IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTIVE SALES ANALYTICS FOR VARIABLE TIMEFRAMES USING MULTIPLE MACHINE LEARNING MODELS

[1]Rushikesh Shinde, [2]Rafe Shamsi, [3]Pratham Nanaware, [4]Savita Lohiya

[1]Student, [2]Student, [3]Student, [4]Assistant Professor

[1]Department of Information Technology,

[1]SIES Graduate School of Technology, Navi Mumbai, Maharashtra, India.

Abstract: Predicting future sales accurately is crucial for businesses to optimize inventory management, plan marketing strategies, and enhance overall profitability. In this paper, we propose a comprehensive analysis of predictive sales analytics for variable timeframes utilizing machine learning models, specifically XGBoost, Linear Regression, Random Forest, and Long Short-Term Memory (LSTM) networks. Through empirical evaluation and comparative analysis, we demonstrate the efficacy of XGBoost in forecasting sales for different timeframes, namely daily, weekly, and monthly. Our findings highlight the superior performance of XGBoost in terms of accuracy and robustness, making it an ideal choice for businesses seeking reliable sales predictions across diverse temporal scales.

Index Terms - Predictive analytics, Sales forecasting, XGBoost, Random Forest, Linear Regression, LSTM, Variable timeframes, Streamlit.

I. INTRODUCTION

Sales forecasting serves as a pivotal element in assisting businesses to make well-informed decisions regarding resource allocation, inventory management, and strategic planning. However, conventional sales prediction methods often struggle to adequately capture the complexities inherent in today's rapidly evolving market landscape. To tackle this challenge, the utilization of machine learning techniques has emerged as a robust solution for generating precise sales forecasts by leveraging historical data patterns. Introducing Sales Prophesy, our research paper presents a sales prediction application designed to harness machine learning models for forecasting future sales trends. The principal aim of Sales Prophesy is to furnish businesses with a dependable and efficient means of projecting sales outcomes, thereby facilitating improved decision-making and strategic planning processes. This study delves into the development, implementation, and evaluation of Sales Prophesy, with a specific emphasis on its efficacy in generating accurate sales predictions. Additionally, we undertake a comparative analysis of several machine learning models, such as LSTM, Linear Regression, Random Forest, and XGBoost, to gauge their performance in terms of sales prediction accuracy and robustness.

Through this comparative analysis, our objective is to pinpoint the most effective approach for sales forecasting while also delineating the advantages of deploying Sales Prophesy within real-world business contexts. Ultimately, our research endeavours to underscore the potential of Sales Prophesy as a valuable asset for businesses seeking to refine their sales strategies and gain a competitive edge in the contemporary marketplace.

II. RELATED WORK

Previous research in sales forecasting has explored a plethora of methodologies ranging from statistical time series models to machine learning algorithms. Traditional techniques such as ARIMA (Autoregressive Integrated Moving Average) have been widely employed but often lack the flexibility to accommodate variable timeframes effectively. Recent advancements in machine learning have introduced more sophisticated approaches capable of capturing nonlinear relationships and temporal dependencies present in sales data. Linear Regression, XGBoost, Random Forest, and LSTM have emerged as prominent contenders due to their ability to handle complex datasets and deliver accurate predictions.

III. METHODOLOGY

In this section, we provide an overview of the dataset used for our experiments and describe the implementation details of the machine learning models considered in our study.

A. Dataset Description and Pre-processing

The dataset comprises historical daily sales data spanning multiple years, which we have pre-processed and segmented into weekly and monthly intervals for the daily data. Each record includes relevant features such as date, product ID, and sales volume. The data is then pre-processed to pivot the dataset and get the sales volume on monthly and weekly basis depending on the analysis required.

