Analysing Walmart Sales

Siddharth Sharma
MSc in Data Analytics, School of Computing
National College of Ireland
Dublin, Ireland
x16148371@student.ncirl.ie

***Abstract* – In today's service-oriented world CRM/ Analytical CRM is important as the customer satisfaction needs to be taken care of, otherwise it may be bad for any company's reputation and growth. CRM - Consumer Relationship Management, it aims at analysing consumer's relationship with a company and build marketing campaigns. Analytical CRM is used for exploring the data of customers activity such as purchasing patterns, behaviour patterns while browsing products etc. Walmart is an American multinational retail corporation which runs departmental stores across the world. Walmart uses machine learning technique's extensively to increase its profits by retaining customers, getting insight into what the customer is looking for or buying, discovering patterns over seasonal shopping etc. This study will use statistical modelling and machine learning algorithm to see the factors responsible for increasing weekly sales at Walmart.**

Keywords—Walmart, Machine Learning, Data mining, Statistical Model;

I. INTRODUCTION

With the escalating usage of information technology in business sector, marketing strategies for companies have drastically changed in the past decade [1]. There is a very competitive market which is growing exponentially and to survive this competitiveness companies have to pay extra attention towards customer satisfaction. Companies like Amazon, Flipkart, Alibaba, Siemens, Philips, Sony, Cannon, Nikon, Adidas, Nike etc. demand feedback from their customers for each and every item that is being purchased. Customer's feedback is highly valued and taken into consideration for gaining profits and improving their services. Data mining utilizes large datasets for finding patterns and analysing the behaviour of customers to provide insightful information such as what the customer is looking for, when the customer is searching for a particular product (e.g. during holiday season, off season etc.), which particular area is being browsed (e.g. electronics, cosmetics, clothing, books, sports gear, home appliances, motor vehicles, toys, games etc.), mode of payment (i.e. by cash, e-wallet, credit card, debit card etc.). After evaluating all the aspects of customer behaviour lucrative offers are made in accordance (e.g. someone might get discount on electronic goods based on the previous shopping trends) to retain a customer. Retaining a customer in such a competitive market is really tough as a customer might change the vendor because a good deal might be offered, ease of accessibility, variety of goods etc. Data mining offers a prospective analysis which helps a company to improve its relationships with customers, customer support services and customer retention [2].

Machine learning algorithms help a company to improve their overall growth by excelling is sectors like sales, marketing, customer support services etc. [3]. In the following sections of the report we will discuss the background of the dataset, research question & the hypothesis followed by a literature review, methodologies employed, conclusion and finally the references used to complete the report.

II. BACKGROUND OF DATASET

In this report we have used the data provided by Walmart for its 45 stores across different locations. The data set is available on Kaggle (https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data). There are 8191 records and 8 attributes which contribute towards the analyses of the sales of different departments. Walmart provided the data to predict the sales of the stores. Modeling of the retail data is always a challenge and here we find the correlation of various attributes that led to the magnanimous sales of 45 stores based in different regions.

The data set contains the following attributes []:

- Store number
- Date: weekly data for 4 years
- Average temperature in the region
- Fuel price in the region
- Markdown 1-5 are Walmart's promotional offers during various seasons
- CPI- Consumer Price Index
- Unemployment Rate
- Is Holiday- indicating whether it was a holiday or not.

III. Research Question and Hypothesis

Which factors impact the most on a stores weekly sale?
How Consumer Price Index affects the Walmart Sales?

Hypothesis: -
Ho: Promotional Activities(Markdowns) affect the weekly sales.
H1: Promotional Activities(Markdowns) do not affect the weekly sales.

