JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Accurate Sales Forecasting: The Impact of Data Preprocessing Techniques on Machine Learning Models

[1]Ahmad Saleh.

1Faculty of Graduate Studies for statistical research Cairo University, Giza, Egypt

Abstract : Accurate sales forecasting in retail is crucial for optimizing inventory management, reducing operational costs, and enhancing customer satisfaction. This research investigates the role of advanced data preprocessing in improving sales forecasting accuracy of 45 retail outlets of Walmart store which are at different geographical locations using various supervised machine learning techniques, including Linear Regression, random forest, k-nearest neighbors (K-NN), support vector regression (SVR), extra tree regression, and XGBoost. The study implements comprehensive preprocessing steps such as filling missing values with medians, addressing negative and zero values, and detecting outliers. Feature engineering enriches the dataset by extracting temporal information and capturing the effects of holidays and markdowns. These preprocessing strategies are pivotal in refining data quality, thereby enhancing the robustness and precision of the predictive models. Our results indicate a significant improvement in model performance following advanced preprocessing. Techniques like Z-score-based outlier removal, normalization, and handling missing values considerably boost prediction accuracy. Models such as Random Forest and Extra Trees demonstrate high efficacy, reflecting their enhanced ability to capture complex patterns in the data. This study highlights the essential role of data preprocessing in sales forecasting within the retail sector, providing a comparative analysis of various machine learning models and their effectiveness. By demonstrating the benefits of advanced data cleaning and transformation techniques, the research offers practical guidance for improving sales prediction accuracy, thereby supporting better inventory management, minimizing costs, and enhancing customer satisfaction.

Introduction

Retail is one of the most significant and rapidly evolving business domains within the data science field, largely due to its huge data volumes and optimization challenges. These challenges include determining ideal prices, generating personalized recommendations, managing discounts, and maintaining optimal stock levels, all of which can be addressed using advanced data analysis methods. In today's dynamic and highly competitive business environment, predicting sales accurately has become increasingly complex. Even minor enhancements in sales prediction can significantly aid retailers in reducing operational costs and boosting sales, ultimately leading to greater customer satisfaction. Prediction of correct sales at every outlet is important for the accomplishment of each retailing company as it aids in management of inventory [1]

Accurate sales forecasting is essential for the success of any retail company. It helps in effective inventory management, ensures the proper distribution of products across various stores, and mitigates the risks associated with overstocking and understocking. By optimizing these aspects, retailers can minimize losses, maximize sales, and enhance overall customer satisfaction. However, predicting sales involves considering a multitude of factors, making it a challenging task for retail companies. External factors such as weather conditions, seasonal trends, local events, and competition from other retail stores and online shopping platforms play a crucial role. Additionally, internal actions like promotions, discounts, and pricing strategies add to the complexity of the problem.

Utilizing Machine Leaning (ML) models to anticipate the sales of items and wares has become a rising area for analysts in present time [2] Various machine learning techniques can be employed for sales forecasting in the retail sector, each with its unique strengths and methodologies. Random Forest Regression is a supervised learning algorithm that utilizes ensemble learning to improve forecasting accuracy. It constructs multiple decision trees during training and uses the average of their predictions as the final output. This method enhances prediction accuracy by reducing overfitting and capturing diverse patterns in the data.

K-Nearest Neighbors (K-NN) Regression is another technique that predicts the value of a new data point based on its similarity to points in the training dataset. It assigns weights to the nearest neighbors, making it effective for capturing local patterns in the data. Support Vector Regression (SVR), derived from Support Vector Machine principles, aims to find the best-fit line that maximizes the number of points lying on it, thereby ensuring robustness and accuracy. Extra Tree Regression, also an ensemble learning model, closely resembles Random Forest but with a few key differences. While Random Forest uses bootstrapping (sampling with replacement) and selects optimal split points, Extra Tree Regression utilizes the entire original sample of data and chooses split points randomly. This approach can lead to faster training times and potentially different insights from the data. Moreover, XGBoost, a powerful implementation of gradient boosting, has gained popularity for its efficiency and accuracy in predictive modeling. XGBoost builds models sequentially, each correcting the errors of the previous ones, resulting in highly accurate and robust predictions. Its ability to handle missing data, regularization techniques, developing accurate Sales forecasting systems [3]

After reviewing the extensive literature on sales forecasting methodologies, a comprehensive approach was adopted. The sales forecasting model was trained using various machine learning algorithms, including Linear Regression, Random Forest Regression, KNN Regression, SVR, Extra Tree Regression, and XGBoost. The results obtained from each model were thoroughly analyzed and compared to determine the most effective approach. This comparative analysis aims to provide valuable insights into the strengths and limitations of each technique, ultimately guiding the selection of the most suitable method for sales forecasting in the retail industry.

I. Related Work

There are many related works from researchers about sales forecasting due to its significant impact on inventory management, supply chain efficiency, and overall business strategy. This section reviews related works that have utilized various machine learning techniques for sales prediction, highlighting comparative analyses and advancements in forecasting models.

Random Forest and Extra Trees: Stuti Raizada, Jatinderkumar R. Saini [1] performed a comparative analysis of supervised machine learning techniques for sales forecasting, providing insights into the performance of Random Forest, Extra Trees and Linear Regression.
S. Yi [16] focused on Walmart sales prediction using Random Forest, demonstrating its application and performance.
Bo Yao [9] use Decision Tree Regressor and Random Forest Regressor to predict Walmart Recruiting - Store Sales data. Using correlation, mean absolute error, and mean square error to evaluate the prediction results of the models. the prediction effect of Random Forest Registrar performs the best of

XGBoost: S. Yi and B. Seaman and J. Bowman [10] focused specifically on Walmart sales prediction, employing various machine learning models including the M5 forecasting model and XGBoost. These studies showcased the practical applications of machine learning in large-scale retail operations, emphasizing the critical role of model accuracy and computational efficiency.
Yiyang Niu [4] propose a sales forecasting model based on XGboost algorithm to solve the Walmart sales tasks which can effectively mine attributes of different dimensions.
Xie Dairu and Zhang Shilong [11] proposed an efficient and accurate sales forecasting model using machine learning, they used XGBoost algorithm to utilize the Walmart features for forecasting the future sales amount
KNN: Dr. C. Shyamala, P. Sabarish [12] proposed a measurement for foreseeing long run bargains of Tremendous Bazaar Companies keeping in see the bargains of past a long time. A comprehensive consider of deals estimate is done utilizing Machine Learning models such K-Neighbors Regressor

Hybrid Models: S. B. Latha. and C. Dastagiraiah et al. [2] developed an adaptive machine learning model for Walmart sales prediction, illustrating the benefits of model adaptation.
Arnab Mitra and Arnav Jain [6] proposed hybrid (RF-XGBoost-LR) model for sales forecasting of a retail chain considering the various parameters of forecasting accuracy It is observed that the hybrid RF-XGBoost-LR outperformed the other models measured against various metrics of performance.
Qianying Li and Mingyang Yu [13] introduced a model tailored for sales forecasting based on a Transformer with encoder– decoder architecture and multi-head attention mechanisms. they have made specific modifications to the standard Transformer model, such as removing the Softmax layer in the last layer and adapting input embedding, position encoding, and feedforward network components to align with the unique characteristics of sales forecast data and the specific requirements of sales forecasting.
Comparative Analysis of Supervised Machine Learning Techniques: Stuti Raizada and Jatinderkumar R. Saini [1] comparative analysis of some of the Supervised Machine Learning Techniques have been done such as Multiple Linear Regression Algorithm, Random Forest Regression Algorithm, K-NN Algorithm, Support Vector Machine (SVM) Algorithm and Extra Tree Regression to build a prediction model and precisely estimate possible sales of 45 retail outlets of Walmart store which are at different geographical locations.
Tianyang Lu [14] used three regression models, the linear regression model, the elastic net model and the SVR models to find out the relationship between the factors and the weekly sales of two Walmart stores and try to use the model to predict its future sales.

