IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 1, August 2023
Impact Factor: 7.301

Predictive Analytics in Business Intelligence for Sales Forecasting

Kshitij Dixit
Independent Researcher
kshitijdixit102@gmail.com

Abstract: *The accurate competitive advantage of businesses depends on accurate sales forecasting. The increased availability of retail sales data and the emergence of machine learning (ML) algorithms offer a chance to improve prediction accuracy. The purpose of the proposed research is to create an efficient ML model to predict weekly retail sales based on the Walmart dataset. To extract meaningful temporal and economic features out of the data, the study uses data preprocessing, exploratory data analysis, and feature engineering. Random Forest and CatBoost as ensemble learning models are used to forecast the pattern of sales every week. The models are evaluated using various measures like R^2, MAE, RMSE, and MSE. Both models are quite predictive of the experimental results. Random Forest has the best R^2 0.9690 and MAE of 0.369 and CatBoost has the best R^2 0.9842 and MAE of 0.317 with low forecasting error and high prediction accuracy. It is also noted that the proposed models are better in the comparison with the base models such as XGB, LSTM, Extra Trees, SARIMA, and Prophet. Overall, the present research work presents a potent ML model, which integrates both temporal engineering of features and ensemble models to make the correct decisions in terms of retail sales. The findings highlight the importance of historical sales trends and extraneous factors in forecasting, and provide realistic insights into how demand planning and inventory use can be improved in the retailing business.*

Keywords: Retail Analytics, Sales Forecasting, Business Analytics, Product Demand, Machine Learning, Walmart Dataset

I. INTRODUCTION

Retail is considered one of the most important and dynamically developing spheres of the data science industry, with its enormous volumes of data and optimization challenges [1]. The modern-day business environment is very dynamic and competitive, and as such, the accuracy of sales prediction has become more complicated. Minor improvements in sales forecasting would be of great help to retailers to cut operational expenses and increase sales, and eventually bring more customer satisfaction [2]. Detection of appropriate sales in each store is significant to the achievement of any retailing firm because it can help to manage stocks [3]. Right sales forecasting is the key to the success of any retail company [4]. It assists in the effective management of inventory, proper distribution of products to different stores, and reduces risks arising due to overstocking and understocking. The optimization of these aspects helps retailers to reduce losses and maximize sales, and increase customer satisfaction, in general [5][6]. Nevertheless, sales forecasting entails a myriad of factors and thus, it is a tricky undertaking to the retail firms. The weather conditions, seasonal trends, local events, and competition with other retail stores and online shopping platforms are some of the external factors. Also, promotions, discounts, and pricing strategies are internal factors that contribute to the problem complexity [7].

The high Artificial Intelligence (AI) and Machine Learning (ML) algorithms to predict sales of products and goods have emerged as an emerging field of interest to analysts in the current times. Sales forecasting in the retail sector can be done using various ML methods, each having its own strengths and weaknesses [8][9]. This is because ML can model advanced patterns of data and it can be used in various fields like Finance, Management, Marketing, and Retailing. Proper forecasts are useful in shaping and improving marketing policies in the market, making them quite useful. With the growth of businesses and the dynamism of markets, these primitive ways were no longer sufficient [10]. The traditional statistical techniques do not tend to be in a position to determine complex relationships between sales and other modifying factors such as economic indicators, seasonal variations, and store-related differences. The need to develop an effective ML-based forecasting model informed the purpose of the current study, which would help to investigate previous selling patterns and confounding variables within the Walmart dataset. Random Forest (RF) and CatBoost ensemble learning algorithms can be employed to boost prediction accuracy and support enhanced inventory and demand planning in retail industry.

A. Paper Organization

The paper is structured as follows: Section II examines theoretical foundation and literature. Section III discusses adopted methodology. The results and comparison are presented in Section IV. In Section V, conclude study by reviewing the key conclusions, pointing out any shortcomings, and future research.

II. LITERATURE REVIEW

The main objective of this investigation is to examine different ML techniques that different researchers use to predict sales.

Thivakaran & Ramesh (2022). To provide more accurate findings, the suggested Big-Mart sales forecasting system makes use of multivariate Poisson distribution techniques, generative adversarial networks, and recurrent multi-level deep reinforcement learning engines. The results and trials demonstrate how well suggested DL-based sales prediction model performs in comparison to rival systems. It demonstrates that suggested DL technique outperforms other current sales prediction models by 5% to 10% [11].

