Sales Forecasting

First **Arnav Singhal**, Second **Naman Tiwari**, and Third **Muskan Garg**

***Abstract*—**Retail markets are the market where the consumer goods or services are sold via multiple channels to consumers. According to Statista, the retail market holds 5.7 % of the total GDP of the United States of America[1]. Walmart is one of the industry leaders in the Retail market. So we are doing the Stock prediction of the Walmart sales weekly. Sales of any store is usually very much affected by socioeconomic factors. Considering this point we have considered employment, CPI and fuel prices as three of the input features to the model. In this project, I am going to predict the sales of different stores and the aggregate sale of Walmart. I am using the vector autoregression models as well as the Long Short Term Memory based neural network too, according to the studies in a lot of cases LSTM tends to perform better[2] specially when the patterns are complex. Vector autoregression model takes multiple series into account for prediction of a particular time series. The main objective is to focus on the prediction of the sales of Walmart and the influence of the other factors like 'Temperature', CPI, Unemployment and Fuel Prices are taken into account. The whole point of using multiple models is to make sure that the final model is lite in terms of computation as well as good enough in terms of performance. Later the model will be deployed locally as a GUI web application so that it can be used in a practical environment as well as the response can be used to retrain it in future.

***Index Terms*—Sales forecasting, Effect of socioeconomic factors on sales of a product, ARIMA model, SARIMA model and LSTM model**

I. Introduction

TODAY the market is not the same as it was centuries ago, the options are a lot and hence there is a lot of competition for survival of the business, the competition is not only for selling the maximum amount of goods but also to optimize the investment in the business, so that maximum revenue can be generated by using minimum investment possible. There are some strategies used by every business, one of them is management of advertisement campaigns and the other is inventory management.

To run effective ad campaigns it is important to know your customer and it is important to know the pattern of the sale of your product. For knowing the customer we generally perform customer segmentation, but to know the patterns in the sale of your product and to understand the highs and lows of the demand forecasting is a very efficient way. In our use case the forecasting is being done for Walmart, the data being explored for the project happens to contain the store wise information of sales. If we explore the data with respect to the sales column, we notice that the performance of each store is not the same, so we can use the sales forecasting to potentially increase the sales of selective stores as well.

Another use case related to sales forecasting is to manage the stock of the goods in the inventory so that the waste of goods can be minimised is a very crucial factor. The businesses who are lacking in inventory management may not make a good profit or they may even struggle to survive in the market. To make sure you always have a sufficient amount of goods in your inventory, you have to predict the possible demand in the future. The time period of forecasting may vary for different businesses based on the type of goods they sell and the type of market they are in. For a bakery store the timestamp is smaller in comparison to the grocery store. One of the ways to predict the demand is by sales forecasting.

Apart from the fact that we can optimise the sale itself and boost the effect of ad campaigns, we can exploit the other features like CPI, Unemployment and Temperature to make better offers and deals for the customers to make them attached to the business. But in this case we will stick to the sales forecasting only and try to utilise these features in predicting sales.

II. Data Description

The Walmart dataset consists of three different csv files namingly “train.csv”, “features.csv” and “stores.csv”. Train.csv file contains “Store”, “Dept”, “Date”, “Weekly_Sales” and “IsHoliday”. The description of the attributes in the train.csv data file are described below;

- ➔ Store: store numbers,
- ➔ Dept: The department numbers,
- ➔ Date: The weekly dates,
- ➔ Weekly_Sales: The weekly sales aggregated on corresponding dates in the region.
- ➔ IsHoliday: The True and False values based on the fact that if it is a week of special holiday or not.

“features.csv” contains "Store", "Date", "Temperature", "Fuel_Price", "MarkDown1", "MarkDown2", "MarkDown3", "MarkDown4", "MarkDown5", "CPI", "Unemployment", "IsHoliday". The detailed description of the attributes mentioned above are provided below;

- ➔ Store: ID of the stored
- ➔ Date: Weekly dates
- ➔ Temperature: Temperature aggregated on the corresponding date in the region
- ➔ Fuel_Price: Fuel price aggregated on the corresponding date in the region
- ➔ MarkDown1-5: Data related to promotional markdowns, these columns are not very much described they are quite anonymous
- ➔ CPI: The Customer Price Index
- ➔ Unemployment: The unemployment rate
- ➔ IsHoliday: If it is a special holiday week or not

The “store.csv” file contains features like “Store”, “Type” and “Size”. The Store column contains the store numbers, type and size columns contain the type and corresponding size of the store.

III. Related Work

The most popular techniques for current time series forecasting are using ARIMA (statistical domain technique) or Neural Networks (machine learning domain). The fact that in cases of stock prices and sales forecasting the current values are related to the past stock prices and sales respectively, makes ARIMA models quite suitable for these situations. Taking the advantage of the relationship of the current data point to the past data point ARIMA models can come up with very reasonable and good results. For example the Nokia stock price prediction done in a paper by Adebiyi [5] in 2010 shows a very accurate forecasting. On the other hand, the ability of neural networks to represent and learn complex relationships in nonlinear data makes them an even better choice for time series analysis involving nonlinear patterns. For example in many scenarios where the time series is assumed to be linear and ARIMA is applied the results come to be quite unsatisfactory, because in such cases there are other socioeconomic factors like recession, unemployment affect the prices a lot and ARIMA can not represent these anomalies precisely. Another work of the same author [5] shows the forecasting of Dell stock prices where both the models are used and the Neural network clearly outperforms the ARIMA model. In our case we are going with three different models assuming that our data might have seasonality and other socioeconomic factors like CPI, Unemployment and fuel prices might be affecting the sales.

