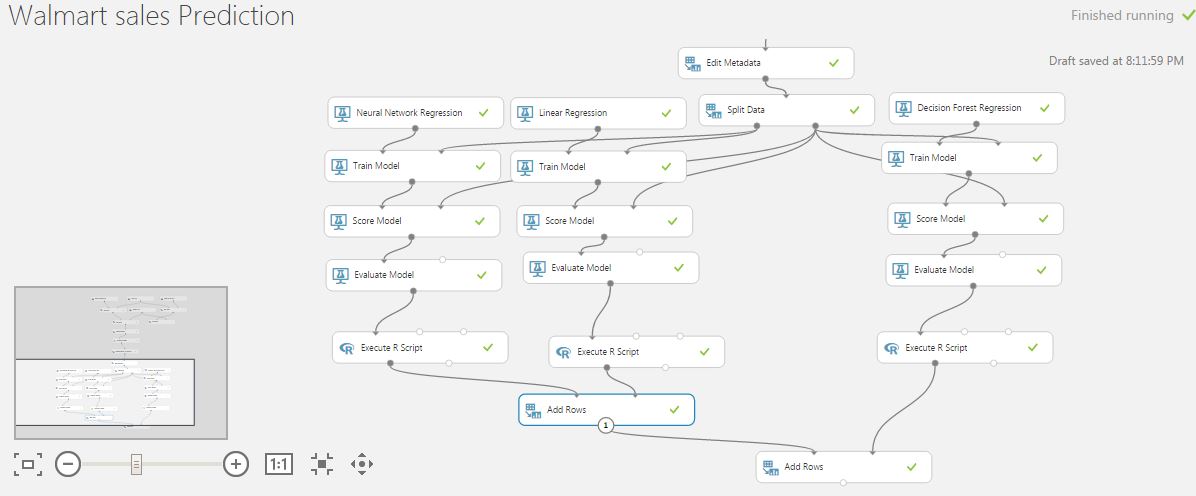
**4.1 Model built n Azure Machine Learning Studio**

Predict the weekly sales depending on the parameters provided.

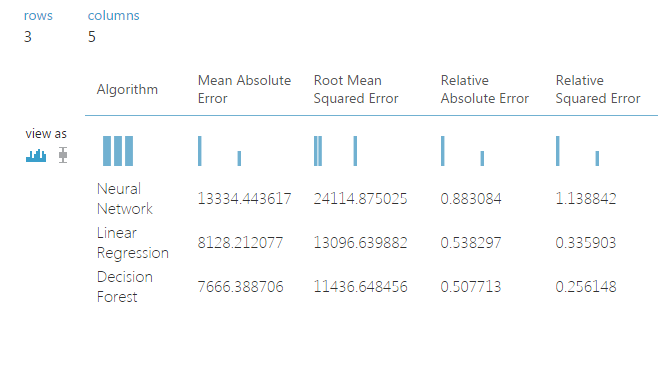
This helps in yearly or quarterly planning of the financials to achieve the set goals for the period. The revenue for this company is then estimated. We can include the increase in revenue for holiday season which compensates for the lag in revenue collection during regular days.



We are using three models for comparison viz: Neural Network Regression, Linear Regression and Decision Forest Regression.



4.3 Comparative Analysis based on Performance Matrix:-



**5 Classification**

Find if the department in the store will have sales above the average sales of the department in a month or lower. To decide which department needs more attention or refinement or change in products or costs.

This helps in analyzing the inventory stock if it is actually preferred by the customers or other products are to be introduced.

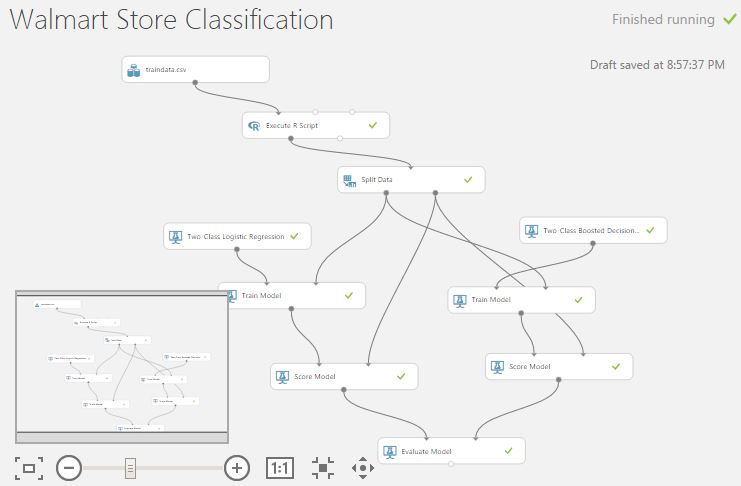
A store might be doing well but there needs to be analysis of each department in the store to decide if these products are actually preferred by customers in that area.

