Walmart Sales Prediction Using Rapidminer

Prepared by : Nagarjun Singharavelu

I. Introduction:

Wal-Mart Stores, Inc is an American Multinational retail corporation that operates a chain of discount department stores and Warehouse Stores. Headquartered in Bentonville, Arkansas, United States, the company was founded by Sam Walton in 1962 and incorporated on October 31, 1969. It has over 11,000 stores in 27 countries, under a total 71 banners. Walmart is the world's largest company by revenue, according to the Fortune Global 500 list in 2014, as well as the biggest private employer in the world with 2.2 million employees. Walmart is a family-owned business, as the company is controlled by the Walton family. Sam Walton's heirs own over 50 percent of Walmart through their holding company, Walton Enterprises, and through their individual holdings. The company was listed on the New York Stock Exchange in 1972. In the late 1980s and early 1990s, the company rose from a regional to a national giant. By 1988, Walmart was the most profitable retailer in the U.S. Walmart helps individuals round the world economize and live better.

The main aim of our project is to identify the impact on sales throughout numerous strategic selections taken by the corporate. The analysis is performed on historical sales data across 45 Walmart stores located in different regions. The foremost necessary is Walmart runs many promotional markdown events throughout the year and we have to check the impact it creates on sales during that particular period. The markdowns precede prominent holidays, the four largest of which are the Labor Day, Thanksgiving and Christmas. During these weeks it is noted that there is a tremendous amount of change in the day-to-day sales. Hence we tend to apply different algorithms which we learnt in class over this dataset to identify the effect of markdowns on these holiday weeks.

II. Information about dataset:

We had taken four different datasets of Walmart from Kaggle.com containing the information about the stores, departments, average temperature in that particular region, CPI, day of the week, sales and mainly indicating if that week was a holiday. Let us explain each dataset in detail.

Stores:

- The no. of attributes in this dataset is 3.
- They are store number, type of store and the size of store.
- Output attribute is the size of store.
- There are 45 stores whose information is collected.
- Stores are categorized into three such as A, B and C, which we assume it to be superstores containing different types of products.
- The store size would be calculated by the no. of products available in the particular store ranging from 34,000 to 210,000.

Train:

- This is the historical training data, which covers to 2010-02-05 to 2012-11-01.
- It consists of the store and department number.
- Date of the week.
- Weekly sales in USD for the given department in the given store.
- Also data denoting if the week is a special holiday week by displaying true or false.
- The department no. ranges from 1-99.
- The dataset consists of many records that are blank. Hence they were not considered for analysis.
- The output attribute is sales and the data denoting special holiday week.

Test:

- This dataset is similar to train data, except that the sales column has been eliminated.
- It is used to predict the sales for each triplet of store and department.

Features:

- This file contains regional activity for the given dates such as average temperature in that region which is measured which ranges from -7 to 100 F.
- Cost of fuel in that particular region is available in the dataset from which we can identify that if the fuel is higher in that particular region, whether it affects the sales as customer use automobiles to reach the store.
- Customer price index (CPI) which is used to measure the changes in price level in that particular store.
- Un-employment rate in that particular region.
- Data denoting if that particular week is a special holiday week.
- 5 Markdowns are available in this dataset which denotes the promotional markdowns which Walmart is running. There are 5 markdowns in that particular duration in the dataset.
- The output attribute is Markdowns and data denoting if it's a holiday week.

Among these 5 files, Features and train files contain every necessary elements to do the predication. So we are going to take advantage of them.