The dataset dataset we are working on has been explored and worked on by Stojanović, Nikola, Marina Soldatović, and Milena Milićević in 2014, [6] they have done the forecasting on various models like SVM, Neural Network, W-Isotonic regression, Linear regression and KNN model. Although the performance achieved a good accuracy in the use case but the work was not application oriented, the performance achieved by the paper is mentioned below;

Model	Absolute error
SVM (Support Vector Machine)	11.517,09
Neural Network	11.899,511
W-Isotonic Regression	14.1807,728
Linear Regression	14.528,238
K-NN (k-Nearest Neighbor)	20.633,996

Table 1: Performance of various models in paper[6] published in 2014

The above result clearly shows that the performance of the Neural network is quite promising, to make improvements on the model we are going to include the other parameters like Unemployment, CPI, Temperature etc. Assuming this will add value to the model and we might achieve better results. Another

difference is that we are also using the ARIMA model to create another version of the model. The work we have encountered so far are not application based, our work will focus on the application perspective of the forecasting.

III. Block Diagram of the Work

The Block diagram of the tentative work is shown below;

A. *ARIMA and SARIMA*

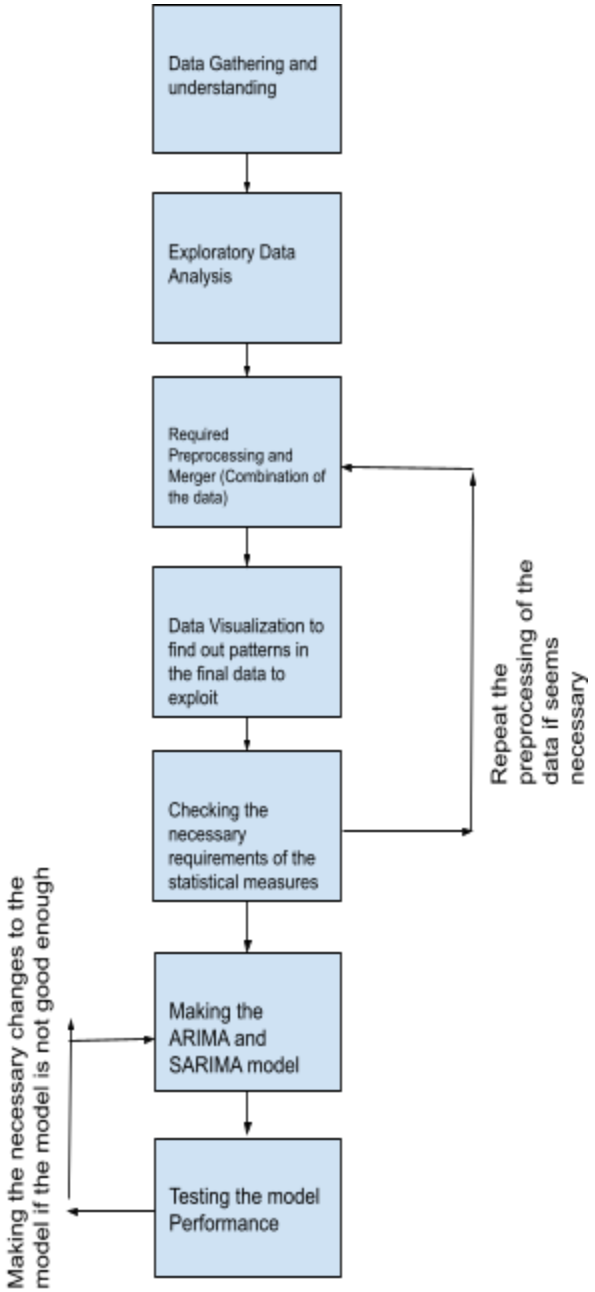

Figure 1.1: Block Diagram for ARIMA and SARIMA models

B. *Neural Network Model*

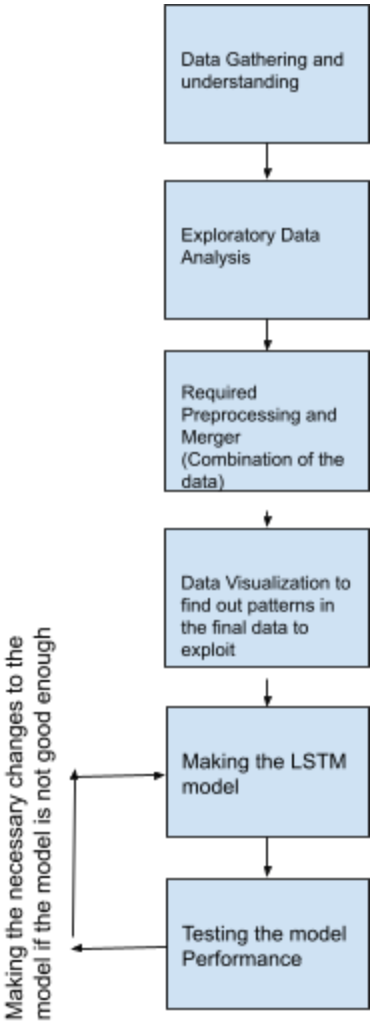

Figure 1.2: Block Diagram forNeural Network models

IV. Exploratory Data Analysis

Before making the model we need to know each and every possible patterns and relationships between each and every feature. The difference between classical machine learning methods and forecasting is that, in forecasting, the most important column is the time or date column. Date column is used to maintain the sequence of the data points so that we can relate the current sale with the sale of past days. Apart from that the analysis of weekly sales will be done with respect to the date column. We will perform the exploratory data analysis on the raw data to find out if there is any visible pattern in the columns.

