Optimizing Walmart Sales Predictions with Machine Learning Techniques

D. Mohammad Abdulla
mohammadabdulla20march@gmail.com
Artificial Intelligence,
Amrita School of Engineering,
Bengaluru, India

G. Tejdeep Reddy
BL.EN.U4AIE21048@bl.students.amrita.edu
Artificial Intelligence,
Amrita School of Engineering,
Bengaluru, India

G. Mukesh Venkata Sai
BL.EN.U4AIE21050@bl.students.amrita.edu
Artificial Intelligence,
Amrita School of Engineering,
Bengaluru, India

S. Srihemanth
BL.EN.U4AIE21123@bl.students.amrita.edu
Artificial Intelligence,
Amrita School of Engineering,
Bengaluru, India

Dr. Manju Venugopalan
v_manju@blr.amrita.edu
Department of Computer Science and Engineering,
Amrita School of Engineering,
Bengaluru, India

***Abstract*—**This study examines the application of data mining in retail sales for sales forecasting and demand prediction. Sales prediction is a critical aspect that significantly influences the long-term success of any organization. Various techniques, including Time Series Algorithm, Regression Techniques, and Association Rule, can be employed for supermarket sales prediction. This paper conducts a comparative analysis of several Supervised Machine Learning Techniques, including Multiple Linear Regression Algorithm, Random Forest Regression Algorithm, K-NN Algorithm, Support Vector Machine (SVM) Algorithm, XGboost Model, and Extra Tree Regression. The objective is to build a prediction model for estimating sales at 45 different geographical locations of Walmart stores, a prominent global retail chain. The study acknowledges that sales are influenced by events and holidays, sometimes on a daily basis. The prediction model incorporates features such as previous sales data, promotional events, holiday weeks, temperature, fuel prices, Consumer Price Index (CPI), and state unemployment rates. Data is collected from 45 Walmart outlets, and various Supervised Machine Learning Techniques are applied to predict Walmart's sales accurately. The paper aims to assist business owners in selecting an appropriate approach for sales prediction, considering different scenarios like temperature variations, holidays, and fuel prices. Business owners can choose the best marketing and promotion plans for their grocery goods by being aware of these variables.

***Keywords*—Sales forecasting; linear regression; random forest regression; KNN regression algorithm; SVM algorithm; Random Forest, XGboost Model, CNN classifications, supervised machine learning techniques, MAE, MSE, RMSE, R2.**

I. INTRODUCTION

The retail sector stands out as a prominent and rapidly expanding domain within the field of data science due to its vast data volumes and numerous optimization challenges, including determining ideal prices, recommendations, discounts, and stock levels. Addressing these challenges through diverse data analysis methods is essential. Predicting commodity sales in today's dynamic business environment poses significant challenges, but even slight improvements in sales prediction can lead to reduced operational costs, improved sales, and heightened customer satisfaction. Accurate sales predictions for every retail outlet are crucial for the success of retail companies. This accuracy aids in inventory management, ensures proper product distribution among stores, mitigates issues of over and under stocking to minimize losses, and maximizes customer satisfaction. Numerous factors contribute to sales prediction complexity, including external factors like weather, seasonal trends, competition, and online shopping, as well as internal factors such as promotions, discounts, and pricing. Supermarket sales forecasting heavily relies on machine learning techniques, and one supervised learning algorithm that uses ensemble learning is Random Forest Regression. During training, decision trees are built, and predictions are made using the mean output of those trees. This method enhances prediction accuracy by averaging the predictions from multiple trees. K-NN Regression predicts new values based on feature similarity to existing data points, while Support Vector Regression (SVR) determines the best-fitting line with the maximum number of points lying on it. Extra Tree Regression, another ensemble learning model, closely resembles Random Forest but selects cut points differently. In this study, a review of literature related to sales forecasting is conducted, and the adopted methodology is defined. The model is trained on various machine learning algorithms, including Linear Regression, Random Forest Regression, KNN Regression, SVR, and Extra Tree Regression. The obtained results from each model are discussed, leading to the conclusion of the study.

II. LITERATURE SURVEY

A. Paper-I (Author : GV Aditya ,Published 2023)

Linear Regression is a linear equation to observed data, the technique known as "linear regression" can be used to model the relationships between two variables. The dependent variable (target) is the other variable, while the explanatory variable (predictor) is the first one. It involves determining which line of fit best suits testing and training data. This method has been applied to forecast commodity demand through the examination of retail sales data. A crucial component of production and supply chain management is sales forecasting.