III. Data transformation & Preprocessing:

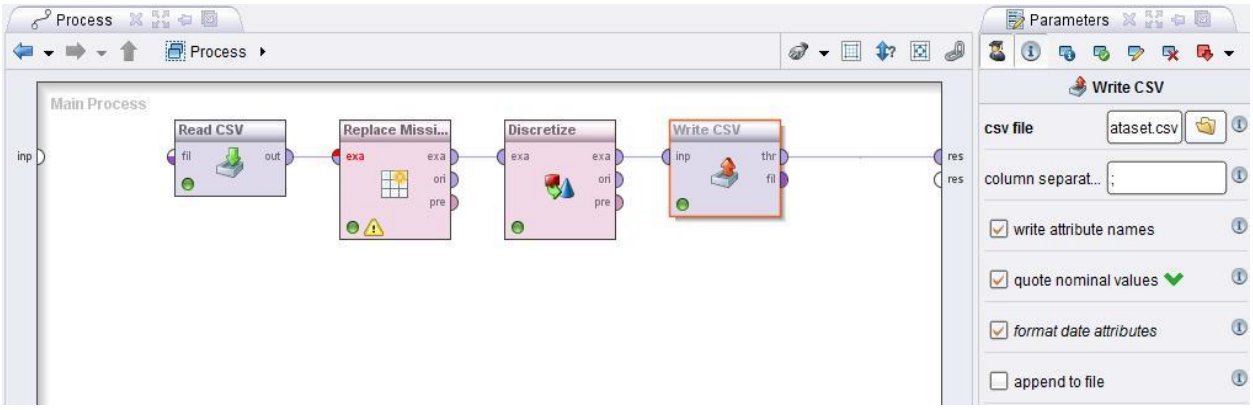
Our aim is to find Walmart's weekly sales. We have all the necessary data for predicting the weekly sales in Feature data. But the previous weekly sales report is in Train data set. So we decided to combine all the data sets into a single data set. After combining, we now have the sales data of 45 stores with up to 99 departments which consists of totally 421,000 records. Now our final data set has the following attributes.

- Store
- Date

- Temperature
- Fuel_Price
- MarkDown1-5
- CPI
- Unemployment
- IsHoliday
- Weekly_Sales
- Dept
- Type

Now we have 13 attributes, in which attributes such as department number (Dept), Date and Type are not required for our prediction, since it doesn't have any impact on its weekly sales.

Hence we are removing these attributes from our final data set. After removing all the unwanted data and attributes in our dataset we now have 6,435 records on total.With our huge dataset, it is difficult to get exact result about future sales from models because of the uncertainty in so many disturbing elements. So we decided to group the data and then implementing with classifiers or regression models. By using Discretize by Size operators, we grouped the data based on Weekly_Sales into 5 categories and then exported the output as CSV file to desktop.



After processing, now check the data of weekly sales, we found that the ranges are as follows;

range1 [-? - 9144.495]

range2 [9144.495 - 14213.040]

range3 [14213.040 - 22267.010]

range4 [22267.010 - 39499.755]
range5 [39499.755 - ?]

Now we manually label the whole 6,435 records into 5 groups each contain 1,287 records.

Groups	Mark as	Description
Group 1	A	range1 [-? - 9144.495]
Group 2	B	range2 [9144.495 - 14213.040]
Group 3	C	range3 [14213.040 - 22267.010]
Group 4	D	range4 [22267.010 - 39499.755]
Group 5	E	range5 [39499.755 - ?]

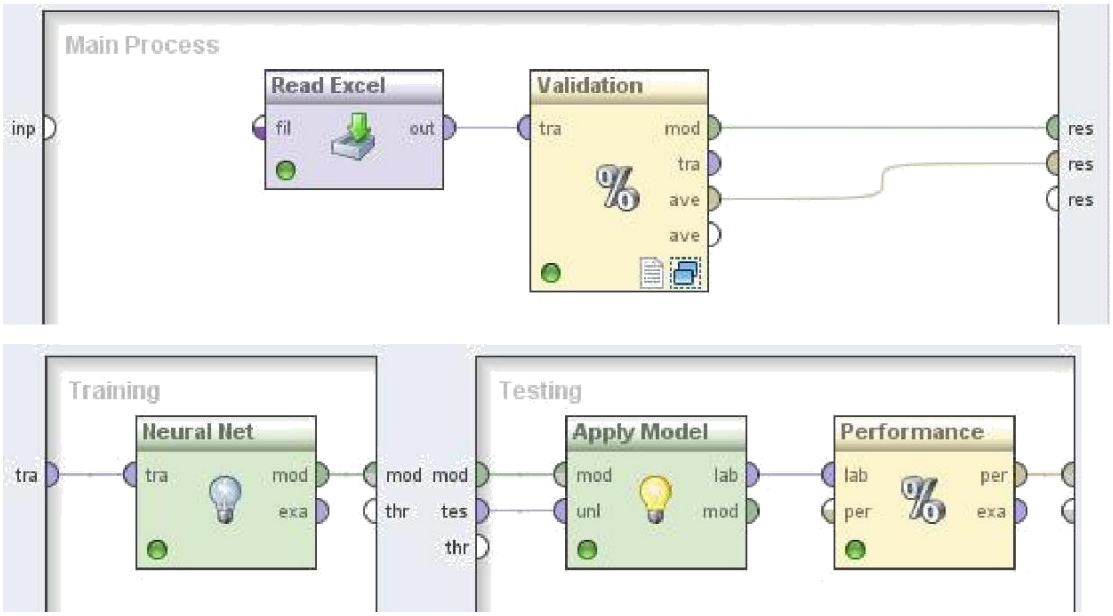
IV. Mining models & Performance:

Our purpose is to build a model to predict Walmart's weekly sales in future, so we would like to select the most suitable model for predication. Look into the table, we find out the data is structured. However, the parameters are complex and do not show its rules obviously. Even more, we have no idea about whether it is linear or not.

