International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Retail Chain Sales Analysis and Forecasting

Prof. Vijay Jumb*, Saikumar Kandakatla[1], Shantanu Sawant[2], Chandan Soni[3]

*Assistant Professor, Department of Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India

[1,2,3]B.E student, Computer Engineering, Xavier Institute of Engineering, Mumbai, Maharashtra, India

---

Abstract - *Most organizations like to assess the up and coming exchanges. A better determining can maintain a strategic distance from them from over-evaluating or under-assessing the more extended term exchanges which brings about a phenomenal damage than the organizations. Utilizing dependable exchange estimation, organizations could allocate their properties all the more reasonably and make improved revenues.But anticipating the deals is very confounded gratitude to a few inner outer elements from the neighboring environment. Consequently inside the present commercial center, there's an earnest necessity for the advancement of a reasonable guaging framework which is quick, adaptable and may give high precision. We expect to put on different AI procedures to assemble and modify a business foreseeing show and perform estimation on bargains data to return over this necessity.We additionally will furnish a simple to utilize result with different representation instruments for the ease of clients. Our project additionally takes a shot at examination of the different parameters which could impact the deals.*

Key Words: **Weekly Sales, Forecasting, Seasonal, Trend, Remainder, Analysis.**

1. INTRODUCTION

The aim of this project is to empower class administrators of Walmart to check week by week deals of divisions. Investigation incorporates the impact of the markdown on deals, and furthermore the degree of impact on deals by fuel costs, temperature joblessness, CPI and so forth has been broke down utilizing basic and various direct relapse models.

One test of displaying retail information is the need to settle on choices dependent on constrained history. On the off chance that Christmas comes yet once every year, so does the opportunity to perceive how key choices affected the base line.In this task, Walmart organization are given recorded deals information for 45 Walmart stores situated in various locales. Each store contains numerous divisions, and we should extend the deals for every office in each store. To add to the test, chose occasion markdown occasions are remembered for the dataset. These markdowns are known to influence deals, however it is trying to anticipate which divisions are influenced and the degree of the effect.

1.1 Flow of the Project

1) Input and Output. 2) Analysis of Parameters.3) Forecasting through different Algorithm. 4) Selecting Best Algorithm and Forecasting its output.

Information of Output is for the clarification of each factor which are introduced in the datasets. Investigation of Parameters is for comprehension and clarification of leading examination process over the all dataset's sections. Choosing the factors which really shows connection, covariance or reliance over the objective factors. At that point we could process through different calculations and have nearly study them. Thus, the algorithm which gives better exact outcome is chosen as the yield.

2. Input and Output

Information gave chronicled deals information for 45 Walmart stores situated in various locales. Each store contains various divisions, and you are entrusted with foreseeing the office wide deals for each store.In expansion, Walmart runs a few limited time markdown occasions consistently. These markdowns go before conspicuous occasions, the four biggest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these occasions are weighted multiple times higher in the assessment than non-occasion weeks. Some portion of the test introduced by this opposition is demonstrating the impacts of markdowns on these occasion a long time without complete/perfect authentic information.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:

1) Store - the store number
2) Dept - the department number
3) Date - the week
4) Weekly_Sales - sales for the given department in the given store
5) 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:

1) Store - the store number
2) Date - the week
3) Temperature - average temperature in the region
4) Fuel_Price - cost of fuel in the region
5) 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.
6) CPI - the consumer price index
7) Unemployment - the unemployment rate
8) IsHoliday - whether the week is a special holiday week.

For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):

1) SuperBowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13.
2) Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13.
3) Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13.
4) Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13.

Output contains the overall forecasted sales graph with confidence intervals. With providing the accuracy and analysing the output using residuals. As we need to process each department of each store independently, we need to combine the forecasted sales of all the applied forecasting algorithms independently. Thus we can provide an single output graph which shows the average forecasted sales of Walmart.

3. Analysis of Parameters

Fig -1: CPI vs Weekly_Sales

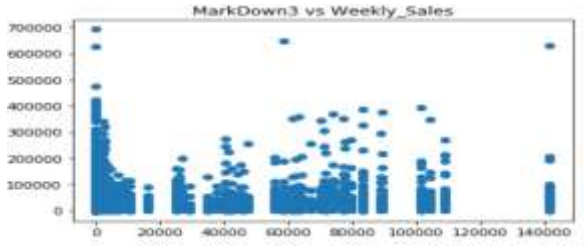

Fig -2: MarkDown3 vs Weekly_Sales

As we can see in graphs there isn't any variable which shows normal distribution. Similarly, other variables also show same status. And it could be seen clear to have the variables are mostly independent on the output variable. Hence, we could further check the correlation matrix.