A. Train.csv

The plot of Weekly_Sales against Date column is being represented in the following column.

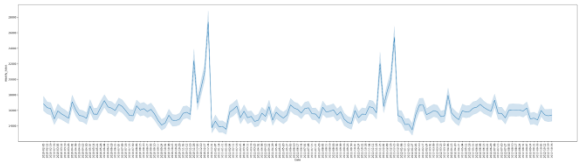

Figure 2: Variation of weekly sales with respect to date

From the above figure the pattern of the sale is clearly visible, we can notice that the sales have a tendency to repeat over the year, and there are abrupt increases in sales over a specific time duration for each year. According to the data description of the source this abrupt change can be explained by the fact that the sales increase on the time of festivals, the festivals mentioned in the following list;

- Super Bowl: 12-Feb-2010, 11-Feb-2011, 10-Feb-2012, 8-Feb-2013
- Labor Day: 10-Sep-2010, 9-Sep-2011, 7-Sep-2012, 6-Sep-2013
- Thanksgiving: 26-Nov-2010, 25-Nov-2011, 23-Nov-2012, 29-Nov-2013
- Christmas: 31-Dec-2010, 30-Dec-2011, 28-Dec-2012, 27-Dec-2013

We see the shaded region because of the different stores and their corresponding sales, the hard line shows the mean value of the sales from all the stores on that date.

B. Features.csv

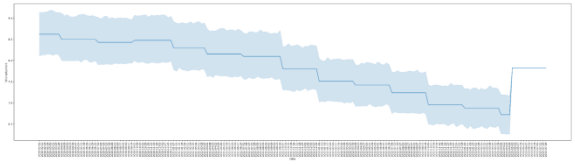

Figure 3: Variation of Unemployment with respect to the date

The above figure shows the variation of Unemployment with respect to date, we can witness that the unemployment rate is decreasing over the years. In the second half of the year 2013 the unemployment increased very rapidly and stayed the same after that point. Other than the abrupt increase in 2013 there is not other abrupt change.

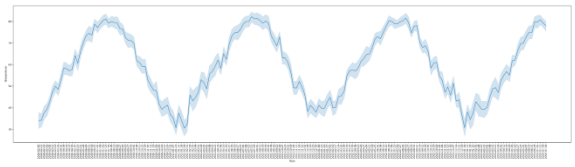

Figure 4: Variation of Temperature with respect to date

The above figure shows the variation of temperature over the dates, and we can clearly notice that the temperature increases and decreases in a year, and the same pattern repeats every year. But the interesting fact is that the lowest value of temperature is not the same for each year it has increased overall. The shaded region represents the different temperature in different regions of storage, but the pattern is the same for all of them.

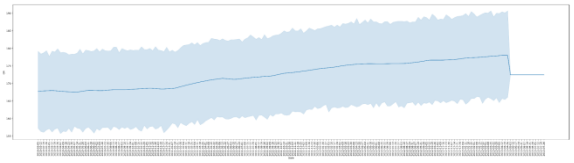

Figure 5: Variation of CPI with respect to date

The above figure shows the variation of CPI over the years and we can notice that opposite to unemployment CPI is increasing over the time. In year 2013 just like Unemployment, CPI is also going through abrupt change, the change in case of CPI is not in positive but opposite direction. The possible reasons of inflation in CPI are addressed in various papers over the year, some of the main factors are corruption, economic growth and foreign investment.[3]

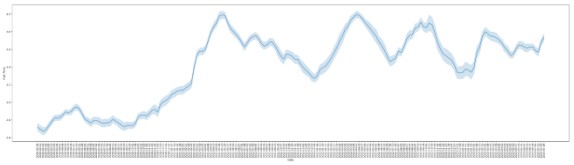

Figure 6: Variation of Fuel Prices with respect to date

The fuel price is appearing to increase over the year, there are ups and downs in each year but overall the price is increasing, this trend can be backed by the real reasons from the real world. Fuel price affect almost every business in many direct or indirect ways[4], the manufacturing and transport majorly depend on the fuel, so fuel price must potentially affect overall sale of a business.

C. *Stores.csv*

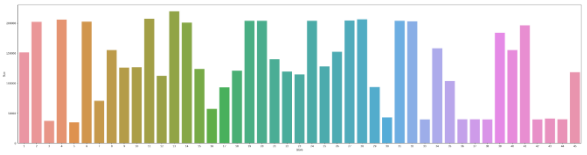

Figure 7: Variation of Fuel Prices with respect to date

Stores file does not contain a lot of significant features apart from store number and size. The plot in figure number 6, x axis holds the store number and y axis holds the size of the store.

D. *Combined Data*

After merging the train, feature data frames and aggregating the data over dates, we get the average Weekly sales and other features. The following table holds the first five rows of the final data frame (aggregate version).

Date	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Size
2010-02-05	16836.12199	FALSE	33.27794	2.717869	-999.0	-999.0	-999.0	-9999.0	-9999.0	167.398405	8.576731	137430.535364
2010-02-12	16352.05603	TRUE	33.36181	2.696102	-999.0	-999.0	-999.0	-9999.0	-9999.0	167.384138	8.567309	137622.780785
2010-02-19	16216.65899	FALSE	37.03831	2.673666	-999.0	-999.0	-999.0	-9999.0	-9999.0	167.338966	8.576351	137278.637219
2010-02-26	14899.54988	FALSE	38.62956	2.685642	-999.0	-999.0	-999.0	-9999.0	-9999.0	167.691019	8.561375	137346.344629
2010-03-05	15921.01727	FALSE	42.37398	2.731816	-999.0	-999.0	-999.0	-9999.0	-9999.0	167.727351	8.572689	137577.841033

Table 2: Aggregated (over Date) merged(features + train) dataset

Apart from this aggregate version we have created a similar data frame for each and every store of Walmart. Now we will try to make some generalisations so as to finalise the attributes which are going to be used in the modelling. Our primary motive will be to include as many features as we can but, the final decision will be dependent on the fact that if the feature is somehow related to the Weekly Sales feature or not.