	A	B	C	D
1	date	store	item	sales
2	01-01-2013	1	1	13
3	02-01-2013	1	1	11
4	03-01-2013	1	1	14
5	04-01-2013	1	1	13
6	05-01-2013	1	1	10
7	06-01-2013	1	1	12
8	07-01-2013	1	1	10
9	08-01-2013	1	1	9
10	09-01-2013	1	1	12
11	10-01-2013	1	1	9
12	11-01-2013	1	1	9
13	12-01-2013	1	1	7
14	13-01-2013	1	1	10
15	14-01-2013	1	1	12
16	15-01-2013	1	1	5

B. Model Training

The equations are an exception to the prescribed. The dataset is split into training and testing sets using the train-test split method from the scikit-learn library in Python. The training set comprises a certain percentage (e.g., 80%) of the data, while the testing set contains the remaining samples. This partitioning ensures that the models are trained on a subset of the data and evaluated on unseen data to assess their generalization performance.

C. Cross-Validation

To ensure the robustness of our findings, we employ cross-validation techniques such as k-fold cross-validation. This involves partitioning the dataset into k subsets and training the models on k-1 folds while validating on the remaining fold. This process is repeated k times, with each fold serving as the validation set exactly once. The average performance across all folds provides a more reliable estimate of the model's generalization performance.

D. Experimental Setup

All experiments are conducted in Python programming language utilizing widely-used libraries such as scikit-learn, XGBoost, pandas, matplotlib and Keras. The experiments are performed on a standard computing environment with sufficient computational resources to ensure timely execution.

E. Statistical Analysis

Statistical tests may be employed to assess the significance of observed differences in model performance. Techniques such as paired t-tests or ANOVA can help determine whether the performance differences between models are statistically significant.

F. Machine Learning Models

Three machine learning models are considered for this study: XGBoost, Linear Regression, Random Forest, and LSTM. These models are implemented using popular libraries in Python, including XGBoost, scikit-learn, and Keras.

- XGBoost: XGBoost is a scalable and efficient gradient boosting algorithm known for its superior performance in classification and regression tasks. It constructs an ensemble of weak learners in a sequential manner, optimizing a differentiable loss function at each iteration.

Fig 3.1 Prediction using XGBoost Model

- Random Forest: Random Forest is a powerful ensemble learning technique used for classification and regression tasks. It constructs multiple decision trees during training and aggregates their predictions to make final decisions. Each tree is trained on a random subset of the data and features, reducing overfitting. Random Forest provides robust predictions, handles large datasets well, and is widely used in fields like finance, healthcare, and bioinformatics for its versatility and accuracy.

Fig 3.2 Prediction using Random Forest Model

- LSTM: Long Short-Term Memory networks are a type of recurrent neural network (RNN) designed to model temporal sequences and capture long-range dependencies. LSTMs excel at learning patterns in time-series data and are well-suited for forecasting tasks. Long Short-Term Memory networks are implemented using Keras, a popular deep learning library in Python. The LSTM model is trained on sequential data representing the time-series nature of the sales data. Hyper parameters such as the number of LSTM units, dropout rate, and learning rate are tuned using techniques such as grid search or random search.

Fig 3.3 Prediction using LSTM Model

- Linear Regression: Linear regression is a statistical technique used to understand the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to fit a straight line to the data points that minimizes the difference

between observed and predicted values. This method facilitates prediction and inference, making it valuable in various fields such as economics, finance, and social sciences.

Fig 3.4 Prediction using Linear Regression Model

IV. PROPOSED SYSTEM

Sales Prophesy is designed as a web-based application utilizing Streamlit, a Python library for creating interactive web applications. The architecture of the application comprises several components, each serving a distinct role in facilitating the sales prediction process.

- **Data Upload Module**

The Data Upload Module allows users to upload a CSV file containing historical sales data. Upon file upload, the module performs data validation and preprocessing to ensure compatibility with the subsequent analysis steps. Users are provided with feedback on the uploaded file's format and integrity.

Fig. 4.1 Data Upload Module

- **Product Selection Interface**

Once the data is uploaded, users are presented with a Product Selection Interface where they can choose the specific product for which they seek sales predictions. This interface displays a list of available product IDs extracted from the uploaded CSV file, enabling users to make informed selections based on their business requirements.

