International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Leveraging Machine Learning for Superior E-Commerce Performance: Advanced Sales Forecasting Techniques for Business Growth

Dhaval Gogri[1], Niketu Dave[2], Ashish Korpe[3]

[1]Fremont, USA
Email: *dhaval.gogri17[at]gmail.com*

[2]Fremont, USA
Email: *niketudave[at]gmail.com*

[3]Dublin, USA
Email: *ashish.korpe[at]gmail.com*

Abstract: ***In today's highly consumer-oriented business environment, any enterprise in search of high sales results must find an optimal relationship between supply demand and costs of storage. Adequate sales forecasts are especially important when it comes to helping firms enhance their operations. The advancement in Artificial Intelligence has resulted in an increase in the methods used in the solution of the forecasting problem. Concerning essential algorithms for sales forecasting, this paper employs RF, XGBoost, ARIMA, and KNN. These algorithms were applied to a Walmart real-world sales data set, which is illustrated in Figure 4. The performance of the algorithm results was further compared with each to make use of the R2 score, MAE, and RMSE. The results depicted below show that the RF and XGBoost algorithms offered higher accuracy compared to the other algorithms. The results recommended that the RF model presents the best performance with R^2=96% and MAE=19.85. Furthermore, the XGB model presents a nearly comparable performance with an R^2 score of 94% and an MAE of 55.34. In contrast, ARIMA and KNN models demonstrate comparatively lower accuracy. As a result, the study could be valuable in illustrating how Assuming, the appropriate utilization of the results of machine learning approaches could enhance the actuality of sales forecasts aimed at e-business realization. Future work may thus look at how more data could be incorporated, including customer satisfaction survey results and current market trends, for the purpose of improving the credibility of the model.***

Keywords: Sales forecasting, machine learning, e-commerce, sentiment analysis, business sector, retail

1. Introduction

In today's global business environment where information technology is rapidly becoming the driving force, more and more companies from different industries have been adopting analytical solutions approaches as a way of coping with the increasing pace. Sales forecasting is one of the fields that benefit greatly from data analytics, especially among retail companies. [1][2][3]. By leveraging historical sales data, demographic information, economic indicators, and other relevant factors, retailers can anticipate consumer demand more accurately, optimize inventory management and ultimately improve profitability.

Each company requires sales to be effective, which means that sales must be planned for within any retail center. Thus, to pull higher traffic to each and every store, the mall will endeavor to have cornered and time-sensitive promotions. Many companies and organizations pay special attention to statistical data concerning their products with a view to predicting the customer demand[4]. A sales forecast helps every company to make better business decisions [5][6]. It supports overall business strategy, cost and value, as well as strategic and operating risk management. Forecasting for the calculation of business income at the product level, at the level of an individual company, or at the overall level of the enterprise.

Greater accuracy in prediction is always valuable, in the creation as well as improvement of methods of sale in the trade. Forecasting is important to organizations with new markets, new products and services, and organizations that are experiencing high growth. Sales forecasting is mainly used by businesses in assessing the following two things. First of all, to depict more clear the state present needs of the consumer for the product and service. Second, it is used to estimate future demand for certain products and services within a specific firm. In order to predict the future demand for products, retailers are devoting much time to research. Currently, the long-term outlook for the sales demand of an organization is determined by complex machine learning equations [7]. In order to make very accurate estimations about the next occurrences, an ML model is produced with data to discern cycles. Machine learning models have emerged as indispensable tools in grocery sales analysis [8]. Its simplicity and interpretability make it an attractive choice for modelling sales trends [9].

a) Motivation and Contribution of Study

The development of competition in the context of e-commerce tends to have a significant rate due to the availability of demands for driving business growth by using data analysis. Another key subtopic under sales forecasting includes its importance as it is a vital tool used in management decision-making processes since it helps in determining the right combination of inventory levels, distribution of resources, and even the right prices to charge for an organization's products. The classical methods of forecasting are not able to solve the problems that occur when working with a large amount of data

characteristic of modern e-shops. Machine learning offers advanced techniques that can handle this complexity, making it possible to generate more accurate and actionable sales predictions. This study is motivated by the need to improve sales forecasting accuracy in e-commerce, which directly impacts business growth by enhancing operational efficiency, reducing costs, and maximizing revenue opportunities. This study makes several key contributions to the field of e-commerce sales forecasting.