Now we will perform the exploratory analysis on the merged dataset, this is important to find out the relationship of the other columns with the Weekly_Sales column.

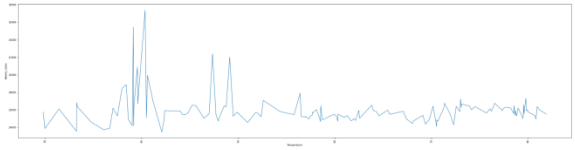

Figure 8: Variation of Weekly_Sales with respect to temperature

The value of sales is higher when the temperature is lower, but the difference is not very significant.

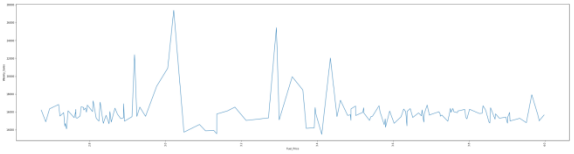

Fig 9: Variation of Weekly_Sales with respect to Fuel_price

The graph of Weekly sales with respect to fuel price has a similar pattern as of temperature, the sale amount is higher when the fuel price is lower. But again the pattern is not very significant.

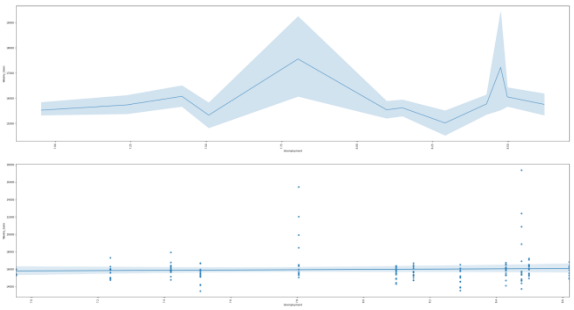

Fig 10: Variation of Weekly_Sales with respect to Unemployment rate (lien plot and regression plot)

The graph of Unemployment rate vs Weekly_Sales does not show any significant patterns in the sale variation. Although the Sales appear to be higher when the Unemployment rate is higher. The regression plot confirms the hypothesis that there is no very significant relationship between Unemployment rate and Weekly_Sales The regression line is horizontal.

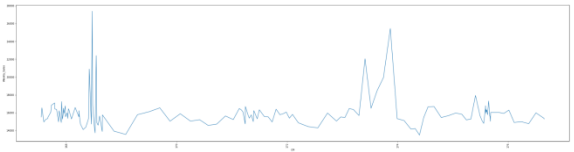

Fig 11: Variation of Weekly_Sales with respect to CPI rate)

The barplot of IsHoliday column with respect to the Weekly_Sales mean column shows us that the Sale on the day of holidays is higher than the normal days, this pattern can also be noticed in the overall plot of Weekly_Sales with respect to time in *Figure 1*.

V. Stationarity of the Data

For the autoregressive models, it is a requirement for the time series to be stationary, otherwise the results might not be of the required goodness. So in our project we have also used various techniques to check if the data is stationary or not.

Autocorrelation and partial autocorrelation are measures of the connection between current and prior series values that tell which previous series values are best for projecting future values. You may use this information to determine the order of processes in an ARIMA model.

More particularly, **The autocorrelation function** is a method of calculating the correlation between two variables (ACF). At lag k, this is the correlation between a sequence of data separated by k periods.

The partial autocorrelation function is a function that calculates the correlation between two variables (PACF). This is the correlation between series values at lag k that are k intervals apart, accounting for the interval values in between.

The autocorrelation and and partial autocorrelation plots are shown below.

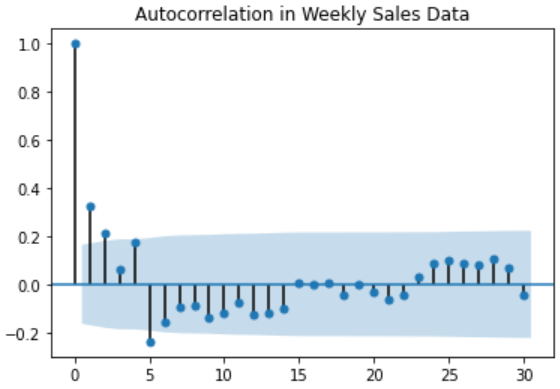

Fig 12: Autocorrelation of weekly sales according to a lag of 30

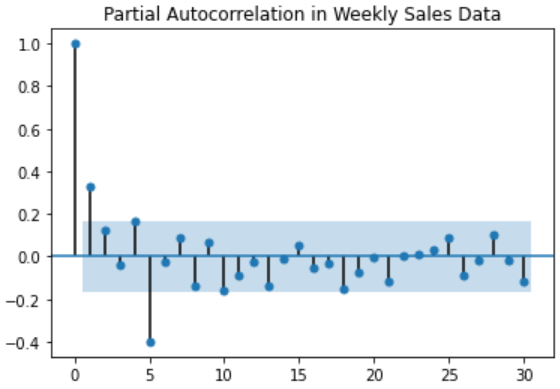

Fig 13: Partial Autocorrelation of weekly sales according to a lag of 30

Now it is the time to check if the data is having any trends, seasonality or residuals, the decomposition of the weekly sales with respect to the trend, seasonality and residual lead to the generation of the following graphs. In the graphs shown below we can

clearly see that the data is usable and we don't need to impose a lot of transformation on it, to verify this hypothesis we will also use the dikkifuller test [8]. But before that lets check out the graphs below.