IV. Review of Literature

Data Mining techniques have been used by a number of retailers around the world to have successful business. Data driven decisions are being made by retailer's like Walmart, Tesco, Metro etc. which are being used for taking advantage in the marketing sector [4]. In 2018, authors Anastasia Griva, Cleopatra Bardaki, Katerina Pramatari, Dimitris Papakiriakopoulos studied customer segmentation using market basket data and implemented clustering techniques, adjusted the product taxonomy to provide an insight into the customers satisfactory level as per their visits to the store [4]. In the year 2016, Manpreet Kaur, Shivani Kang also did a market basket analysis for identifying the trends changing in the market data with the help of association rule mining as it discovers the hidden relationship with large datasets[5]. To understand the customer behavior and help the retailer to make an informed decision is market basket analysis [5]. Authors Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael E. Welge used data mining and knowledge management while facing issues like real-time interactive marketing, cross organisational management and customer relationship management for decision support in marketing sector [1]. The authors used trend analysis, deviation analysis & customer profiling to see how customer knowledge leads to pattern extraction, as the competition in the market has intensified and is increasing constantly pressure on decision makers is also on rise hence long-term customer relationships surfaced, later known as customer relationship management [1].

In 2015 Chieh-Yuan Tsai and Sheng-Hsiang Huang utilized data mining techniques to "optimize shelf space allocation in consideration of customer purchase and moving behaviors" [6]. Here the author has deployed the process of optimizing the shelf space in three stages i.e. in the first stage the customers purchasing behavior and moving behavior has been analyzed, in the second stage UMSPL algorithm has been applied in addition to priori algorithm, in the third stage product items have been used to set the criterion , a ;location preference evaluation has also been carried out [6]. The final outcome of the process undertaken results in helping the retail stores to precisely place the products as per customer patterns over time using classification for each & every product instead of relocating the products from their original positions [6].

V. Methodoloy and Tools Used

Preprocessing of Data:
First the data source was selected from a wide range of datasets then the relevant data was extracted.
The dataset was then cleaned and prepared for applying various models.

Model Framework:
Training and testing data was prepared for the model.
The statistical method was applied on the model for processing the data prepared.
Finally, the model was evaluated, and outcome was published.

Tools Used:
RapidMiner Studio was used for building and processing our model. Where every task was carried out with efficiency and the outcome was obtained.

VI. Implementation and Resul t

1) Linear Regression:
Linear regression has been used in the model to analyze the correlation of the independent variables with the dependent variable. The model was implemented through RapidMiner Studio. Liner regression has been used because it explains the linear relation between the dependent and independent variables better than other techniques such as PCA.

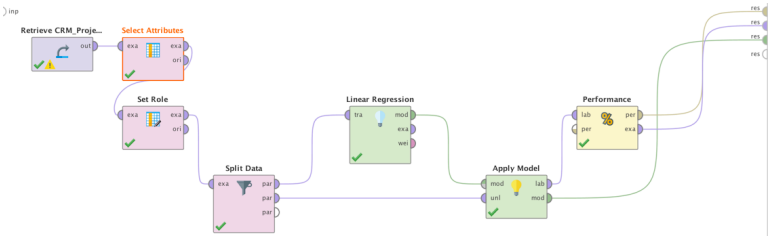

Figure 1: Linear Regression

Attributes were obtained from the data and then dependent variable was selected. Training and testing data was prepared (75%,25%). Performance of the model was tested.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Temperature	116.204	42.181	0.116	0.976	2.755	0.006	***
Fuel_Price	-9744.373	2486.447	-0.165	0.984	-3.919	0.000	****
MarkDown1	0.084	0.054	0.069	0.904	1.558	0.120	
MarkDown5	-0.682	0.219	-0.135	0.994	-3.112	0.002	***
(Intercept)	43797.049	8892.330	?	?	4.925	0.000	****

Figure 2: Coefficient Matrix

As seen in the coefficient matrix the tolerance values are less than 1 which proves that there is no multi collinearity in the model and the independent variables are not correlated with each other. The t-Stat value tells us the impact of the independent variable on the dependent variable e.g. markdown 5(-3.112) has a negative impact on the dependent variable. The p-Value depicts the correlation between fuel price, markdowns, temperature and weekly sales. Fuel price is highly correlated with weekly sales at Walmart stores.