Table 1. Summary of existing work using Walmart sales datasets

Author	Approach	Measure Model Accuracy (R2)
Chenghao Yu (2024) [15]	Compares Random Forest, Linear Regression, and Lasso Regression to determine which model better predicts Walmart sales.	Linear regression: 0.65 Random Forest: 0.95 Lasso regression: 0.65
Tian Yang (2023) [16]	Compares the performance of OLS, and Random Forest models in predicting Walmart store sales.	Random Forest: 0.97 OLS: 0.72

Bo Yao (2023) [9]	Use Decision Tree Regressor, Random Forest Regressor, and KNN to predict Walmart sales and finds Random Forest Regressor to be the most effective model.	Decision Tree: 0.90 Random Forest: 0.93 KNN: 0.95
Dr. Neetu Anand, Dr. Tripti Sharma (2023) [17]	Addresses key retail issues such as client segmentation, inventory management, sales forecasting, and promotion strategies using statistical and machine learning models.	Linear Regression: 0.98
Pathum R. Ranathunga (2022) [3]	Compare between XG Boost, Extra Tree, Random Forest, Decision Tree, KNN and Linear Regression best performance for forecasting sales in Walmart supermarkets.	Extra Tree: 0.97 XGboost: 0.97 Random Forest: 0.96 Decision Tree: 0.96 KNN: 0.38 Linear Regression: 0.08
Arnab Mitra1, Arnav Jain1, Avinash Kishore· Pravin Kumar (2022) [6]	Comparing six machine learning regression techniques with a proposed hybrid (RF-XGBoost-LR) model for sales forecasting, finding the hybrid model to outperform the others.	Random Forest: 0.93 XGBooat 0.94 Gradient boosting 0.91 AdaBoost 0.93 Artificial neural network 0.39 RF-XGBoost-LR (hybrid) 0.95
Akande, Yetunde Faith Idowu,Joyce Misra (2022) [18]	Exploring XGBoost to forecast sales, demonstrating its effectiveness	XGboost: 0.97
Stuti Raizada, Jatinderkumar R. Saini (2021) [1]	Comparative between linear Regression, Random Forest Regression, K-NN, Support Vector Machine (SVM) and Extra Tree Regression to build a prediction model	Extra Tree: 0.98 Random Forest: 0.97 KNN: 0.59 SVR: -1.85 Linear Regression: 0.14
Dr. C. Shyamala, P. Sabarish, T. Vignesh, (2021) [12]	Predicting future sales of large retail companies using machine learning models, including Random Forest, K-Neighbors Regressor, XGBoost, and Linear Regressor.	Linear Regression:0.72 XG boost 0.81 Random Forest: 0.89 KNN gives 0.92
Xie dairu1, Zhang Shilong1 (2021) [11]	Exploring feature engineering and XGBoost for accurate sales forecasting using historical Walmart retail goods dataset.	Linear Regression: 0.78 Ridge: 0.77 XGBoost: 0.65

II. METHODOLOGY

The objective of this research is to compare the performance of several supervised machine learning algorithms in predicting retail sales. This study employs a quantitative approach, with the collection of historical sales data from Walmart. Following data collection, the next step involves preprocessing the data, which includes cleaning and preparing it for analysis. Subsequently, various machine learning models are implemented. The performance of these models is then evaluated using standard evaluation metrics as shown in Fig1. Finally, the results are analyzed to draw meaningful conclusions about the efficacy of the different models in predicting retail sales.

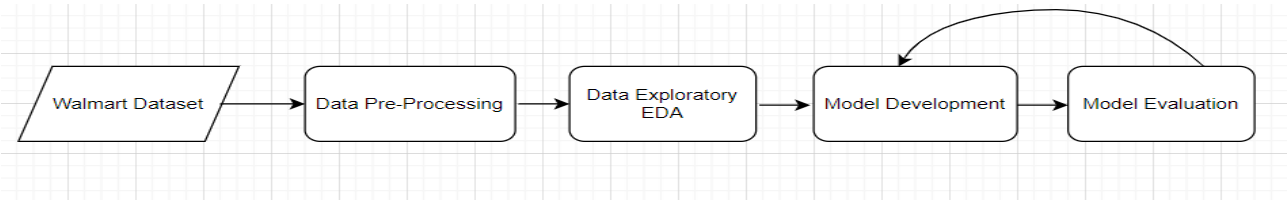


Fig.1 Proposed Methodology.

1. Selection of the dataset

There are many reasons behind my selection of the Walmart dataset. First, it includes a large and diverse set of data, covering 45 store outlets all over the United States. It has 98 items categories. The dataset provides many internal and external factors like holidays, promotions, economic indicators, and weather conditions, providing an opportunity to analyze how these factors influence sales and inform strategic decision-making. Internal factors, such as store size, location, and inventory levels, are also included

Furthermore, the data allows for benchmarking store performance against each other, helping to identify best practices and areas needing improvement. Analyzing sales patterns also offers insights into customer preferences and behavior, informing marketing strategies by revealing which products are popular in different regions and times of the year. From a research and academic perspective, the dataset serves as a valuable resource for studies in retail management, economics, and data science.

Walmart Dataset:
Period: February 5, 2010, to November 1, 2012.
Source: https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast
Variables:

- Store: The store number.
- Date: The week of sales.
- Weekly_Sales: Sales for the given store.
- Holiday_Flag: Indicates whether the week is a holiday week (1) or a non-holiday week (0).
- Temperature: Temperature on the day of sale.
- Fuel_Price: Cost of fuel in the region.
- CPI: Consumer price index.
- Unemployment: Unemployment rate.

2. Data Preprocessing:

Data preprocessing is very crucial for obtaining accurate machine learning models. multiple steps have been applied to ensure that datasets are clean and suitable for analysis. Missing values in the markdown columns ('MarkDown1' to 'MarkDown5') are filled with their respective medians, and negative values and “NaNs” are replaced with 0. The unique values are identified, and the training data is merged with relevant features for completeness. The 'Date' column is converted to datetime format, sorted, and set as the index, facilitating time-based analysis. The consistency of holiday indicators is checked, and redundant columns are dropped to consolidate the 'IsHoliday' feature. From the 'Date' column, year, month, and week numbers are extracted and added as new columns to enhance temporal analysis. For outlier detection, weekly sales data is grouped by store and department to compute various statistical measures, and the resultant table is merged with the main dataset. Rows with missing values in store_data are removed to ensure data integrity. A new DataFrame, data_numeric, containing only the specified numeric columns, is created. Outliers are then identified and removed using Z-scores, filtering out data points with absolute Z-scores greater than 2.5 to prevent extreme values from distorting the analysis. Additionally, outliers with negative weekly sales are eliminated. Feature engineering is applied to capture the effects of holidays and markdowns, including the creation of interaction terms and lagged variables, further enhancing the dataset's quality for model training.

```
features.isna().sum()

Store              0
Date               0
Temperature        0
Fuel_Price         0
MarkDown1       4158
MarkDown2       5269
MarkDown3       4577
MarkDown4       4726
MarkDown5       4140
CPI              585
Unemployment     585
IsHoliday          0
dtype: int64
```