- To collect and use real-world e-commerce sales data for advanced ML applications.
- To implement label encoding for converting categorical data into numerical format.
- To perform feature selection for identifying the most relevant variables to enhance forecast accuracy.
- To apply machine learning models like Random Forest, XGBoost, ARIMA, and KNN, and evaluate them using MAE, RMSE, and R^2.
- To establish a matrix-based comparison of models, providing insights on their performance for sales forecasting.

b) Structure of this paper

The structure used for this paper is provided. Section 1 Introduction to sales forecasting. In section two, literature review is presented. In section 3, the approach to the practical application of the research. Section 4 In the results and discussion section, a comparative assessment of developed models, according to their performance. Lastly, Section 5 - Future Directions comprises the conclusion.

2. Literature Review

This section gives the literature study of earlier work on Performance Sales Forecasting using machine learning Techniques are as follows:

In the research work presented in this paper, Spuritha et al., (2021) the authors use a model of XGBoost in which the learners are calibrated toward the best store-product subsets for the improvement of the overall accuracy in sales estimation. The proposed model forecasted sales for ten stores with fifty products each, thereby producing overall average MAPE, RMSE and R2 at 11.98 %, 6.63 and 0.76 respectively. Further, the dynamic pricing is practiced on the forecasted result which shows the appropriate price for a product with reference to the demand[10].

In this paper, Abubaker and Ala'khalifeh, (2023) the approach of SF is introduced with the help of a recurrent neural network (RNN) in order to arrive at forecast of a market product. RNN is a modern type of neural networks which have been used in this paper kind due to its usage for model-based predictions in numerous applications. One market only has been taken into consideration to be a typical market for which the data is the prices of 800 products for 49 weeks. Using several trials of the training, the key parameters of the RNN were optimized to yield quite reasonable predictions in terms of RMSE of 0.039[11].

In this paper, Chen et al., (2021) have developed the sale prediction Neural Network sale prediction model for Walmart in this paper. Additionally, we assess the performance of the proposed. The results have applied the employ of the NN model on the benchmarks situated at the Kaggle platform. Studies and varying tests have revealed that NN model has a better performance than any other machine learning model. The optimal RMSE in this research is 2.92, which is lower by 2.58 than the RMSE of the Linear Regression Algorithm and the SVM Algorithm. Furthermore, for the mining attributes of different dimensions to run the prediction efficiently, we use SHAP to interpret the NN model[12].

In this paper, Ding et al., (2020) have developed and used a novel sales forecast model with Cat Boosting. The presented algorithm is verified on the Walmart sales dataset which is the largest dataset on this subject. In order to increase the prediction rate as well as the efficiency of the forecasts, we employed great techniques of feature engineering. In experiments, our model is better than some basic machine learning approaches like Linear Regression and SVM, and the RMSE is equal is 0. 605. Our method does not require the finetuning Confocal as other methods, and this makes it more generalizable to other customization datasets, further broadening its functionality[13].

In this work, Ensafi et al., (2022) a public record of a retail store's sales record is used including the steps below in the prediction of furniture sales. For this purpose, several forecasting models are employed. Firstly, simple techniques of time series such as Seasonal Autoregressive Integrated Moving Average Model (SARIMA) and Triple Exponential Smoothing Method are used. Following that, Prophet, LSTM, and CNN models are utilized. Further explanation of these models can be found in the next section. In this case, different accuracy assessment methods such as the RMSE and MAPE are used to assess the performances of the respective models. Thus, in determining the acceptable level of errors, the model with RMSE of 128.51 is acceptable. The findings depict that the use of Stacked LSTM method can predict the demand better that the rest of the methods. Moreover, regarding the performances the Prophet and CNN models using the SAR and MAPE are good[14].

Taking the Walmart supermarket dataset from Kaggle competition, this paper, Niu, (2020) measures the performance of XGBoost sale forecast model in sales data sets. Algorithms of experiment show that our method outperforms all the other machine learning methods for the particular problem. The RMSSE obtained by this paper is 0.141 and 0.113 better than Logistic regression algorithm and Ridge algorithm, respectively. Moreover, this paper also researches the importance of ranking features and obtains some constructive guidance. Table 1 organizes each study with its respective method, dataset, results, and limitations or suggestions for future research[15].

