Data-Driven Sales Insights and Forecasting Using Machine Learning to Promote Decision Making.

Sheetal
Department of Computer Applications
Presidency College, Bengaluru,
Karnataka, *India*
sheetal.sunil@gmail.com

Alli A ,
Department of Computer Applications,
Presidency College, Bengaluru,
Karnataka, India,
alliarul@yahoo.co.in

Vasantha Kumari N,
Department of Computer Applications,
Presidency College, Bengaluru,
Karnataka, India,
vasantha.kn@gmail.com

***Abstract*— Sales is very important factor in business. Supply and demand is mainly liable on the sales forecasting. The research is based on extracting critical information to support strategic decisions through data science approaches. This research work was carried out by gathering and preparing a large amount of sales data, where analysis was carried out in solving data quality concerns and producing a cleaned dataset. Our analysis and findings promises the business the most crucial information about the factors that affects consumer behavior, sales success, and regional differences. These insights are made available to stakeholders promoting data-driven decision-making. The sales forecasting is done using Machine Learning Algorithms such as ARIMA, SRIMA and Prophet. Compared to all three Prophet Out performed. We all propose ensemble model that empowers businesses to improve accuracy and make data-driven decisions and achieve better accuracy in predicting sales demand.**

Keywords—Forecasting, Machine Learning, Sales, Insights, Time Series analysis, data-driven decisions

I. INTRODUCTION

In today's fast-paced business environment, companies are constantly striving to gain a competitive advantage and refine their sales strategies. Companies are dependent on sales forecasting to support them in making informed decisions regarding financial budgeting, production planning, inventory management, and supply chain coordination. Having a balance between supply and demand, avert stockouts, and prevent overproduction all rely on correct sales forecasts. The utilization of advanced analytical methods can significantly enhance the accuracy of forecast in the current data-driven era, providing organizations with a competitive advantage.

The aim of this research is to apply data science methods to assist strategic decision-making through extracting valuable information from sales data. Preparation and collection of large-scale datasets enabled extensive data analysis to address issues with data quality, ensuring a clean and reliable dataset for modeling. Identifying the factors influencing sales success, consumer behavior, and geographical differences was the primary objective.

Three popular machine learning models Auto Regressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA), and Prophet were employed to generate accurate forecasts. Prophet was notable among them due to its capability of addressing data anomalies and identifying complex seasonal patterns. Still, an ensemble model was recommended in order to make it even more accurate and reliable. To generate reliable and credible sales estimates, the ensemble methodology leverages the strengths of individual models. The insights acquired from this research provide stakeholders the capacity to optimize resource allocation, make informed decisions, and enhance operational efficiency. Organizations are able to formulate more efficient marketing strategies, forecast market directions, and realize sustainable growth through the understanding of primary determinants of sales.

The proposed ensemble model and its benefits to sales forecasting are addressed in detail in the subsequent sections that also encompass the research methodology, data analysis, model implementation, and performance evaluation. This research highlights the importance of applying machine learning algorithms to address real business issues and make strategic decisions.

II. RELATED WORK

Taryney, P. D. (2002) work “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Sales” research paper introduced the use of unsupervised sentiment analysis techniques to classify Sales. Turney's work primarily focused on using techniques like pointwise mutual information to extract data from sales[1]. Pang and Lee's survey paper provides an in-depth overview of time series techniques up to 2008. It covers various aspects of sentiment analysis, including document-level and sentence-level sentiment classification, sentiment lexicons, and the challenges associated with sentiment analysis[2]. Bing Liu's book is a comprehensive resource that covers the fundamentals of sentiment analysis and opinion mining. It provides a detailed discussion of techniques and methods for Regression classification, feature selection, and sentiment lexicon creation[3]. This study uses five alternative machine learning (ML) regression algorithms—Linear Regression, Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, and Poisson Regression—to examine in depth how to estimate the price of cotton. metrics for assessing the effectiveness of ML models. The data set includes cotton prices for several Indian states. In this study, the cotton price 2019 data set is used. Prediction values are compared to 2020 data values, and it is found that the predicted prices are getting closer to the actual values. When compared to all other algorithms, Boosted Decision Tree performs well, followed by Bayesian and Linear Regression[4]. The author introduces to these potent methods in the domain of renewable energy power forecasting employing several Deep Learning and Artificial Neural Network algorithms, such as Deep Belief Networks, AutoEncoder, and LSTM. In research studies, they combined these algorithms to demonstrate how well they predicted the energy output of 21 solar power plants when compared to a normal MLP and a physical forecasting model. According to results, Deep Learning algorithms outperform Artificial