Fig.3 Missing Values

3. Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is an important step in understanding the underlying patterns, relationships, and characteristics of the dataset. In this section, various visualizations and summary statistics are employed to explore the Walmart dataset. Data Overview: The dataset consists of three files: train.csv, stores.csv, and features.csv. These are loaded into dataframes named data, stores, and features respectively. The initial inspection of the datasets includes checking their shapes, displaying the first few rows, and summarizing the information about missing values.
The data dataframe is merged with stores and features dataframes on the 'Store' and 'Date' columns to create a consolidated dataset. Fig. 4 Exploring the sales performance over time. By aggregating weekly sales into monthly averages, it highlights trends that are essential for decision-making in retail management

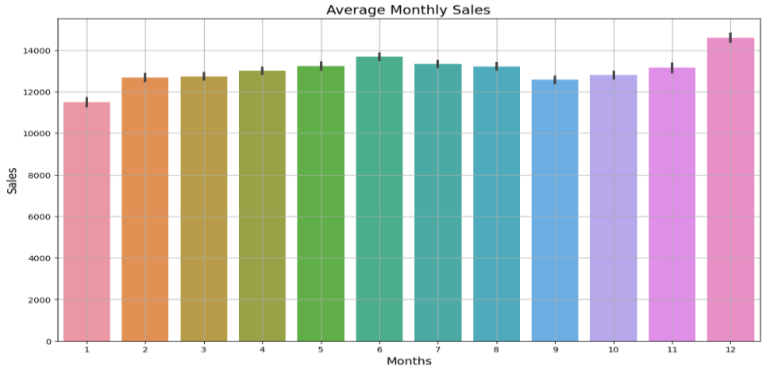


Fig.4 Average Monthly Sales

Fig. 5 provides valuable insights into the performance of individual Walmart stores. It highlights the variation in sales across stores, helping management to identify high and low performers and develop targeted strategies to optimize overall sales performance.

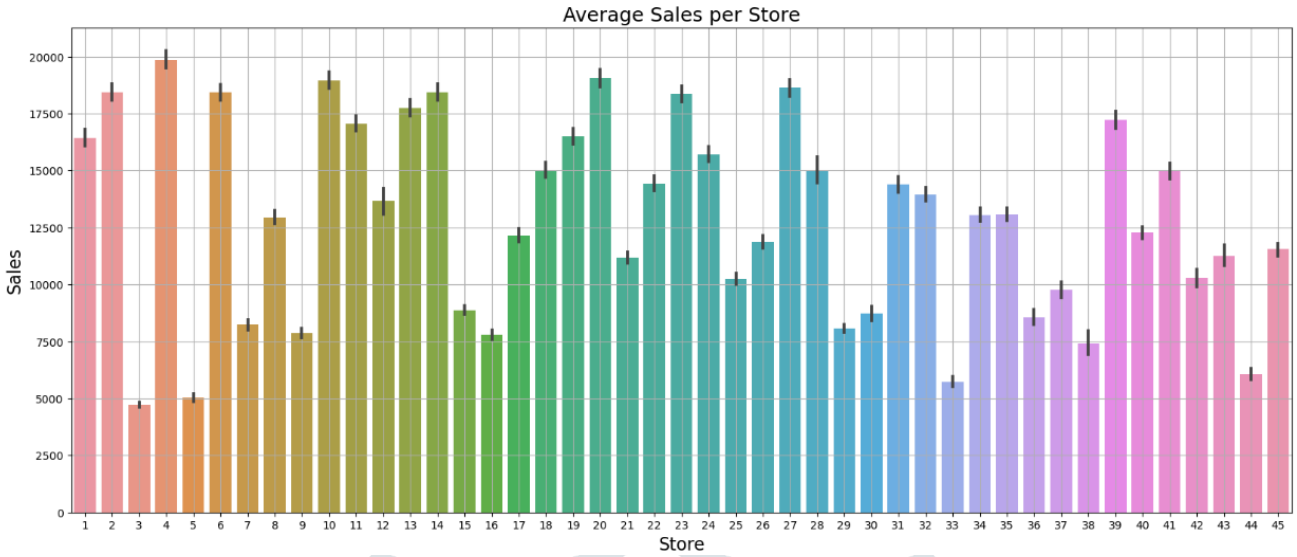


Fig.5 Average Sales Per Store

Fig. 6 Shows the impact of temperature as external factor on weekly sales, Warmer temperatures generally boost sales, especially for products related to outdoor activities and weather-specific needs. This analysis helps in making data-driven decisions for inventory management, promotional strategies, and understanding consumer behavior in response to temperature changes.

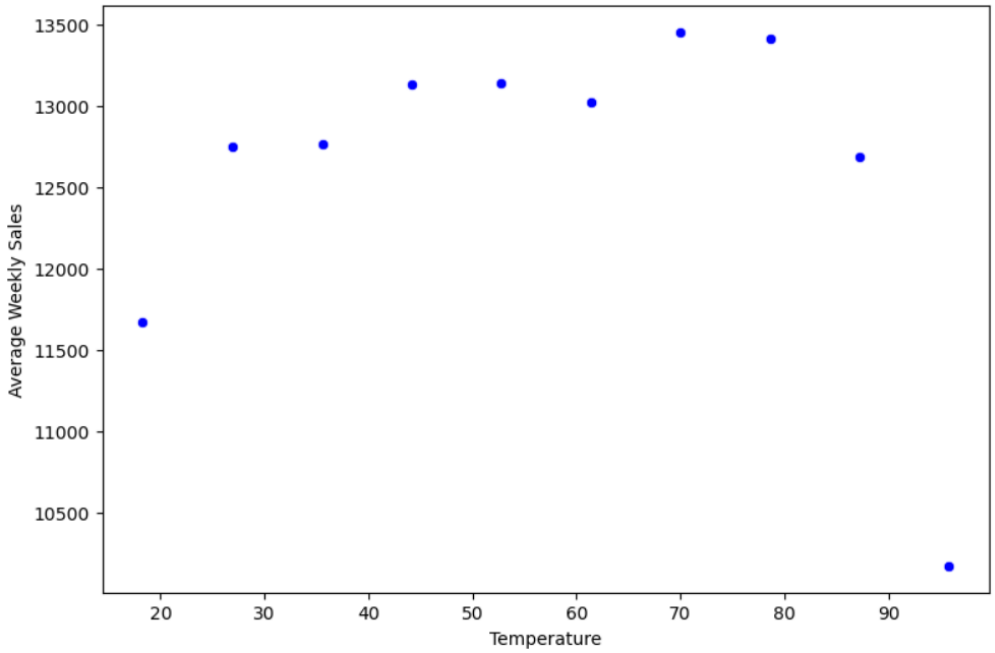


Fig.6 Impact of Temperature on Weekly sales

Fig. 7 concludes that the unemployment rate significantly impacts weekly sales, with higher rates generally leading to reduced sales due to lower consumer spending power. Retailers must adapt their inventory management and promotional strategies to address the shifts in consumer behavior resulting from changes in unemployment rates. Understanding this relationship helps in better forecasting and planning to mitigate the adverse effects of economic downturns on sales performance.