Correlation of the variables and VIF values are shown as follows:

CPI	-0.020930
Dept	0.148077
Fuel_Price	-0.000124
IsHoliday	0.012762
MarkDown1	0.047161
MarkDown2	0.020703
MarkDown3	0.038554
MarkDown4	0.037457
MarkDown5	0.050452
Store	-0.085213
Temperature	-0.002318
Unemployment	-0.025838
Weekly_Sales	1.000000
store_b	-0.131224
store_c	-0.095393
store_size	0.243856

Fig -3: Correlation of Variables

```
for  CPI  value of vif is: 1.28
for  Dept  value of vif is: 1.0
for  Fuel_Price  value of vif is: 1.17
for  IsHoliday  value of vif is: 1.16
for  MarkDown2  value of vif is: 1.11
for  MarkDown3  value of vif is: 1.09
for  MarkDown4  value of vif is: 1.14
for  MarkDown5  value of vif is: 1.19
for  Store  value of vif is: 1.39
for  Temperature  value of vif is: 1.21
for  Unemployment  value of vif is: 1.2
for  store_b  value of vif is: 2.46
for  store_c  value of vif is: 2.75
for  store_size  value of vif is: 3.1
```

Fig -4: VIF values of Variables

Correlation of the variables are clearing showing that most off variables are not correlated to the output variable. Also they are also not correlated with each other as well.

Except MarkDown3 shows some correlation with MarkDown5. Second figure contains VIF values.

VIF (Variance Inflation Factor) which could be commonly used for checking of the any variable whose variance is affected by any other variable. Here as no covariance could also be obtained.

Thus, we couldn't extract any variable which could be use of for further processing. We could process with the Time Series algorithms. Which could be useful for the data which contains patterns of seasonal and trends.

4. Time Series Algorithms

4.1 Seasonal Naive:

Accepting gauge to be equivalent to the last watched an incentive from a similar period of the year [1] (e.g., that long stretch of the earlier year). Officially, the figure for time T+h is composed as:

$$Y_T = Y_{T-m}$$

Where m, represent same season value of previous year. Applying this model, we can get an unmistakable thought regarding how much impact is going on in the everyday deals. More the change the information less will be precision of the model. Consequently, on the off chance that there is genuine less changes happens during earlier year, at that point this model could be the best and straightforward model.

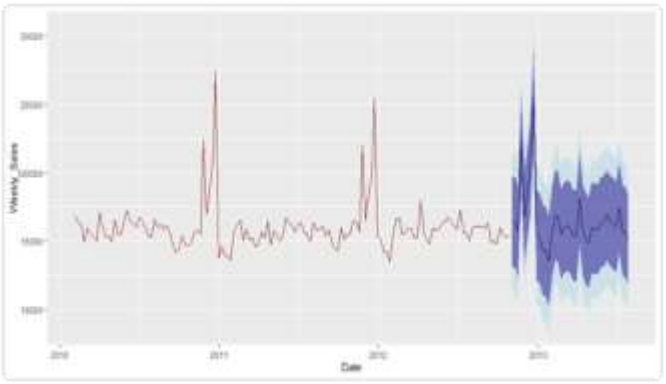

Fig -5: Forecasting through snaive

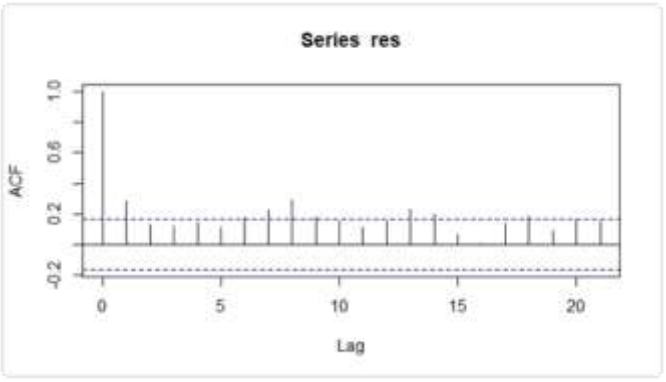

Fig -6: ACF plot of snaive

4.2 Time Series Linear Model:

Time Series Linear Model is only a direct relapse model acquired by applying occasional and pattern variable to subordinate variable.

$$y = B_0*\text{Trend} + B_1*\text{Season} + B_2 + E$$

Here, y would be Weekly_Sales, where B_0, B_1 are the coefficient of the pattern and occasional part, B_2 is the intercept and E will be the error rate related with it.
Trend – Its, only an expanding number as recurrence continues expanding. Season – It is cyclic variable which turn over the recurrence through length of the preparation information. Showing that each present worth could connect with its past worth[1].

