Understanding Time Series: Utilizing sales data to understand how the Time Series algorithms work.

Abhinav Kumar[1] and Neha Vora[2]

[1,2] Department of Information Technology, SVKM'S Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, India
E-mail: [1]abhinavkumar2000@hotmail.com, [2]nehavora2501@gmail.com

Abstract

We as a species have always been interested to know what will take place with us in the future. And with technology growing rapidly now we have started using various methods to try to predict the future with the use of data. One such method is Time series Analysis. So, in this experiment, we saw how to use Time Series and how it can be utilized with the help of sales data of a company that works on seasonality. It will also analyze how to improve the prediction with the help of making required changes along these lines. Line using ARIMA models, seasonality components, and formulas, having space for margin of errors, as well as using the multiplicative and well as additive model for the same. From this process, we found out how easy the process of analyzing data can be with the help of time series. We can represent it with the help of graphs and pictorial representation techniques to make it easy to understand and easy to develop with very little effort and it can be expanded into various pattern-related prediction domains.

Keywords: Sales, predictions, time series, patterns, ARIMA, analysis, graphs, seasonality.

1. Introduction

Data collected over time is often more than just a collection of isolated points; it forms a time series pattern which is a sequence of observations made at successive time intervals. Time series data is ubiquitous and spans various domains, such as financial markets (e.g., stock prices), environmental monitoring (e.g., atmospheric CO2 levels), engineering (e.g., machinery performance metrics), and healthcare (e.g., patient monitoring data). Unlike static cross-sectional data, which provides a snapshot at one

point in time, time series data captures the temporal evolution of phenomena, revealing patterns and trends over extended periods.

As Giuseppe Nunnari; Valeria Nunnari (2017) say the analysis of time series data involves understanding and modeling these temporal dynamics to gain insights and make forecasts. This typically includes identifying and interpreting key components such as trends (long-term movements), seasonal effects (regular, periodic fluctuations), and irregular variations (random noise). Techniques for analyzing time series data range from simple methods like moving averages and exponential smoothing to more sophisticated approaches such as autoregressive models and spectral analysis.

Despite its utility, traditional time series analysis often faces challenges when applied to complex real-world data. For example, many time series models assume that the data is stationary meaning its statistical properties cannot change over time. However, in practice, time series data can exhibit non-stationarity, where patterns and trends evolve. Additionally, these models may struggle with abrupt changes or shifts in the underlying process generating the data.

To address these challenges, there has been significant progress in developing advanced methods for time series analysis. Emerging techniques, including machine learning and deep learning algorithms, offer new ways to model complex patterns and adapt to changing data dynamics. These modern approaches can potentially improve the accuracy and robustness of forecasts by capturing intricate temporal relationships that traditional methods may overlook.

This paper aims to go deep into the realm of time series analysis to find more about, examining current methodologies and exploring opportunities for enhancement. By integrating innovative techniques and addressing existing limitations, this study seeks to advance the field of time series analysis, providing valuable insights for both academic research and practical applications.

2. Related Work

Time Series has been one of the most interesting and got to aspects of prediction models as well as machine learning models as it has a lot of utility in predictions as we can see similar patterns in research papers.

In their research, the authors B. Singh, P. Kumar, N. Sharma and K. P. Sharma(2020) conduct a case study focused on forecasting monthly retail time series data obtained from the US Census Bureau, covering the period from 1992 to 2016. The modeling process is approached in two phases. Initially, the original time series undergoes de-trending through a moving window averaging method. Following this, the residual time series is analyzed using Non-linear Auto-Regressive (NAR) models, employing both the Neuro-Fuzzy systems and the Feed-Forward Neural Networks. To evaluate the effectiveness of the forecasting models, the authors calculate various error metrics, bias, mean absolute error (MAE), and root mean square error (RMSE). Additionally, they compute the model skill index, referencing a traditional persistent model for comparison. The findings indicate a clear advantage in using the proposed methodologies over the conventional approach.