Jiang, Ruan and Sun (2021) examine whether it is possible to estimate Walmart sales using ML models, hybrid models that combine traditional time series models with ML techniques. Walmart supermarket sales data are trained and assessed using the Prophet model, which differentiates between season, holiday, and trend. Next, this model is used for empirical analysis and forecasting. With RMSE of 0.694 and 0.617 for the Prophet and LightGBM models, respectively, the results indicate that the ML models are effective for forecasting sales at retail locations. This offers retailers a novel method for forecasting sales by category and region [12].

Mia, Yousuf and Ghosh (2021) To employ machine learning to estimate future product sales, a business forecasting system is developed and implemented. Before rendering a decision, ML algorithms utilize input variables to generate a pattern. Big data technology is being used more and more in research. Strong forecasting systems may be developed with the use of big data and ML. This work has prepared data for the training of the proposed system using big data processing technologies. The recommended system's expected accuracy for different items varied from 99% to 75%, with a MAPE of 7.32%. Therefore, the suggested approach is highly effective in forecasting future sales [13].

Chen et al. (2021) provide a model for forecasting Walmart's sales using neural networks. Additionally, assess the NN model utilizing datasets from the Kaggle platform. Experiments demonstrate how the NN model performs better than other ML models. The RMSE values are 2.58 and 2.92 lower than those of the SVM and LR methods, respectively. Furthermore employ SHAP for multi-dimensional feature extraction in NN models to get high accuracy predictions [14].

Ding et al. (2020) proposed a sales prediction method based on CatBoost. To train the model, we used the Walmart sales dataset, the largest in this category. Did optimized feature engineering to increase training time and quality. This model shows a significant gain over classical machine learning models like LR and SVM with 0.605 RMSE. Its ability to be parameter-tuned is less with respect to other classical method increasing the its potential utility on other custom datasets [15].

Mortensen et al. (2019) Multiple models were developed by the team to forecast win likelihoods for various sales opportunities. The most accurate and most insightful model was chosen to serve as the client-facing model. The company's Salesforce.com customer relationship management system provides a range of organized and unstructured data to do this. The group experimented with a number of strategies, including random forests, gradient boosting, binomial logistic regression, and other decision-tree methods. Long-term performance was shown to be most influenced by customer characteristics, opportunity, and internal documentation practices. The best model outperformed the

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 3, Issue 1, August 2023 Impact Factor: 7.301

current sales prediction accuracy by forecasting win propensity with 80% accuracy, 77% precision, and 86% recall, respectively [16].

Table 1: Summary of literature study on sales forecasting using machine learning

Author (Year)	Data	Method	Findings	Limitation	Future Work
Thivakaran and Ramesh (2022)	Big Mart sales dataset	Multi-Level GAN, Deep RL Engines, and Multivariate Poisson Distribution	The proposed deep learning approach improved prediction accuracy by 5–10% over existing models.	High computational complexity and model training cost.	Optimize model efficiency and test on diverse retail datasets.
Jiang, Ruan, and Sun (2021)	Walmart sales data (2011–2016)	Prophet model, LightGBM	LightGBM RMSE = 0.617, Prophet RMSE = 0.694; ML models effective for sales forecasting	Limited to one store chain and specific timeframe	Explore hybrid models combining ML and time series methods across multiple regions
Mia, Yousuf and Ghosh (2021)	Product sales data processed using big data techniques	Machine Learning with Big Data Processing	Achieved prediction accuracy between 75%–99% with MAPE of 7.32%.	Performance varies across different product categories.	Investigate ensemble learning methods for consistent accuracy.
Chen et al. (2021)	Kaggle's Walmart sales dataset	Neural Network with SHAP interpretation	Comparing the NN model to SVM and Linear Regression, the RMSE was lowered by 2.58 and 2.92.	Limited interpretability of deep learning models despite SHAP use.	Improve explainable AI techniques and evaluate on additional retail datasets.
Ding et al. (2020)	Walmart sales dataset	CatBoost, feature engineering	CatBoost RMSE = 0.605; faster training and better generalization	Needs effective feature engineering; limited model comparison	Apply method to custom retail datasets and automate feature selection
Mortensen et al. (2019)	Salesforce.com CRM data	Binomial logit, decision trees, gradient boosting, random forest	The best model has an 80% accuracy rate in predicting win propensity, 86% precision, and 77% recall	Focused on CRM data; domain-specific	Incorporate more unstructured data and explore ensemble approaches

A. Research Gap

The majority of the existing sales forecasting research is specific to certain datasets and, therefore, the models are not applicable in general. Most of them primarily use historical sales data and time-series functionality and pay less attention to the outside world, such as promotions, weather, or even economic indicators. Deep learning models can require huge datasets and high-level computing resources, which might be inappropriate for smaller retailers. The interpretability of the model is also restricted, and it becomes difficult to extract actionable information for the managers. Moreover, the majority of studies focus on short-term forecasting, and there is a gap in long-term and multi-store prediction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 1, August 2023
Impact Factor: 7.301

III. METHODOLOGY

Figure 1 presents the overall framework of the proposed sales forecasting approach. It starts with the Walmart dataset then proceeds to data preprocessing and EDA to learn about data patterns. Useful features are generated through feature engineering and one-hot encoding. After splitting data into training and test sets, it is normalized using Minmax scaling. The models being trained are Random Forest and Cat Boost, both trained on processed data. Lastly, explainable AI was further examined in terms of Feature Importance, and the model's performance was assessed using RMSE, MAE, MSE, and R2.