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 14862.710 +/- 0.000
squared_correlation: 0.035
```

For the linear regression model, we have obtained the root mean squared error as ±14862.710 and the squared correlation comes out to be 0.035.

ExampleSet (188 examples, 2 special attributes, 6 regular attributes)

Row No.	Weekly_Sal...	prediction(...	Temperature	Fuel_Price	MarkDown1	MarkDown5	CPI	Unemploy...
1	45025.020	15858.903	50.320	3.243	6520.240	4001.250	224.202	6.525
2	44319.150	17495.590	49.660	3.475	72937.290	6310.180	224.276	6.525
3	48754.470	12226.991	48.010	3.711	10610.740	2747.590	224.565	6.525
4	13275.870	14596.206	47.420	3.243	6722.950	5383.200	223.834	6.237
5	11724.980	12161.142	48.920	3.475	16105.390	7043.580	223.911	6.237
6	8653.620	16091.960	58.130	3.417	2782.180	2046.530	224.803	6.112
7	35000.300	17263.661	49.830	3.243	2503.670	1366.370	227.758	6.108
8	46564.140	15521.694	59.960	3.597	2005.310	527.750	227.960	6.108
9	38489.630	13916.437	54.190	3.711	3805.930	490.130	228.106	6.108
10	33780.200	18001.021	71.690	3.451	976.830	851.010	228.730	5.999
11	33613.910	17734.987	65.230	3.417	539.280	572.620	228.730	5.999
12	18089.610	14639.769	38.920	3.235	4749.650	3743.860	131.957	3.921
13	17615.280	19137.772	49.070	3.232	27134.980	1663.480	132.154	3.921
14	18232.460	15915.095	49.860	3.401	56705.090	7730.260	132.215	3.921
15	18236.150	10034.704	53.210	3.647	23306.040	9316.710	132.415	3.921
16	20413.830	16576.727	63.280	3.436	3578.120	2039.260	132.716	3.896
17	20029.100	16516.314	45.080	3.237	2752.630	1769.010	224.831	5.494
18	19178.360	16358.616	59.050	3.417	21533.520	4109.090	224.828	5.494
19	17821.710	13430.877	53.090	3.475	3763.730	4380.830	224.869	5.494
20	16309.440	14122.089	54.100	3.597	3229.380	1731.250	225.013	5.494
21	34442.290	15085.814	40.460	3.161	2958.800	4190.080	225.711	5.372
22	32832.060	16113.670	44.940	3.237	3311.950	2403.960	225.865	5.372
23	32827.420	18389.584	57.350	3.227	1541.930	1107.600	225.863	5.372
24	36767.220	10298.672	50.080	3.711	5999.240	5362.390	226.188	5.372
25	34063.230	12958.094	54.230	3.606	6162.800	3690.590	226.632	5.372
26	35582.390	15028.621	61.400	3.583	13894.830	3152.220	226.717	5.285
27	13332.750	14320.129	-7.290	2.889	6630.250	1513.310	200.738	7.107
28	30503.300	9396.396	11.940	3.638	3789.540	959.830	201.155	7.107
29	33136.940	11791.177	34.790	3.642	1731.110	1032.440	201.071	6.953
30	26285.270	17091.323	44.340	3.227	6534.460	1406.030	227.834	5.212

Figure 3: Prediction Value

Prediction on the weekly sales was made using the given attributes based on the linear regression model. Our model predicts the weekly sales with a root mean squared error value of ±14862.710.

2) Random Forest

Random Forest comes under supervised learning algorithm which is used for classification and regression. The number of forests is directly proportional to the precision of the results. We have used random forest for regression purpose. Random Forest has been applied despite linear regression technique has been used earlier because random forest can interpret higher complexity in less amount of time.