Various researchers have attempted using SVM in time series prediction and outlined its uniqueness in the machine learning subfield which is SVM. Their characteristic is more pronounced in situations where forecasting is needed especially when the system processes are non-linear, non-static and imprecise in nature. For that reason, these techniques have been successfully applied instead of other nonlinear approaches, including neural networks and multi-layer perceptrons. The eventual goal of this paper is to review the present research work focused on the SVM for forecasting time series and hence provide availability for its usage. Moreover, it acts as a short introduction on SVM towards time series analysis with its pros and cons described with respect to this method.

As Horváth, Csilla & Wieringa, Jaap. (2003). Say various researchers have attempted using SVM in time series prediction and outlined its uniqueness in the machine learning subfield which is SVM. Their characteristic is more pronounced in situations where forecasting is needed especially when the system processes are non-linear, non-static and imprecise in nature. For that reason, these techniques have been successfully applied instead of other nonlinear approaches, including neural networks and multi-layer perceptrons. The eventual goal of this paper is to review the present research work focused on the SVM for forecasting time series and hence provide availability for its usage.

Moreover, it acts as a short introduction on SVM towards time series analysis with its pros and cons described with respect to this method.

3. Proposed Work

3.1 DATASET FORMULATION

The data is from a secondary source. The data originally is from University of Westminster (27-04-2021). So, we going to look for seasonality and repeating patterns in the data.

No cleaning or preprocessing is required as the data set is already clean. If it was needed, we could have used various preprocessing methods like removing null values. And removing outliers

The file has 2 sheets:

In sheet 1- we have sales data for a company in which we have the data of sales in number of sales made in millions per month for a total of nine years. From the year 2001 to the year 2009. And we can see that in Fig 1.

Month	2001	2002	2003	2004
January	139.7	165.1	177.8	228.6
February	114.3	177.8	203.2	254
March	101.6	177.8	228.6	226.7
April	152.4	203.2	279.4	342.9
May	215.9	241.3	317.5	355.6

Fig 1

In sheet 2- we have exactly the same data available to us but here we will use only two columns to represent the data in one row we will have sales by date in date-time format and the other column will have the sales data. It is similar to the previous sheet but different representation format so that if we need it to makes graphs in easier way the it can be utilized. The fig below represents the start of the sheet.

Years	Sales
01-01-2001	139.7
01-02-2001	114.3
01-03-2001	101.6
01-04-2001	152.4
01-05-2001	215.9
01-06-2001	228.6
01-07-2001	215.9

Fig 2

We also have sheet 3 which has the same data but only up to December of 2006 and we are doing this to make the process of this prediction model faster so we will be using 72 columns in this prediction model for time series

Data Cleaning and Preprocessing

As the data is already clean so any kind of data cleaning is not required in this data set. And we can check that by df.info(). Even we needed we could use various methods to clean data using delete remove etc.

Data Splitting

We split the data of 72 rows into training data and testing data. And will be giving 48 rows to train the data, the remaining 24 for testing the data. This means the data will be split into 66-33 patterns. 1/3 to test the data and the remaining 2/3 to train it before.

Ideally, the data is split into 80-20 or 70-30 patterns for optimal use. But here we have 6 years in this data set so it is logical to split it year-wise, henceforth the first four years were given to the training and the remaining 2 years for testing the data. This takes us to the almost ideal percent difference required for prediction algorithms.

3.2 Models used

Plots

Here are different kinds of plots to first see the data and analyze whether the data is a time series-worthy pattern or not. We can also see by different kinds of graphs how the data available to us can be divided into different patterns and how many different ways we can have to represent the data. It can show us what is inside it.

Linear Regression

A linear regression model is a type of ML model that is applied to represent data linearly. It is one of the supervised machine learning algorithms used to calculate the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data. There is simple when only one independent feature is there and when more than one feature are there, then it is called Multiple Linear Regression. Similarly,

when there is only one dependent variable, then it is considered Univariate Linear Regression, whereas when there are more than one dependent variables then known as Multivariate Regression.

