International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi - https://doi.org/10.32628/CSEIT206485

Predicting Total Business Sales using Time Series Analysis

Navya Sri Kalli[1], Harsha Teja Pullagura[2]

[1]CSE Department, VVIT, Namburu, Andhra Pradesh, India & University Innovation Fellow, Fall 2018 cohort in VVIT

[2]Scientist in Allps, University Innovation Fellow , Spring 2018 Cohort in VVIT, Namburu, Andhra Pradesh, India

Article Info
Volume 6, Issue 4
Page Number : 475-482
Publication Issue :
July-August-2020

Article History
Accepted : 10 Aug 2020
Published : 16 Aug 2020

ABSTRACT

Economic activity undergoes 4 phases (expansion, peak, contraction, trough/recession) in which recession is a period of lowest activity and peak indicates the highest activity. Total Business sales is one of the key factors that influence the economic activity of a country. Total sales or gross sales is the grand total of all sales revenues a business generates from normal activities. The frequency of time series sales data can be monthly, quarterly, or annually. Prediction of business sales is highly important as it determines various factors in the market including Gross Domestic Product (GDP). The algorithms or models required for prediction of time series data are different from other machine learning models. Since sales is affected by time, a time series data should be stationary. Only when the data is stationarized, we can apply the algorithms on them. In this paper, monthly sales data is collected and predictions are done using moving average, simple exponential smoothing, Holt's model, ARIMA, and SARIMAX. Root Mean Square(RMS) is the accuracy metric of time series models and lower RMS indicates higher accuracy. In this paper, a lower value of RMS is obtained for the SARIMAX model.

Keywords : RMS, SARIMAX, Time series data, Total Business sales

I. INTRODUCTION

Technology has developed some powerful methods using which we can see things ahead of time. One such method, which deals with time-based data is Time Series Modeling. As the name suggests, it involves working on time (years, days, hours, minutes) based data, to derive hidden insights for decision making.

Time series data is a collection of quantities that are assembled over even intervals in time and are ordered chronologically. The time interval at which data is collected is generally referred as the time series frequency. A time series is a sequence of information that attaches a time to each value. The values can be pretty much anything measurable that depends on time in some way, like prices, humidity, or a number of people. As long as the values recorded are unambiguous, any medium can be measured with time series. There is no minimum or maximum

amount of time that must be included, provided the data that is gathered is in a way that provides the information that can be sought by the investor or analyst examining the activity.

A time series can be taken on any variable that changes over time. Time series data is data that is collected at different points in time. Because data points in time series are collected at adjacent periods there is potential for correlation between observations. The statistical characteristics of time series data often violate the assumptions of conventional statistical methods. Because of this, analyzing time-series data requires a unique set of tools and methods, collectively known as time series analysis. Time series data can be found in economics, social sciences, finance, epidemiology, and the physical sciences and there are no specific boundaries for applications of time series cause data is everywhere.

To express Time-Series efficiently, a notation is followed.

Time-series variables are represented with capital letters of the Latin alphabet like X or Y. For example, we can label the prices of the S&P 500 over some period of time as X. We describe the entire period covered by a time-series with a capital “T”, while we use lower-case “t” to describe a single period within the interval. For example, consider the daily closing prices for the S&P 500 for the entire 2008. Given the uppercase “T” represents the entire year, the lower-case “t” would represent a single day. Then, to denote the closing price on a specific day is done using “X of t”. Alternatively, the precise date, time, or year can be written as a subscript.

Since “t” represents the order of the period of present interest, the previous period is represented as “t minus 1”. Similarly, the next period is denoted as “t plus 1”.

Time series analysis can be useful to see how a given asset, security, or economic variable changes over time. It can also be used to examine how the changes associated with the chosen data point compare to shifts in other variables over the same time period. For example, to analyze a time series of daily closing stock prices for a given stock over a period of one year, a list of all the closing prices for the stock from each day for the past year and list them in chronological order can be obtained. This would be a one-year daily closing price time series for the stock.

Delving a bit deeper, time-series data can be analyzed with technical analysis tools to know whether the stock's time series shows any seasonality. This will help to determine if the stock goes through peaks and troughs at regular times each year. Analysis in this area would require taking the observed prices and correlating them to a chosen season. This can include traditional calendar seasons, such as summer and winter, or retail seasons, such as holiday seasons.