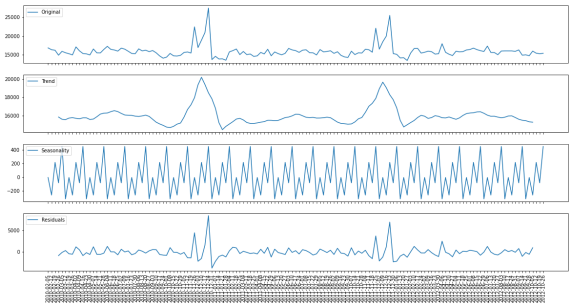

Fig 14: Figure showing the trend, seasonality and residual in the data

Now let's check for the p-value of the dikkifuller test as well [7], the test results are shown below.

```
Test Statistic           -5.930803e+00
p-value                   2.383227e-07
Lags Used                 4.000000e+00
Observations Used         1.380000e+02
Critical Value (1%)      -3.478648e+00
Critical Value (5%)      -2.882722e+00
Critical Value (10%)     -2.578065e+00
```

To support the results shown above we will also use the graph showing the rolling mean and rolling standard deviation as well.

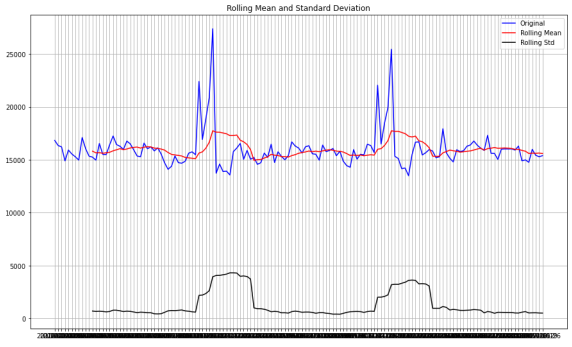

Fig 15: Figure showing the trend, seasonality and residual in the data(Weekly_sale column)

The change in the rolling mean and rolling standard deviation is only occurring when the holidays occur, so we can assume that the goodness of the dataset is sufficient for the model making. But for the sake of excellence we will look for the similar information of the weekly sale difference as well.

```
Test Statistic           -6.684617e+00
p-value                   4.256972e-09
Lags Used                 7.000000e+00
Observations Used         1.340000e+02
Critical Value (1%)      -3.480119e+00
Critical Value (5%)      -2.883362e+00
Critical Value (10%)     -2.578407e+00
```

The graph supporting the above values is shown below;

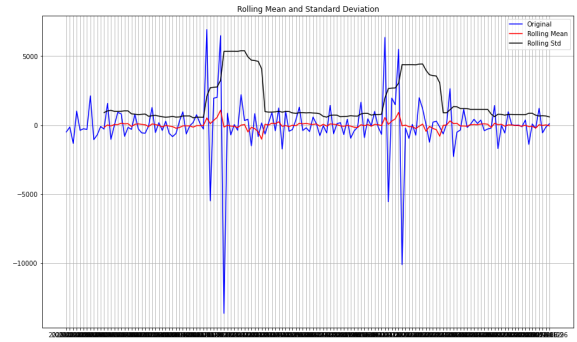

Fig 15: Figure showing the trend, seasonality and residual in the data(diff_1 column)

Now since we are highly familiar with the data and have ensured that the data is good to use with the SARIMA model, we will proceed to train the model now. In the following section we will look at the trained model specifications of the model.

VI. Make and train Sarima Model

The following standard inputs will be encountered when calculating our ARIMA model:

Order(p,d,q):

- p denotes the number of AR words.
- d is the number of times we would change our data.
- q denotes the number of MA words.

When working with SARIMA models, the letter 'S' stands for ‘seasonal’, and we have the following standard inputs:

order by season (p,d,q):

- p represents the number of AR words in terms of seasonal lag.
- d is the number of times our seasonal lag would be different (as seen above)
- In terms of seasonal lag, q is the number of MA periods.
- s denotes the number of seasons in a year.

There are some rules we have to follow while training the sarima models;

1. A larger degree of differencing is likely to be required if the series has positive autocorrelations out to a large number of lags.

2. If the lag-1 autocorrelation is zero or negative, or if the autocorrelations are all minuscule and patternless, the series does not require a greater degree of differencing. If the lag-1 autocorrelation is -0.5 or above, the series may be overdifferenced. WARNING: DO NOT EXCEED IN DIFFERENTIATION!!
3. A model with no differencing orders assumes that the original series is stationary (mean-reverting). A one-order differencing model assumes that the original series has a constant average trend (e.g. a random walk or SES-type model, with or without growth) [9]. A two-order total differencing model presupposes that the original series has a time-varying trend (e.g. a random trend or LES-type model)

The outcome of the training of the SARIMA model is shown below;

Dep. Variable:	Weekly_Sales	No. Observations:	143
Model:	SARIMAX(4, 0, 0)x(0, 1, 1, 12)	Log Likelihood	-1194.809
Date:	Mon, 15 Nov 2021	AIC	2403.617
Time:	13:02:59	BIC	2423.744
Sample:	02-05-2010	HQIC	2411.796
	- 10-26-2012		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	140.9895	103.135	1.367	0.172	-61.152	343.131
ar.L1	0.4593	0.082	5.583	0.000	0.298	0.621
ar.L2	0.1524	0.120	1.275	0.202	-0.082	0.387
ar.L3	-0.0484	0.110	-0.442	0.659	-0.263	0.166
ar.L4	0.1690	0.073	2.306	0.021	0.025	0.313
ma.S.L12	-0.7973	0.173	-4.600	0.000	-1.137	-0.458
sigma2	5.667e+06	0.008	7.18e+08	0.000	5.67e+06	5.67e+06