B. *Paper-2 (Author : Ashok Kumar ,Published 2023)*

Sales forecasting for retail chains using Supervised Machine Learning Techniques We have 16 attributes Weakness.....it is Binary Classification Methods LR, DNN, SVM, RF. RF has accuracy of 98%.KNN SVM have 70% nearly.DNN is 90%.

C. *Paper-3 (Author : M. Zaharia ,Published 2021)*

Three different datasets are used to evaluate the ML model. First is from N.K.P. Fast and interactive analytics over Hadoop data with Spark Methods: LR, Naive Bayes, SVM, RF.RF has high precision Recall, Accuracy with 99%.,Kappa index max was RF with 97%.

D. *Paper-3 (Author : M. Lawrence et al ,Published 2000)*

This study show that the corporate forecasts were not always more accurate than the basic naive value for the companies under investigation. Moreover, the conclusion appears harsh at first glance since the naive measure was not modified for the effects of seasonality. Due to the fact that there was a serial correlation in the mistakes, the majority of the sources of error appeared to result from both forecast bias and inefficiency. Contextual information appeared to have little effect on accuracy because of these two factors. Bias, however, can be deliberate; that is, motivated by organizational behavioral characteristics and the requirement to foster adherence to the estimate. This implies that the company's forecasting process may have goals beyond forecast accuracy.

E. *Paper-3 (Author: Siddhaling Urolagin ,Published 2020)*

Since KNN methods are nonparametric, they are mostly applied to forestry-related issues such as remote sensing. However, the statistical characteristics of KNN regression are not as thoroughly studied as those of parametric regression analysis, or linear regression, which has the advantages of being simple to fit, requiring the estimation of fewer coefficients, and being straightforward to understand. When comparing KNN regression to linear regression, we find that the gap between the training and testing errors is larger. This suggests that the KNN regression model is overfitting, as it performs better on training data than testing data. Based on this analysis, we can conclude that linear regression is a better model for the dataset mentioned above because it reduces training and testing errors for RMSE and

III. METHODOLOGY

To conduct this research, we followed a systematic procedure outlined in

Figure-1. The key steps involved in our study are detailed below:

A. Dataset and Features Description:

We utilized sales data from 45 Walmart stores for our prediction model, covering historical sales from February 5, 2010, to November 1, 2012. The dataset comprises three separate files corresponding to each year, and model accuracy is evaluated for each.

The dataset includes the following fields:
Store: Number of stores (45 stores considered).Date: First date of the week for time series forecasting.Weekly Sales: Sales for the given store on a weekly basis. Holiday Flag: Indicator for special holiday weeks (1 for Holiday week, 0 for Non-holiday week).Temperature: Recorded temperature on the day of sale. Fuel Price: Cost of fuel in the store's region. CPI: Dominant Consumer Price Index. Unemployment: Dominant unemployment rate in the store's region.

Figure-2 provides a snapshot of the Walmart store dataset.

B. Dataset Preparation:

Imported necessary libraries (e.g., numpy, pandas,matplotlib, seaborn). Loaded the dataset for the years 2010, 2011, and 2012 into the IDE.

Prepared the data by converting dates to datetime format. Checked for missing or null values. Split the date column, creating day, month, and year columns. Identified outliers in Temperature, Fuel Price, CPI, and Unemployment by plotting them on the X-axis. Removed outliers and confined the data to a reasonable range. Reassessed plots to ensure their integrity without outliers.

C. Applying Correlation:

Investigated correlations among variables to identify relationships. Analyzed the correlation matrix to guide feature selection.

D. Applying Machine Learning Techniques:

Imported the sklearn library for model development. Selected features and target for X and Y axes. Split the data into training and testing sets in an 80:20 ratio.

E. Developing Classification Model:

Utilized Linear Regression, Random Forest Regressor, KNN Regressor, SVM, and Extra Tree Regressor for sales prediction. Conducted a comparative analysis of the prediction models.

F. Model Performance Analysis:

Calculated errors in the prediction models using Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error.

By following these steps, our research aimed to provide insights into sales prediction models for Walmart stores.