According to problems above, we determined to use Neural Network Model with our project because it has the characteristics below:

- It is for complicated prediction problems.
- Visualization or understanding of the rules are not needed.
- Accuracy is very important.

The design for implementing the Neural Network model in rapid miner is as follows:



The output we obtained is as follows,

Accuracy = 77.61%, when Learning Rate = 2000 & Training Cycles = 0.05

	true A	**true B**	**true C**	**true D**	**true E**	**class precision**
pred. A	939	49	83	36	95	78.12%
pred. B	17	1114	58	75	101	83.61%
pred. C	160	102	923	148	29	68.77%
pred. D	60	95	93	780	187	65.20%
pred. E	23	276	43	161	788	63.04%
class recall	78.32%	68.09%	77.92%	65.00%	67.67%	

We tried to test our result by changing learning rate and training cycle values and noted down the changes too. The result obtained in this process is as follows;

Accuracy = 72.69%, when Learning Rate = 1000 & Training Cycles = 0.05

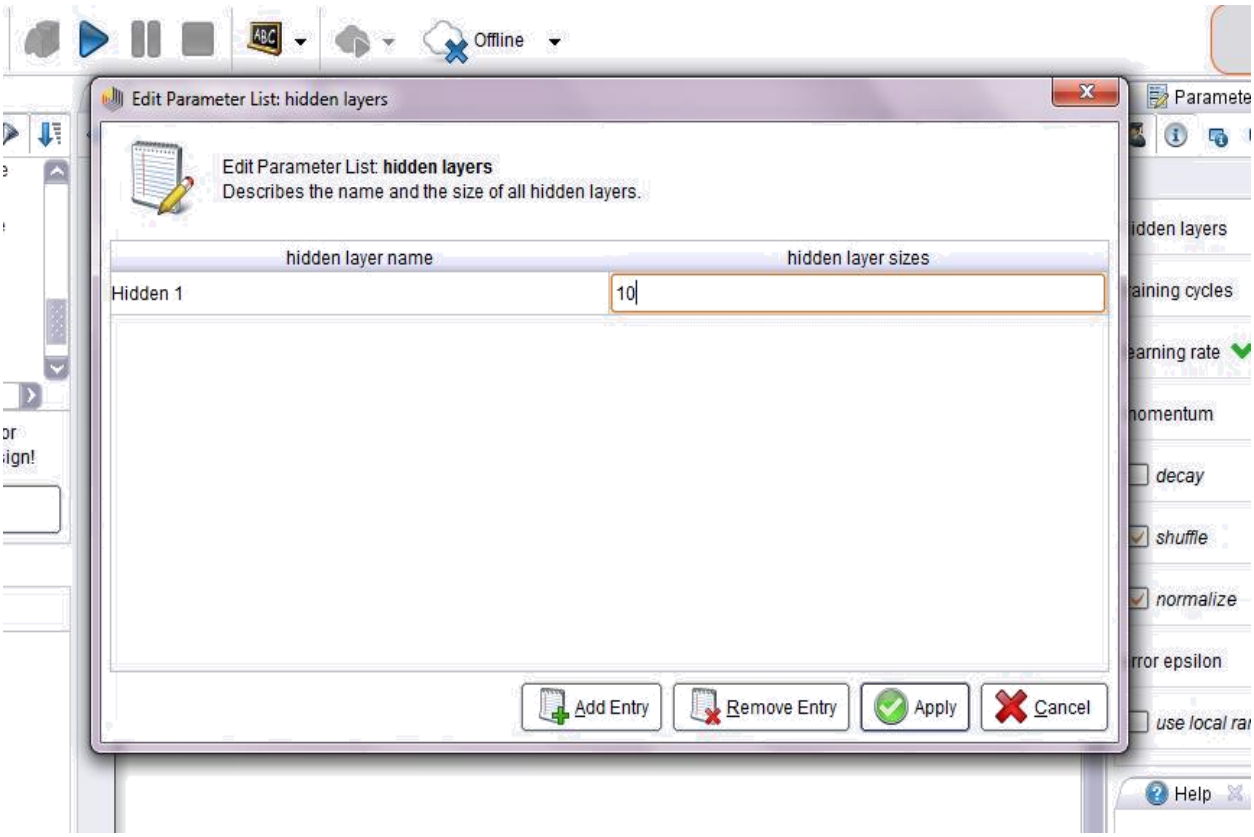
	true A	**true B**	**true C**	**true D**	**true E**	**class precision**
pred. A	1012	62	96	42	102	77.62%
pred. B	23	1193	58	76	120	84.61%