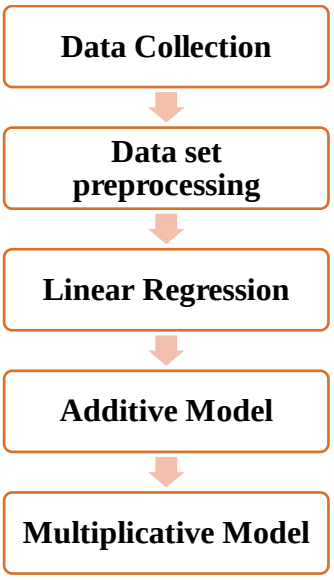
Additive model

The additive model can be expressed with the formula O = T + S + C + I, where O stands for the original data. In this equation, T represents the trend component, S accounts for seasonal variations, C captures cyclical changes, and I denotes irregular fluctuations. This model is particularly useful for analyzing data that exhibits seasonality, helping us identify and understand underlying patterns and trends effectively.

Multiplicative model

The multiplicative model is represented by the formula Y = T × S × C × I, where Y indicates the overall result derived from four key components. In this model, the seasonal, cyclical, and irregular variations (S, C, and I) are expressed as decimal percentages. The trend (T) reflects the long-term movement within a time series, allowing for a deeper understanding of how these factors interact to influence the outcome.

No. of rows	73
No of columns	2
Models Use	3
Test data	24 rows (33.33%)
Train data	48 rows (66.66%)
Aim	forecasting



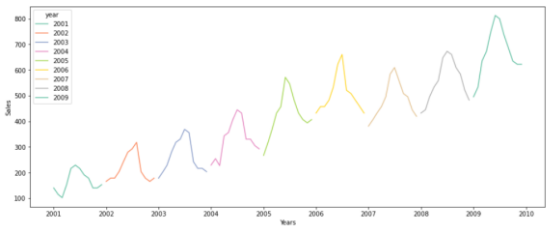
3.3 EXPERIMENTAL SETUP

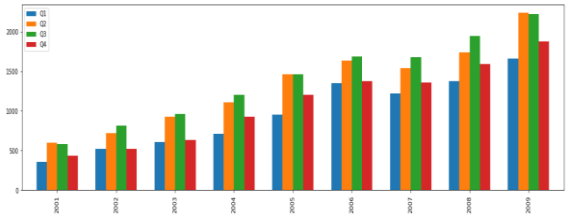
Language	Python 3
Processor	Intel 11 den
Ram	8 GB
Storage	120 GB
Operating System	Windows
Environment	Jupiter
Frameworks & Libraries	Maplotlib, Numpy,,Pandas, Sklearn

4. Results and Discussion

In this process initially, we understood The data shape by using df.(info). Then we plotted the graph to check for the appearance of cyclic patterns, which is very important for this experiment.

Then we plotted different kinds of graphs on the data to show how can we represent the data. Line graph, area plot, bar graph, we also divided the sales into 4 quarters per year and dew heat mat as well as the bar graph accordingly.





So after understanding the data linear regression model was used to train the data. And then we used this result output to

Plot the graph with a linear regression line.

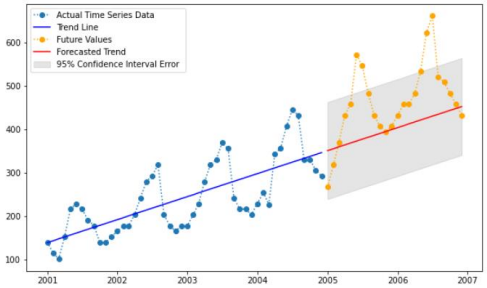
For the sake of this experiment, we use margin of error as the prediction cannot be 100% accurate so we have a margin of 5% making it 95%. The formula for margin of error is.