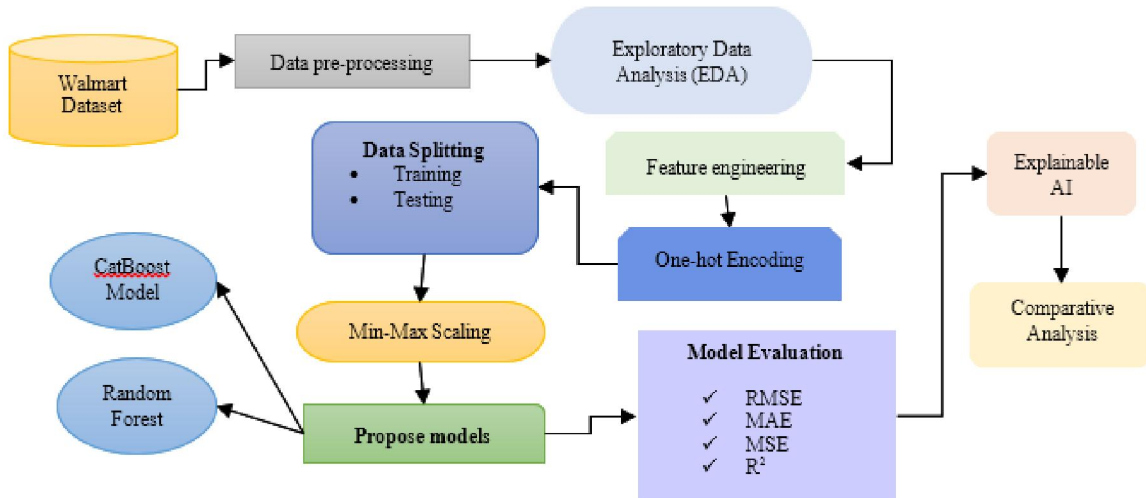

Figure 1: Proposed Flowchart for Sales Forecasting

Discuss each step of propose methodology in detail in the next section.

A. Data gathering and Pre-processing

The Walmart dataset consists of 6,435 rows and 8 columns, containing the weekly sales information on several stores. Some of the main columns are: Date, Store, and Weekly Sales as well as economic and external factors like CPI, Fuel_Price, Temperature, and Unemployment. The rows are defined by the stores and week, so they contain both the time-related and categorical data to predict the sales pattern. The pre-processing of data is done to make the dataset ready to train model. The data is initially loaded and checked for its format and to confirm the absence of missing or repeated values. The dataset is clean, with no missing or duplicate values, and is ready for feature engineering and modeling.

B. Exploratory Data Analysis (EDA)

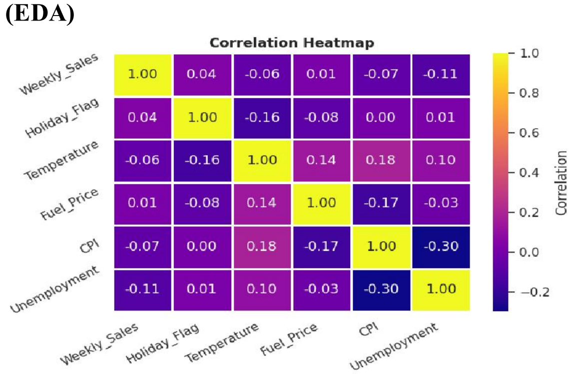

Figure 2: Correlation Heatmap of Numerical Features

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 1, August 2023 Impact Factor: 7.301

A crucial first step in comprehending underlying patterns in the dataset, connections, and features is exploratory data analysis, or EDA. This section uses a variety of visualizations and summary statistics to examine the Walmart dataset. Figure 2 displays the correlation heatmap for the dataset's numerical variables. The findings show that most variables are weakly correlated with one another. Temperature (0.06), fuel price (0.01), CPI (0.07), and unemployment (0.11) have relatively little link with weekly sales. There is a moderate negative relationship between CPI and Unemployment (-0.30). Overall, the heatmap indicates that multicollinearity among the features is low, which is advantageous for predictive modeling.