Ljung-Box (Q):	43.43	Jarque-Bera (JB):	89.16
Prob(Q):	0.33	Prob(JB):	0.00
Heteroskedasticity (H):	0.27	Skew:	0.72
Prob(H) (two-sided):	0.00	Kurtosis:	6.77

The diagnostics of our model is shown in the following graphs;

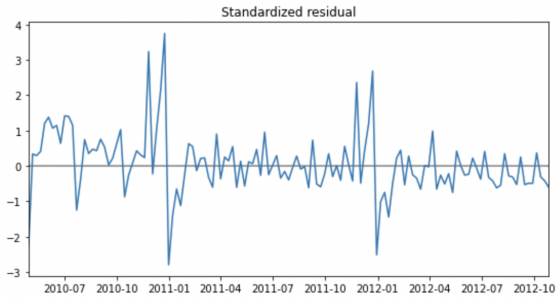

Fig 16: Figure showing the standard residual in the model

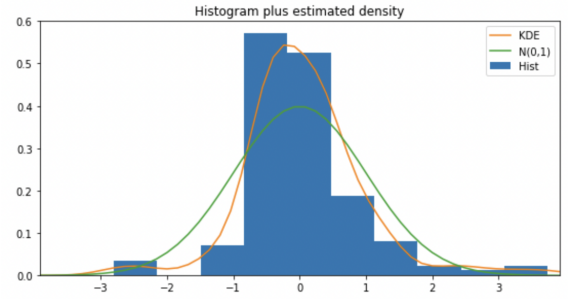

Fig 17: Figure showing the estimated density of the model

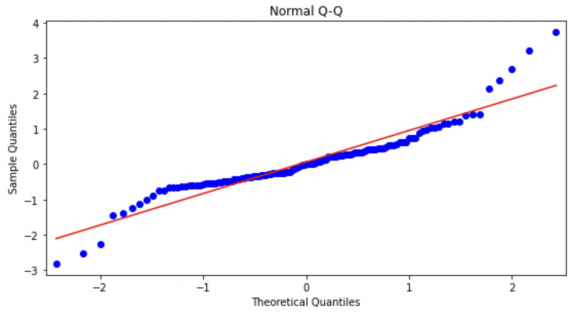

Fig 18: Figure showing the graph of sample quantile vs theoretical quantile of the graph

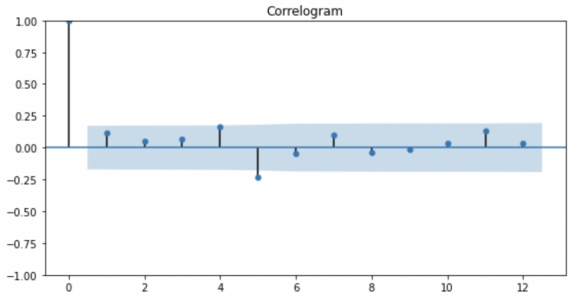

Fig 19: Figure showing the autocorrelation of the residual errors

In *Figure 16* figure, the residual errors appear to be uniformly distributed with a mean of zero.
An almost normal distribution with a mean of zero is suggested by the density diagram in *Figure 17*.
In *Figure 18* the red line should be perfectly aligned with all of the dots. A skewed distribution would be indicated by substantial variances.
The explanation of *Figure 19* says that the ACF plot, also known as the Correlogram, shows that the residual errors are not autocorrelated. Any autocorrelation would imply that the residual errors have a pattern that the model does not account for[10]. As a result, you'll need to increase the number of Xs in your model.

VII. Performance of the Autoregressive model

The model we have trained in the above section has the following performance;

- The Mean Absolute Error value of the model is 916.17

```
mean_absolute_error(actual, pred)
```

```
916.1721475850277
```

Following graph shows the prediction using the model

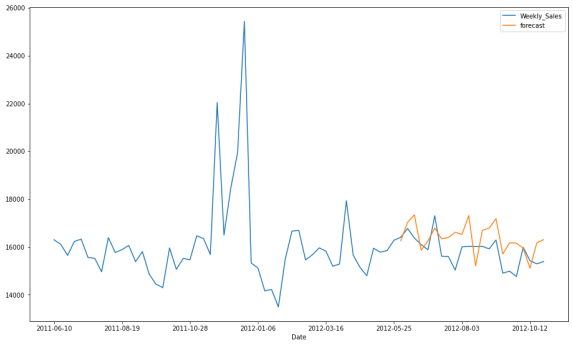

Fig 20: Figure showing the prediction using the SARIMA model

The plot clearly shows that the model is good but it is not good enough yet. If we wish to deploy this model in the production we would need to get more data points to retrain the model in the future. That was the motive of the whole project. We wish to make a model usable in production, so In the last stage of the project which will be deployment we will try to make the model retrainable. For the time being we will build another forecasting model.

