Sales Forecast for Walmart

Introduction

This report discusses our key messages and findings from the analysis that we have done for Walmart. Walmart is one of the largest retailers in the world and it is very important for them to have accurate forecasts for their sales in various departments. Since there can be many factors that can affect the sales for every department, it becomes imperative that we identify the key factors that play a part in driving the sales and use them to develop a model that can help in forecasting the sales with some accuracy.

For this project, we have used the dataset available from 'Walmart Store Sales Forecasting' project that was available on Kaggle. In this dataset, we have weekly sales data for 45 stores and 99 departments for a period of 3 years. In addition, we had store and geography specific information such as store size, unemployment rate, temperature, promotional markdowns etc. Using these factors, we needed to develop a regression model that can forecast the sales and is also computationally efficient and scalable. The key issues that we have faced in this analysis is the large dataset that resulted into several computational challenges because of which we had to modify our approach in addressing the problem. We also faced significant challenges in identifying the right variables on which the analysis could be conducted

In this project, we conducted multiple linear regression to predict the future sales. There were several different factors that we analyzed in our regression model starting with a full model with all the variables and then moving towards a reduced model by eliminating insignificant variables. We used several different exploratory analyses to identify the key variables for our regression equation such as correlation plots, heatmaps, histograms etc.

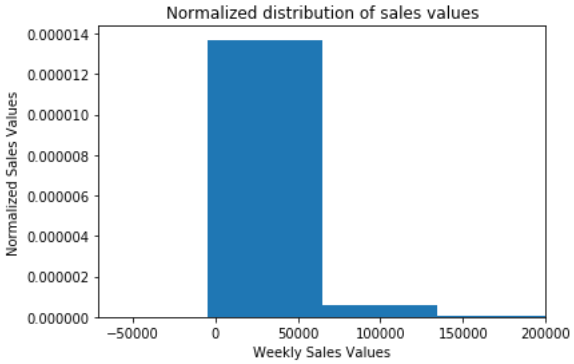
Few other time series forecasting models could have been used as the weekly sales is highly dependent on the past year. Moreover, ARIMA modelling techniques like exponential smoothening and holt winters could have helped us capture the seasonality in the model in a better way. Furthermore, ARIMAX model would have enabled us to have an accurate time series model based on previous weeks of data as well as factor in few important variables like holiday and department type to get an even better accuracy

Computational Setup / Steps

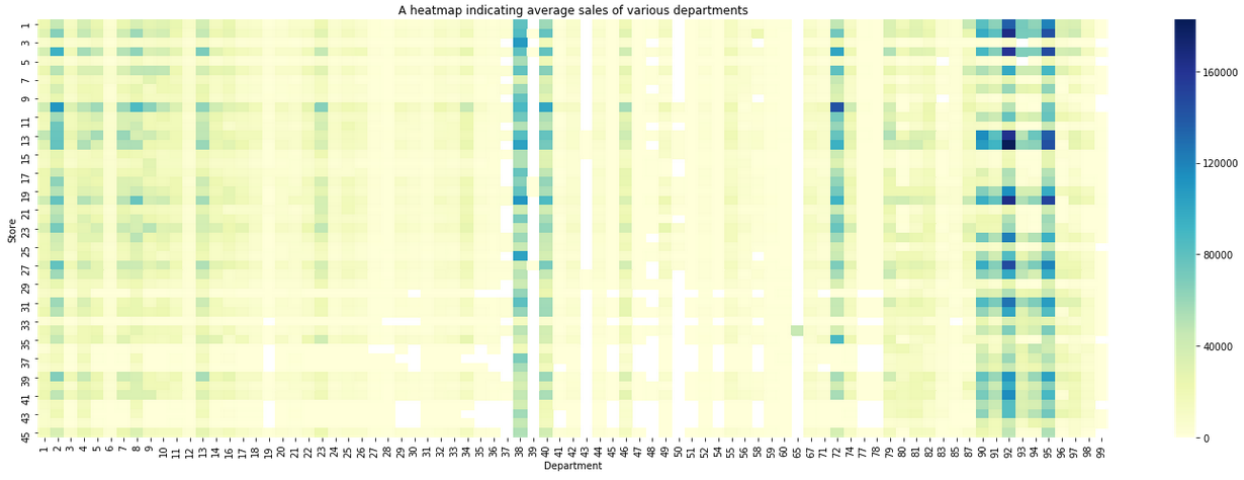
The dataset was picked up from a Kaggle competition. It basically housed three major files, the test data, the train data and the features data set. We had a host of parameters listed in it e.g. Temperature, Fuel Price, Promotional Markdowns 1-5, Inflation(CPI), Unemployment rate etc. There were around 420,000 rows for the training set which made processing it very hard. Also, as we did not have the real results of the test data in the file, we leveraged our training set and divided it into 80:20 ratio to enable us to get a prediction accuracy for the model
As our first step we focused on merging the data sets to get all the factors which could help us in predicting the weekly sales. Post that we spent some time identifying the erroneous entries in data, descriptive statistics indicated towards the possibility of negative sales value in the data. To