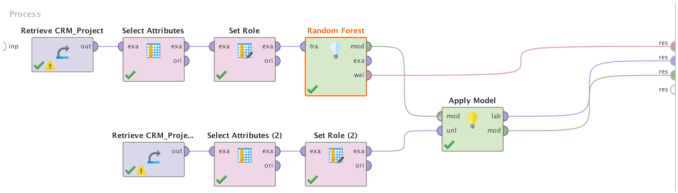

Figure 4: Random Forest

Random forest was applied using RapidMiner Studio. Data was extracted from the dataset, attributes were selected for the preparation of the model and weekly sales was assigned to label for the role of dependent variable. Finally, random forest was applied and attribute weights and labeled data were used to produce the final result.

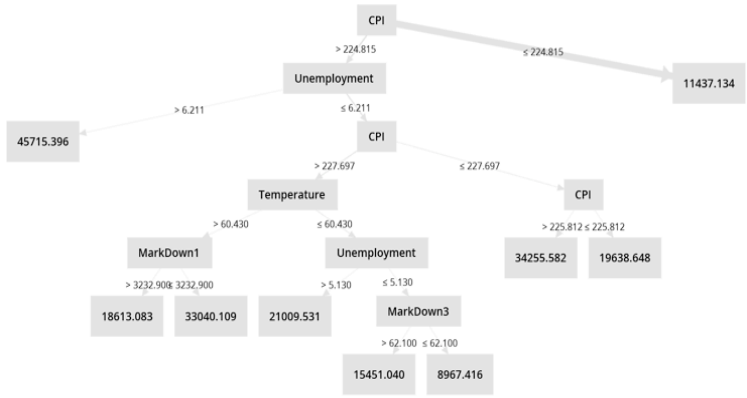

Figure 5: Regression Tree

Random Forest generated a regression tree which has the root node as Consumer Price Index (CPI) which then further bifurcates to unemployment rate, then temperature followed by markdowns and then finally the prediction attribute weekly sales.

attribute	weight
Temper...	0.041
MarkDo...	0.208
MarkDo...	0.262
CPI	0.351
Unempl...	0.138

Figure 6: Attribute Weights

Random forest generates an attribute weight table which depicts importance of the attributes according to the weight values.

Row No.	id	Weekly_Sal...	predictionL...	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemploy...
1	1	46549.730	12808.306	41.730	3.161	1214.080	25366.330	15.010	72.360	3940.020	224.081	6.525
2	2	45025.020	12808.306	50.320	3.243	6520.240	16134.600	12.170	774.550	4001.250	224.202	6.525
3	3	44418.110	12808.306	42.920	3.237	3772.690	3559.460	3.880	246.620	1900.400	224.236	6.525
4	4	45971.300	12808.306	53.370	3.227	965.890	1097.910	0.100	225.360	1831.880	224.236	6.525
5	5	47903.010	12808.306	56.460	3.244	9290.910	1359.900	265	20657.820	972.610	224.235	6.525
6	6	43675.610	12808.306	56.670	3.417	32355.160	729.800	280.890	20426.610	4671.780	224.235	6.525
7	7	44319.150	12808.306	49.660	3.475	72937.290	6665.520	47.210	13014.670	6310.180	224.276	6.525
8	8	44619.520	12808.306	50.250	3.597	20107.750	3163.890	42.200	15657.300	5812.860	224.420	6.525
9	9	48754.470	12808.306	48.010	3.711	10610.740	261.460	2.800	25.540	2747.590	224.565	6.525
10	10	47089.540	12808.306	50.810	3.658	5000.580	290.460	78.770	606.150	3697.110	224.709	6.525
11	11	44428.710	19663.958	55.330	3.622	3808.130	0	15.650	2616.600	1909.170	224.836	6.525
12	12	45299.920	19663.958	63.420	3.611	12553.980	0	495.100	6787.750	2545.660	224.919	6.525
13	13	47077.720	19663.958	51	3.606	13067.460	0	384.900	122.930	3903.800	225.003	6.525
14	14	46752.120	19663.958	58.590	3.583	12872.340	5687.860	485.970	478.040	5092.330	225.087	6.314
15	15	44339.200	19663.958	62.720	3.529	3672.430	932.580	52.860	949.070	2836.640	225.170	6.314
16	16	44763.330	19663.958	67.100	3.451	3530.360	3.950	107.500	458.760	2062.050	225.170	6.314
17	17	45833.770	19663.958	59.230	3.417	2387.720	0	98.340	516.280	1421.630	225.170	6.314
18	18	15808.150	12808.306	39.120	3.161	2045.130	34381.530	26.360	159.760	3863.310	223.714	6.237
19	19	13275.870	12808.306	47.420	3.243	6722.950	12098.170	13.680	1067.360	5383.200	223.834	6.237
20	20	12532.940	12808.306	40.980	3.237	3737.160	3047.900	9.260	49.140	2758.100	223.869	6.237