ISBN: 978-93-343-1044-3

Neural Networks and other reference models, such as physical models, in terms of forecasting.[5]. These issues have been solved using the Natural Language Processing (NLP) method, together with machine learning (ML) and deep learning (DL) techniques. The corpus used to train the model affects the correctness of the job. Support Vector Machine (SVM) of ML produced greater accuracy among the articles that were reviewed than other ML approaches. Researchers are experimenting with various DL algorithms as the use of DL approaches for NLP grows. Most of the NLP with Review Discussion in this paper will point researchers working on NLP in Tamil in the right direction and help them select the best Deep Learning and Machine Learning algorithms to produce reliable results.[7]. Significant features have been chosen using the supervised Boruta feature selection process. Then, on the basis of the processed feature set, cutting-edge machine learning and deep learning algorithms such as Gradient Boosting (GB), Extra Tree Regression (ERT), Deep Neural Network (DNN), and Long Short Term Memory Network (LSTM) are used to meticulously assess the degree of predictability of the aforementioned assets. In order to draw conclusions, a variety of numerical and statistical tests have been performed on the integrated prediction frameworks. Also used to investigate the nature and influence of various traits are Explainable AI frameworks. The results do indicate that, despite having highly volatile characteristics, both India VIX and historic volatility can be accurately forecasted using the suggested designs and provide useful, practical information.[9].

III. PROPOSED WORK

This study proposes a comprehensive approach that utilizes time series analysis, time series forecasting, and demand prediction techniques to enhance sales insight and optimize sales forecasting processes. Preprocessing, model choice, execution, assessment, creation of the ensemble model, exploration of the information (EDA), and assortment compilation are all component parts of careful research method. The proposed work consists of the following key components:

A. Time Series Analysis: Time series analysis will be utilized to develop an in-depth insight into the past sales history. This will include investigating trends, seasonality, and any underlying patterns in the data. Decomposition, autocorrelation function analysis, and stationarity testing will be used to reveal the temporal pattern in the sales data..

B. Feature Engineering: The features that are pertinent will be engineered from the historical sales data in order to extract useful information that can be helpful in demand prediction. The most impactful variables for forecasting precision will be identified using feature selection techniques.

C. Time Series Forecasting Models: Different time series forecasting models will be developed and compared to forecast future sales demand. Some of the methods that will be tried include classical methods such as ARIMA (Auto Regressive Integrated Moving Average) and advanced machine learning methods such as Prophet. The models will be compared on the basis of their performance regarding forecasting accuracy and stability.

D. Demand Prediction:. The central emphasis of this project is to formulate a demand forecast model that exploits the findings from time series forecasting and forecasting models. This model will allow firms to predict customers' demand properly, which can facilitate them to plan inventory levels, production levels, and selling strategies optimally.

E. Model Evaluation and Validation: The developed models will be tested stringently with suitable performance measures, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Cross-validation methods will be utilized to determine model generalizability to new data to guarantee their robustness in real-world applications.

IV. METHODOLOGY

A. Data Collection:

Gather historical sales data, which typically includes transaction records, product information, customer data, and relevant external factors. Ensure data quality through data cleaning and validation procedures. Store the collected data in a structured and accessible format for analysis.

B. Exploratory Data Analysis (EDA):

Conduct initial exploratory data analysis to gain insights into the dataset's characteristics. Visualize the data to identify trends, seasonality, and potential outliers. Perform statistical tests for stationarity to assess the need for differencing.

C. Time Series Decomposition:

Decompose the time series data into its constituent components, such as trend, seasonality, and residual (error).Decomposition helps in understanding the underlying patterns and removes noise from the data.

D. Feature Engineering:

Engineer relevant features from the data that may impact sales demand. Create lag features to capture historical dependencies. Encode categorical variables and perform one-hot encoding or label encoding as needed.