Table 1: Summary of previous study on sales prediction using different approaches

Author	Method	Dataset	Results	Limitations/Future Work
Spuritha et al., 2021	XGBoost	10 stores with 50 products	MAPE: 11.98%, RMSE: 6.63, R^2: 0.76	Limited scope in number of stores/products; future work may involve applying the model to larger datasets or integrating external factors like seasonality.
Abubaker and Ala'khalifeh, 2023	Recurrent Neural Network (RNN)	800 product sales over 49 weeks	RMSE: 0.039	The model relies on short-term past sales (3 weeks) as inputs; future work could involve more advanced feature engineering or external factors like promotions.
Chen et al., 2021	Neural Network (NN)	Walmart sales dataset from Kaggle	RMSE: 2.92 (compared to LiR), RMSE: 2.58 (compared to SVM)	Model interpretation is limited despite SHAP use; future work could enhance explainability and explore more complex neural network architectures.
Ding et al., 2020	CatBoost	Walmart sales dataset	RMSE: 0.605	Limited to Walmart dataset; future work might include testing on a wider variety of datasets or comparing performance with deep learning models.
Ensafi et al., 2022	SARIMA, Triple Exponential Smoothing, Prophet, LSTM, CNN	Retail store's furniture sales dataset	RMSE: 128.51	The study's focus was on furniture sales, which may limit generalisation; future work could examine broader retail categories or hybrid models.
Niu, 2020	XGBoost	Walmart sales dataset from Kaggle	RMSSE: 0.141 lower than Logistic Regression, 0.113 lower than Ridge Regression	Future work could explore optimising feature selection and enhancing the scalability of the model to different retail sectors.

3. Methodology

The methodology for Leveraging Machine Learning for Sales Forecasting Techniques for Business Growth begins with the collection of a large, unstructured dataset from Kaggle, specifically containing historic weekly sales data from Walmart stores. The data undergoes extensive preprocessing, which includes filling in missing values, removing noisy data through outlier detection, and applying label encoding to convert categorical variables into numerical format. Min-max normalization is used to rescale data, ensuring uniformity and preventing any feature from disproportionately influencing the models. Feature selection follows, where relevant features are chosen based on domain knowledge to optimize model performance. The dataset is then split into training and testing sets using an 80:20 ratio for the evaluation of the outperforming capacity of the model. In order to answer the second research question, many of the different machine learning algorithms like Random forest, XG Boost, ARIMA, and KNN are implemented for sales forecasts. Efficiency of these models is determined by certain parameters such as Mean Absolute Error, Root-mean square error and R^2 score. The following flowchart for research with each step and phase are displayed in Figure 1.

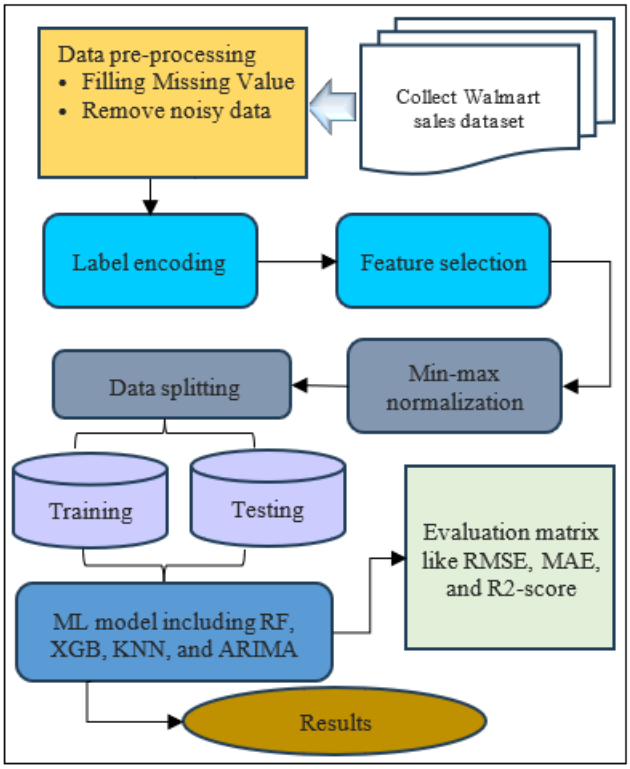


Figure 1: Flowchart for sales forecasting in Business Growth

a) Data Collection

The dataset for study can be from the Kaggle website hosted by Walmart, this dataset is large entries and it's unstructured data. It also holds data of weekly sales volume of nearly 45 Walmart stores along with the departmental data of diverse regions of the country. The instinct and illustration for information are provided below blow:

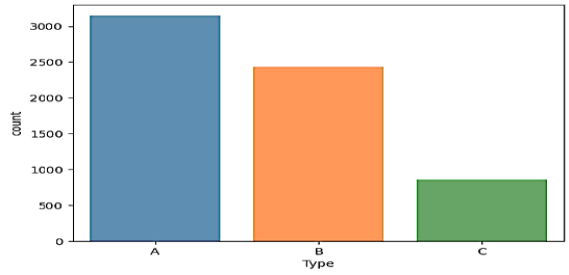


Figure 2: Count plot for Product Type

Figure 2 shows sales distribution across three product types A, B, and C where A has the highest sales, close to 3000 units, followed by Product Type B with just over 2500. Product Type C has the lowest sales, slightly above 500, indicating A is the most popular and C the least.