TABLE I
DATASET

	Store	Date	Temperatu	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployn	IsHoliday
2	1	05-02-2010	42.31	2.572	NA	NA	NA	NA	NA	211.09636	8.106	FALSE
3	1	12-02-2010	38.51	2.548	NA	NA	NA	NA	NA	211.24217	8.106	TRUE
4	1	19-02-2010	39.93	2.514	NA	NA	NA	NA	NA	211.28914	8.106	FALSE
5	1	26-02-2010	46.63	2.561	NA	NA	NA	NA	NA	211.31964	8.106	FALSE
6	1	05-03-2010	46.5	2.625	NA	NA	NA	NA	NA	211.35014	8.106	FALSE
7	1	12-03-2010	57.79	2.667	NA	NA	NA	NA	NA	211.38064	8.106	FALSE
8	1	19-03-2010	54.58	2.72	NA	NA	NA	NA	NA	211.21564	8.106	FALSE
9	1	26-03-2010	51.45	2.732	NA	NA	NA	NA	NA	211.01804	8.106	FALSE
10	1	02-04-2010	62.27	2.719	NA	NA	NA	NA	NA	210.82045	7.808	FALSE
11	1	09-04-2010	65.86	2.77	NA	NA	NA	NA	NA	210.62286	7.808	FALSE
12	1	16-04-2010	66.32	2.808	NA	NA	NA	NA	NA	210.4887	7.808	FALSE
13	1	23-04-2010	64.84	2.795	NA	NA	NA	NA	NA	210.43912	7.808	FALSE
14	1	30-04-2010	67.41	2.78	NA	NA	NA	NA	NA	210.38955	7.808	FALSE
15	1	07-05-2010	72.55	2.835	NA	NA	NA	NA	NA	210.33997	7.808	FALSE
16	1	14-05-2010	74.78	2.854	NA	NA	NA	NA	NA	210.33743	7.808	FALSE
17	1	21-05-2010	76.44	2.826	NA	NA	NA	NA	NA	210.61709	7.808	FALSE
18	1	28-05-2010	80.44	2.759	NA	NA	NA	NA	NA	210.89676	7.808	FALSE
19	1	04-06-2010	80.69	2.705	NA	NA	NA	NA	NA	211.17643	7.808	FALSE
20	1	11-06-2010	80.43	2.668	NA	NA	NA	NA	NA	211.4561	7.808	FALSE
21	1	18-06-2010	84.11	2.637	NA	NA	NA	NA	NA	211.45377	7.808	FALSE
22	1	25-06-2010	84.34	2.653	NA	NA	NA	NA	NA	211.33865	7.808	FALSE
23	1	02-07-2010	80.91	2.669	NA	NA	NA	NA	NA	211.22353	7.787	FALSE
24	1	09-07-2010	80.48	2.642	NA	NA	NA	NA	NA	211.10841	7.787	FALSE
25	1	16-07-2010	83.15	2.623	NA	NA	NA	NA	NA	211.10039	7.787	FALSE
26	1	23-07-2010	83.36	2.608	NA	NA	NA	NA	NA	211.23514	7.787	FALSE
27	1	30-07-2010	81.84	2.64	NA	NA	NA	NA	NA	211.3699	7.787	FALSE
28	1	06-08-2010	87.16	2.627	NA	NA	NA	NA	NA	211.50466	7.787	FALSE