E. Time Series Forecasting Models:

Implement a variety of time series forecasting models, including but not limited to:

1) ARIMA (Auto-Regressive Integrated Moving Average)

2) SARIMA (Seasonal ARIMA)

3) Prophet

4) Ensemble model

F. Model Evaluation:

Evaluate models using appropriate metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and possibly others relevant to business objectives. Compare the performance of different models to select the most accurate one.

ISBN: 978-93-343-1044-3

V. EXPERIMENTS AND RESULTS

To validate the proposed methodology for sales insight and forecasting using time series analysis, time series forecasting, and demand prediction, a series of experiments were conducted. These experiments were carried out using real-world sales data from a retail company. Here is an overview of the experiments and their results:

a) Exploratory Data Analysis (EDA):

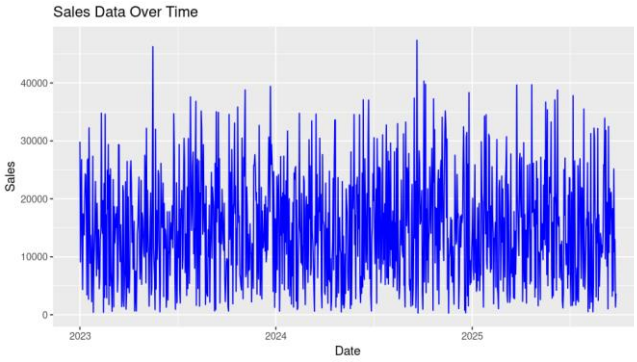

Figure - 1: Sales over time

The figure 1is a graph that illustrates the total sales figures for different years. The X-axis displays the years, while the Y-axis represents the total sales valueThe graph provides a visual comparison of sales performance across the specified years, allowing for easy identification of trends, fluctuations, and relative sales volumes for each year.

The analysis in fig .2 presents a time series representation of monthly sales data. The X-axis signifies time, with each point or interval corresponding to a month. The Y-axis represents the sales values, often in a specific currency or unit. The data points or lines on the graph illustrate the sales figures for each month over a defined period, allowing for the observation of sales trends, seasonality, and fluctuations.

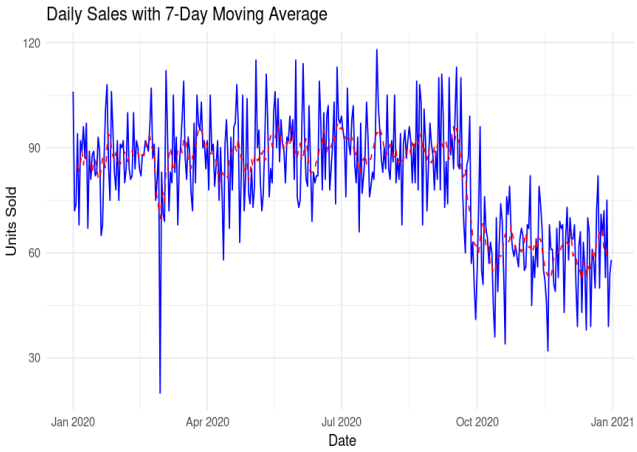

Figure – 2 Daily sales with 7 days moving average

b) Time Series Decomposition:

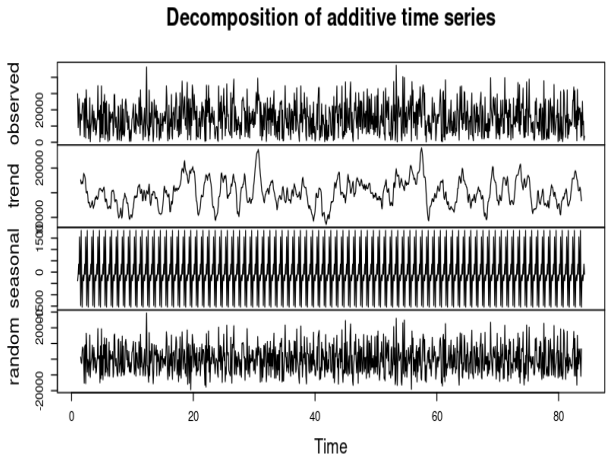

Figure – 3: Decomposition of additive time series

By examining this time series, one can identify patterns, such as monthly peaks or valleys in sales, and assess the overall performance and growth of the business across different months and years. The analysis may include additional information, such as trendiness, moving averages, or annotations to highlight key insights or events within the time series as shown in figure.4 Decomposition of additive time series.

c) Feature Engineering:

Figure – 4: sales based on features

The image 4 is a bar graph that illustrates the total sales figures for different years. The X-axis displays the years, while the Y-axis represents the total sales value. Each year is represented by a separate vertical bar, with the height of each bar indicating the corresponding sales amount. The graph provides a visual comparison of sales performance across the specified years, allowing for easy identification of trends, fluctuations, and relative sales volumes for each year.