21	21	11983.340	12808.306	51.330	3.227	1599.610	1605.810	0.600	161.210	1963.720	223.869	6.237
22	22	11012.520	12808.306	54.750	3.244	13078.900	1642.020	545.200	12725.350	5002.740	223.869	6.237
23	23	11428.250	12808.306	56.080	3.417	63622.340	687.560	358.290	44824.980	8167.670	223.869	6.237
24	24	11724.980	12808.306	48.920	3.475	16105.390	6019.520	11.330	5982.790	7043.580	223.911	6.237
25	25	10473.780	12808.306	48.160	3.597	6907.350	4354.200	16.580	1054.970	3019.840	224.055	6.237
26	26	11039.120	12808.306	46.080	3.711	6023.680	439.600	4.400	56.860	4416.820	224.199	6.237
27	27	10689.970	12808.306	51.120	3.658	24134.430	54.430	136.700	7235.690	6441.670	224.343	6.237
28	28	8148.690	12808.306	55.140	3.622	7640	0	29.430	1461.600	1824.170	224.469	6.237
29	29	9049.020	12808.306	59.970	3.611	11184.280	0	853.300	3319.200	3870.640	224.553	6.237
30	30	8734.190	12808.306	50.540	3.606	8248.470	0	769.510	774.080	3513.620	224.636	6.237

Figure 7: Prediction Model

Prediction using random forest was made for weekly sales.

VII. CONCLUSION

In this paper, we have analysed the Walmart data through a machine learning method. As we can see in the research conducted that our null hypothesis has failed i.e. all promotional offers do not have an impact on the weekly sales of any store across different locations. With the help of linear regression, we have the necessary proof that promotional offers don't affect the sales i.e. p-value of the markdowns are not significant. A blog has been maintained from the start of the project:

Week Number	Task Undertaken
1	Research for topic and dataset
2	Search for Research paper related to topic
3	Proposal Submitted
4	Received feedback for the proposal
5	Started reading research papers
6	Pre-processing of data
7	Model development
8	Start of Final Project
9	Project completed and reviewed

VII. REFRENCES

[1] M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, "Knowledge management and data mining for marketing," Decis. Support Syst., vol. 31, no. 1, pp. 127–137, 2001.

[2] S. R. Ahmed, "Applications of data mining in retail business_31," Int. Conf. Inf. Technol. Coding Comput. 2004. Proceedings. ITCC 2004., pp. 455–459, 2004.

[3] A. K. Muhammed, "Application of Data Mining In Marketing Application of Data Mining In Marketing," IJCSN Int. J. Comput. Sci. Netw., vol. 2, no. 5, pp. 2277–5420, 2013.

[4] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, "Retail business analytics: Customer visit segmentation using market basket data," Expert Syst. Appl., vol. 100, pp. 1–16, 2018.

[5] M. Kaur and S. Kang, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," Procedia Comput. Sci., vol. 85, no. Cms, pp. 78–85, 2016.

[6] C. Y. Tsai and S. H. Huang, "A data mining approach to optimise shelf space allocation in consideration of customer purchase and moving behaviours," Int. J. Prod. Res., vol. 53, no. 3, pp. 850–866, 2015.

[7] "Walmart Recruiting - Store Sales Forecasting" Available Online: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data [Accessed On: 03/02/2018]