Fig 1: Features of Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8190 entries, 0 to 8189
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         8190 non-null   int64
 1   Date          8190 non-null   object
 2   Temperature   8190 non-null   float64
 3   Fuel_Price    8190 non-null   float64
 4   MarkDown1     4032 non-null   float64
 5   MarkDown2     2921 non-null   float64
 6   MarkDown3     3613 non-null   float64
 7   MarkDown4     3464 non-null   float64
 8   MarkDown5     4050 non-null   float64
 9   CPI           7605 non-null   float64
 10  Unemployment  7605 non-null   float64
 11  IsHoliday     8190 non-null   bool
dtypes: bool(1), float64(9), int64(1), object(1)
memory usage: 712.0+ KB
```

Fig 2: Dataset Description

A. Dataset Preparation

Dataset preparation is a critical step in the data mining process, emphasizing data pre-processing. Our primary focus was on the training dataset, where we executed essential transformations. To facilitate model development, we converted categorical data into numerical values. Specifically, we assigned 0 for females and 1 for males, ensuring appropriate representation. After these transformations, the dataset was carefully processed and made ready for analysis. In the final steps, we opted to allocate 75% of the data for training purposes and reserved the remaining 25% for testing. This decision ensures a balanced and effective approach to model evaluation.

Flow Diagram

The objective of this study was to anticipate sales at Walmart stores using a variety of supervised machine learning techniques, such as Support Vector Machine, Random Forest, K-NN, Linear, and Extra Tree regression (Fig. 3). The dataset used to train the predictive models included data from 45 Walmart retail locations. To guarantee data quality, the dataset underwent extensive cleaning using Python. Important variables that were taken into account in order to estimate sales included prior sales, holidays, gasoline prices at particular times of the year, and unemployment rates. After that, the dataset was divided into training and testing sets while keeping a predetermined ratio.

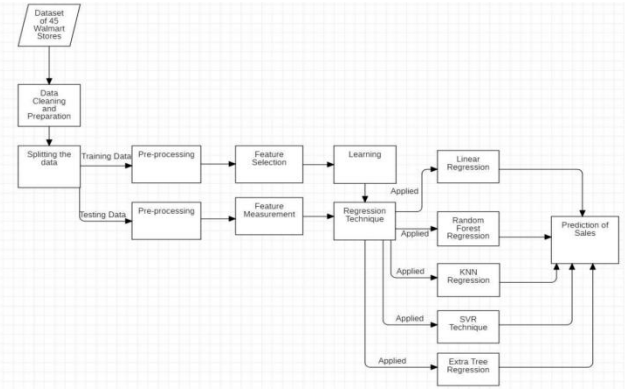

Fig 3: Flow Diagram

B. Correlation

Understanding the relationships among variables is a crucial aspect of our analysis, A key component of our study is figuring out how the variables relate to one another, therefore we concentrated on looking for correlations within the dataset. In this step, patterns and connections between various features were found by looking at the correlation matrix. We evaluated the direction and strength of correlations between variables as part of our correlation analysis process. We learned a great deal about the interactions between different components in the dataset by doing this. Our feature selection procedure was informed by this data, which made sure that the most pertinent variables for our research were taken into account. To summarise, our research's correlation step was crucial in influencing the later phases of model construction and provided a basis for defensible