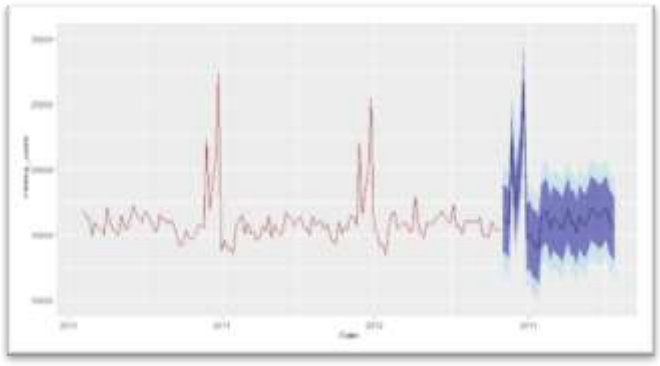

Fig -7: Forecasting through tslm

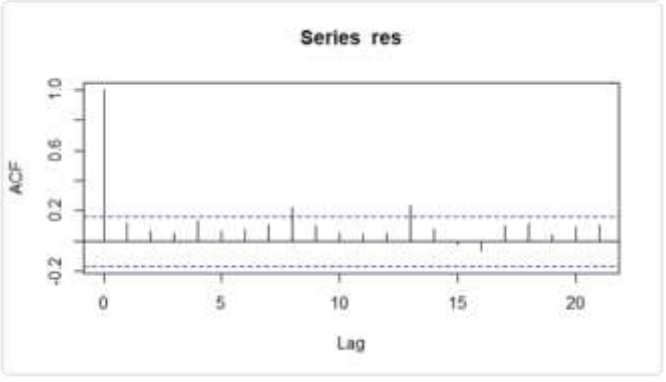

Fig -8: ACF plot of tslm

4.3 ARIMA (p, d, q):

Arima model is a blend of 2 regressive models with differencing included they are AR (autoregressive model) and MA (moving normal model).

Autoregressive model is only a relapse applied on the present an incentive as reliant variable with past p – number of past lagged value as autonomous factors.[2]

AR(p)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t$$

Moving Average model is additionally a relapse applied on the present an incentive as reliant variable with past q – number of past lagged errors as free factors.[2]

MA(q) --

$$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

Thus, ARMA will be obtained as:[2]

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

Consequently, Arima model will got by consolidating ARMA model with differencing(d)[2].

Here, we have to assess the three parameters they are p, q, and d. This, parameters can really be assessed naturally through auto. arima() work present in R[4]. Likewise, it can also deal with applying the seasonal Arima model as well.

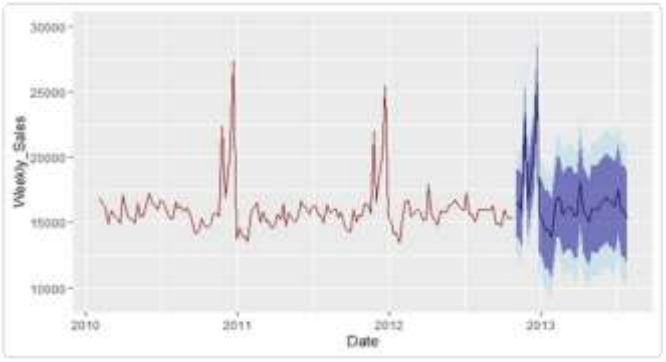

Fig -9: Forecasting through ARIMA

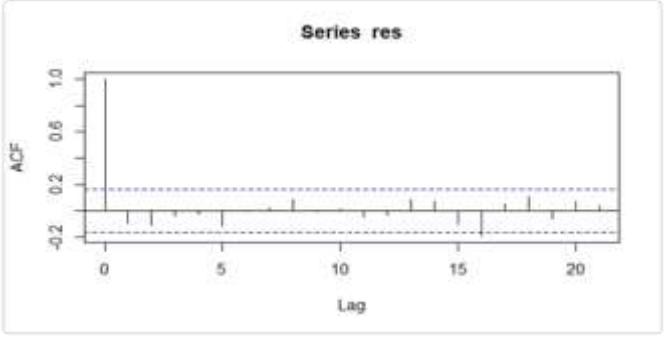

Fig -10: ACF plot of ARIMA

4.4 STLF (Seasonal and Trend decomposition using Loess):

STL is a flexible and vigorous strategy for breaking down time arrangement. STL is an abbreviation for "Regular and Trend decay utilizing Loess", while Loess is a strategy for evaluating nonlinear connections. The STL technique was created by Cleveland, Cleveland, McRae, and Terpenning (1990)[3].