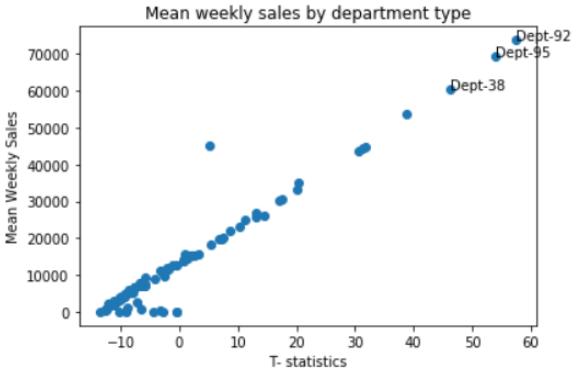
confirm and revalidate the same we plotted a normalized histogram, if we look closely we can see that there are some values which lie on the negative side of zero indicating negative sales. Therefore, we sanitized our data to drop these erroneous values to minimize the impacts of them on our analysis.

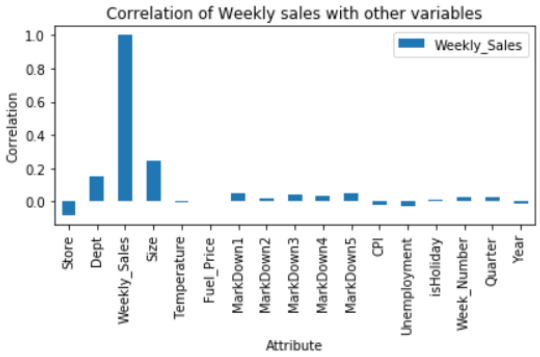
.



As our next step we created a full model consisting of majority of variables available with us, to facilitate this we created lagged variable based on timeframe (weekly lag sales), as we believe that this would help us capture the inherent attributes of particular store type and department. Moreover, we created dummy variables out of the categorical values given in the data set (e.g., department and type). The main intent behind creating them was to factor in some departments which perform very differently from the rest. If we take a closer look at the heat map and the scatter plot of mean weekly sales vs T-Statistics we can safely conclude that dept 38, 95 and 92 play a pivotal role in predicting future sales. Consequently, we decided to include them in our regression variables