ISBN: 978-93-343-1044-3

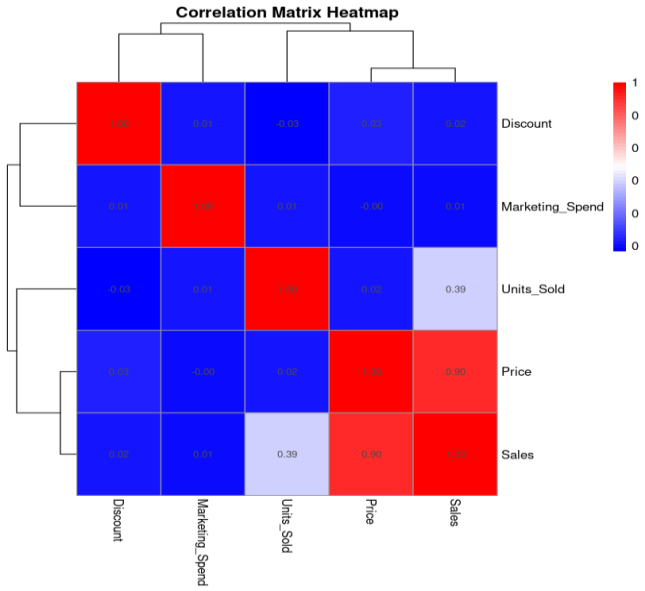

Figure – 5: Correlation Matrix Heatmap

The correlation matrix heatmap is a graphical representation of the relationships between variables within a dataset. It typically takes the form of a grid, where rows and columns correspond to the variables being analyzed. Each cell in the grid contains a color-coded value that represents the strength and direction of the correlation between the two variables it connects.

d) Time Series Forecasting Models:

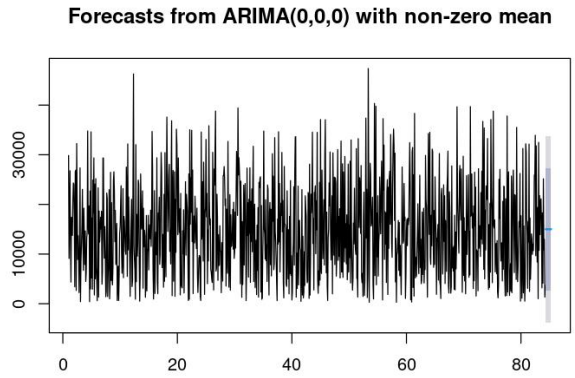

Figure – 6: ARIMA MODEL

The ARIMA (AutoRegressive Integrated Moving Average) model was used to identify linear relationships in the time series. It performs well with non-stationary data with trends and seasonality.

The SARIMA (Seasonal ARIMA) model is an extension of ARIMA that adds seasonal patterns and thus performs better with data that has evident seasonal trends.

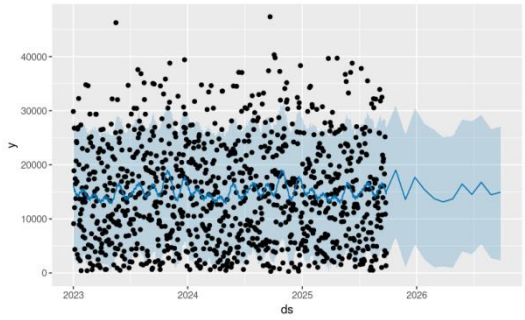

Figure – 7: SARIMA MODEL

e) Evalution and performance matrix

The performance of the models was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results are summarized in the table below:

Model	MAE	MSE	RMSE
ARIMA	7994.297	92051625	9594.354
SARIMA	7918.080	91719693	9577.040
Prophet	7910.179	90489920	9512.619

Table 1: Performance metrics

Model	MAE Improvement	MSE Improvement	RSME Improvement
SARIMA	0.95	0.36	0.18
PROPHET	1.05	1.70	0.85

Table 2: Comparison of models

The table 2 highlights the improvements in MAE, MSE, and RMSE for SARIMA and Prophet models compared to ARIMA.