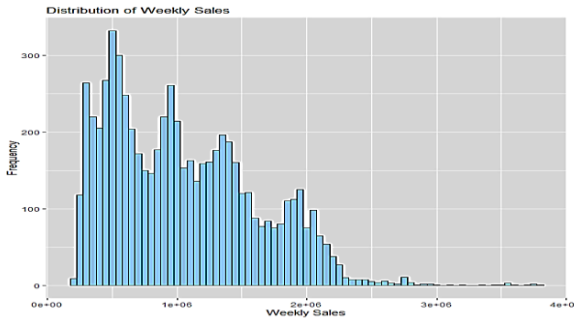


Figure 3: Histogram of Walmart's weekly sales

The histogram in Figure 3 provided shows the distribution of Walmart's weekly sales. Most of the sales figures fall between 0 and 1 million dollars, with several peaks indicating that certain sales amounts occur more frequently. This visual representation can help in analysing sales patterns and making informed business decisions.

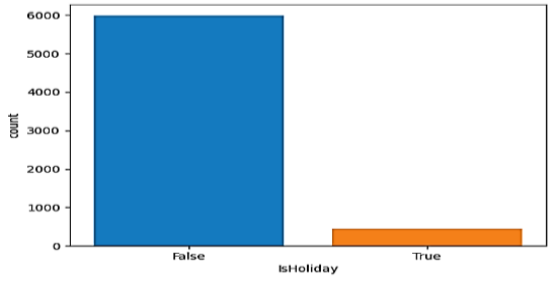


Figure 4: Holiday Sales Analysis

Figure 4 shows the Holiday Sales Analysis compares sales during holiday and non-holiday periods. Non-holiday sales, represented by the blue bar, reach up to 6000, indicating a much higher count compared to holiday sales, which are represented by the orange bar and barely exceed 1000. This suggests that sales are significantly higher during non-holiday periods.

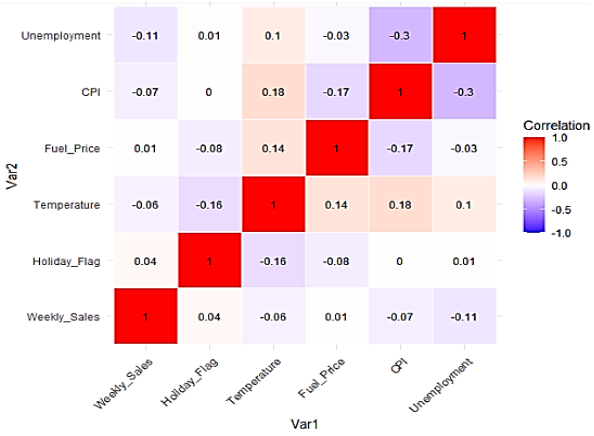


Figure 5: Correlation plot between variables

The following Figure 5 shows a correlation matrix heatmap. It shows the correlation coefficients between several variables, including Unemployment, CPI (Consumer Price Index), Fuel_Price, Temperature, Holiday_Flag, and Weekly_Sales. The colors range from red to blue, indicating the direction and degree of the association white means no correlation while blue denotes negative correlation and red vertical bar opposes it with positive correlation.

b) Data Preprocessing

Data preprocessing, which entails organizing, cleansing, and modifying databases to extract valuable information, is a crucial component of data mining. There are several procedures to perform before processing may start. Among them are transformation, integration, purification, and reduction of datasets. Several processes were employed in this project to collect the data in the correct format. The essential preprocessing actions are listed below:

Filling Missing Value: There are several ways to address missing values. Want to apply a sophisticated imputation approach, delete cases with missing data, or fill in values that are missing with zeros or the mean [16].