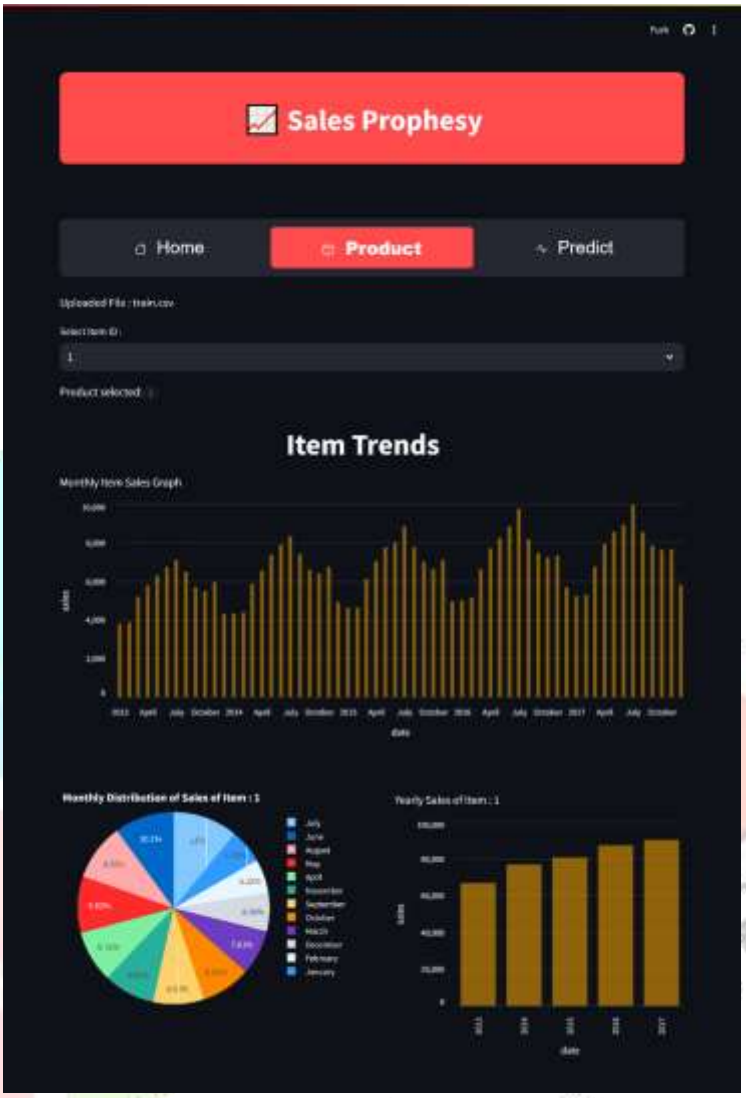

Fig 4.2 Product Selection and Visualization Dashboard

- **Visualization Dashboard**

Following product selection, users are directed to a Visualization Dashboard showcasing various graphical representations of the selected product's sales data. The dashboard includes interactive visualizations such as line charts, bar graphs, and histograms, allowing users to explore trends, patterns, and seasonality within the sales data. These visualizations serve to enhance users' understanding of the product's sales dynamics and inform their decision-making process.

- **Prediction Module**

Upon reviewing the sales data visualizations, users can proceed to the Prediction Module to generate sales forecasts for the selected product. Leveraging the trained XGBoost model, this module processes the historical sales data to produce predictions for future sales periods. Users have the option to specify the forecasting horizon, enabling them to tailor predictions according to their planning horizon and business objectives.

- **Results Presentation**

Once the sales predictions are computed, users are presented with the forecasted sales figures in tabular format. Additionally, the Prediction Module generates various graphical representations, including line

charts illustrating predicted sales trajectories and confidence intervals. These visual aids facilitate the interpretation of the forecasted sales trends and assist users in making informed decisions regarding inventory management, resource allocation, and strategic planning.