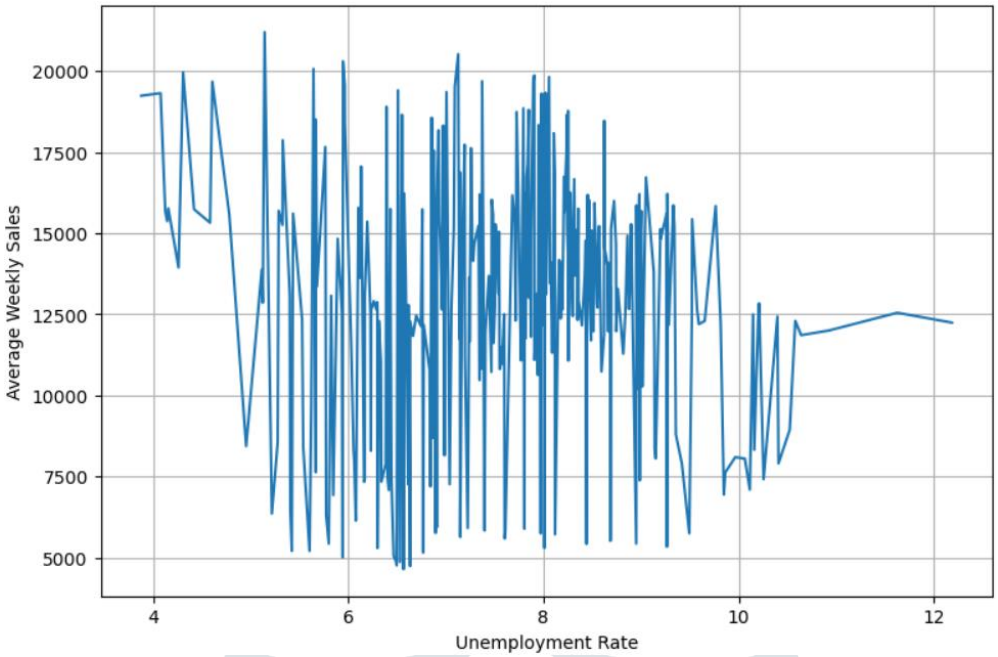


Fig.7 Impact of Unemployment Rate on Weekly sales

Fig. 8 show that fuel prices impact weekly sales, with higher fuel costs generally leading to decreased consumer spending on non-essential goods. This relationship necessitates that retailers adapt their inventory management, pricing, and promotional strategies to accommodate changes in consumer behavior driven by fuel price fluctuations.

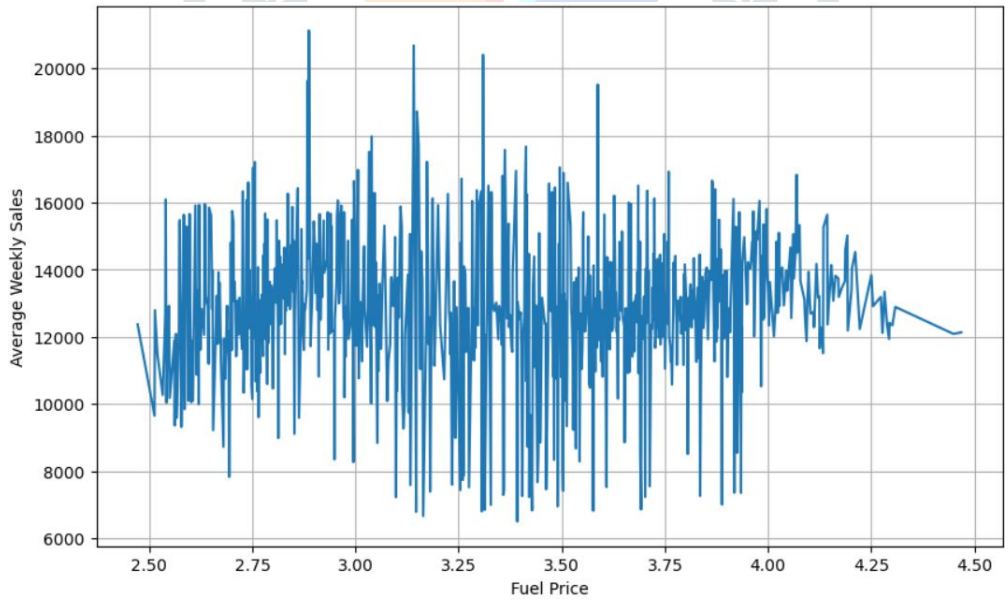


Fig.8 Impact of Fuel Price on Weekly sales

Fig. 9 shows the impact of Consumer Price Index (CPI) on weekly sales as CPI increases, indicating rising inflation, consumer spending generally shifts towards essential items, reducing overall weekly sales for discretionary goods. Retailers must adapt their inventory management, pricing, and operational strategies to navigate the challenges posed by inflation, ensuring they can meet changing consumer demands and maintain profitability during periods of high CPI

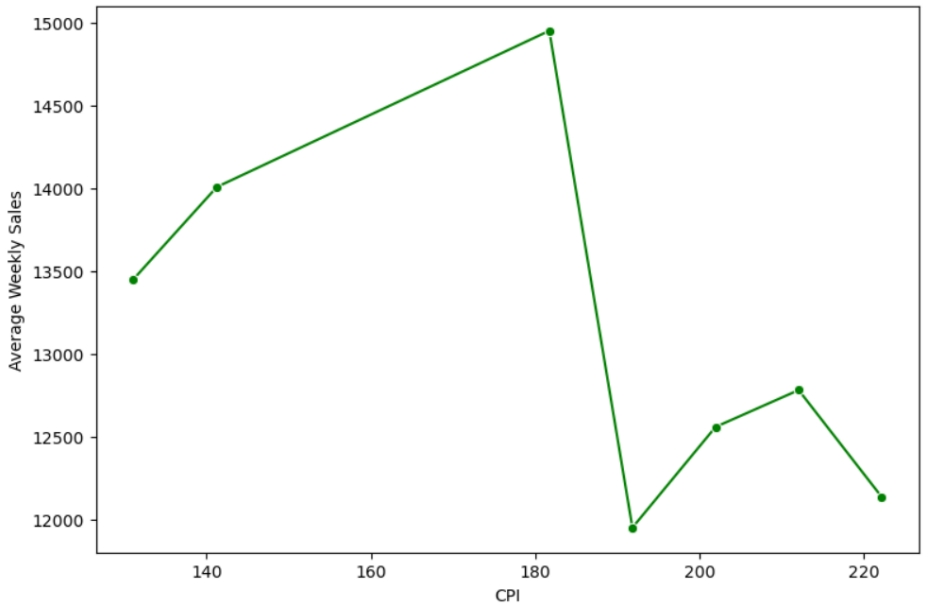


Fig.9 Impact of CPI Indicator on Weekly sales

Fig. 10 shows provides insights into the relationships between various features affecting weekly sales as below:
Weekly_Sales: Dependent variable representing weekly sales.
Size: Store size.
Temperature: Average weekly temperature.
Fuel_Price: Weekly fuel prices.
CPI: Consumer Price Index, indicating inflation.
Unemployment: Unemployment rate.
Total_MarkDown: Total markdowns or discounts applied.
According to the correlation matrix statistical matrix, Total Markdown1 and store size have a heights relation.