The analysis is done for each month of each department over all 45 stores to see if the department of a store will generate expected revenue that will help in giving expected profit to the investment.

If a department is not doing well then removal of the department can be an option for that store to maintain its financial stability.

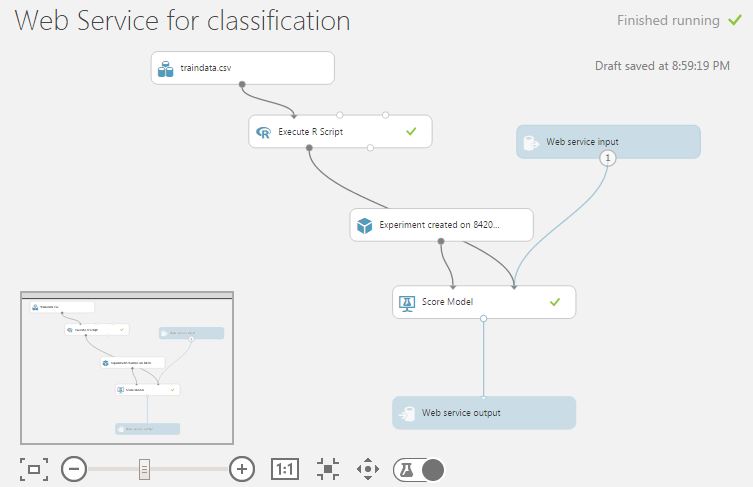
**5.1 Classification models in Azure Machine Learning Studio**

We are using two models for classification viz: Two Class Logistic Regression and Two Class Boosted Decison