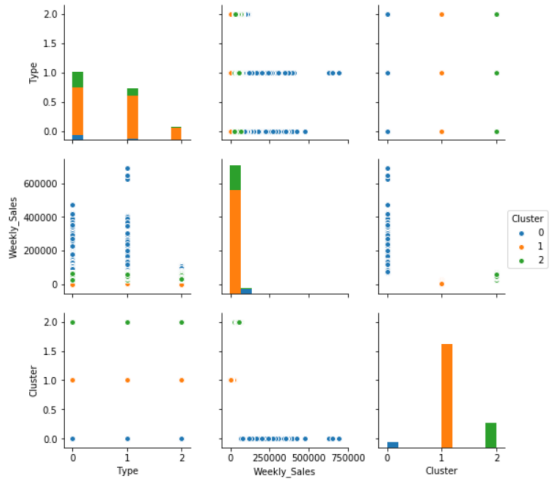


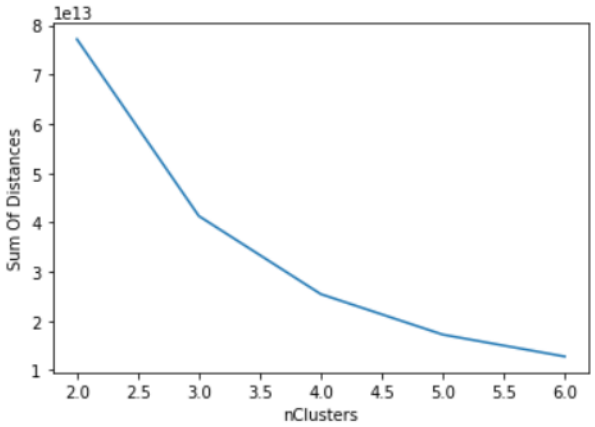




We also did various descriptive statistics. We saw the correlation matrix and found that size and department are correlated with weekly sales.

We did clustering based on Type of store and Weekly sales and found 3 clusters (based on elbow plot we decided the number of clusters). With these insights, we did the reduced model of linear regression.





COMPUTATIONAL CHALLENGES

Loops took lot of time since the dataset is huge (420 K records). Hence, we used minimal number of loops to reduce the computation time.

We faced computational challenges in deriving sales lag as explained below.

SLOWEST PART OF CODE

We created master Data Frame with all the relevant variables and added variables like weekly sales lag (previous week sales), monthly sales lag, etc. Even when we did it without loops, we found that it took lot of time because of time taken for running through all the indices. To overcome this challenge, we sorted by store, department, year and week number. After this we shifted the data by one row. Then we merged these shifted sales (i.e., previous week sales) with the master Data Frame. In order to avoid the errors that happened during this process, we used if condition and corrected the values. This gave us considerable computational challenge as well.

Clustering takes time because of the sheer size of the data. Moreover, when trying to get the elbow plot and running iterations for a range of clusters it takes much more time. The algorithm tries to zero in on the minimum distance for all the rows and repeats the process for different clusters hence demanding a lot of computational processing.

Time-analysis also takes time. We have overcome this by splitting the batches into mini-batches. This has reduced the time-taken but still it requires sufficient amount of time to run **Scaling the data**

We found that the scaling of data makes the running time go up considerably. So, we have run the OLS batch-wise. We split the batch into number of mini-batches containing 10000 records each. Then we run the OLS on each and keep changing the beta values till it stabilizes. In this way, we have overcome the time-taken for running the code. However, still it takes considerable time to run this part of the code. This time taken will increase when the data doubles, triples or quadrupled. However, the time taken is less than what it would have taken without batch-wise regression.

RESULTS

We found the regression equation when we don't use promotional markdown variables as:

$$\begin{aligned} Weekly\ sales = {} & 247.396 - (845.53 * holiday) + 9.979 * Temperature \\ & + 0.9067 * Weekly\ Sales\ lag + 698.93 * Type\_A + 5.613 * Type\_B \\ & - 457.146 * Type\_C + 2363.138 * Week\_50 + 6783.568 * Week\_51 \\ & + 5074.345 * Dept\_95 + 5462.166 * Dept\_92 + 4289.939 * Dept\_38 \end{aligned}$$