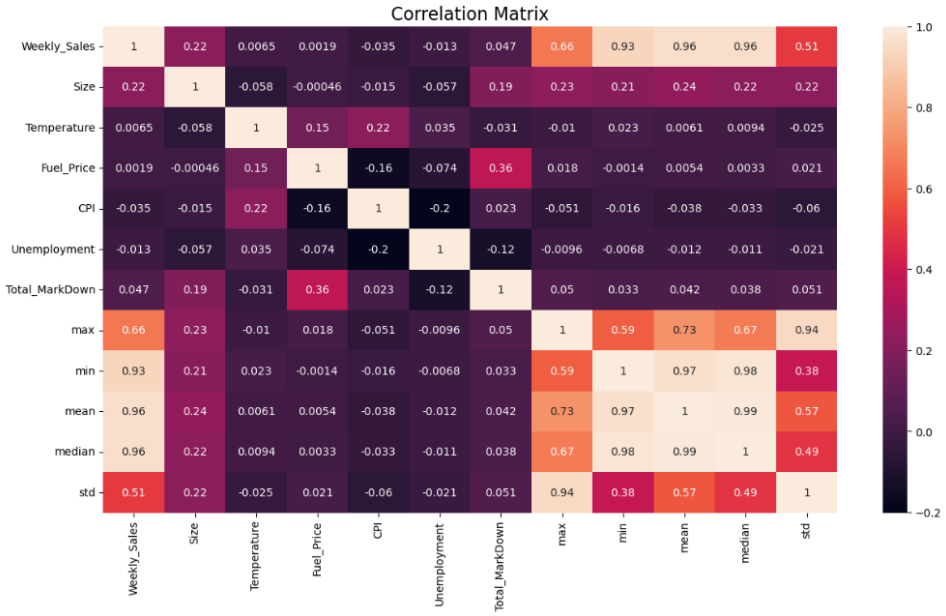


Fig.10 Correlation Matrix

4. Feature Selection:

Feature selection is an important step in building an efficient and accurate predictive model.
insights into the most influential features affecting the target variable, Weekly Sales. The Random Forest Regressor was trained using all available features, and the importance of each feature was extracted and ranked. The top features, based on their importance scores, were then selected for further modeling.
This selection streamlined the data, retaining only the most significant variables, thereby enhancing model performance and efficiency. Extracting time-related features and preparing the data for modeling is well-handled through a series of steps involving date conversion, sorting, and feature extraction to ensure all features were on a similar scale, normalization was performed using MinMaxScaler, which is especially important for distance-based algorithms like K-Nearest Neighbors (KNN). The final dataset included only the selected top features along with the target variable, facilitating more efficient training and potentially improving the model's accuracy. This approach not only reduced the complexity of the model but also made it more interpretable and easier to maintain.

5. Model Implementation:

The data analysis involved implementing and evaluating various supervised machine learning models. These models included Linear Regression, which served as a baseline to understand the linear relationship between features and sales; Random Forest Regression, an ensemble learning method using multiple decision trees to improve prediction accuracy; KNN Regression, a non-parametric method predicting the target value based on feature similarity, Extra Trees Regression, an ensemble method similar to Random Forest but with random selection of cut points; and XGBoost, an advanced boosting algorithm known for its high performance and efficiency in handling large datasets. The process starts by splitting the preprocessed and feature-selected dataset into training and testing sets, using an 80-20 split ratio. Each model is instantiated and trained using the training dataset. The training phase involves fitting the models to the data, allowing them to learn the underlying patterns and relationships between the features and the target variable. This multi-model approach provides a diverse set of predictions, leveraging different algorithmic strengths and offering a comprehensive analysis of the predictive capabilities for weekly sales forecasting.

6. Model Evaluation:

The model evaluation involves assessing the performance of various regression models using multiple metrics. After training each model on the training dataset, predictions are made on the test dataset. The evaluation metrics used include Mean Absolute Error (MAE) which measures how accurate the predicted
values are compared to the observed ones. [3]

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i | \quad (6)$$

Root Mean Squared Error (RMSE) this index measures the error between the model's output (ŷ) and the target values (y).[3]

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (7)$$

Mean Squared Error (MSE), and the R-squared (R^2) the coefficient of determination, reflects the proportion of the total variation of the dependent variable that can be explained by the independent variable through regression. The closer R2-Score is to 1. [19] the better the model. Its formula is as follows:

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y2)^2} \quad (8)$$

These metrics provide a comprehensive understanding of the models' accuracy and predictive capabilities. Our research also includes visual comparisons of the predicted values against the actual values, helping to illustrate the models' performance. Additionally, the accuracy of each model is calculated as a percentage score, allowing for straightforward comparison between different models. These evaluations ensure that the strengths and weaknesses of each model are clearly understood, guiding further refinement and selection of the best-performing model for predicting weekly sales.

III. RESULTS AND DISCUSSION

It has been observed that the good implementation of data preprocessing and effective handling of outliers is very important for enhancing the accuracy and reliability of predictive models.
the initial run implemented basic data preprocessing techniques and lacked thorough handling of outliers. The initial steps included joining the tables before importing the dataset, extracting simple features like year, month, and day, resulting in relatively low model performance and relatively higher error as shown in table 2

Table 2. Summary of model's performance before handling outliers

Model	R2	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Random Forest	0.971	1391.4	14650559.2	3827.6
Extra Tree	0.965	1403.2	17556415.6	4190.0
XGBoost	0.932	3043.8	34345552.7	5860.5
Linear Regression	0.088	14498.4	467198316.1	21614.7
KNN Regression	0.312	11609.5	352269934.9	18768.8
SVR	-0.083	12931.3	555377995.1	23566.4

Table 3. Summary of model's performance after handling outliers

Model	R2	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Random Forest	0.979	0.0154384986829464	0.00093234844348225	0.030534381334526
Extra Tree	0.981	0.0148672511617376	0.00084471973581328	0.0290640626171443

XGBoost	0.974	0.019548307057361	0.00115476733573999	0.033981867749433
Linear Regression	0.923	0.0298729099394516	0.0033887990614299	0.0582133924576636
KNN Regression	0.923	0.031937716767884	0.0033860463032977	0.058189743970030
SVR	0.012	0.144007865015135	0.044011459323073	0.209789082945402

1. Extra Tree Regressor:

The Extra Tree Regressor has the best performance on this dataset, with an R^2 score of 0.981, indicating that approximately 98.1% of the variance in weekly sales can be explained by the model. The low MAE, MSE, and RMSE values further emphasize the accuracy of predictions. Fig.11 illustrates a high correlation between predicted and actual sales since actual values represented by red line and prediction values represented by blue line.

Reasons of extra tree algorithm performance: The model can well handle non-linear relationships and interactions among features, which is critical for sales prediction. Also it uses diverse and relevant features, capturing various aspects that influence sales. Extra Trees Regressor leverages ensemble learning, improving robustness and accuracy compared to single decision trees.

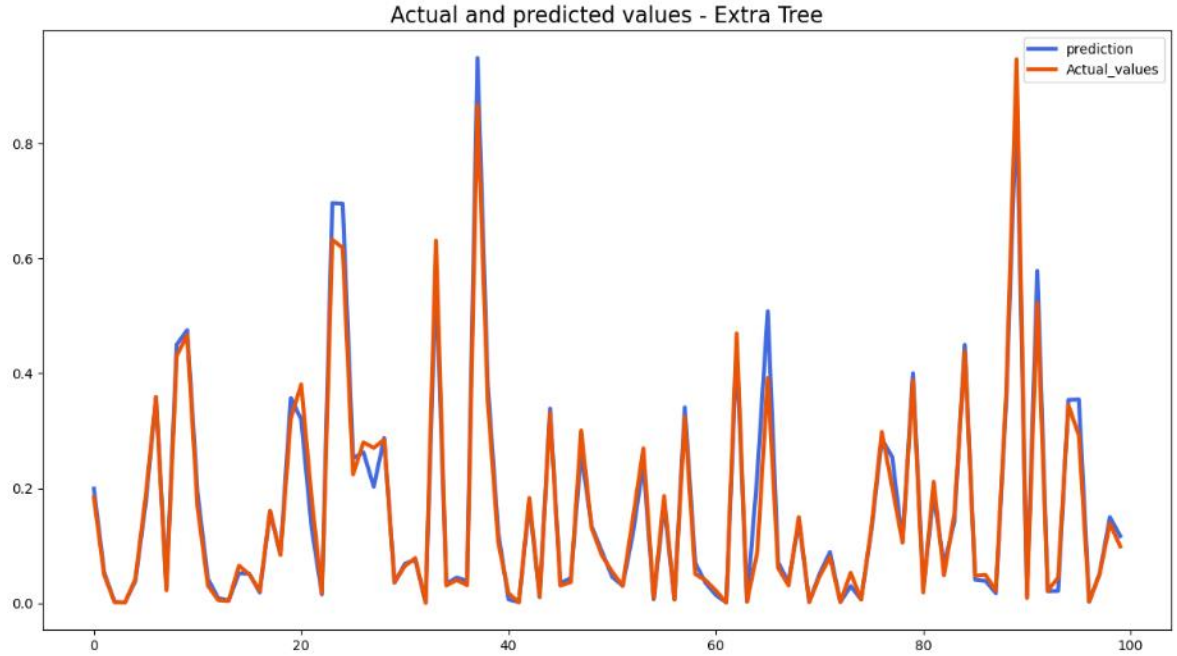


Fig.11 Extra Tree Prediction Line Chart

2. Random Forest Regressor:

The Random Forest Regressor is very similar to the Extra Tree it's achieves an R^2 score of 0.979, meaning it can explain 97.9% of the variance in weekly sales. The performance metrics (MAE, MSE, and RMSE) are within a perfect range for a high-accuracy regression model. Its ability to capture complex relationships between various influencing factors makes it a reliable model for sales prediction.