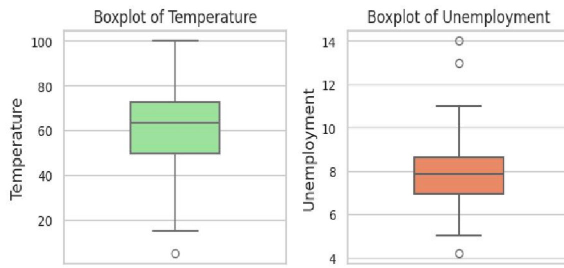

Figure 3: Boxplot of Temperature and Unemployment

The distribution of temperature and unemployment data is shown in Figure 3. The temperature boxplot (green) indicates that temperature variations are between 20 and 100, with the median of approximately 70, and only one low outlier exists below 20, which indicates a moderate variation. The boxplot of unemployment (orange) has a range of about 4-14, median of about 8 and three outliers two higher points above 12 and one lower point below 5. The range of unemployment is more dispersed and irregular than the temperature. Collectively, these plots make a clear visual summary of the central tendency, spread and outlier behavior in both data sets.

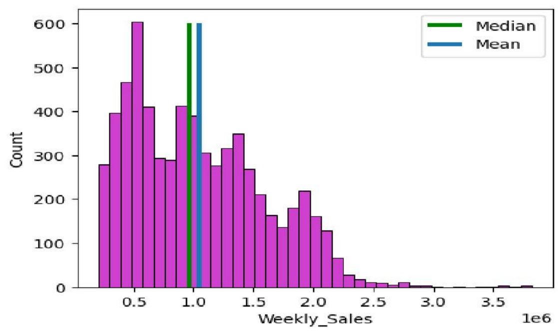

Figure 4: Histogram of Weakly Sales Trends

Figure 4 depicts the weekly sales distribution with the majority of the values ranging between 0.4 million and 1.5 million. The distribution is positively skewed as the few weeks with many higher sales are well above 2 million. The median (blue line) is relatively lower than the mean (green line), which shows that there are higher values of sales that affect the mean. Generally, the plot indicates that average weekly sales are approximately 1 million, with some occasional surges in sales.

C. Feature Engineering

Increasing the data's capacity for prediction is the aim of feature engineering. The Date column is turned into a datetime format, and attributes like year, month and day are extracted, with month and day being turned into cyclical sine and cosine data to capture seasonality. The Store variable is encoded using one-hot encoding, and time-series features including lag variables (lag_1, lag_2, lag_4) and rolling statistics (rolling_mean_4, rolling_std_4) are created from Weekly_Sales. Following the operation of creating features, rows of missing values due to these operations are eliminated.

D. Data Splitting

To assess model's performance, processed data is split into training and testing sets. There is a chronological 80/20 split, with 80% being used as training and 20% testing. As a result, the training set has 5, 144 observations and 59 features, and the testing set has 1,287 observations and 59 features. The methodology guarantees that model is evaluated with uncertain future data after being trained on historical data, and that it maintains the data's time sequence.

E. Min-Max normalization

$$\breve{z}_i = \frac{z_i - z_{min}}{z_{max} - z_{min}} \quad (1)$$

In this notation, the values of the i^{th} observation are limited to the interval [0,1] and are represented as $\breve{z}_i$. The sign z_i represents the original, unnormalized value; $z_{\max}$ and $z_{\min}$, respectively, denote the maximum and minimum values of the same in the corresponding feature.

F. Propose CatBoost Model

Cat Boost is a technique that combines a decision tree approach with gradient boosting. Cat Boost's training speed on a graphics card, user-friendliness, and support for categorical variables are the features that are driving its popularity. Some of the key parameters used in training of CatBoost model include following. A typical value of the number of boosting iterations is iterations = 1000, which determines the number of trees. The learning rate = 0.03 determines the rate of step taken in updating models. The depth = 6 is used to identify the maximum depth of each tree. The regression tasks are usually designed to use the loss function = RMSE. The l2_leaf_reg = 3 is useful in minimizing the overfitting through regularization. Additionally, randomized = 42 is utilized to guarantee that the results are repeatable.

G. Propose Random Forest

An example of an ensemble strategy made up of several DT is RF. While individual trees in these ensembles might not be interpretable, the combination of several trees' predictions improves accuracy. RF model is implemented with the help of the RF function that trains many decision trees and consolidates their estimates to achieve better accuracy. The default hyperparameters are the popular ones that are used in training in R. The number of trees in environment is n_estimators = 500. The value of the max_features argument is square root of total number of features. Because the max_depth is NULL, trees grow until their leaves are either pure or contain very few samples. Min samples split, which is set to min samples split 2, is the fundamental minimum number of samples required to split a node. Min samples leaf indicates that a leaf node's minimum samples are 1. Each tree is trained using a random sample of data, and bootstrap option is set to TRUE.