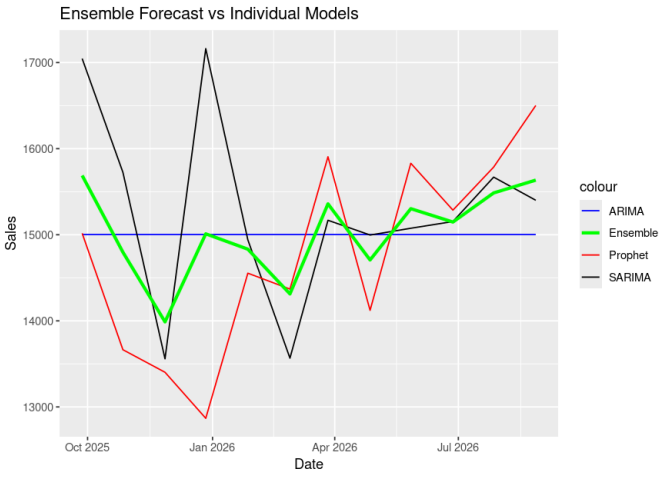

Figure – 8:Comaparison of models with Ensemble model

ISBN: 978-93-343-1044-3

The ensemble model integrates predictions from the individual models (ARIMA, SARIMA, and Prophet) to provide improved accuracy. The comparison chart in Figure 8 shows the effectiveness of the ensemble method with reduced error rates and overall better performance.

VI. CONCLUSION

The effectiveness of time series models in sales forecasting is demonstrated through experiments conducted using real sales data. The Prophet model was more accurate than the ARIMA and SARIMA models, despite both models having robust forecasts. The performance was improved even more by the application of an ensemble model, which reduced errors across all criteria for evaluation.

The results indicate that seasonal as well as trend-based patterns are better captured by an integrated approach that utilizes a number of models. Enterprises can utilize this comprehensive forecasting procedure to prepare for future demand, streamline inventory control, and take informed decisions.

VII. REFERENCES

[1] Taryney, P. D "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Sales" National Research Council of Canada, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, (2002), Philadelphia, Pennsylvania, 417-424

[2] Pang, B., & Lee, L. (2008)" Time Series Analysis on Concurrent Data"

[3] Liu, B. (2012)"Time Series Forecasting Models for Regression"

[4] Deepa.S, Alli, A., Sheetal, Gokila, S. "Machine learning regression model for material synthesis prices prediction in agriculture"Materials Today: Proceedings, 2021, 81(2), pp. 989–993

[5] Gensler, J. Henze, B. Sick, N. Raabe Deep Learning for solar power forecasting—An approach using Auto Encoder and LSTM neural network.

[6] Brownlee, J. (2018a). How to develop LSTM models for time series forecasting.

[7] Gokila, S., Rajeswari, S., Deepa, S." TAMIL- NLP: Roles and Impact of Machine Learning and Deep Learning with Natural Language Processing for Tamil, Proceedings of 8th IEEE International Conference on Science, Technology, Engineering and Mathematics, ICONSTEM 2023, 2023.

[8] Corrius, J. (2018). Simple stationarity tests on time series - bluekiri - Medium. [online] Medium. https://medium.com/bluekiri/simplestationarity-tests-on-time-series-ad227e2e6d48

[9] I. Ghosh, M.K. Sanyal. Introspecting predictability of market fear in Indian context during COVID-19 pandemic: An integrated approach of applied predictive modelling and explainable AI

[10] Alli, A., Vijay Fidelis, J., Deepa, S., Karthikeyan, E. "One-Dimensional Chaotic Function for Financial Applications Using Soft Computing Techniques". Lecture Notes in Networks and Systemsthis link is disabled, 2021, 127, pp. 463–468

ISBN: 978-93-343-1044-3