Fig.12 illustrates a high correlation between predicted and actual sales since actual values represented by red line and prediction values represented by blue line.

Reasons of random forest algorithm performance:

The Random Forest Regressor is effective in capturing complex relationships between features due to its ensemble nature, where multiple decision trees are averaged to improve accuracy and reduce overfitting.

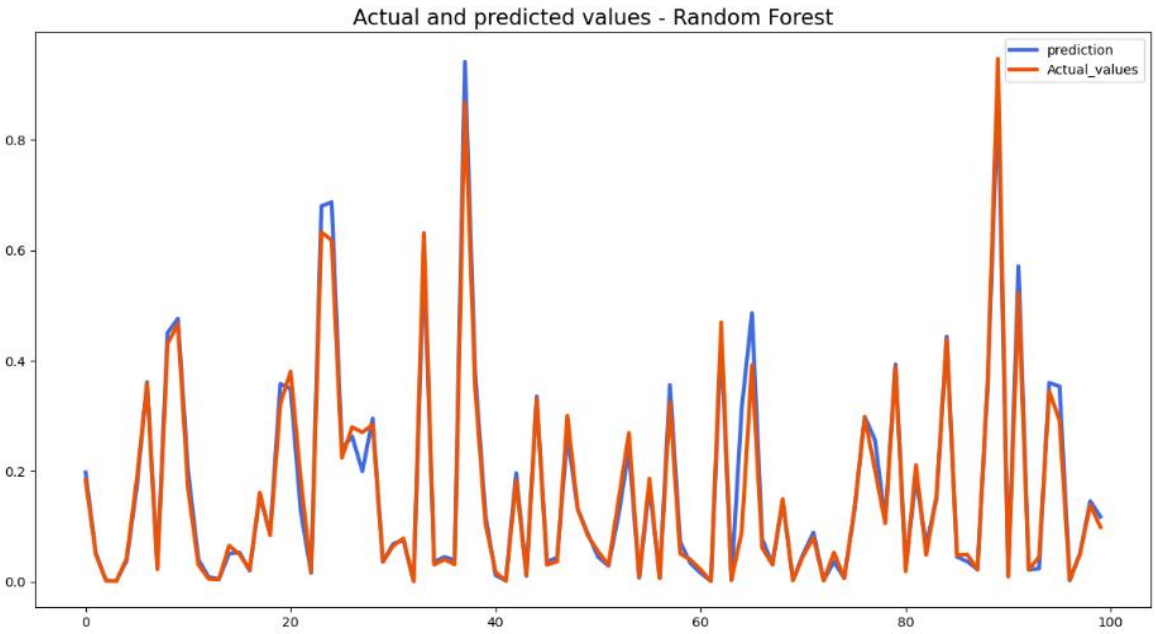


Fig.12 Random Forest Prediction Line Chart

3. XGboost Regressor:

The XGBoost Regressor provides highly accurate predictions with excellent performance metrics it's achieves an R^2 score of 0.974. The low MSE, MAE, and RMSE values indicate minimal prediction errors,
Reasons of Performance Fig.13 shows the high correlation between predicted and actual sales for XGBoost algorithm since predicted Values represented by below color and actual values represented by red colors.

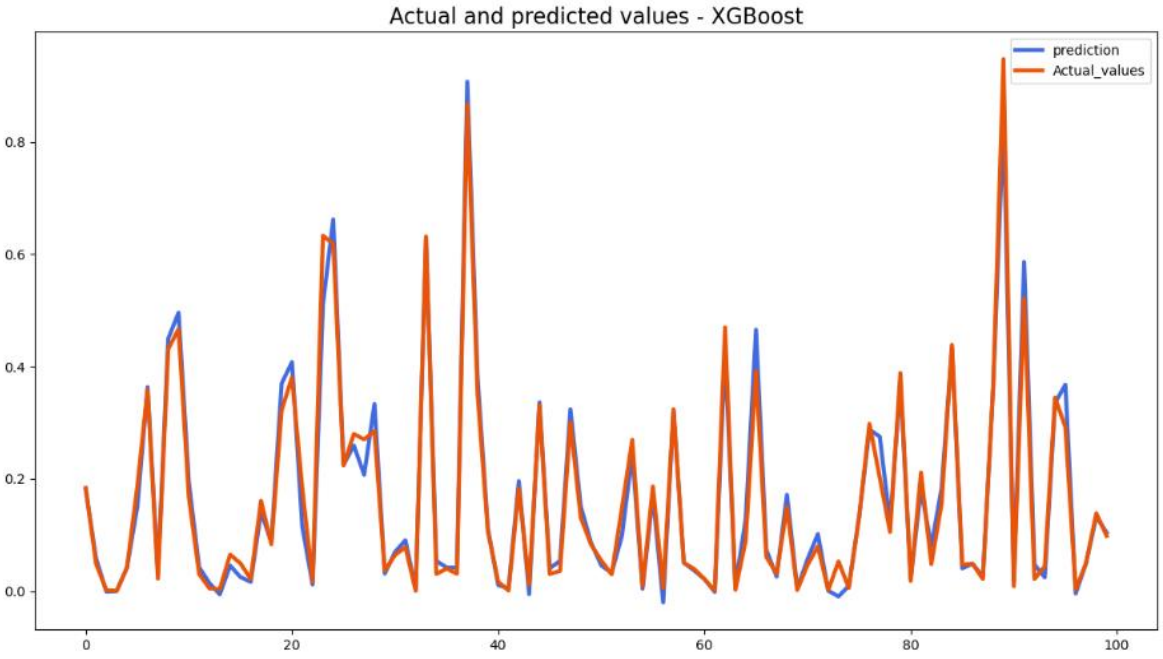


Fig.13 XGBoost Prediction Line Chart

Reasons of XGBoost algorithm Performance: XGBoost can capture complex non-linear relationships and interactions between features due to its ensemble nature and the boosting mechanism
The regularization parameters in XGBoost help to prevent overfitting by penalizing overly complex models. This balance between model complexity and performance contributes to its high R^2 value and low error metrics. XGBoost's optimization techniques (e.g., parallel processing, handling of sparse data) make it computationally efficient, allowing for better performance on large datasets and complex problems.

4. Linear Regression:
Linear Regression provides a good level of accuracy but is outperformed by more sophisticated models like XGBoost and Random Forest. Its R^2 score of 0.924 shows that it can capture a significant portion of the variance in the target variable, but its higher MSE, MAE, and RMSE values compared to XGBoost indicate less precision. Fig.14 shows the correlation between predicted and actual sales values for linear regression algorithm since predicted Values represented by below color and actual values represented by red colors.