The two primary parameters to be picked when utilizing STL are the pattern cycle window (t.window) and the regular window (s.window). These control how quickly the pattern cycle and occasional parts can change. Littler qualities take into account increasingly fast changes. Both t.window and s.window ought to be odd numbers; t.window is the quantity of back to back perceptions to be utilized while evaluating the pattern cycle; s.window is the quantity of sequential years to be utilized in assessing each an incentive in the occasional segment. The client must determine s.window as there is no default. Setting it to be vast is comparable to constraining the occasional segment to be intermittent (i.e., indistinguishable across years). Indicating t.window is discretionary, and a default worth will be utilized in the event that it is discarded[3].

$$y_t = T_t + S_t + R_t$$

STL function could be used for the forecasting function as well. The decomposition obtained through STL could be used for the various other algorithms. These algorithms could be ARIMA or ETS.

STLF – ETS: ETS is an exponential smoothing. ETS contains Seasonal, Trend and remainder could be none, additive or multiplicative. Also, trend could be applied as damped or non – damped. Thus, in this way, the various stable combination of all model will be obtained as 21 models. Also, best model among those is selected using the BIC, AICC, $AICC_c$. [3]

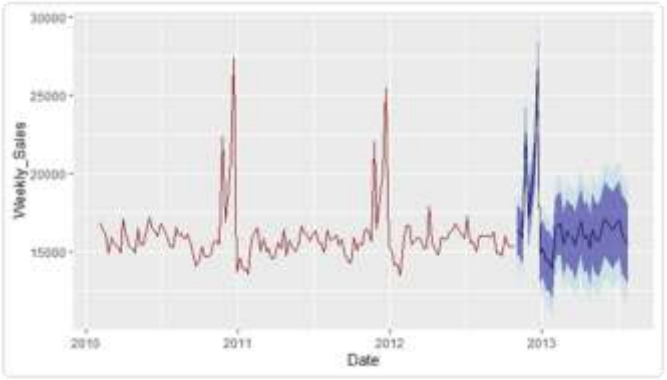

Fig -11: Forecasting through STLF - ETS

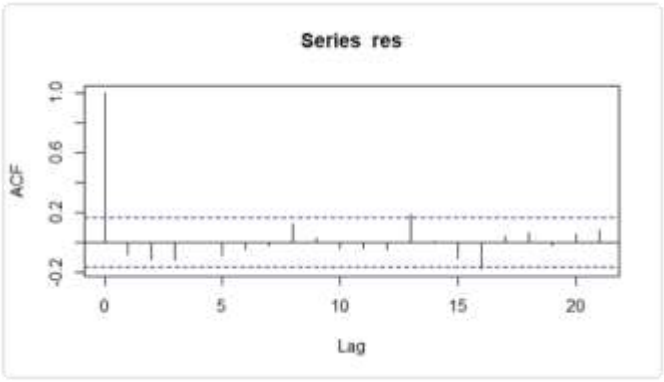

Fig -12: ACF plot of STLF - ETS

STLF – ARIMA: STLF when added to the arima the seasonal component would be handled by the STL itself. Then, ARIMA would be processed on the seasonally adjusted data. The confidence intervals and accuracy measure would be obtained through the arima model[3].