CI = μ+−MEμ+−ME (Confidence Interval)

where ME = z * SE

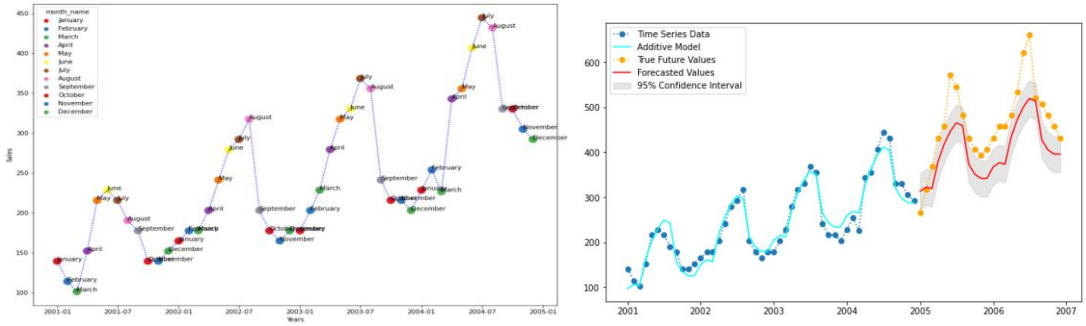
where SE = Standard deviation Error.

z = 1.96 (for 95% Confidence Interval)



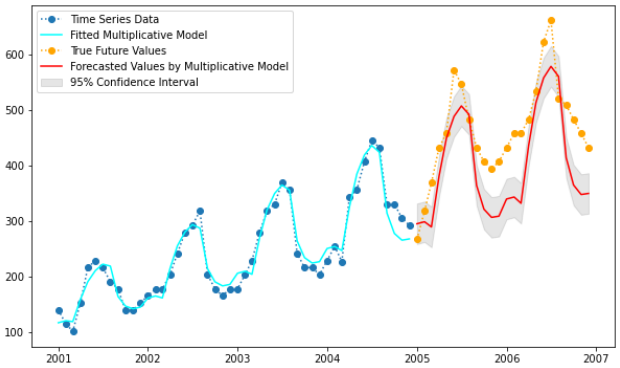
As we can see in the graph the points plotted are in no way near the required accuracy.

The additive model can be expressed with the formula O = T + S + C + I, where O stands for the original data. In this equation, T represents the trend component, S accounts for seasonal variations, C captures cyclical changes, and I denotes irregular fluctuations. So, as we can see this graph does a lot better than the graph of the previous model, and it is possible because we used the different components of time series itself in the graph such as seasonality which divides the graph into month-wise patterns. As we can see below in line graph.



As we saw adding the element of time series worked so well so what if instead of adding the components we multiply it by the components of the time series. We know multiplication has a bigger impact than addition. And that is exactly what we did next. We applied a multiplicative model for the same data.

The formula Y = T × S × C × I represent the multiplicative model, where Y indicates the overall result derived from four key components. In this model, the seasonal, cyclical, and irregular variations (S, C, and I) are expressed as decimal percentages. The trend (T) reflects the long-term movement within a time series.



$$\hat{Y}_t = a + b_1 * t + b_2 * x_1 + b_3 * x_2 + \cdots + b_{n-1}x_{n-1}$$

Where,
$x_1, x_2, x_3 \ldots x_n$ are seasonal components

As we can see from above us this model seems pretty accurate. Almost perfect with almost all lines being present in the region of 95% percent confidence interval. This shows how much of an important role seasonality and trends play in time series applications.

Forecasting was the ingrained part of the code with the help of which we were able to predict and draw graphs for the same.

5. Conclusion

In wrapping up this analysis, it's clear that choosing the right model can make a significant difference in understanding time series data. Our investigation revealed that the multiplicative model was the most effective, adeptly capturing the complex patterns and interactions within the data. This superiority arises from the model's ability to handle varying levels of seasonality and trend through its multiplicative nature, which allows it to adapt to changes in the amplitude and frequency of these components more flexibly. Although the additive model scored well, it did not perform as well as the multiplicative model because it forges a constant additive effect, which is not the case for seasonal-type data with varying and often changing trends. For instance, the linear regression model was basic, which its lack which is why it did not quite cut it because of its linear assumption. That is, linear relationships were not enough since our dynamics were not linear.