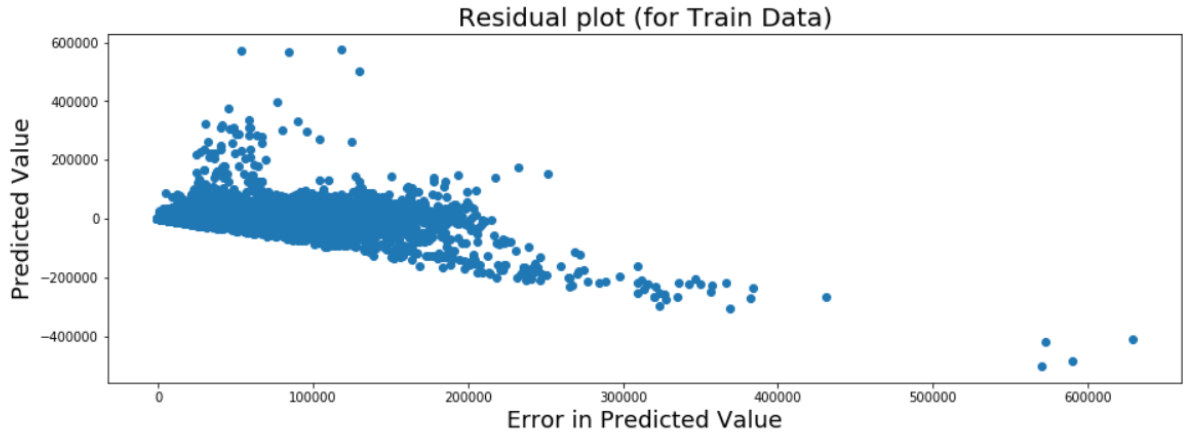
The accuracy of this model is found to be 84.56%.

When we use promotional markdown in the regression equation, the regression equation is:

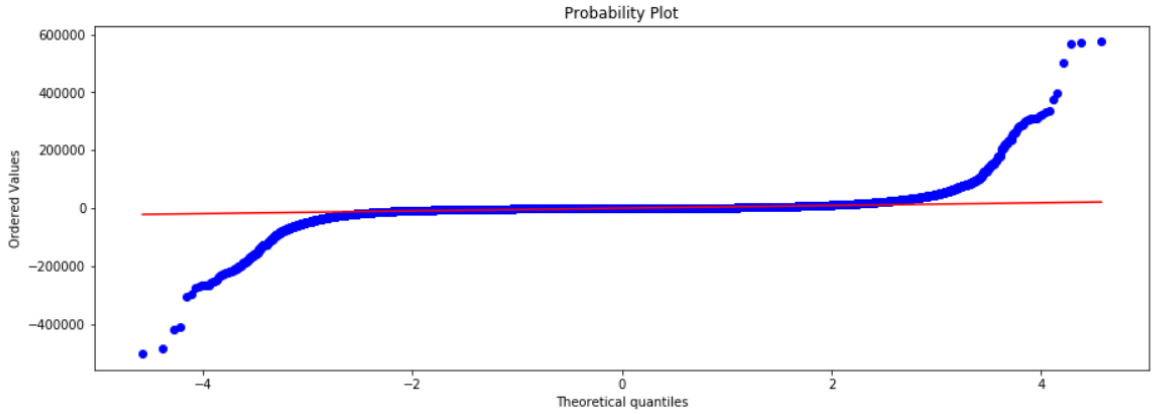
$$\begin{aligned} Weekly\ sales = {} & 37.195 - (852.324 * holiday) + 9.827 * Temperature + 0.929 \\ & * Weekly\ Sales\ lag + 523.328 * Type\_A - 72.719 * Type\_B - 413.414 \\ & * Type\_C + 2256.694 * Week\_50 + 6673.006 * Week\_51 + 4280.914 \\ & * Dept\_95 + 4809.748 * Dept\_92 + 3575.849 * Dept\_38 + 0.0938 \\ & * Markdown - 3.881e-6 * Markdown * weekly\_sales\_lag \end{aligned}$$

We get accuracy as 84.53% in this model. We find that the regression co-efficient for interaction variable (markdown $*$ weekly_sales_lag) is miniscule. The regression co-efficient for markdown is also low at 0.0938. These both again confirm the fact that Markdowns given at Walmart are not significantly impacting the sales. They will have to overhaul the Markdowns given and try new markdowns.

The residual plot shows that the errors are random and there is no pattern in that. Hence there is no heteroscedasticity.