H. Evaluation Criteria

An essential step in determining the process's quality is assessed by looking at the outcomes. The effectiveness of the computations that were evaluated using the specified parameters:

R-Squared (R^2)

A regression line's ability to fit dispersed data is assessed using the R^2 score. For comparable datasets, higher R^2 values indicate a smaller discrepancy between expected and actual data. It determines correlation between expected and actual data on a scale of 0 to 1. It is provided by Equation (2):

$$R^2 = 1 - \frac{SSR}{TSS} \quad (2)$$

Mean Absolute Error (MAE)

It is average absolute difference between actual observations and expectations for test sample. The definition of MAE is as follows Equation (3):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_1| \quad (3)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 1, August 2023 Impact Factor: 7.301

Root Mean Squared Error (RMSE)

The RMSE is the square root of the mean of squares of all errors [17]. Once more, RMSE displays the degree to which line of best fit resembles collection of points shown in Equation (4)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_1)^2} \quad (4)$$

Mean Squared Error (MSE)

It is mean squared of variation between expected and actual values [18]. A better-performing model is indicated by a lower MSE; if it is 0, it indicates flawless prediction with no error, which is very uncommon and challenging but feasible. It is calculated using the Equation (5 and 6):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2 \quad (5)$$

$$MSE\ \% = \left(\frac{MSE}{Mean\ of\ the\ actual\ values}\right) \times 100 \quad (6)$$

These measures are used to evaluate and contrast how well various categorization methods detect different sorts of performance.

IV. RESULTS AND DISCUSSION

A local computer running an Intel Core i7-12700H CPU @ 2.70 GHz and 32.0 GB of RAM is used for the experiment. This setup offered sufficient processing power to efficiently train and test a model. Table II presents experimental results of the proposed model, Random Forest, and CatBoost to predict sales in terms of evaluation parameters R^2, MAE, RMSE, and MSE. Performance metrics show that both models are predictive though CatBoost is slightly higher on all metrics than RF. CatBoost has a higher R^2 of 0.9842 compared to the RF 0.9690, indicating that the former has a superior fit and predictive accuracy. CatBoost has lower MAE (0.317) and RMSE (0.488) than Random Forest, with MAE of 0.369 and RMSE of 0.684. Overall, propose models that demonstrate superior capability in minimizing prediction errors and improving forecasting reliability.

Table 2: Experiment Result of propose models for sales forecasting

Measures	Random Forest	CatBoost
R2	0.9690	0.9842
MAE	0.369	0.317
RMSE	0.684	0.488
MSE	4.680	2.386

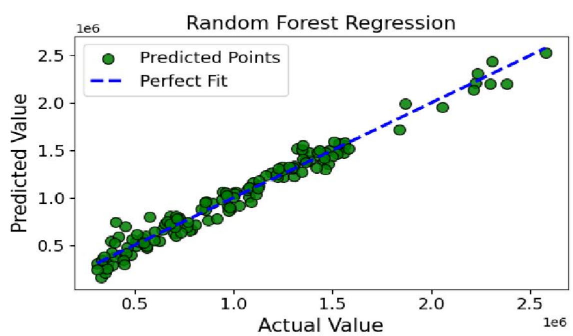

Fig 5 Scatter plot for Actual and predicted sales with Random Forest

The link between actual and expected sales, as estimated by the RF regression model, is shown in Figure 5. In a scatter plot, the ideal scenario of perfect prediction is represented by a blue dashed line that contrasts actual sales on y-axis with anticipated numbers on x-axis. The green data points are very close to this reference line, meaning that the model

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 3, Issue 1, August 2023
Impact Factor: 7.301

has good predictive accuracy. The model accurately captures sales patterns, as shown in this visualization, and the ideal prediction line shows only slight variation, demonstrating that the RF method is very strong at predicting sales outcomes.

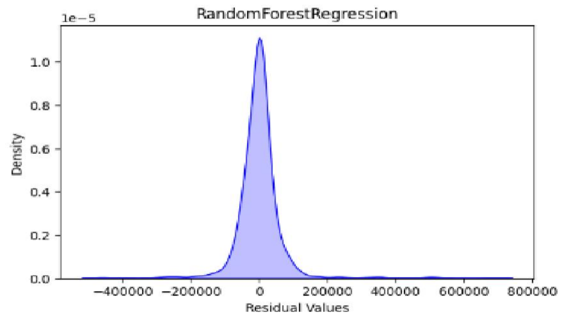

Fig 6 Residual Distribution Plot of Random Forest

The residual distribution map for the RF regression model shows Figure 6 illustrates the difference between expected and True values. The density curve shows that most residuals cluster around zero, indicating that the model predictions are often quite correct. The sharp peak around zero indicates few prediction errors when a large number of observations is taken. However, the presence of wider tails indicates that a few predictions deviate more from the actual values, highlighting some variability in model performance.