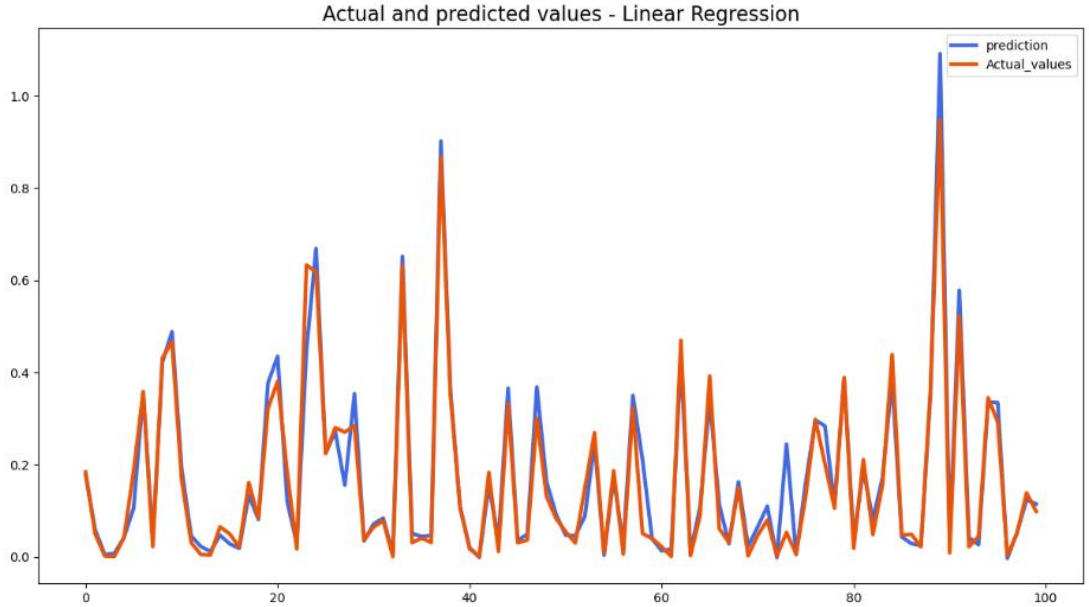


Fig.14 Linear Regression Prediction Line Chart

For tasks requiring higher accuracy and the ability to model complex interactions, XGBoost and Random Forest are preferable. However, for simpler problems or as a baseline, Linear Regression remains a valuable tool.

Reasons of linear regression performance: Linear Regression assumes a linear relationship between the input features and the target variable. This assumption can lead to suboptimal performance when the actual relationship is non-linear or involves complex interactions between features. Feature Scaling: Linear Regression is sensitive to the scale of features. Proper scaling of features can improve performance, but it still may not handle interactions between features effectively compared to ensemble methods.

5 KNN:

KNN Regressor provides a reasonably accurate prediction similar to linear regression but is less precise than the XGBoost, random forest and extra tree. Its R^2 score of 0.924 suggests that it captures a significant portion of the variance in the data. In practical applications, KNN can be a good baseline model when simpler predictions are preferred. Fig.15 shows the correlation between predicted and actual sales values for KNN algorithm since predicted Values represented by below color and actual values represented by red colors.

Reasons of KNN algorithm Performance:

KNN regression predicts the value of a point based on the average of its nearest neighbors. This approach works well when the target variable has a local structure but can struggle with more complex patterns where non-local interactions are significant. With high-dimensional data, distances between points can become less meaningful, reducing KNN's effectiveness.

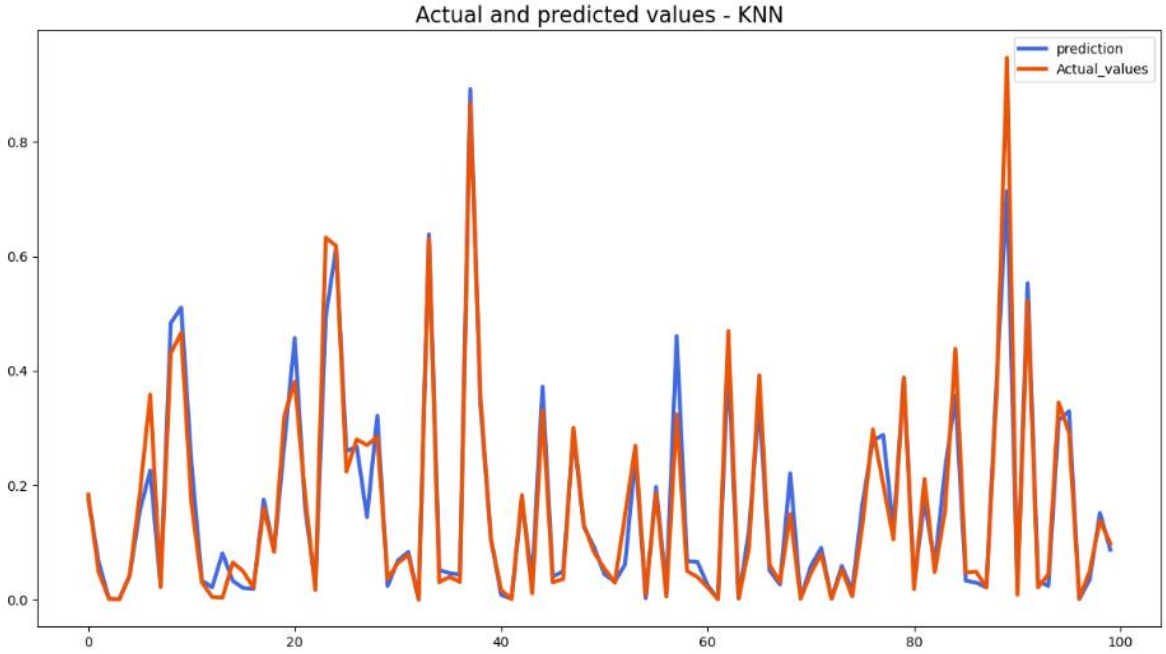


Fig.15 KNN Prediction Line Chart

6. SVR:

The SVR model underperforms significantly in comparison to other models. With an R^2 value of 0.012, it fails to capture the variance in the target variable effectively, and its high MSE, MAE, and RMSE indicate substantial errors in predictions. For practical applications where higher accuracy and better fit are required, SVR may not be the most appropriate mode. Fig 16 shows

the low correlation between predicted and actual sales values since predicted Values represented by below color and actual values represented by red colors.