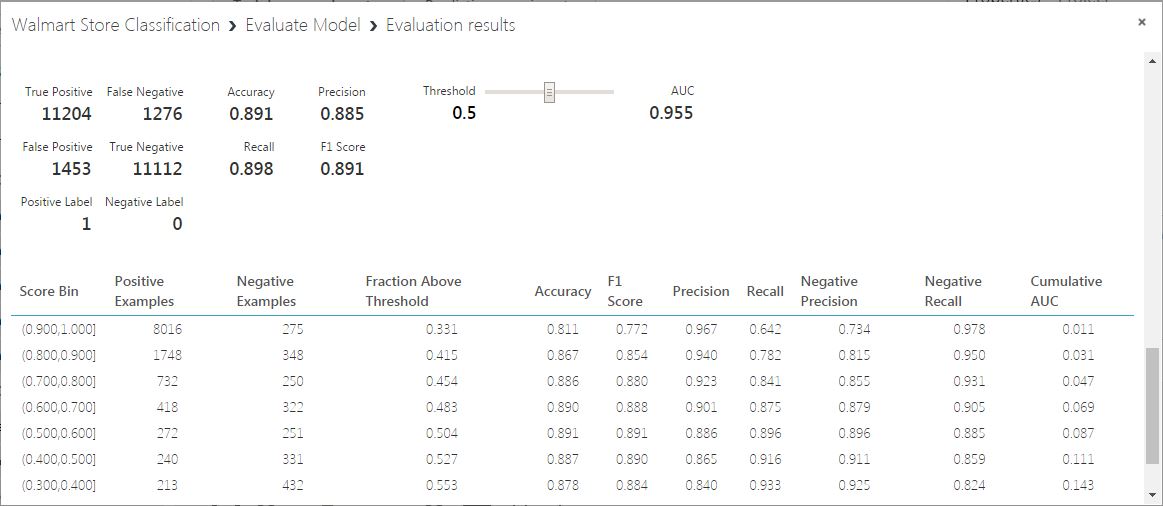
****

**Web service:**

The above model is deployed as a web service

****

**Evaluation Results:**

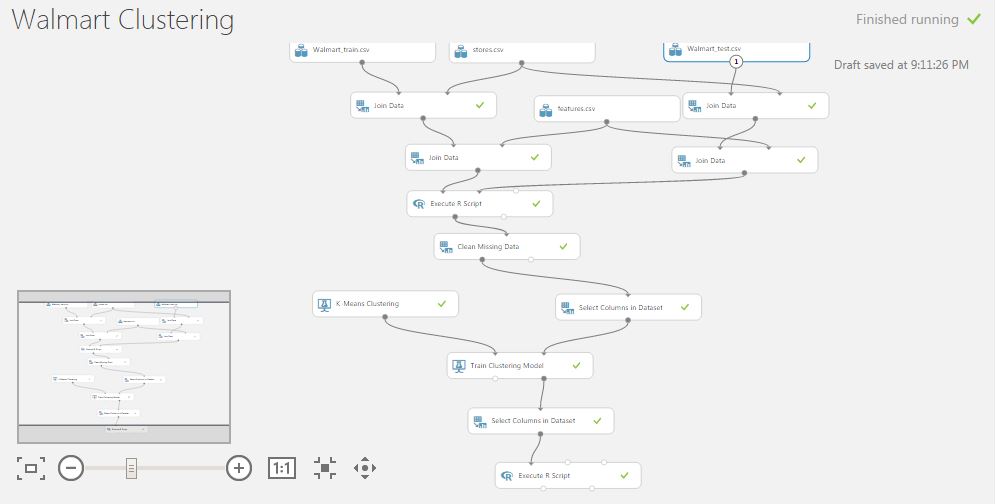
****

**6 Clustering Model**

Set financial sales goals for a given store by finding the cluster it falls in depending on its external conditions. This information is mainly useful if you are planning to open a new store in an area. This will give you the goals you need to set for revenue of the store. Thus estimate the cash flow and help in taking business decisions to establish its customer base and then aim for growth in finances to expand its business area.

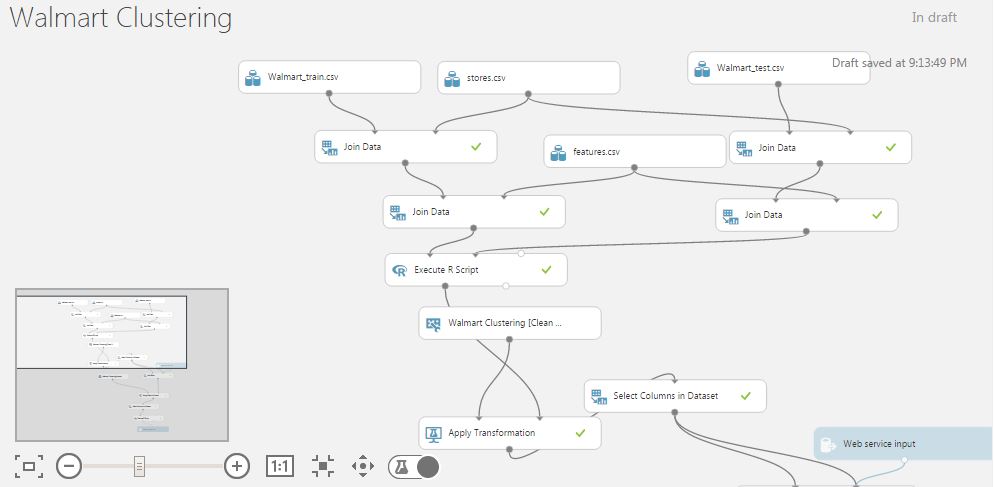
**6.1 K- means Clustering using Azure Machine Learning Studio**

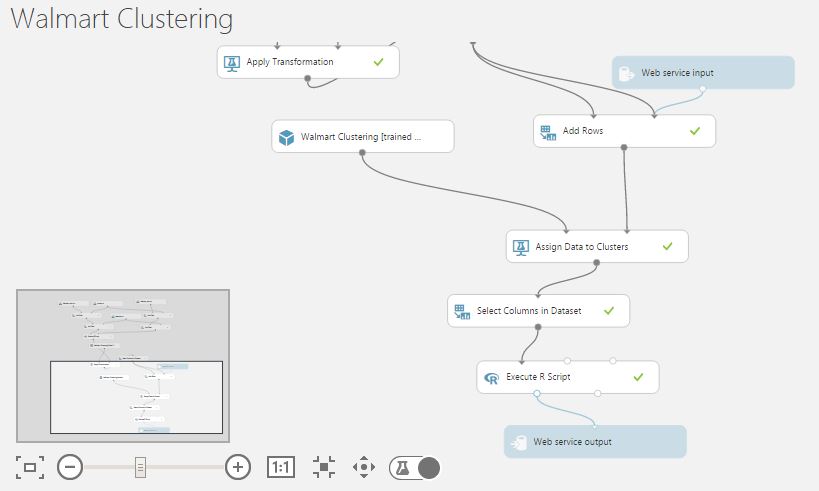
We use the external factors such as: Weekly\_Sales, Size, Temperature, Fuel\_Price, CPI, Unemployment to predict the cluster in which the store falls.



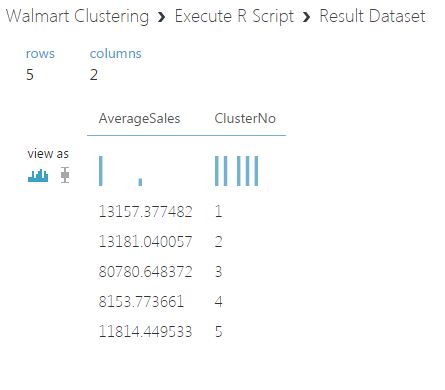
**Web service:**

The above model is deployed as a web service

****

****

**Results:**

****