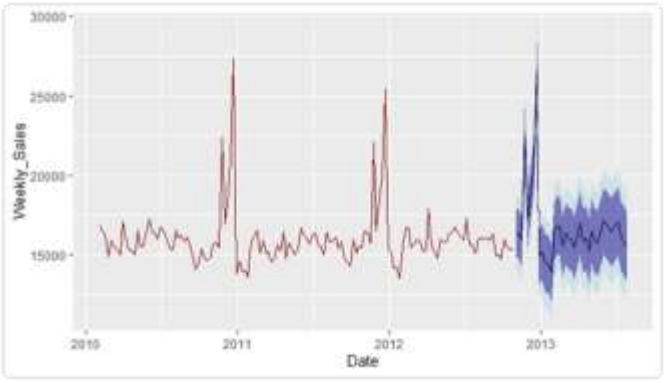

Fig -13: Forecasting through STLF - ARIMA

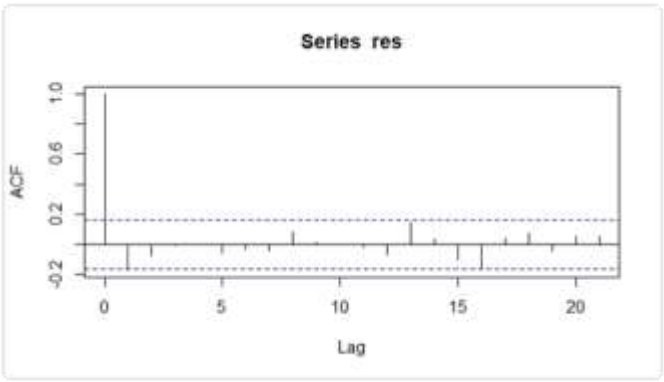

Fig -14: ACF plot of STLF - ARIMA

5. Selecting the Best Algorithm

Accuracy Measures: Here, Accuracy measure is taken by MAE (Mean Absolute Error).
MAE is calculated as :

$$T^{-1}\sum_{t=1}^{T}|y_t - \hat{y}_{t|t-1}|$$

The purpose of taking MAE over RMSE – We need to calculate the average accuracy of all the forecasting models iterated over each department of each store. Thus, information carried through MAE would not be changed by averaging them. Larger the MAE value lowers the accuracy of the model.

The ACF Residual plot obtained of every algorithm suggest there any presence of information present in the residuals. Thus, checking the ACF plot with their given bounds we can have an intuition of information presence[1].

Comparison for every algorithm:

Seasonal Naïve shows a very large MAE value. Suggests that there would be a more variability in the data. Thus, the data changes in day to day. The ACF Residual plot also shows information presence in residuals at lag 1 and at higher lags. Also, Seasonal Naïve takes less time approx. 3 mins.

MAE value of snaive -- 1765.8599

TSLM works properly over the data. Much better than the seasonal naïve and other models, but it couldn't get extract all the information present in data. Which could be seen in the residual plots at 8 and 13 respectively. It also takes only 4 mins to run the process.

MAE value of tslm -- 860.569

ARIMA actually takes a lot of time as it runs as auto.arima() which select p, q and d variables automatically according to the data. This leads to lot of computations and it takes approx. 17 mins to run the algorithm. Varies accuracy of the algorithm is also lesser than the TSLM.

MAE value of ARIMA -- 875.32

STLF – ETS could be better process. But Ets mostly works for short term forecasting. As we need to forecast for longer term period there would be losing of the accuracy. Hence it has lower accuracy then the STLF – ARIMA model. The residual plot also suggest that lags 13 and 16 there is an small amount of information would be lost during process. This model takes lesser time ARIMA model only 3 mins also accuracy better than TSLM model.

MAE value of STLF – ets - 782.749
MAE value of STLF – ARIMA - 760.665

STLF – ARIMA is actually could be selected as an best model. It extracted all the information and presenting the residuals as white noise which makes the model to be more accurate. But it takes higher time then STLF -- ETS model up to 7 mins.

Hence, it would be better to select STLF – ARIMA model as the best model with accuracy of 760.665 and 7 mins to process whole dataset.

6. CONCLUSION

The finish of our undertaking is to differentiate the effect on sales throughout various vital choices taken by the company . The examination is performed on authentic deals information across retail locations situated in various areas. Examination incorporates the impact of the markdown on deals, and furthermore the degree of impact on deals by fuel costs , temperature, joblessness , CPI then on has been broke down utilizing basic and various straight relapse models Analytical apparatuses utilized in venture are RStudio. Predictive analytics is at the guts of supply chain process that helps Wal-Mart reduce overstock and stay properly stocked on the foremost in-demand products.

Providers to retail organization are required to utilize the continuous seller stock administration framework that causes them limit the stock for a specific item if there are no huge deals for it. This causes retailers to spare assets to purchase items that have more noteworthy request and have expanded likelihood for more prominent benefits.

REFERENCES

[1] Forecasting: Principles and Practice second edition by Rob J Hyndman and George Athanasopoulos, University of Western Australia.
[2] Forecasting: Principles & Practice first edition by Rob J Hyndman 23-25 September 2014 University of Western Australia.
[3] STL : A Seasonal - Trend Decomposition procedure based on Loess by Robert B. Cleveland, William S. Cleveland, Jean E. Mcrae, Irma Terpenning.
[4] Forecast v8.11 package documentation by Rob J Hyndman, University of Western Australia.