VIII. Make and train LSTM model

Demand data sets contain a variety of seasonalities. The LSTM can capture both long-term seasonalities, such as a yearly pattern, and short-term seasonalities, such as weekly patterns.
It's only natural that events impact demand on the day of the event, as well as the days preceding and after it. The LSTM can categorise impact patterns from a wide range of events. The LSTM's multiple gates increase its ability to collect non-linear correlations for prediction. In general, causal variables have a non-linear impact on demand. The LSTM may learn about these parameters if they are present in the input variable.
The data needs to be changed for the LSTM model we are going to incorporate. We will generate 5 new columns called lag_1, lag_2, lag_3, lag_4 and lag_5 respectively. The lag columns are nothing but the old difference in sale values. We have created two different models of LSTM the structure of the first model is shown below;

A. *LSTM forecasting model version one*

The first model is quite simple as our dataset is also very small

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm (LSTM)                  (None, 128)               74752

 dense (Dense)                (None, 64)                8256

 dense_1 (Dense)              (None, 1)                 65

=================================================================
Total params: 83,073
Trainable params: 83,073
Non-trainable params: 0
_________________________________________________________________
```

But the performance of the first model was not enough; the error we received was ~494.6. The training and validation curve of the model is shown below. We can clearly see that the lines of training and validation are very smooth and the best accuracy is 494.6 approximately.

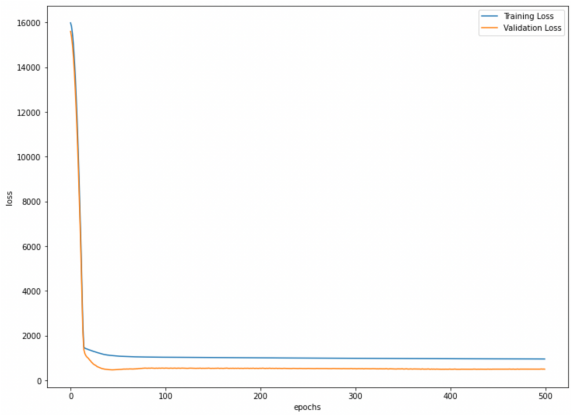

Fig 21: Figure showing the variation of training and validation loss over the training

On performing the prediction using this model we found out the performance as shown in the following figure.

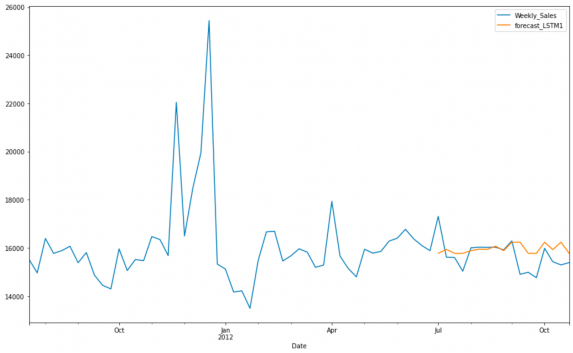

Fig 22: Figure showing the prediction using the first LSTM model

Now as the first attempt was not good enough so we will now make some upgrades to the model so that we can improve the performance of the model. The prediction graph shows that the model is predicting the highs better than that of lows but the figure of the graph is quite close.
The analysis of the graph also suggests that we need to increase the complexity of the model a bit so that it can learn better than this time.

B. LSTM forecasting model version two

For the second version of the model we made a very slight modification to the model, we made the model more complex than that of the last model, a summary of the model is printed below;

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_3 (LSTM)                (None, 1, 128)            74752

 lstm_4 (LSTM)                (None, 50)                35800

 dense_6 (Dense)              (None, 128)               6528

 dense_7 (Dense)              (None, 1)                 129

=================================================================
Total params: 117,209
Trainable params: 117,209
Non-trainable params: 0
_________________________________________________________________
```

The current model has a better loss of ~357.1 on the test set which is lower than the last LSTM model, the lower loss will lead to a better model and hence a better performance. The variation of the loss on the training and validation is shown below;

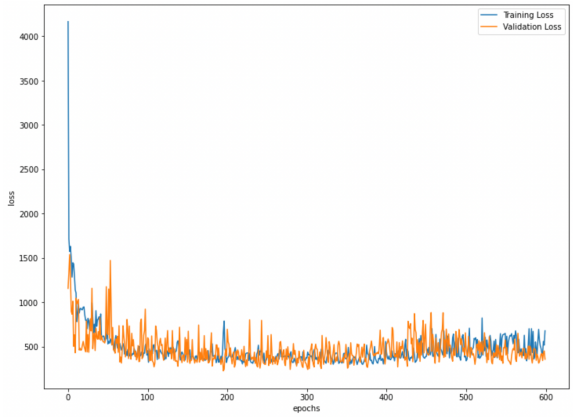

Fig 23: Figure showing the variation of training and validation loss over the training

This time we noticed some fluctuations while training the model unlike the last time. The struggle of a stronger mind can be seen in the training error.
Now If we proceed further for the model prediction on the test data we get the following graph. If we compare the model performance with the last two models we can clearly see that the current model performs way better than the others.