Alternatively, a stock's share price changes as it relates to an economic variable, such as the unemployment rate can be recorded. By correlating the data points with information relating to the selected economic variable, patterns in situations exhibiting dependency between the data points, and the chosen variable can be observed.

A time-series graph plot observed values on the y-axis against an increment of time on the x-axis. These graphs visually highlight the behavior and patterns of the data and can lay the foundation for building a reliable model. More specifically, visualizing time series data provides a preliminary tool for detecting if data:

- ✓ Is mean-reverting or has explosive behavior;
- ✓ Has a time trend;
- ✓ Exhibits seasonality;
- ✓ Demonstrates structural breaks.

This, in turn, can help guide the testing, diagnostics, and estimation methods used during time series modeling and analysis.

1. Mean Reverting Data

Mean reverting data returns, over time, to a time-invariant mean. It is important to know whether a model includes a non-zero mean because it is a prerequisite for determining appropriate testing and modeling methods.

For example, unit root tests use different regressions, statistics, and distributions when a non-zero constant is included in the model.

A time series graph provides a tool for visually inspecting if the data is mean-reverting, and if it is, what the data is centered around. While visual inspection should never replace statistical estimation, it can help you decide whether a non-zero mean should be included in the model.

For example, the data in the figure below varies around a mean that lies above the zero line. This indicates that the models and tests for this data must incorporate a non-zero mean.

2. Time Trending Data

In addition to containing a non-zero mean, time series data may also have a deterministic component that is proportional to the time period. When this occurs, the time series data is said to have a time trend.

Time trends in time series data also have implications for testing and modeling. The reliability of a time series model depends on properly identifying and accounting for time trends.

A time series plot that looks like it centers around an increasing or decreasing line, like that in the plot above, suggests the presence of a time trend.

3. Seasonality

Seasonality is another characteristic of time-series data that can be visually identified in time series plots. Seasonality occurs when time series data exhibits regular and predictable patterns at time intervals that are smaller than a year.

An example of a time series with seasonality is retail sales, which often increase between September to December and will decrease between January and February.

4. Structural Breaks

Sometimes time series data shows a sudden change in behavior at a certain point in time. For example, many macroeconomic indicators changed sharply in 2008 after the start of the global financial crisis. These sudden changes are often referred to as structural breaks or non-linearities. These structural breaks can create instability in the parameters of a model. This, in turn, can diminish the validity and reliability of that model.Though statistical methods and tests should be used to test for structural breaks, time series plots can help for preliminary identification of structural breaks in data.

Structural breaks in the mean of a time series will appear in graphs as sudden shifts in the level of the data at certain breakpoints. For example, in the time series plot above there is a clear jump in the mean of the data which around the start of 1980.

Time series is said to have the following features in it.

1. Time Period

There aren't any limitations regarding the total time span of a time series. It could be a minute, a day, a month, or even a century. All that's needed is a starting and an ending point. There are usually numerous points in-between and the interval of time separating two consecutive ones is called a "time period". For example, if the data was recorded once per day from 1/1/2000 to New Year's Eve 2009, a single time period would be a day, while the entire time span would be a decade.

2. Frequency

The "frequency" of the dataset tells us how often the values of the data set are recorded. To be able to analyze time-series in a meaningful way, all time-periods must be equal and clearly defined to maintain a constant frequency. The two features are related. This frequency is a measurement of time and could range from a few milliseconds to several decades. However, the ones that are most commonly encountered are daily, monthly, quarterly, and annual.

3. Patterns

Patterns we observed in time-series to persist in the future. That is why it is often tried to predict the future by analyzing recorded values. They repeat over a certain period of time in the whole interval.

II. TECHNIQUES USED

1. Moving Average Smoothing:

$$MA_{t+1} = \frac{\sum_{i=t-n}^{t} x_i}{n}$$

The optimal number of observations are to be used in the forecast for effective results. It can be found by checking the square error mean of multiple of n observations. The minimum starts at 3 observations and can go up to half of the data set size + 1. In this paper, the rms value is calculated with 5 observations.