Remove noisy data: To remove noisy data, identify outliers and irrelevant features using statistical methods and visualizations.

c) Label encoding

Fast-forward to Label encoding which is an extraction activity in the machine learning system where labels or strings are encoded numerically. Where each category is given an integer and the values are made discrete through a process of making the categorical variables into numerical ones. The main purpose is to lead the data into machine learning algorithms which only accept numerical data. Ordinal data is the data which is arranged in the category and an improved order such that one category is more significant or inferior to the other. However, one has to be a little careful when working with label encoding and always pay a lot of attention to the surrounding environment.

d) Data Normalization with min-max

That is the reason why Normalization is the pre-processing of the methodology. Quantization flips the data set from the original quantitative range to another quantitative range that is pre-designated. [17]. Normalization is applied to standardize

the data range within a fixed level in order to reduce the impact of any specific value. In this work, Min-Max Normalization is used which is carried out using the following formula: (1)

$$\hat{x}_n = \frac{x_n - x_n(min)}{x_n(\max) - x_n(\min)} \quad (1)$$

where is the minimum of the shown elements in the given data of the range and Likewise max.

e) Feature selection

Other important step in designing the model is feature selection for the selection of the required features from the data set depending on the knowledge. The data set used here has numerous characteristics from which we select the required characteristics necessary for enhancing the measurement of performance. It is the process of selecting optimal features for your machine learning model based on the type of problem attempting to solve it including or excluding some of the features while others remain invariant.

f) Data Splitting

The Walmart sales data is split in an 80:20 ratio, a dataset is split, 80% of the data is applied for training the machine learning model and the remaining 20% is used for model evaluation.

g) Models Implementation

For the purpose of Sales Forecasting Techniques for Business Growth use multiple machine learning models that are described in below:

1) Random Forest (RF)

Utilizing an ensemble approach and a well-liked statistical-based learning strategy, Random Forest forecasts utilizing The bootstrapping method and at the same time, decision trees are done in combination, therefore extracting valuable information from the original samples[18][19]. The final findings of the RF regression approach are obtained by taking the mean of the predictions [20]. It may be described as a type of additive model that uses judgements from other learning models to generate predictions. Conventionally, this is stated as (2):

$$g(x) = f_0(x) + f_1(x) + f_2(x) \quad (2)$$

In order to produce more accurate and dependable prediction results, RF builds many decision trees and merges them. RF increases the model's unpredictability as the decision trees develop.

2) XGBoost

An ensemble learning system called XGBoost makes predictions using decision trees. XGBoost is an ensemble learning system that can work with regression problems Decision trees are used to make predictions. The mathematical model for XGBoost regression can be expressed as (3):

$$y = f(x) \quad (3)$$

where y is the predicted property price, The input feature vector is denoted by x (such as the quantity of bedrooms, square footage, etc.), and f(x) is the XGBoost model that predicts y based on x. To compute f(x), XGBoost builds an group of decision trees with the training objective of minimizing the MSE loss function. To get a final prediction, the model integrates the forecasts from several decision trees. The XGBoost reversion model's general form is represented by (4):

$$y = \sum(k = 1\ to\ K)\ fk(x) \quad (4)$$

Where K is the total number of decision trees in the ensemble and fk(x) is the forecast of the k-th decision tree. The prediction of each tree is a weighted sum of the leaf values of the tree, which are learned during training [21][22]. The prediction of the XGBoost model for a given input x is obtained by summing the predictions of all the decision trees in the ensemble.

A. Evaluation Parameters

An essential stage of the machine learning system is model assessment. It enables us to evaluate the efficacy of ML models and acquire a deeper understanding of their strengths and deficiencies RMSE, MAE, and R^2 value are utilized for the forecasting models comparison, as shown below:

Mean Absolute Error: It is a low measure of the errors of a set of forecasts. Since absolute doesn't consider the error's sign – positive or negative – it attributes the same value to each error. MAE is very easy to compute, as shown in Equation (5). These amounts are summed up and divided by the total number of observations giving the "total error." The absolute values of the mistakes refer to the amount by which the actual measurement differs from the one that was expected.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (5)$$

where the number of observations is n, the forecast value is y_i and the real value is, $\hat{y}_i$.

Root Mean Square Error (RMSE): It is the average value of the square discrepancy among forecast and observation, or the square root of the mean square error. Reduced error indicates increased model correctness and has the following definition (6):

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (6)$$

In this instance, "n" denotes the number of samples, "y_i" denotes the real values, and "$\hat{y}_i$" is the value that is anticipated or forecast.

R² Score: It is sometimes referred to as the coefficient of determination, and it indicates the extent to which a model can account for change in a dependent variable. A fitted regression line's dispersed data is assessed using the R^2 score. In the same way, as for the comparing of comparable data sets, a higher coefficient of determination also mean less variability between the real data and the projection data. It is imaginable on a basic from 0 to 1scale in relation to the real and projected data. As provided by Eq. (7)

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}} = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(\hat{y}_i - \mu)} \quad (7)$$

where $\hat{y}_i$ is the real value, $\hat{y}_i$ is the forecast value, μ is the average value, and SSres is the sum of squares of the residuals.