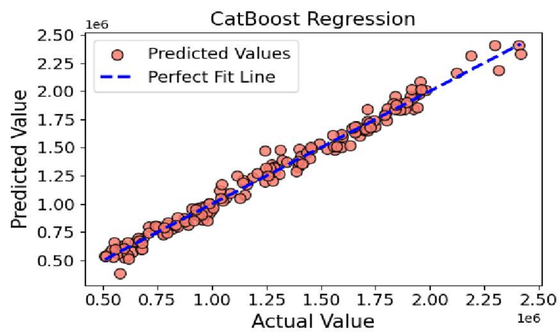

Fig 7 Scatter plot for Actual and predicted sales with CatBoot

Figure 7 displays a scatter plot of actual sales compared to sales numbers predicted by the CatBoost regression model. For the blue line, the anticipated and actual values are identical, indicating a perfect match. The majority of the data points are close to this line, so it is very predictive. The minor deviations are just a minor inaccuracy in prediction, and general viability of model in sales forecasting.

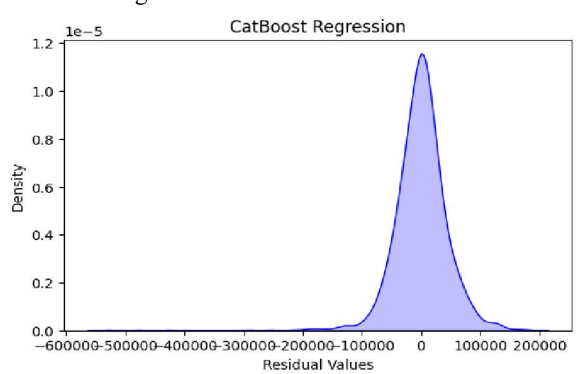

Fig 8 Residual Distribution Plot of CatBoost Regression Mode

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 1, August 2023 Impact Factor: 7.301

Figure 8 presents distribution of residuals for the CatBoost regression model. Residuals are the differences between predicted and real sales, and residual distribution is employed to estimate prediction errors. The model projections are often close to real ones, as chart shows that most residues are concentrated around zero. The bell-shaped density curve indicates that errors are approximately normally distributed, confirming that the CatBoost model performs well with minimal bias and a small deviation in predictions.

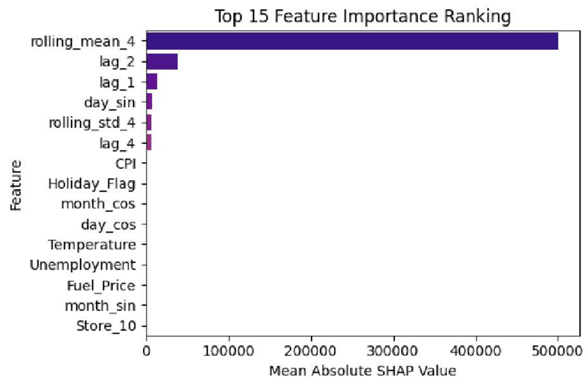

Fig 9 Top 15 Feature Importance Score with RF

The random forest model's top 15 feature significance scores, expressed as mean absolute SHAP values, are shown in Figure 9. The most significant one is rolling_mean_4, then lag_2, lag_1, and lag4, which means that the trends in previous sales contribute significantly to the prediction. Time-related factors also behave in a moderate way with the presence of external factors like CPI, Temperature and Fuel Price having a lower impact on the model.

A. Comparative Analysis and Discussion

Table III presents a comparison of the recommended and baseline sales forecasting models based on RMSE, MAE, and R^2 metrics. The traditional models like SARIMA demonstrate very high error values with RMSE of 1812.38. DL–based LSTM improves performance with RMSE of 13.2629 and MAE of 9.6883. Ensemble methods like Extra Trees further reduce the error and achieve an R^2 of 0.9166, while Prophet records an R^2 of 0.8243. The proposed models have the highest performance with RF yielding RMSE of 0.684, MAE of 0.369, and R^2 of 0.9690, and CatBoost with the highest accuracy with RMSE of 0.488, MAE of 0.317 and R^2 of 0.9842, which depicts that the forecasting ability is better than that of the baseline approaches.