2. Simple Exponential Smoothing:

In this technique, weights are assigned to the observations in such a way that, the most recent values receive larger weights than that of distant past. The weights go in a decreasing order exponentially and the smallest weights are given to the farthest observations.

$$\hat{y}_{T+1|T} = \sum_{j=0}^{T-1} \alpha(1-\alpha)^j y_{T-j} + (1-\alpha)^T \ell_0.$$

It is an easily learned and easily applied procedure for making some determination based on prior assumptions by the user, such as seasonality. This method can be used for data without a trend in it. Two parameters, smoothing level and optimized are used to forecast with values '0.7' and 'false' respectively. The fit has reduced the RMS value with this technique when compared with the moving average technique.

3. Holt's Linear Model:

This method is an extension of exponential smoothing which can forecast data with the trend. Before applying this method directly its suggestible to decompose the series and observe the results. Time series data can be decomposed into four parts. They are observed, trend, seasonality, and residuals.

$$\begin{aligned} \text{Forecast equation} \quad & \hat{y}_{t+h|t} = \ell_t + hb_t \\ \text{Level equation} \quad & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\ \text{Trend equation} \quad & b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}, \end{aligned}$$

where bt denotes trend of time, lt is is used as estimate level of trend. Alpha and Beta are the smoothing parameters for level and trend respectively.

A method names seasonal decompose is used to this. After the decomposition is done, graphs are observed and forecasting is done. Here, smoothing level and

smoothing slope are the parameters that are used with values' 0.2' and '0.2'. This method reduced the RMS to a greater extent and also gave the least RMS value when compared with all the techniques used in this paper.

4. Holt's Winter Model:

This is one of the methods that can give better results to datasets that suffer from seasonality. All the above-mentioned models don't take into the seasonality of data into consideration. As this method uses exponential smoothing to seasonal components along with trend in the data, this method is preferred for data with seasonality. The seasonal component is expressed as follows:

$$s_t = \gamma^*(1-\alpha)(y_t - \ell_{t-1} - b_{t-1}) + [1-\gamma^*(1-\alpha)]s_{t-m},$$

Unlike Holt's Linear model where we use Holt().fit() method, here we use ExponentialSmoothing().fit(). The parameters also differ. In this paper, seasonal periods and trend are used with values '4' and 'add' respectively. The RMS value has increased for this model compared to the above model.

5. ARIMA

It stands for Auto-Regressive Integrated Moving Average model. It is applicable only for stationary time series. The mean, variance, and covariance of the of respective terms should not be a function of time i.e., they should not increase or decrease with time and instead should be a constant. There is one more reason for making the time series data stationary. It is, variables are independent when data is stationary and we get more information from data once its stationary. A data is said to be stationary when the trend and seasonality are removed from the data. In this paper, dickey fuller test is used to check the stationarity of the data using window as parameter with value '4' as there are four quarters a year and seasonality changes for each quarter in a year. In the test, the test statistic obtained is greater than the critical value. In compliance with the null hypothesis of Dickey fuller test, this indicates that the data is not stationary.

To make the data stationary, trend is removed from the data and test statistic has fallen below the critical value making the data stationary. Seasonality is also removed to get effective results and data is made stationary.

Results of Dickey-Fuller Test before:

Test Statistic -0.595108
p-value 0.872083
#Lags Used 3.000000
Number of Observations Used 307.000000
Critical Value (1%) -3.451831
Critical Value (5%) -2.871001
Critical Value (10%) -2.571811
dtype: float64

Results of Dickey-Fuller Test after:
Test Statistic -6.988385e+00
p-value 7.860135e-10
#Lags Used 1.300000e+01
Number of Observations Used 2.790000e+02
Critical Value (1%) -3.454008e+00
Critical Value (5%) -2.871956e+00
Critical Value (10%) -2.572320e+00
dtype: float64

When the data is made stationary, forecasting of data can be done by finding optimistic values for its parameters that are p,q, and d. ACF and PACF plots are used to find these values. AR model and MA model are combined and ARIMA model is formed.

For AR model, the formula is given as:

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

This is extended in further stage using white noise.
For MA model, it is represented as:

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \cdots$$

The values of p,d,q are found to be 1,2,1 respectively.

The full ARIMA model is the integration of both AR and MA model and it is represented as shown:

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

With this technique, the RMS value is reduced when compared with Holt's Winter model but the value is greater than all other methods.