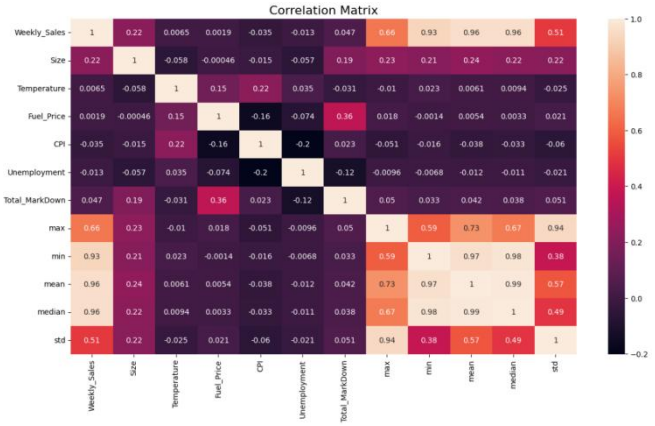

Fig 4: Correlation between features of dataset

C. Applying Machine Learning Techniques

For the model-building phase, we employed four classification algorithms: Logistic Regression, Random Forest, and XGBoost. Our analysis aimed at predicting sales and understanding the factors contributing to attacks on humans by products.

1) Linear Regression: Linear Regression is a technique to model the relationships between two variables by fitting a linear equation By fitting a linear equation to observed data, one can model the relationships between two variables using

the technique known as linear regression. The dependent variable (target) is the other variable, while the explanatory variable (predictor) is the first one. It involves determining which line of fit best suits testing and training data. This method has been applied to forecast commodity demand through the examination of retail sales data. A crucial component of production and supply chain management is sales forecasting.

2) Random Forest: Random Forest, an ensemble classifier, leverages decision tree algorithms in a randomized manner. Developed by Leo Breiman, it excels in both regression and classification tasks within supervised machine learning. Notably, Random Forest avoids tree pruning, demonstrating randomness in creating a bootstrap dataset and constructing decision trees from it. The algorithm yields fast results with high prediction accuracy, making it suitable for handling diverse input data effectively. The combination of subsurface randomization and bagging enhances its performance by replacing the training dataset for each new tree.

3) KNN Regression: is a regression technique that, after examining data from previous cases, predicts values using a similarity measure. It takes the features out of the data and predicts the values of fresh data points by using ‖feature similarity. The average of the new data point's closest neighbors is used to determine the value of new data. Finding an inverse distance weighted average of the K-nearest neighbors is the alternative method for using KNN. The distance functions employed are the same as those used for classification.