Table 3: Comparative analysis for sales forecasting between base and proposed models

Models	RMSE	MAE	R2
XGB [19]	21.77	-	0.946
LSTM [20]	13.2629	9.6883	-
Extra Trees [21]	1.1227	1.2605	0.9166
SARIMA [22]	1812.38	1521.32	-
Prophet [23]	0.0941	-	0.8243
RF	0.684	0.369	0.9690
CatBoost	0.488	0.317	0.9842

The research describes certain advantages of retail sales forecasting. The proposed solution is more accurate in predictions by depending on advanced ensemble software such as Random Forest and Cat Boost. Engineered features effectively represent the pattern of sales in history and time and strengthen the forecasting power of the pattern. The strategy also guarantees a stable model performance by means of normalization of data. The predictions made by experiments reveal that Cat Boost has a higher accuracy in its predictions, and the value of its errors is lower than the values of the errors made by the model of Random Forest

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 3, Issue 1, August 2023 Impact Factor: 7.301

V. CONCLUSION AND FUTURE WORK

Sales forecasting is essential for improving planning and decision-making in the retail industry. The paper gives the solution to the problem of predicting weekly sales using the Walmart data using ML and explains the impact of economic and time-related variables on the trends in sales. Exploratory analysis provides details of the data distribution and the correlation between the variables, whereas feature engineering determines meaningful temporal features with historical sales data. The experimental results confirm that the models suggested are very predictive. CatBoost model performs best with a 2 value of 0.9842, which means extremely high levels of prediction. It also captures fewer values of errors with an MAE of 0.317 for CatBoost and an RMSE of 0.684 for RF, which also captures good forecasting performance. Overall, the results prove that ensemble learning models can be effectively applied to recreate sales trends and increase the level of predictions. The suggested system enables sales forecasting and leads to enhanced demand forecasting, inventory control, and strategic decision-making in retail organizations.

A. Key Contributions

The following is a summary of this work's primary contributions:

- Provides a viable ML-based model of retail sales prediction with the Walmart data.
- Demonstrates strong predictive performance using ensemble models such as RF and CatBoost.
- Highlights the significance of historical sales data where lag and rolling characteristics play an important role in prediction.
- Shows improved performance compared with several baseline forecasting models.
- Supports accurate sales prediction, It enhances inventory control and demand planning in retail establishments.

The paper proposes a machine learning-retail sales forecasting framework model that integrates the temporal feature engineering using ensemble learning models. The study is innovative, and the lag and rolling statistical properties used along with RF, and CatBoost algorithms are efficient to understand previous dynamics of sales and external economic variables. The proposed approach improves the accuracy of the prediction and it is more predictive than other traditional and benchmark models. These results support the application of ensemble learning approaches to improve retail sales forecasting and facilitate data-driven decision-making in the retail industry.

B. Limitation and future work

The present study is based on a single dataset in which the number of extraneous variables is also small, which restricts the generalization of the results. The analysis does not include such important factors as promotions, holidays, and customer behavior. More comprehensive retail data can be utilized in the future, and additional economic or promotional elements can be taken into account. Further research can also be done on the developed DL or hybrid models to improve accuracy of forecasting.

REFERENCES

[1] N. Kourentzes, J. R. Trapero, and D. K. Barrow, “Optimizing forecasting models for inventory planning,” Int. J. Prod. Econ., vol. 225, p. 107597, Jul. 2020, doi: 10.1016/j.ijpe.2019.107597.

[2] C. Patel, “Effect of Digital Transformation on Customer Engagement in Retail Industries : A Comparative Review,” J. Technol., vol. 10, no. 1, pp. 165–174, 2022.

[3] L. Baardman, I. Levin, G. Perakis, and D. Singhvi, “Leveraging Comparables for New Product Sales Forecasting,” Prod. Oper. Manag., vol. 27, no. 12, pp. 2340–2343, Dec. 2018, doi: 10.1111/poms. 12963.

[4] K. M. R. Seetharaman, “Analyzing the Role of Inventory and Warehouse Management in Supply Chain Agility: Insights from Retail and Manufacturing Industries,” Int. J. Curr. Eng. Technol., vol. 12, no. 06, pp. 583–590, Jun. 2022, doi: 10.14741/ijcet/v.12.6.13.

[5] X. Bi, G. Adomavicius, W. Li, and A. Qu, “Improving Sales Forecasting Accuracy: A Tensor Factorization Approach with Demand Awareness,” INFORMS J. Comput., vol. 34, no. 3, pp. 1644–1660, Nov. 2020, doi: 10.1287/ijoc. 2021.1147.

[6] T. K. Thivakaran and M. Ramesh, "Exploratory Data analysis and sales forecasting of BigMart dataset using supervised and ANN algorithms," Meas. Sensors, vol. 23, p. 100388, Oct. 2022, doi: 10.1016/j.measen.2022.100388.