6. SARIMAX:

This model takes seasonality of data into consideration. Order and seasonal order parameters are used in this model to plot the graph with values (1,2,1) and (0,1,2,4) respectively. This model reduced the RMS value to a great extent and gave the second lowest RMS value for test data among all the techniques used.

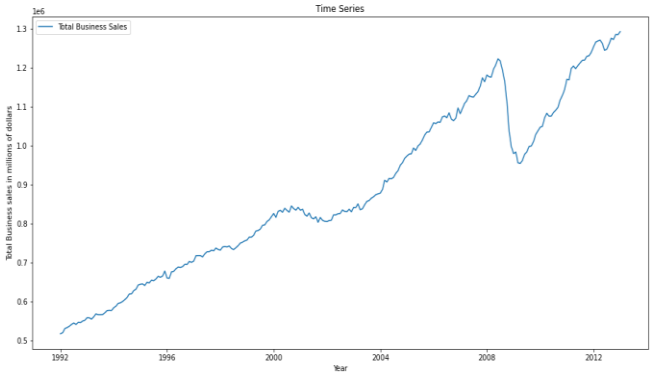
III. RESULTS

The data when first observed, It initially contains two columns where a column represents the month in date format and another column represents sales in millions of dollars in numeric format. The data is in seasonally adjusted format.

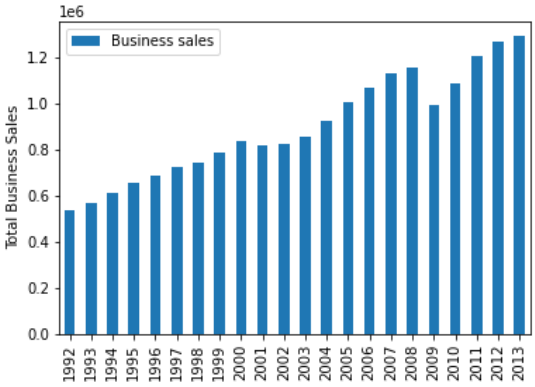
Sales value kept rising from the start and has seen a decline just thrice in the mentioned period. One is between 2001 - 2002, 2008 - 2009, and the other is between 2016 - 2017.

The 2008 - 2009 period is under recession and the decline rate is very high during this period. The other decline periods are not very notable and took a rise immediately. Data has an increasing trend in it.

The total data is split into training and testing data where 75% of data is given to the training data set and 25% is used for testing. When training data is observed in graphical format, it shows as:

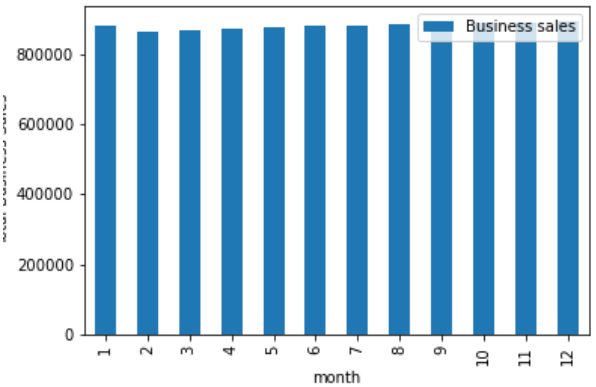


When data is grouped using years in a bar plot, it is displayed as shown



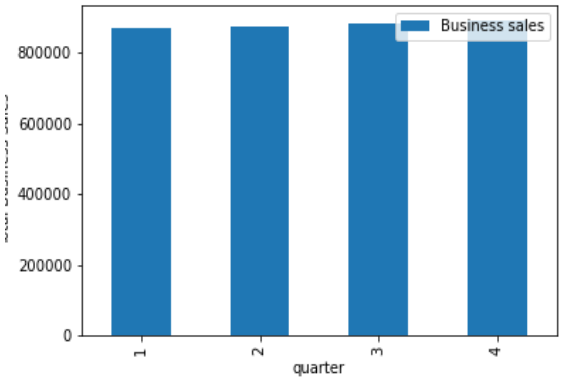
Among all the years, 2009 is observed to face a greater decrease followed by 2001.

Similarly, when data is grouped with months, it shows the following pattern in the plot.