4. Experiments and Discussion

The experimental results of the machine learning models that have been implemented on the Walmart sales dataset for sales

forecasting. In order to facilitate comprehension of the experimental outcomes, the subsequent section implemented an assortment of graphs, charts, and tables. The following Table 2 provides the ML model performance on the same dataset.

Table 2: ML model performance for Sales forecasting in business sector

Matrix	Random forest	XGBoost
R2-score	96	94
MAE	19.85	55.34
RMSE	43.02	33.38

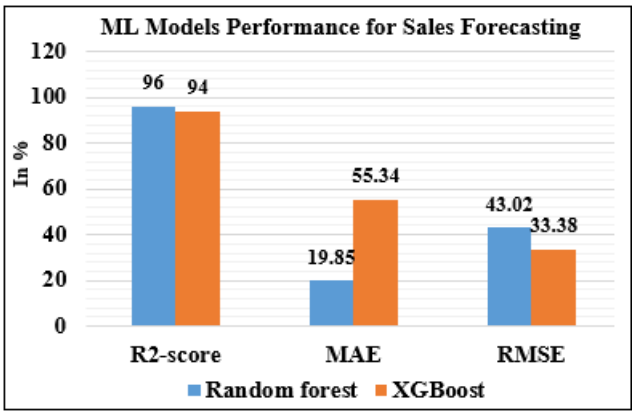


Figure 6: Bar Graph for models' performance

Bar graph for ML model performance shows in Table 2 and Figure 6. The results indicate that the RF model has an R^2 score of 96, demonstrating a strong fit to the data. It also exhibits a low MAE of 19.85, suggesting accurate predictions. The RMSE for RF is 43.02, indicating that the model has a moderate level of error in its predictions. XGBoost model achieves an R^2 score of 94, indicating a good fit to the data. The MAE is 55.34, suggesting a moderate level of deviation in the model's predictions from real ethics. Additionally, the RMSE is 33.38, demonstrating the model's ability to manage larger errors effectively.

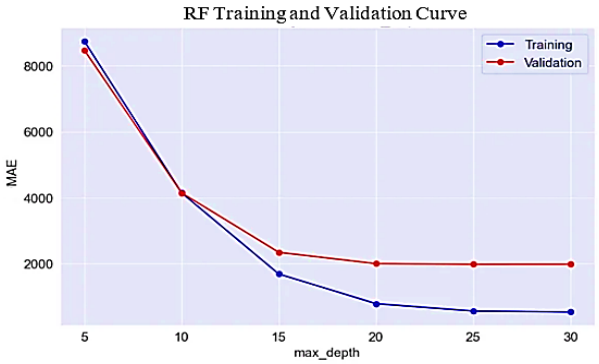


Figure 7: Training and validation curve for RF model

Figure 7 shows the training and validation curves for a RF model, with the X-axis representing 'max_depth' (ranging from 5 to 30) and the Y-axis showing MAE (from 0 to 8000). The training curve sharply decreases in MAE as 'max_depth' increases and then flattens, while the validation curve declines more gradually, eventually running parallel to the training curve but at a higher MAE. This indicates that while increasing 'max_depth' initially reduces error, it reaches a point where further complexity doesn't significantly improve validation performance.

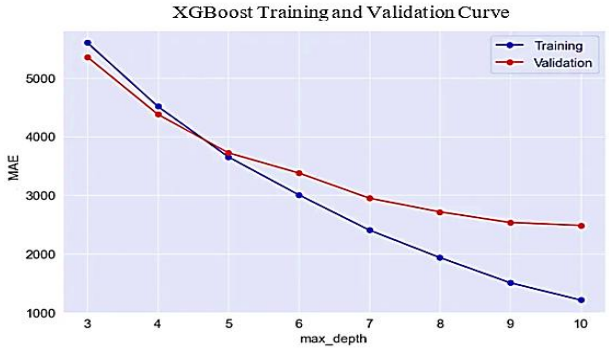


Figure 8: Training and validation curve for XGBoost model

This is further depicted in the Figure 8 below training and validation curve for the XGBoost model. The MAE against the 'max_depth' parameter. As 'max_depth' increases from 3 to 10, both the training (blue line) and validation (red line) errors decrease, with the training error consistently lower, indicating better performance on the training data. The curves suggest that increasing 'max_depth' improves model performance, but caution is needed to avoid overfitting if the validation error starts to rise at higher depths.