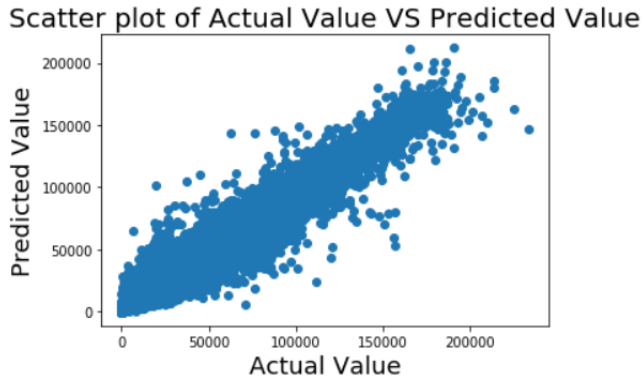


The above plot shows that the residuals have fairly similar variance. Hence we can say that the data is not having heteroscedasticity.



From the above plot, we find that the residuals are normally distributed. Hence both normality and homoscedasticity are there. Therefore, we can use the regression equation without any transformations.

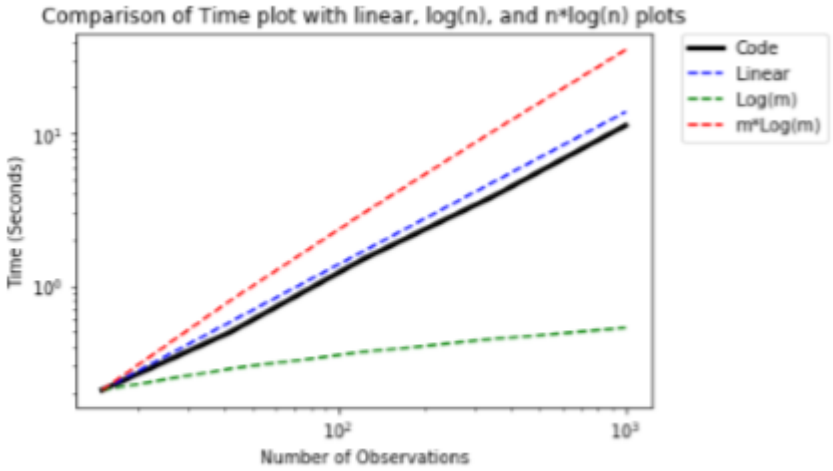
Our regression is having a high accuracy and the scatter plot confirms this:



The next major hurdle is the K-Means algorithm. As we know, adding new data points to existing cluster won't require a lot of iterations. Also, now that we have calculated the important departments, we will monitor the same departments for a long time.

The most important factor in time complexity is the running time of the linear regression. Linear regression takes $O(c^2n)$ time complexity as can be seen from the graph below. This might create a problem when the data size doubles. So, we do use iterative linear regression. In this process, use the current coefficients as the latest estimate of beta and we update them based on each new incoming row.

We have also taken care of the situation when the data comes in batches. As can be seen in the commented code in the time complexity block, we can find the updated estimate per batch in a single step.



The next analysis we did was profiling. Based on our analysis, length computation and assignment are the two main time-consuming factors. As both of these are internal commands, we conclude that the code is indeed currently optimal. We also tried minimal use of loops in our code and tried to find more efficient alternatives such as shift and map.

CONCLUSION

In conclusion, we find that our regression equation is quite accurate (84.5% accuracy) in predicting the weekly sales. Walmart can use it to forecast the sales better. They need to focus on the inventory planning of key departments like 38,92 and 95. They need to overhaul the Markdowns that are given currently as they are not having the intended impact on sales. They need to focus on the year-end inventory as week 51 and 52 play a crucial part in predicting sales.

INTERESTING PROBLEMS TO INVESTIGATE FURTHER

Walmart can analyze the entire store data across US to arrive at an even more accurate prediction. They can analyze the inventory data as well to optimize their inventory. They can analyze the sales targets and incentives that are given for employees to arrive at achievable sales targets for employees to motivate them better.