pred. C	184	114	1025	181	28	68.91%
pred. D	69	124	104	842	209	64.89%
pred. E	34	306	37	179	861	61.89%
class recall	76.72%	69.31%	78.65%	63.79%	67.23%	

Accuracy = 75.35%, when Learning Rate = 1000 & Training Cycles = 0.3

	true A	**true B**	**true C**	**true D**	**true E**	**class precision**
pred. A	908	86	114	56	108	74.38%
pred. B	25	1033	50	64	83	86.31%
pred. C	164	123	895	176	40	64.02%
pred. D	73	117	94	732	189	62.75%
pred. E	29	277	47	172	780	63.07%
class recall	77.03%	65.24%	77.28%	64.42%	67.02%	

Since weekly sales is the core for any industry. So it is necessary to get the more accurate results, since it have direct impact on companies revenue. To achieve better performance and accuracy results, we have now increased the hidden layer size to 10.

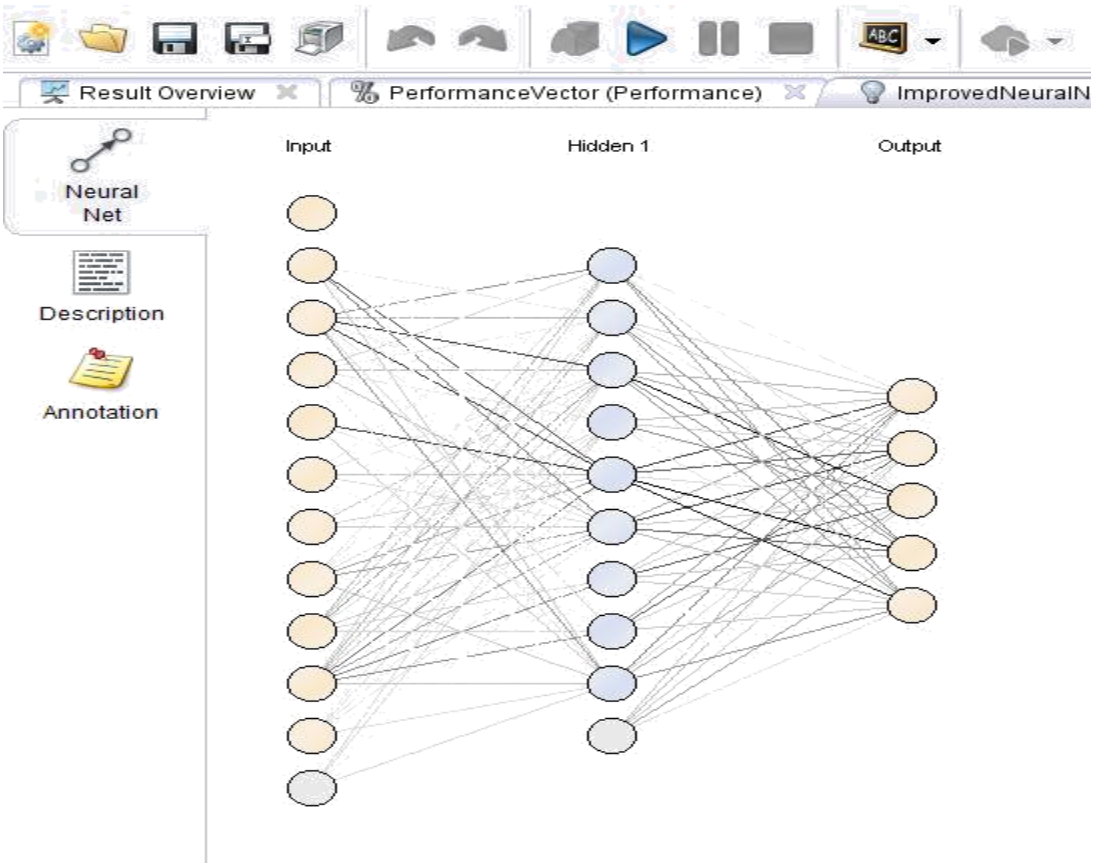


We now combined all the hidden layer results obtained from each node in a single table to check where the weekly sales have a greater impact.