Table 3: Comparison analysis of different ML models on Walmart sales dataset

Models	R2-score	MAE
ARIMA[23]	91.29	76.04
KNN[24]	59.4	81.99
RF	96	19.85
XGB	94	55.34

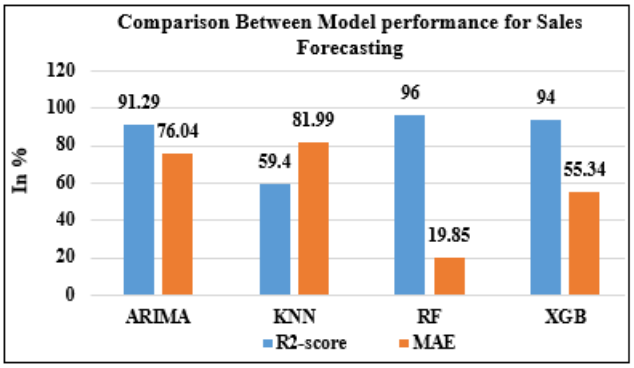


Figure 9: Bar graph for comparison of the model's performance

The above Table 3 and Figure 9 show the comparison of the models' performance. In comparison, RF model stands out with an impressive R^2 score of 96 and a low MAE of 19.85, indicating excellent predictive accuracy and minimal deviation from actual sales figures. The XGB model also performs well, with an R^2 score of 94 and an MAE of 55.34, demonstrating good fit but with a higher error than RF. In contrast, the ARIMA model shows a lower R^2 score of 91.29 and a significantly higher MAE of 76.04, indicating less accuracy in capturing sales trends. The KNN model performs the weakest, with an R^2 score of 59.4 and an MAE of 81.99, reflecting its challenges in effectively forecasting sales compared to RF and XGB. Overall, RF and XGB exhibit superior performance, making them more suitable choices for sales forecasting.

5. Conclusion and Future Work

Sales forecasting is an important determinant of growth and, in some cases, profitability of contemporary retail organizations managing large outlets. This paper convinces the audience that sales forecasting can help businesses manage their money and their merchandise more effectively, and that it can lead to smarter business plans. Provide a reliable and accurate ML -based sales forecasting model in this study. By applying models such as RF, XGBoost, ARIMA, and KNN to a Walmart sales dataset, By applying models such as Random Forest (RF), XGBoost (XGB), ARIMA, and KNN to a real-world dataset, the research highlights the superior performance of RF and XGB in predicting sales trends, with RF achieving R^2 score of 96% and 94% and a low Mean Absolute Error (MAE) of 19.85 and 55.34, indicating excellent predictive accuracy. However, there are limitations to the study, including the reliance on historical data, which may not account for unexpected market shifts or external factors. Additionally, the model's performance could be further enhanced with more advanced feature engineering and tuning techniques. For future work, A more comprehensive understanding of sales trends might be possible through the integration of other data sources, such as social media trends, economic indicators, and consumer sentiment research.

References

[1] M. A. Meitriana, K. R. Suwena, and N. W. A. Santi, "Fintech: Its Effect on Business Sustainability," *J. Pendidik. Ekon. Undiksha*, vol. 15, no. 1, pp. 226–231, Aug. 2023, doi: 10.23887/jjpe.v15i1.64752.

[2] D. Rohaan, E. Topan, and C. G. M. Groothuis-Oudshoorn, "Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider," *Expert Syst. Appl.*, 2022, doi: 10.1016/j.eswa.2021.115925.

[3] J. Thomas, "A Survey of E-Commerce Integration in Supply Chain Management for Retail and Consumer Goods in Emerging Markets," *J. Emerg. Technol. Innov. Res.*, vol. 10, no. 12, 2023.

[4] P. P. Jubin Thomas, "Predictive big data analytics for supply chain demand forecasting," *Int. J. Creat. Res. Thoughts*, vol. 10, no. June 2022, pp. 301–330, 2022.

[5] R. Gustriansyah, E. Ermatita, and D. P. Rini, "An approach for sales forecasting," *Expert Syst. Appl.*, 2022, doi: 10.1016/j.eswa.2022.118043.

[6] M. R. Kishore Mullangi, Vamsi Krishna Yarlagadda, Niravkumar Dhameliya, "Integrating AI and Reciprocal Symmetry in Financial Management: A Pathway to Enhanced Decision-Making," *Int. J. Reciprocal Symmetry Theor. Phys.*, vol. 5, no. 1, pp. 42–52, 2018.