4) XGBoost: XGBoost, an optimized distributed gradient boosting tree model designed to be highly efficient and flexible, is good at classification and regression problem. The experimental results demonstrate that the proposed model based on XGBoost has good performance in improving the running rate and the prediction accuracy. And we achieve better sales forecasting compared to other popular machine learning methods.

D. Developing Regression Model:

Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

IV. RESULTS

Tables I, II, and III, which present the outcomes of the prediction models fed with datasets from the three years 2010, 2011, and 2012, are summarized below It can be shown from Tables I, II, and III that the Mean Absolute Error is lowest for Extra Tree Regression and greatest for Support Vector Regression throughout the course of all three years. The average magnitude of the error in the prediction set is called the Mean Absolute Error. It is the mean of the absolute difference between the observed and predicted values over the test sample.

	MAE	MSE	RMSE	R2
Linear Regression	0.03	0.003	0.05	0.92
Random Forest Regression	0.015	0.009	0.03	0.97
K Neighbors Regression	0.03	0.003	0.06	0.91
XGboost Model	0.019	0.001	0.03	0.97

Out of all the models that have been described thus far, the Extra Tree Regressor Model performed the best at predicting sales. While the graph and Random Forest Technique appear to be quite similar, the accuracy of the former is higher, with all three cases achieving an accuracy of 98% because all data points nearly fall on the best of fir. This is because, in order to deliver superior performance, both of these strategies use an ensemble model of learning that averages the results from multiple decision trees.

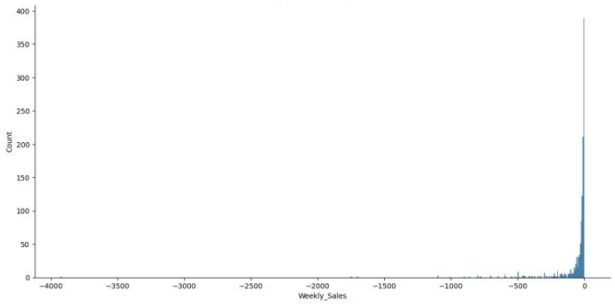

Fig 5: Negative Weekly Sales

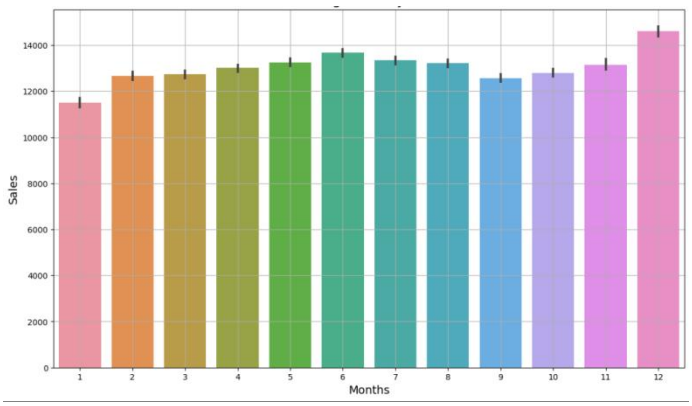

Fig 6: Average Monthly Sales

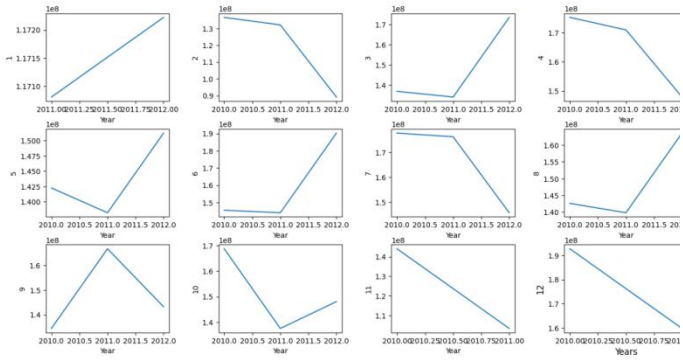

Fig 7: Monthly Sales for each Year

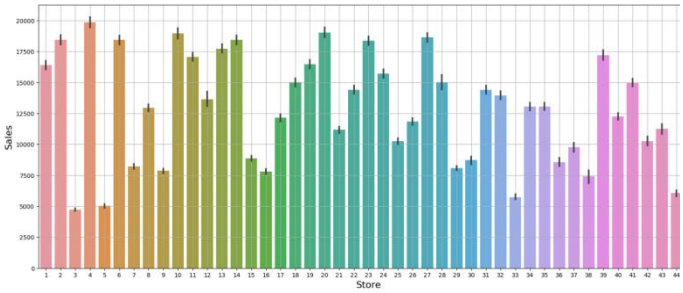

Fig 8: Average Sales per Store

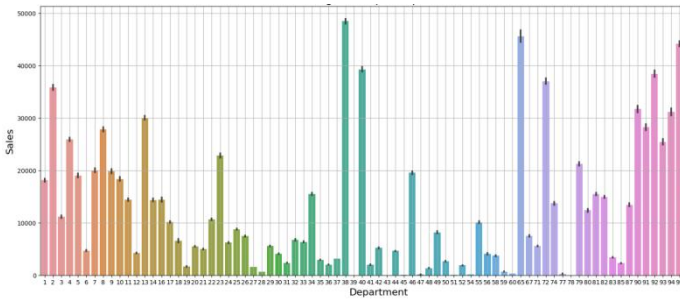

Fig 9: Average Sales per Department

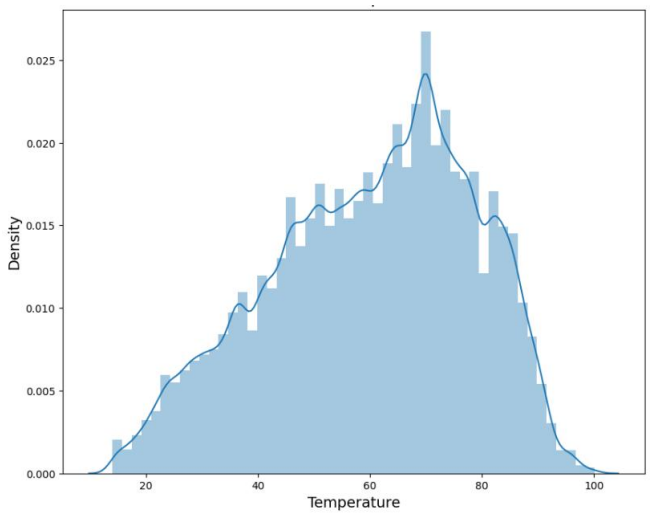

Fig 10: Effect of Temperature

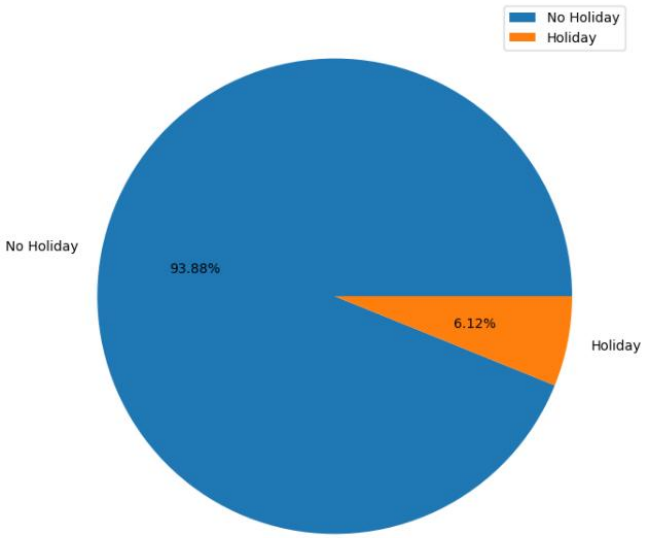

Fig 11: Pie chart distribution

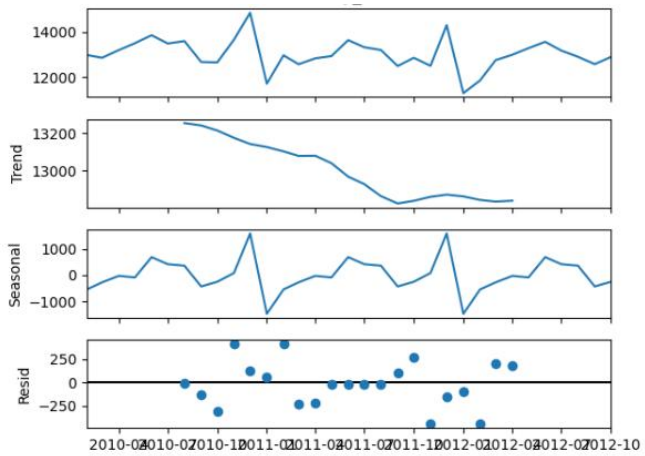

Fig 12: Weekly Sales

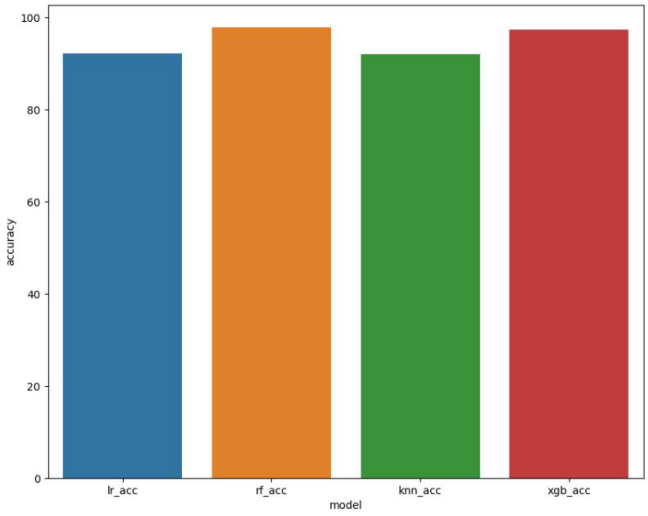

Fig 13: Comparing Models

Model	Accuracy
Linear Regression	92.2
Random Forest Regression	97.8
K Neighbors Regression	91.9
XGboost Model	97.3
CNN	0.32
Support Vector Machine	93.2

V. CONCLUSION