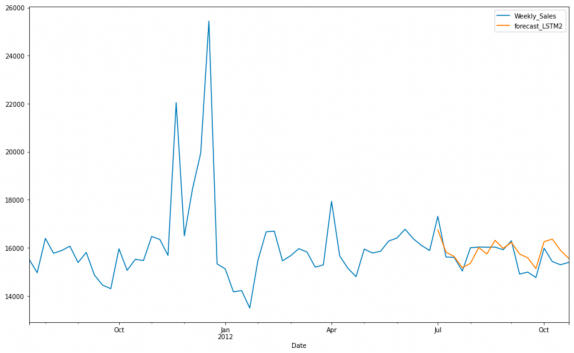

Fig 24: Figure showing the prediction using the second LSTM model

IX. Model Comparison

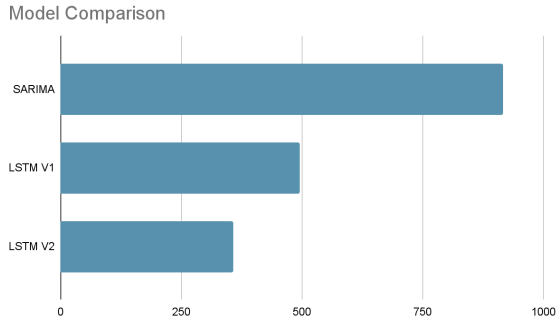

Fig 25: Figure showing the Comparison of the models

It is fairly obvious that a LSTM model performs better than a SARIMA model, reason being the LSTM structure is much more complex than that of SARIMA and that is why it can learn more information in comparison to the normal SARIMA model. If we look into our comparison graph we can clearly see that the loss for both the LSTMs is lower than that of the SARIMA model. But there are scenarios where one might need to pick Autoregressive models over neural networks, especially when the patterns are very simple. In such cases we should not spend a lot of computation costs. The performance of the SARIMA is much worse than the other LSTM models. The Loss of SARIMA is 837.62, the performance of the first LSTM model is 799.06 and the last LSTM model is 481.91.

Even though the complexity of the model was not too much the performance was quite good, it was a restriction as well to take a small model to avoid overfitting. The model can be strengthened in the future when we can have a larger amount of data, but for the time being we can stick with this current model.

X. Future Of the Project

As we have already discussed earlier, the model we have right now is quite simple, but for this particular use case it works like a charm. But in future as the sales data increases we can increase the complexity of the model, this will make sure that the performance of the model is top notch over the time. It is necessary to note that any product is not at its best in the beginning. It needs various tweeking and upgrades over the time, similarity our work has a lot of scope for future improvements.

XI. Acknowledgment

We are truly grateful to our college for providing such great resources and opportunities so that we can accomplish a task like this. We are in forever debt to our teachers for supporting and guiding us over this long span of time period. We have learned a lot as a team by working on this project. And we are truly grateful to our parents for being such a motivation and support for us.

References

[1] Bonanno, Alessandro, and Stephan J. Goetz. "WalMart and local economic development: A survey." *Economic Development Quarterly* 26.4 (2012): 285-297.

[2] Siami-Namini, Sima, Neda Tavakoli, and Akbar SiamiNamin. "A comparison of ARIMA and LSTM in forecasting time series." *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018.

[3] Podobnik, Boris, et al. "Influence of corruption on economic growth rate and foreign investment." *The European Physical Journal B* 63.4 (2008): 547-550.

[4] Maghelal, Praveen. "Investigating the relationships among rising fuel prices, increased transit ridership, and CO2 emissions." Transportation Research Part D: Transport and Environment 16.3 (2011): 232-235.

[5] Adebiyi, Ayodele Ariyo, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. "Comparison of ARIMA and artificial neural networks models for stock price prediction." Journal of Applied Mathematics 2014 (2014).

[6] Stojanović, Nikola, Marina Soldatović, and Milena Milićević. "Walmart recruiting–store sales forecasting." Proceedings of the XIV International Symposium Symorg. 2014.
[7] Manuca, Radu, and Robert Savit. "Stationarity and nonstationarity in time series analysis." Physica D: Nonlinear Phenomena 99.2-3 (1996): 134-161.
[8] Palachy, Shay. "Stationarity in time series analysis." Towards Data Science. Saatavissa: https://towardsdatascience. com/stationarity-in-time-seriesanalysis-90c9 4f27322. Hakupäivä 31 (2019): 2019.
[9] Martinez, Edson Zangiacomi, Elisângela Aparecida Soares da Silva, and Amaury Lelis Dal Fabbro. "A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil." Revista da Sociedade Brasileira de Medicina Tropical 44 (2011): 436-440.
[10] Xu, Shuojiang, Hing Kai Chan, and Tiantian Zhang. "Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach." Transportation Research Part E: Logistics and Transportation Review 122 (2019): 169-180.
[11] Van Houdt, Greg, Carlos Mosquera, and Gonzalo Nápoles. "A review on the long short-term memory model." Artif. Intell. Rev.53.8 (2020): 5929-5955.
[12] Salman, Afan Galih, et al. "Single layer & multi-layer long short-term memory (LSTM) model with intermediate variables for weather forecasting." Procedia Computer Science 135 (2018): 89-98.
[13] Dubey, Ashutosh Kumar, et al. "Study and analysis of SARIMA and LSTM in forecasting time series data." Sustainable Energy Technologies and Assessments 47 (2021): 101474.
[14] Xiong, Caiquan, et al. "Time Series Prediction of Wind Speed Based on SARIMA and LSTM." Conference on Complex, Intelligent, and Software Intensive Systems. Springer, Cham, 2021.

Arnav Singhal is of entrepreneurial mindset. Has also written his algorithm which has been published on medium.com. He passed 12th from Amity School in flying colors and he was an active member of MUN. Some of the project highlights of Arnav are Online Learning Portal, Web based automated campus placement preparation assistant for students etc.

Muskan Garg believes her fast-learning abilities, commitment to succeed, and relevant studies makes her fit for the project.An android app to calculate carbohydrates consumption for diabetic patients, e-commerce website, E-wallet are some of the projects done by her. She completed various MS azure certifications, intrigued in data structures and algorithms and has keen interest in music.

Naman Tiwari is currently pursuing my Bachelor's degree in Computer Science Engineering from Bennett University Greater Noida.
A self-taught programmer with 3+ years of experience in writing and debugging code. My interest moves from Web development to Entrepreneurship, Consultancy, Marketing, and 3rd management field.