[7] D. M. U. Ashraf, "A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques," *Lahore Garrison Univ. Res. J. Comput. Sci. Inf. Technol.*, 2022, doi: 10.54692/lgurjcsit.2022.0604399.

[8] P. Khare and S. Srivastava, "AI-Powered Fraud Prevention: A Comprehensive Analysis of Machine Learning Applications in Online Transactions," *J. Emerg. Technol. Innov. Res.*, vol. 10, pp. f518–f525, 2023.

[9] S. K. R. A. Sai Charan Reddy Vennapusa, Takudzwa Fadziso, Dipakkumar Kanubhai Sachani, Vamsi Krishna Yarlagadda, "Cryptocurrency-Based Loyalty Programs for Enhanced Customer Engagement," *Technol. Manag. Rev.*, vol. 3, no. 1, pp. 46–62, 2018.

[10] M. Spuritha, C. S. Kashyap, T. R. Nambiar, D. R. Kiran, N. S. Rao, and G. P. Reddy, "Quotidian Sales Forecasting using Machine Learning," in *Proceedings of the 2021 IEEE International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2021*, 2021. doi: 10.1109/ICSES52305.2021.9633975.

[11] F. Abubaker and Ala'khalifeh, "Sales' Forecasting Based on Big Data and Machine Learning Analysis," in *9th 2023 International Conference on Control, Decision and Information Technologies, CoDIT 2023*, 2023. doi: 10.1109/CoDIT58514.2023.10284159.

[12] J. Chen, W. Koju, S. Xu, and Z. Liu, "Sales Forecasting Using Deep Neural Network and SHAP techniques," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE 2021*, 2021. doi: 10.1109/ICBAIE52039.2021.9389930.

[13] J. Ding, Z. Chen, L. Xiaolong, and B. Lai, "Sales Forecasting Based on CatBoost," in *Proceedings - 2020 2nd International Conference on Information Technology and Computer Application, ITCA 2020*, 2020. doi: 10.1109/ITCA52113.2020.00138.

[14] Y. Ensafi, S. H. Amin, G. Zhang, and B. Shah, "Time-series forecasting of seasonal items sales using machine learning – A comparative analysis," *Int. J. Inf. Manag. Data Insights*, 2022, doi: 10.1016/j.jjimei.2022.100058.

[15] Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," in *Proceedings - 2020 International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE 2020*, 2020. doi: 10.1109/ICBASE51474.2020.00103.

[16] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, 2021, doi: 10.1186/s40537-021-00516-9.

[17] G. Manikandan, N. Sairam, S. Sharmili, and S. Venkatakrishnan, "Achieving privacy in data mining using normalization," *Indian J. Sci. Technol.*, 2013, doi: 10.17485/ijst/2013/v6i4.16.

[18] N. Agarwal, S. Gupta, and S. Gupta, "A comparative study on discrete wavelet transform with different methods," in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, Mar. 2016, pp. 1–6. doi: 10.1109/CDAN.2016.7570878.

[19] Pranav Khare and Shristi Srivastava, "Data-driven product marketing strategies: An in-depth analysis of machine learning applications," *Int. J. Sci. Res. Arch.*, vol. 10, no. 2, pp. 1185–1197, Dec. 2023, doi: 10.30574/ijsra.2023.10.2.0933.

[20] J. Huo, T. Shi, and J. Chang, "Comparison of Random Forest and SVM for electrical short-Term load forecast with different data sources," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2016. doi: 10.1109/ICSESS.2016.7883252.

[21] Wankar H, Dimble K, Dasgaonkar P, Chavan V, and Sayyad A, "Property Price Prediction Engine Using XGBoost Regression," *Int. J. Creat. Res. Thoughts*, vol.

11, no. 4, pp. 190–194, 2023.
[22] J. Thomas, "Enhancing Supply Chain Resilience Through Cloud-Based SCM and Advanced Machine Learning: A Case Study of Logistics," *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 9, 2021.
[23] D. Wankhede, A. Athanikar, Y. Borude, R. Chougule, and O. Diware, "An Experimental Analysis of Machine Learning Model in the Context of Sales Forecasting : - A Walmart Dataset as A Case Study," *Grenze Int. J. Eng. Technol.*, 2024.
[24] B. Yao, "Walmart Sales Prediction Based on Decision Tree, Random Forest, and K Neighbors Regressor," *Highlights Business, Econ. Manag.*, 2023, doi: 10.54097/hbem.v5i.5100.