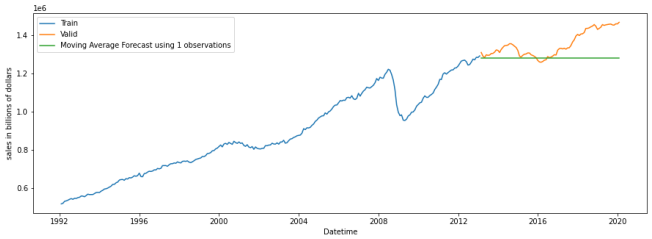
It is observed from the data that eding months in an year have higher sales compared with the initial months of an year.

In the next step, data is observed by grouping it in quarterly manner and it shows the following

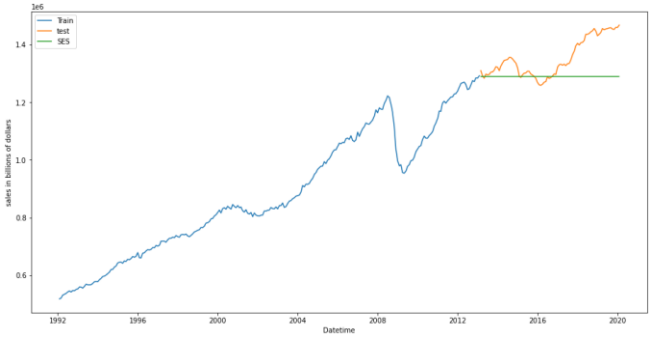


From this graph, it is observed that sales for each quarter kept rising in an year till the end.

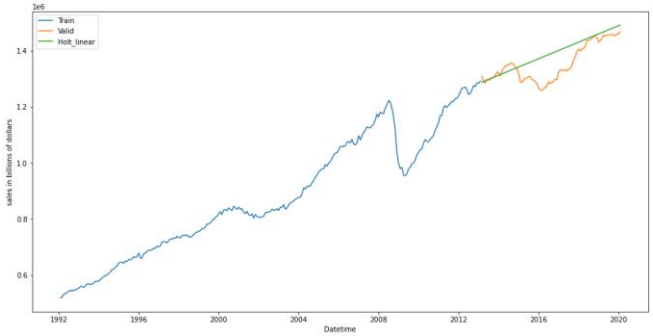
Data is forecasted using Moving Average technique and the data looks as shown:



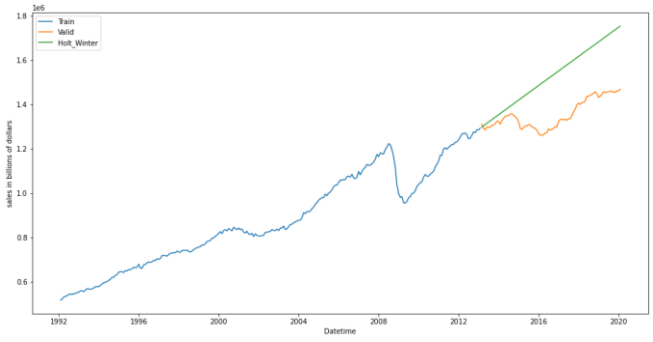
Simple Exponential Smoothing shows better performance in the validation data than that of the above technique and the data looks as shown in the graph:



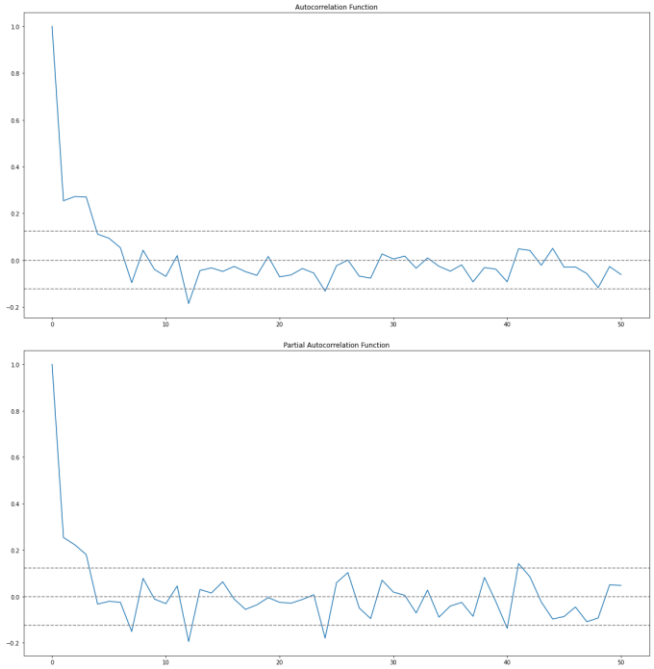
When, Holt's Linear model is used, the results are even better and the graph is shown:



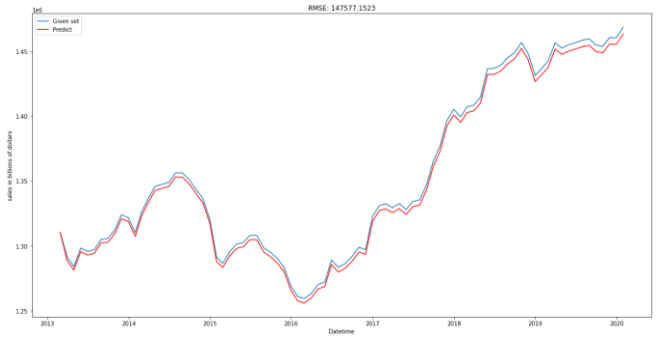
In the similar manner, when Holt's Winter model is considered, the graph looks as shown below:



Similarly, for ARIMA model to be applied, ACF and PACF graphs are plotted first and they look like:

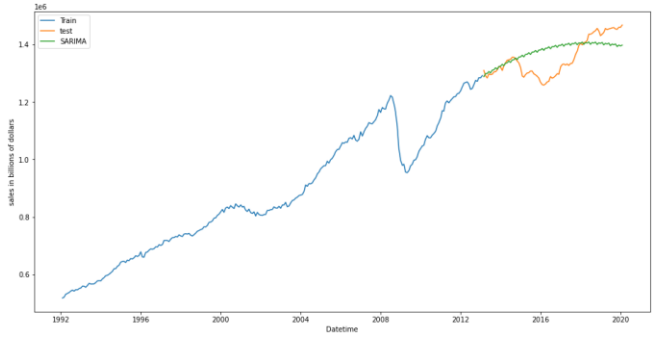


Using, these two plots, AR and MA models are generated which help to generate the ARIMA model graph and it is shown below:



Unlike all other graphs, this graph only shows testing data.

SARIMA model generates the graph in the following pattern:



IV. CONCLUSION

Moving average technique gives a RMS value of 96541.22016295874 for testing data and it is reduced by Simple Exponential Smoothing which gives 90989.03306181119 as RMS. Holt's Linear Model has the RMS value of 12891.348117170583 for validation and 55386.22164094275 for test data where as Holts Winter Model gives 194815.94835705226.ARIMA model gives a value of 147577.1523.SARIMA model reduces the RMS value to 60905.80853896327. This is the second lowest RMS value after Holts Linear Model. It is observed from these values that, for data like Total Business Sales in Time series analysis, among all the methods used Holts Linear model and SARIMAX model gave effective results.

V. REFERENCES

[1]. Yan-ming Yang, Hui Yu and Zhi Sun, "Aircraft failure rate forecasting method based on Holt-Winters seasonal model," 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, 2017, pp. 520-524, doi: 10.1109/ICCCBDA.2017.7951969.

[2]. Yan-ming Yang, Hui Yu and Zhi Sun, "Aircraft failure rate forecasting method based on Holt-Winters seasonal model," 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, 2017, pp. 520-524, doi: 10.1109/ICCCBDA.2017.7951969.

[3]. S. I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou and A. G. Bakirtzis, "Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting," 2016 IEEE International Energy Conference (ENERGYCON), Leuven, 2016, pp. 1-6, doi: 10.1109/ENERGYCON.2016.7514029.

[4]. Etuk, Ette & Mohamed, Tariq. (2014). Full Length Research Paper Time Series Analysis of Monthly Rainfall data for the Gedaref rainfall station, Sudan, by Sarima Methods. International Journal of Scientific Research in Knowledge. 2. 320-327. 10.12983/ijsrk-2014-p0320-0327.

Cite this article as :
Navya Sri Kalli, Harsha Teja Pullagura, "Predicting Total Business Sales using Time Series Analysis ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 4, pp. 475-482, July-August 2020. Available at doi : https://doi.org/10.32628/CSEIT206485
Journal URL : http://ijsrcseit.com/CSEIT206485