our research focused on sales forecasting for Walmart stores, employing a comparative analysis of Supervised Machine Learning Techniques. This study aimed to accurately predict sales for 45 retail outlets across different locations, considering factors like previous sales data, promotional events, holidays, temperature, fuel price, Consumer Price Index (CPI), and unemployment rate.The methodology included data preparation, correlation analysis, application of machine learning techniques, and model development. Various algorithms such as Multiple Linear Regression, Random Forest Regression, K-NN, SVM, and Extra Tree Regression were utilized, and model performance was assessed using metrics like Mean Absolute Error and Mean Squared Error. Our findings offer valuable insights for business owners, assisting them in selecting appropriate approaches for sales prediction under diverse scenarios. This information can inform strategic decisions related to promotions and marketing, ultimately optimizing business outcomes. As the retail landscape evolves, leveraging advanced data mining and machine learning techniques becomes crucial for maintaining competitiveness. Future work may explore integrating more sophisticated models, larger datasets, and real-time analytics to further enhance the accuracy of sales predictions in the dynamic retail environment.

VI. REFERENCES

[1] Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. San Diego, California: UC San Diego Jacobs School of Engineering.

[2] Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship

[3] Wayne, L. (2014). Winston. Analytics for an Online Retailer: Demand Forecasting and Price Optimization.

[4] Mekala, P., & Srinivasan, B. (2014). Time series dataprediction on shopping mall. Int. J. Res. Comput. Appl. Robot, 2(8), 92-97.

[5] M. Singh, B. Ghutla, R. Lilo Jnr, A. F. S. Mohammed and M. A. Rashid, "Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2017.

[6] D. Silverman, ―Interpreting Qualitative Data: Methods for Analyzing Talk Text and Interaction‖, Text and Interaction, Sage Publications Ltd: Methods for Analyzing Talk, 2006.

[7] A. S. Harsoor and A. Patil, "Forecast of sales of walmart store using Big Data application", International Journal of esearch in Engineering and Technology, vol. 4, pp. 6, June 2015.

[8] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, et al., "Fast and interactive analytics over Hadoop data with Spark", U senix - The Advanced Computing Systems Association, 2012.

[9] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters", Association for Computing Machinery, 2008.

[10] Sohrabpour, V., Oghazi, P., Toorajipour, R., & Nazarpour, A. (2021). Export sales forecasting using artificial intelligence. Technological Forecasting and Social Change, 163, 120480.

[11] Vahid Sohrabpour, Pejvak Oghazi, Reza Toorajipour, Ali Nazarpour. (2021). Export sales forecasting using artificial intelligence, Technological Forecasting and Social Change, Volume 16