Attributes	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9	Node 10
Store	-30.224	27.949	-20.9	-31.401	5.072	-17.01	40.862	27.088	18.266	16.821
Temperature	-0.192	0.68	-0.965	-0.601	0.038	-0.847	-0.436	-1.513	-1.094	0.126
Fuel_Price	1.054	1.459	-0.46	0.632	-1.242	-1.713	0.698	1.303	2.589	-0.007
MarkDown1	0.114	-2.679	-0.24	-0.636	-4.109	-1.322	1.004	0.648	-1.147	-2.402
MarkDown2	0.88	0.476	-0.03	2.65	-0.387	-1.306	0.328	0.633	1.983	-0.821
MarkDown3	9.209	-0.635	0.631	7.522	-6.247	-2.194	-6.743	-0.075	1.157	6.285

MarkDown4	-0.613	-0.554	0.131	1.257	-2.967	-0.445	0.753	0.691	-0.115	0.599
MarkDown5	0.41	-3.728	-3.697	1.786	-14.643	-2.111	0.341	-0.999	-7.048	-0.908
CPI	-6.316	29.004	-12.179	-12.975	0.568	11.217	-27.856	-15.442	-5.764	1.045
Unemployment	-12.906	-17.876	4.064	-1.446	8.895	3.908	7.056	12.174	-31.068	1.142
IsHoliday	0.056	0	-0.191	-0.484	-0.068	-0.244	-0.151	-0.27	-0.036	-0.025
Bias	-13.162	-10.766	-9.052	-8.572	-25.596	9.488	-10.875	20.103	-2.328	-9.733

The hidden layer network for this dataset will look like:

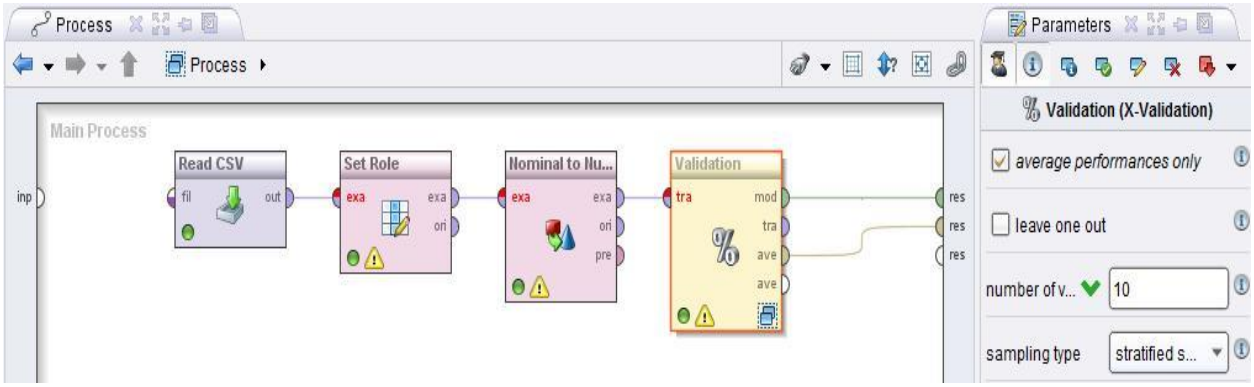


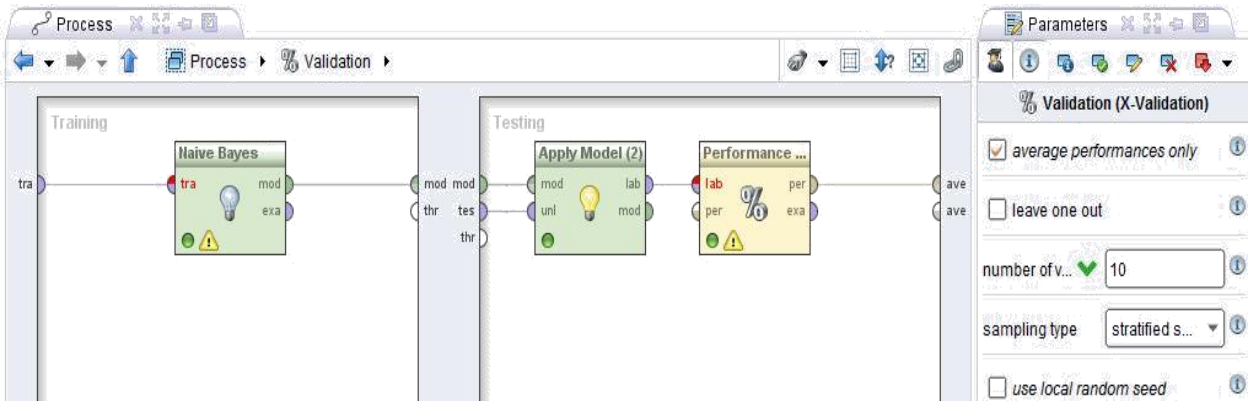
Other Models:

We tried to analyze data with more models other than Neural Network. But we find out that Neural Network model is the most suitable one for Walmart dataset. Other models cannot work on this example.

Naïve Bayes:

Design:





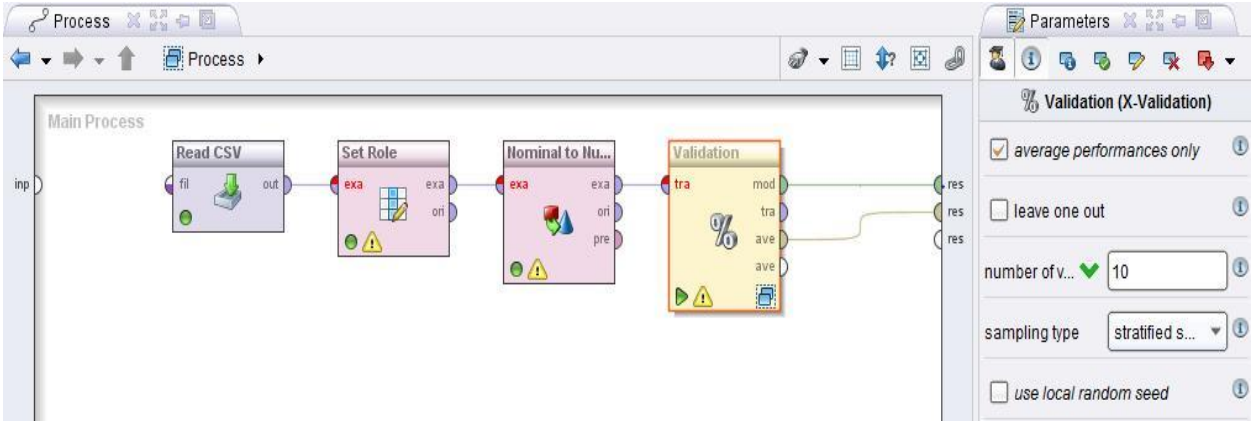
Output:

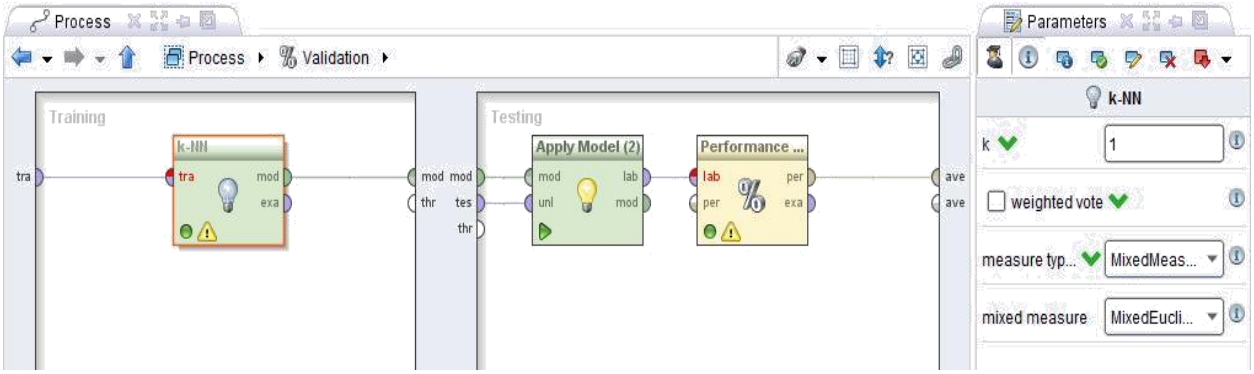
Accuracy = 23.42%

	true A	true B	true C	true D	true D	class precision
pred. A	8190	0	0	0	0	
pred. B	19795	23905	23910	23905	23909	20.71%
pred. C	15895	19128	19124	19128	19126	20.70%
pred. D	0	0	0	0	0	0.00%
pred. E	3917	4782	4781	4782	4780	20.74%
class recall	17.13%	49.99%	40.00%	0.00%	10.00%	

K-NN:

Design:





Output:

When K = 1, Accuracy = 28.76%

	true A	true B	true C	true D	true E	class precision
pred. A	385	207	373	180	199	28.65%
pred. B	190	619	259	311	301	36.85%
pred. C	352	271	187	225	188	15.29%
pred. D	113	267	124	180	210	20.13%
pred. E	159	272	257	304	302	23.34%
class recall	32.11%	37.84%	15.58%	15.00%	25.17%	

When K = 10, Accuracy = 31.06%

	true A	true B	true C	true D	true E	class precision
pred. A	429	234	403	168	206	29.79%
pred. B	272	834	288	421	428	37.18%
pred. C	219	123	186	97	213	24.20%
pred. D	195	266	162	347	281	28.64%
pred. E	84	179	161	167	72	10.86%
class recall	35.78%	50.98%	15.50%	28.92%	6.00%	

Decision Tree:

No matter what the parameters are, the Accuracy is 25.42% and never change.

	true D	true S	true C	true B	true A	class precision
pred. D	0	0	0	0	0	0.00%
pred. S	1199	1636	1200	1200	1200	25.42%
pred. C	0	0	0	0	0	0.00%
pred. B	0	0	0	0	0	0.00%
pred. A	0	0	0	0	0	0.00%
class recall	0.00%	100.00%	0.00%	0.00%	0.00%	

As we show above, Naïve Bayes Model products the lowest Accuracy (18.63%). Abstractly, the Naïve Bayes model is a conditional model like this:

$$P(\text{Sales} \mid \text{Store, Data, Temperature} \ldots \& \text{IsHoliday}) = \frac{P(\text{Sales}) * P(\text{Store, Data, Temperature, } \ldots \& \text{IsHoliday} \mid \text{Sales})}{P(\text{Store, Data, Temperature, } \ldots \& \text{IsHoliday})}$$

Because variable Store, Data to IsHoliday are independent on each other, so:

P (Store, Date, Temperature … & IsHoliday) =P (Store)*P (Date)*…..*P (IsHoliday)

Due to so many numbers in columns Store, Date … IsHoliday that do not repeat, the probability of each Variables is too small. So P (Store)*P (Date)*…..*P (IsHoliday) will be far lower than 1/6435. This means the probability of sales basing such a model is infeasible.

V. Interpretation:

As a result, we find out that how each parameter affects the weekly sales. MarkDown 1 to 5 has the highest weight as 9 to -16 which mean it really makes an enormous impact on the sales. Promotion will increase weekly sales remarkably. Fuel price and temperature also makes a positive impact, higher price and temperature makes higher sales. CPI and Unemployment rate having a heavy negative impact on the prospects of sales. The higher CPI and unemployment rate, the less weekly sales. Holidays affect weekly sales slightly. We think customers don't care whether today is holiday or not, the only reason they buy items is promotion. Therefore, Promotion drives sales and revenue to increase.

VI. Conclusion:

The main objective was to model the effects of markdown events throughout the year and it was successfully measured using various algorithms. From the analysis it is noted that MarkDown 1 to 5 has the highest weight as 16, which means it really makes an enormous impact on the sales. Promotion will increase weekly sales remarkably. Fuel price and temperature also makes a positive impact, higher price makes higher sales. CPI and Unemployment rate has a heavy negative impact on the prospects of sales. The higher CPI and unemployment rate, the less is the weekly sales. Holidays affect weekly sales slightly. We felt customers didn't care whether it was a holiday or not because the only reason they buy items is due to promotion. It is notable that they are attracted to various deals in departmental stores all over the world.

References:

1. https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data
2. http://www.walmart.com/
3. http://en.wikipedia.org/wiki/Walmart