[7] A. Rezazadeh, "A Generalized Flow for B2B Sales Predictive Modeling: An Azure Machine-Learning Approach," Forecasting, vol. 2, no. 3, pp. 267–283, Aug. 2020, doi: 10.3390/forecast2030015.

[8] K. Saraswathi, N. T. Renukadevi, S. Nandhinidevi, S. Gayathridevi, and P. Naveen, "Sales prediction using machine learning approaches," in AIP Conference Proceedings, 2021, p. 140038. doi: 10.1063/5.0068655.

[9] V. S and D. Preethi, "An Analysis of Machine Learning Algorithms to Predict Sales," Int. J. Sci. Res., vol. 11, no. 6, pp. 462–466, jun. 2022, doi: 10.21275/SR22601144946.

[10] V. Varma, "Data Analytics for Predictive Maintenance for Business Intelligence for Operational Efficiency," Asian J. Comput. Sci. Eng., vol. 7, no. 4, 2022.

[11] T. . Thivakara and M. Ramesh, "Sales Data Analysis and Prediction System for Big Mart using Deep Recurrent Reinforcement Principles," in 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, Sep. 2022, pp. 1715–1721. doi: 10.1109/ICIRCA54612.2022.9985752.

[12] H. Jiang, J. Ruan, and J. Sun, "Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast," in 2021 IEEE 6th International Conference on Big Data Analytics, ICBDA 2021, 2021. doi: 10.1109/ICBDA51983.2021.9403224.

[13] A. R. Mia, M. A. Yousuf, and R. Ghosh, "Business Forecasting System using Machine Learning Approach," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), IEEE, Jan. 2021, pp. 314–318. doi: 10.1109/ICREST51555.2021.9331114.

[14] 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), IEEE, Mar. 2021, pp. 135–138. doi: 10.1109/ICBAIE52039.2021.9389930.

[15] J. Ding, Z. Chen, L. Xiaolong, and B. Lai, "Sales Forecasting Based on CatBoost," in 2020 2nd International Conference on Information Technology and Computer Application (ITCA), IEEE, Dec. 2020, pp. 636–639. doi: 10.1109/ITCA52113.2020.00138.

[16] S. Mortensen, M. Christison, B. Li, A. Zhu, and R. Venkatesan, "Predicting and Defining B2B Sales Success with Machine Learning," in 2019 Systems and Information Engineering Design Symposium (SIEDS), IEEE, Apr. 2019, pp. 1–5. doi: 10.1109/SIEDS.2019.8735638.

[17] A. V Tatachar, "Comparative Assessment of Regression Models Based On Model Evaluation Metrics," Int. Res. J. Eng. Technol., vol. 8, no. 9, pp. 853–860, 2021.

[18] M. R. Suyambu and P. K. Vishwakarma, "Improving grid reliability with grid-scale Battery Energy Storage Systems (BESS)," Int. J. Sci. Res. Arch., vol. 13, no. 1, pp. 776–789, Sept. 2024, doi: 10.30574/ijsra.2024.13.1.1694.

[19] A. R. Bilipelli, "End-to-End Predictive Analytics Pipeline of Sales Forecasting in Python for Business Decision Support Systems," Int. J. Curr. Eng. Technol., vol. 12, no. 6, pp. 819–827, 2022, doi: 10.14741/ijcet/v.12.6.17.

[20] K. Vavliakis, A. Siailis, and A. Symeonidis, "Optimizing Sales Forecasting in e-Commerce with ARIMA and LSTM Models," in Proceedings of the 17th International Conference on Web Information Systems and Technologies, SCITEPRESS - Science and Technology Publications, 2021, pp. 299–306. doi: 10.5220/0010659500003058.

[21] D. Agnani, A. Bavkar, S. Salgar, and S. Ahir, "Predicting E-commerce Sales & Inventory Management using Machine Learning," ITM Web Conf., vol. 44, p. 03040, May 2022, doi: 10.1051/itmconf/20224403040.

[22] K. Singh, P. M. Booma, and U. Eaganathan, "E-Commerce System for Sale Prediction Using Machine Learning Technique," J. Phys. Conf. Ser., vol. 1712, no. 1, 2020, doi: 10.1088/1742-6596/1712/1/012042.

[23] A. Ecevit, İ. Öztürk, M. Dağ, and T. Özcan, "Short-Term Sales Forecasting Using LSTM and Prophet-Based Models in E-Commerce," Acta Infologica, pp. 0–0, Apr. 2023, doi: 10.26650/acin. 1259067.

ISSN 2581-9429 IJARSCT