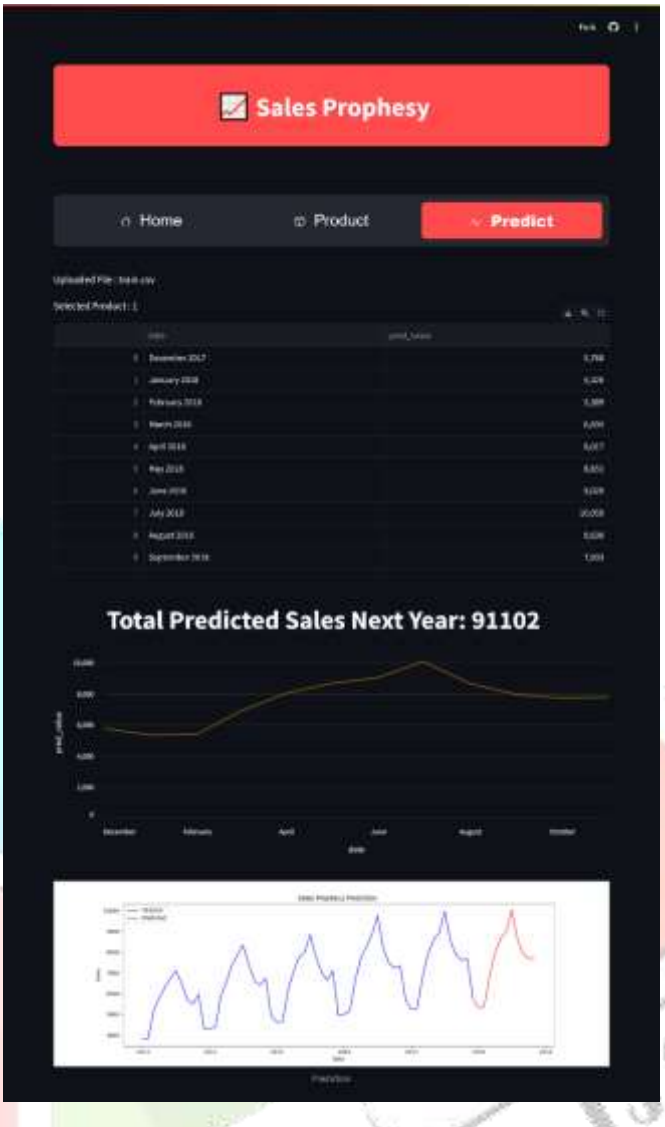

Fig. 4.3 Prediction Module and Result Presentation

V. OBJECTIVES

The objectives of the "Sales Prophesy: Intelligent Sales Forecasting and Inventory Optimization" project can be summarized as follows:

- Accurate Sales Forecasting: Develop a robust machine learning model that utilizes historical sales data to accurately predict future sales trends for individual products, product categories, and the entire inventory.
- Inventory Optimization: Provide real-time inventory management recommendations, including restocking levels and reorder points, to ensure that businesses maintain and optimal balance of inventory, reducing stockouts and carrying costs.
- User-Friendly Interface: Develop an intuitive and user-friendly application interface that allows businesses to easily visualize sales predictions, inventory recommendations, and marketing strategies through dashboards and reports.

VI. EXPERIMENTAL RESULTS

We conducted experiments to evaluate the performance of XGBoost, Random Forest, and LSTM across different timeframes. Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared were used to assess the accuracy of each model. A comparative analysis is conducted to assess the strengths and weaknesses of each model across different timeframes (daily, weekly, and monthly).

VII. RESULTS AND DISCUSSION

Our results indicate that XGBoost consistently outperforms Random Forest and LSTM across all timeframes in terms of prediction accuracy. Table 1 summarizes the comparative performance of the four models.

Model	RMSE	MAE	R2
LSTM	707.561423	540.000000	0.981005
Random Forest	624.097882	470.833333	0.985222
XGBoost	573.957461	380.000000	0.987501
Linear Regression	481.503548	385.166667	0.991204

Table 1: Comparative Performance of XGBoost, Linear Regression, Random Forest, and LSTM