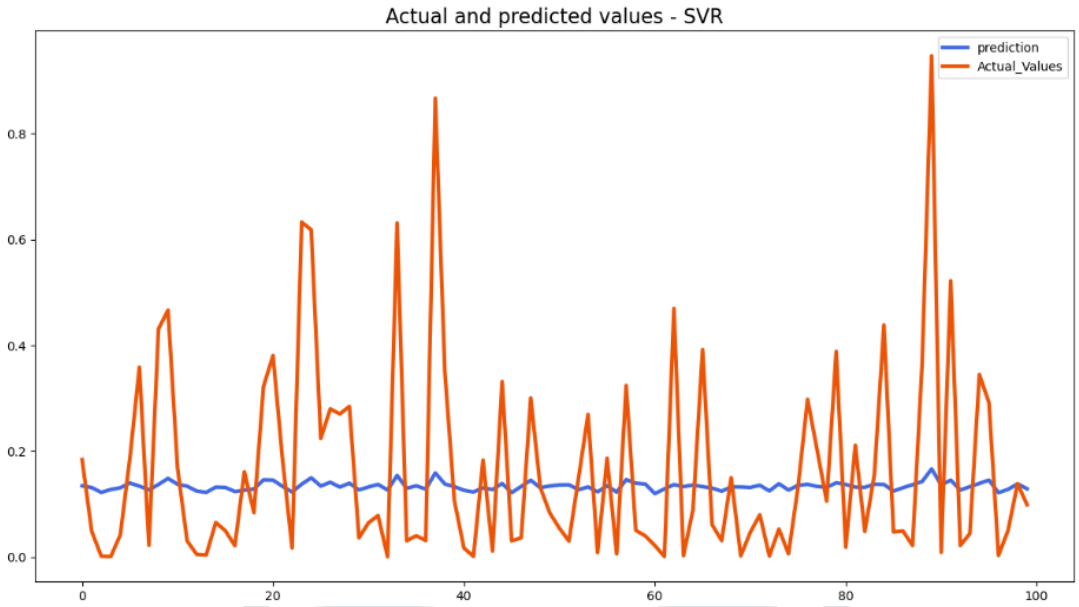


Fig.16 SVR Prediction Line Chart

Reasons of SVR algorithm performance: support vector regression models require an accurate selection of the configuration parameters in order to achieve good generalization performance.[7] it can be sensitive to the choice of kernel and hyperparameters. The performance might indicate inappropriate selection of these parameters, leading to poor generalization. Scaling issues: SVR is sensitive to the scale of the input data. If the data is not properly scaled, the model's performance might be adversely affected.

7. Conclusion and Future Work

The analysis of the initial and enhanced models for the Walmart sales dataset emphasize the importance of perform data preprocessing in machine learning. In the initial approach, basic preprocessing techniques such as filling missing values with zeros were employed, and there was no handling of outliers or negative values. This led to poor model performance, especially for Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) models, which indicating poor fit and predictive capabilities.

On the other hand, the advanced models introduced more improved preprocessing steps, missing values were handled with the column means rather than zeros. these enhanced preprocessing steps resulted in a significant improvement in model performance across the board. For instance, the Random Forest model achieved an R-squared value of 0.979, and the extra trees Regression model reached 0.981, reflecting their improved ability to explain the variance in the data. The XGBoost model also exhibited a notable R-squared increase to 0.974, demonstrating enhanced predictive accuracy.

Moreover, the mean squared error (MSE), Mean Absolute Error (MAE), and root mean squared Error (RMSE) were markedly lower in the enhanced notebook for all models, confirming the positive impact of advanced preprocessing on model accuracy and reliability. The dramatic reduction in error rates is likely due to effective preprocessing, such as handling outliers and normalizing data. These steps can significantly

In future, we aim to enhance the accuracy and improve our sales forecasting methodology by integrating more comprehensive datasets and use advanced hybrid models. we will continue to explore various approaches to systematically evaluate critical data and variables. Additionally, our investigation will extend into the realm of deep learning algorithms to address the complexities and obstacles in sales prediction.

IV. Acknowledgment

I would like to express my sincere gratitude to all those who have supported me in completing this thesis.
First and foremost, I am deeply grateful to my supervisor, Dr. Ahmed Hamza, from the Department of Computer and Information Sciences at the Institute of Statistical Studies and Research, University of Cairo. His extensive knowledge and logical thinking has been invaluable assets throughout this project. His understanding, encouragement, and personal guidance provided a strong foundation for this thesis.

REFERENCES

[1] J. R. S. Stuti Raizada, "Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting," *2021 Int. Conf. Adv. Comput. Technol. Appl. ICACTA 2021*, vol. 12, no. 11, pp. 102–110, 2021, doi: 10.1109/ICACTA54488.2022.9753636.

[2] D. E. and P. C. S. R. S. B. Latha, C. Dastagiraiah, A. Kiran, S. Asif, "An Adaptive Machine Learning model for Walmart sales prediction," *2023 Int. Conf. Circuit Power Comput. Technol. (ICCPCT)*, 2023, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10245029

[3] P. Rangika, "MACHINE LEARNING BASED SALES FORECASTING SYSTEM Pathum R . Ranathunga Department of Computing," no. July, 2022.

[4] Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," *Proc. - 2020 Int. Conf. Big Data Artif. Intell. Softw. Eng. ICBASE 2020*, pp. 458–461, 2020, doi: 10.1109/ICBASE51474.2020.00103.

[5] S. K. Udgata and S. Sethi, *Intelligent Systems*. Springer, 2021.

[6] A. Mitra, A. Jain, A. Kishore, and P. Kumar, "A Comparative Study of Demand Forecasting Models for a Multi - Channel Retail Company : A Novel Hybrid," *Oper. Res. Forum*, pp. 1–22, 2022, doi: 10.1007/s43069-022-00166-4.

[7] P. Tsirikoglou, S. Abraham, F. Contino, C. Lacor, and G. Ghorbaniasl, "A hyperparameters selection technique for support vector regression models," *Appl. Soft Comput. J.*, vol. 61, pp. 139–148, 2017, doi: 10.1016/j.asoc.2017.07.017.

[8] S. Malik, M. Khan, M. K. Abid, and N. Aslam, "View of Sales Forecasting Using Machine Learning Algorithm in the Retail Sector," *J. Comput. Biomed. Informatics*, vol. 06, no. 02, 2024, [Online]. Available: https://www.jcbi.org/index.php/Main/article/view/370/299

[9] S. Yi, "Walmart Sales Prediction Based on Decision Tree, Random Forest, and K Neighbors Regressor," *Highlights Sci. Eng. Technol.*, vol. 47, pp. 87–94, 2023, doi: 10.54097/hset.v47i.8170.

[10] B. Seaman and J. Bowman, "Applicability of the M5 to Forecasting at Walmart," *Int. J. Forecast.*, vol. 38, no. 4, pp. 1468–1472, 2022, doi: 10.1016/j.ijforecast.2021.06.002.

[11] I. I. Conference, C. Electronics, and C. Engineering, "Machine Learning Model for Sales Forecasting by Using XGBoost," no. Iccece, pp. 480–483, 2021.

[12] P. Sabarish, T. Vignesh, and S. Yogeendran, "Walmart Sales Prediction Using Machine Learning Algorithms," vol. 25, no. 4, pp. 7872–7882, 2021.

[13] Q. Li and M. Yu, "Achieving Sales Forecasting with Higher Accuracy and Efficiency: A New Model Based on Modified Transformer," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 4, pp. 1990–2006, 2023, doi: 10.3390/jtaer18040100.

[14] T. Lu, "Weekly Sales Prediction of Walmart Stores Based on Three Regression Models," vol. 88, pp. 29–38, 2024.

[15] C. Yu, "Walmart Sales Forecasting using Different Models," *Highlights Sci. Eng. Technol.*, vol. 92, pp. 302–307, 2024.

[16] T. Yang, "Sales Prediction of Walmart Sales Based on OLS, Random Forest, and XGBoost Models.," *Highlights Sci. Eng. Technol. 49*, 2023.

[17] N. Anand, T. Sharma, K. Gaurav, and K. Kharbanda, "RESEARCH IN ENGINEERING MANAGEMENT RETAIL ANALYSIS — WALMART ' S TREND ASSESSMENT," pp. 363–373, 2023.

[18] Y. F. Akande, J. Idowu, A. Misra, S. Misra, O. N. Akande, and R. Ahuja, "Application of XGBoost Algorithm for Sales Forecasting Using Walmart Dataset," *Lect. Notes Electr. Eng.*, vol. 881, no. August, pp. 147–159, 2022, doi: 10.1007/978-981-19-1111-8_13.

[19] Z. Chen, "Sales Forecast of Walmart on Account of Multivariate Regression and Machine Learning Methods," 2023, doi: 10.4108/eai.28-10-2022.2328455.