5.1 Scope for Future Work

These results imply that even though simple models may permit the acquisition of straw man stylization, diving into more complex analysis such as multiplicative models appear beneficial since it harnesses the essence of the data better. Looking ahead, options for the

continuation of this study are rather thick on the ground. In particular, the application of machine learning techniques such as RNNs or LSTMs could improve the capacity to handle many types of dependencies over time which are also nonlinear. Moreover, the research on hybrid models where elements of adding and multiplying techniques are utilized is also worth pursuing, especially on datasets that have broad seasonal variations. Finally, extending the application of such models to various datasets irrespective of domains would, in turn, test the model's versatility. To summarize, such a strategy underlines the necessity of conformable approaches with the data used for analysis, and provides an opportunity for further progression of time series analysis.

References

[1] "P. Ghosh, O. Samanta, T. Goto and S. Sen, "Sales Forecasting of Overrated Products: Fine Tuning of Customer's Rating by Integrating Sentiment Analysis," 2024
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10531712&isnumber=10380310"

[2] "S. Rahman and M. S. U. Zaman, "Time Series Sales Forecasting: A Hybrid Deep Learning Regularization Approach," *2024*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10561831&isnumber=10561632"

[3] "G. A. Sampedro, "Predicting Pre-Order Sales Using Time Series Algorithm, Forecasting, and ARIMA Model in Python for Small Businesses," *2024*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10457267&isnumber=10457087"

[4] "T. Saravanan, T. Sathish and K. Keerthika, "Forecasting Economy using Machine Learning Algorithm," 2022
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10060725&isnumber=10059582"

[5] "R. Al-Omoush, S. Fraihat, G. Al-Naymat and M. Awad, "Design and Implementation of Business Intelligence Framework for a Global Online Retail Business," *2022*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10009688&isnumber=10009661"

[6] “Gupta, K. Singh, N. Sharma and M. Rakhra, "Machine Learning For Detecting Credit Card Fraud," *2022.*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10126861&isnumber=10126139”

[7] “N. Lili, "Superstore Sales Forecasting Based on Elastic net Regression and BP Neural Networks," *2021.*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9988373&isnumber=9987731”

[8] “S. Kumar Singh, K. Pal Sharma and P. Kumar, "Analytical study for Price Prediction of Bitcoin using Machine Learning and Deep Learning," *2022.*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9848077&isnumber=9847665”

[9] “P. Kaunchi, T. Jadhav, Y. Dandawate and P. Marathe, "Future Sales Prediction For Indian Products Using Convolutional Neural Network-Long Short Term Memory," *2021*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9587668&isnumber=9586796”

[10] “S. K. Sogi and S. Kumar Mittal, "A Comprehensive Review and Analysis for forecasting Industrial Data," *2021*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9478082&isnumber=9478075”

[11] “A. Sharma, H. Liu and H. Liu, "Best Seller Rank (BSR) to Sales: An empirical look at Amazon.com," *2020*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9282620&isnumber=9282607 “

[12] “B. Singh, P. Kumar, N. Sharma and K. P. Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling," *2020*
URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9071463&isnumber=9071428”

Author's biography

Abhinav Kumar holds a bachelor's degree in information technology (BSc.IT) with a CGPA of 8.34 and is currently pursuing his master's in information technology (MSc.IT).

He is currently working in the industry as a Sports Analyst. And works on data analysis and towards creating better machine learning models in general.

Neha Vora is currently pursuing her Ph.D. in Computer Science and holds a master's in computer applications (MCA). She is qualified in NET, SET, and GATE, and brings over 9 years of teaching experience, along with 1 year of industry experience. Her primary research areas include computer vision, image processing, machine learning, object detection, and artificial intelligence.