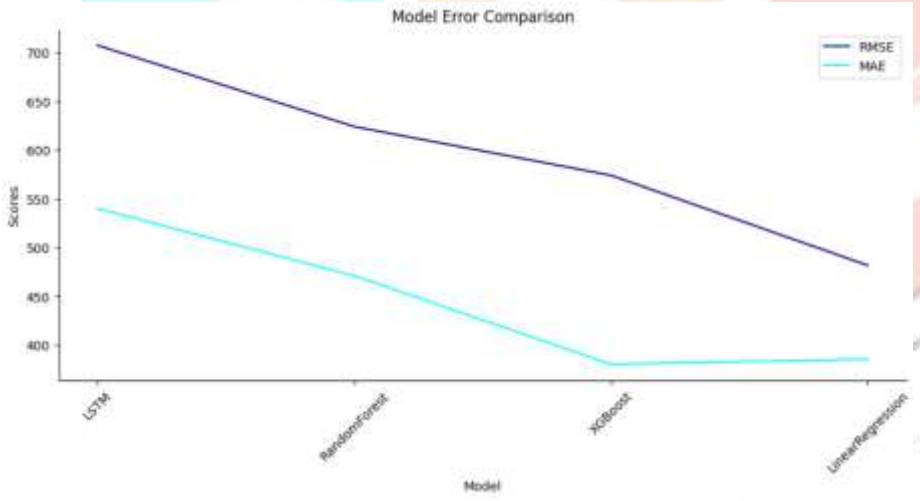

Fig. 7.1 Line graph of Comparative Performance of XGBoost, Linear Regression, Random Forest, and LSTM

VIII. CONCLUSION

This paper offers a thorough examination of predictive sales analytics employing XGBoost, Random Forest, and LSTM across variable timeframes. Our analysis highlights XGBoost's superior accuracy in forecasting sales over daily, weekly, and monthly intervals. The robust performance of XGBoost underscores its efficacy in handling the intricacies of sales data, presenting valuable insights for businesses requiring dependable sales predictions across diverse temporal scales. Additionally, our research supports the implementation of Sales Prophesy, leveraging the strengths of XGBoost to enhance sales forecasting accuracy and facilitate strategic decision-making in real-world business scenarios.

IX. FUTURE WORK

Future research could explore additional features and pre-processing techniques to further enhance the predictive performance of machine learning models in sales forecasting. Additionally, investigating ensemble approaches that combine the strengths of multiple algorithms may yield even more accurate predictions.

REFERENCES

[1] Bohdan M. Pavlyshenko, "Machine-Learning Models for Sales Time Series Forecasting," IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2018.

[2] Youness Jouilil and Driss Mentagui, "Comparing the Forecasting Accuracy Metrics of Support Vector Regression and ARIMA Algorithms for Non-Stationary Time Process," Mathematics and Statistics, Vol.11, No.2, pp. 294-299, 2023.

[3] Praveen K B, Pradyumna Kumar, Prateek J, Pragathi G, "Inventory Management using Machine Learning," International Journal of Engineering Research and V9(06), 2020.

[4] Frank M. Thiesing and Oliver Vornberger, "Forecasting Sales Using Neural Networks," Proceedings of International Conference on Neural Networks (ICNN'97), Springer, Berlin, Heidelberg, 2005.

[5] Tora Fahrudin, Nelsi Wisn, Patrick Adolf Telnoni, Dedy Rahman Wijaya, "Sales Forecasting Web Application in Small and Medium Enterprise," International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), Jakarta, Indonesia, 2022.

[6] Liliya Demidova and Mariya Ivkina, "Development and Research of the Forecasting Models Based on the Time Series Using the Random Forest Algorithm," 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), Lipetsk, Russia, 2020.

[7] Zixuan Huo, "Sales Prediction based on Machine Learning," 2nd International Conference on E-Commerce and Internet Technology (ECIT), Hangzhou, China, 2021.

[8] M.Krishna Satya Varma, N.Sai Durga Manikanta, M.Hemanth, M.Vinay and K.Bharath Sri Sai Pradeep, "Sales Forecasting Using Xgboost," International Journal of Creative Research Thoughts (IJCRT), 2023.

[9] Xie dairu and Zhang Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2021.

[10] Choujun Zhan, Jianbin Li, Wei Jiang, Wei Sha and Yijing Guo, "E-commerce Sales Forecast Based on Ensemble Learning," IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN), Chongqing, China, 2020.