**Comparative Analysis of Machine Learning Models for Walmart Sales Forecasting**

|  |  |  |
| --- | --- | --- |
| Farhana Elias  *Department of ECE*  *North South University*  Dhaka, Bangladesh farhana.elias@northsouth.edu | Md. Shihab Reza  *Department of ECE*  *North South University*  Dhaka,Bangladesh shihab.reza@northsouth.edu | Md. Faizullah Farhan  *Department of ECE*  *North South University*  Dhaka,Bangladesh faizullah.farhan@northsouth.edu |

***Abstract*— Sales forecasting plays a crucial role in enabling businesses to make informed decisions by utilizing AI, machine learning, and data analysis to predict future sales based on historical data and market trends. This paper conducts a comprehensive analysis and comparison of various machine-learning techniques for Walmart sales forecasting. The objective is to quantify the relative performance of twelve distinct machine learning models, including Random Forest, Linear Regression, SVR, Decision Tree, KNN, Gradient Boosting, XGBoost, LightGBM, AdaBoost, Ridge, LinearSVR, and MLP. These models were rigorously trained and tested on a consistent dataset, ensuring a fair and reliable comparison of their performance metrics. The findings reveal XGBoost as the most accurate model, achieving an impressive accuracy of 98%. LightGBM and Random Forest closely follow, both attaining 97% accuracy. However, the LinearSVR model, despite its popularity in other domains, demonstrates the least effective performance, achieving only 49% accuracy. This research provides valuable insights for businesses in selecting the most suitable and accurate model for sales forecasting, particularly in the context of the retail industry.**

***Index Terms*—Walmart sales forecasting, Machine learning, Random Forest, Linear Regression, SVR, Decision Tree, KNN, Gradient Boosting, XGBoost, LightGBM, AdaBoost, Ridge, LinearSVR, MLP, performance metrics, accuracy, model selection, retail industry.**

I. INTRODUCTION

Accurately predicting future sales based on historical data is of paramount importance for retailers. Reliable sales forecasts enable businesses to adapt their investment strategies and sales tactics promptly. Regardless of the business size or the number of salespeople involved, sales forecasting plays a critical role in various scenarios. It directly influences planning, budgeting, and other aspects of the business operations. While creating precise sales estimates can be challenging, their significance cannot be overstated. Relying solely on past performance for sales predictions carries inherent risks. In numerous domains, including scientific, industrial, commercial, and economic activities, accurate forecasts are indispensable [1]. The

planning and execution of trade and business activities heavily rely on forecasting future sales demand. Effective forecasting provides valuable insights for inventory management and enables commercial organizations to make informed decisions and implement necessary adjustments [2].

II. LITERATURE REVIEW

To effectively address the sales forecasting problem, it is essential to comprehend the technologies explored in this field, which have been extensively studied by scholars and experts. Among the numerous machine learning models employed for predictions, the Prophet model and machine learning model have gained significant attention [1]. Previous research has shown that the machine learning model proposed in this paper yields smaller prediction errors and provides interpretable results. In comparison, the LightGBM model exhibits higher accuracy when compared to the Prophet model, while the Prophet model demonstrates improved accuracy over the SARIMAX model [1]. In the domain of retail sales forecasting, both the Prophet and LightGBM models have demonstrated strong performance, achieving RMSE values of 0.694 and 0.617, respectively [1].

Gumus et al. [11] utilized XGBoost to analyze factors influencing crude oil prices and forecast future oil prices. Another study [10] leveraged a combination of random forest and XGBoost to develop a data-driven framework for wind turbine fault detection. The approach involved ranking the features using random forest and training ensemble classifiers with XGBoost for each specific fault. Pavlyshenko [2] conducted extensive research on linear models, machine learning algorithms, probabilistic models, logistic regression, and stacking approaches for time series forecasting and regression problems. The study concluded that regression approaches and machine learning algorithms, particularly tree-based algorithms such as Random Forest and Gradient Boosting Machine, offer promising results in sales prediction compared to traditional time series methods. Furthermore, in [3], the experimental results highlight the superior performance of an XGBoost-based model in sales forecasting when compared to traditional machine learning models. Notably, the XGBoost model achieved the lowest RMSSE score of 0.655, outperforming the classic Linear Regression model by 19.5% (with a score of 0.783) and the Ridge Regression model by 13.6% (with a score of 0.774). These findings provide strong evidence that the proposed XGBoost-based model significantly outperforms other regression models in sales forecasting.

Niu [5] proposed an XGBoost-based model that addresses algorithmic challenges by incorporating detailed feature engineering processes, including memory compression, temporal feature extraction, statistical features, and important feature selection. The XGBoost model achieved an RMSSE value of 0.652, outperforming comparison approaches such as Logistic Regression and Ridge algorithm, which yielded RMSSE values of 0.765.

In [6], various regression methods and time series analysis techniques were explored for sales forecasting using Walmart sales data in Microsoft Azure Machine Learning Studio. Boosted Decision Tree Regression outperformed other methods with a coefficient of determination of 0.97.

DSF, a novel Seq2Seq framework, was introduced for E-Commerce product sales forecasting in [8]. DSF utilizes heterogeneous features and incorporates relevant features proactively. The sales residual network architecture captures the impact of substitutable products and leverages complex competing relations. Notably, DSF outperforms the current deep learning-based solution deployed on the Tmall2 platform.

Junhang Chen [7] proposes a Neural Network (NN) model for predicting Walmart's sales. Experimental results demonstrate that the NN model outperforms other machine learning models, with significantly lower RMSE values compared to the Linear Regression and SVM algorithms. SHAP is employed to interpret the NN model and analyze features across different dimensions, contributing to improved prediction accuracy. In [4], a GBDT-based LightGBM model is used for sales forecasting. This model demonstrates a Root Mean Square Error (RMSE) of 2.09, highlighting its superior predictive capability in comparison to both the linear regression model (RMSSE = 3.35) and the SVM model (RMSSE = 2.88). These findings provide strong evidence supporting the effectiveness of our LightGBM-based sales forecasting model in generating accurate predictions.

# III. METHODOLOGY

A. Aim of this study

To compare different regression machine learning approaches, such as Random Forest, Linear Regression, SVR, Decision Tree, KNN, AdaBoost, XGBoost, Ridge, LinearSVR, MLPRegressor, and Gradient Boosting, this study will undertake a comparative analysis of them. The goal is to assess how well these models perform at properly predicting Walmart sales. This study attempts to establish the best method for sales forecasting in the retail sector by determining the model that yields the most accurate results.

Additionally, the study intends to create a web platform that incorporates the top-performing model. The platform's goal is to change Walmart's sales forecasting process and improve the experience of managing retail operations as a whole. The online application's best model will be used to generate precise and trustworthy sales estimates, supporting Walmart with decision-making and improving inventory control.

B. Dataset

The dataset utilized in this study consists of historical sales information from Walmart stores and was obtained from Kaggle, a publicly accessible platform[12]. The dataset used in this study covers the time period from 2010-02-05 to 2012-11-01 and consists of 6,435 rows and 9 columns, with no null values. The dataset includes essential details such as store numbers, sales during specific weeks, holiday flags indicating special holiday weeks, temperature on the day of sale, regional fuel prices, the prevailing consumer price index (CPI), and unemployment rates. Additionally, the dataset incorporates significant holiday events such as the Super Bowl, Labor Day, Thanksgiving, and Christmas, along with their corresponding dates. These variables enable a comprehensive exploration and analysis of the data, facilitating the development and evaluation of accurate machine learning models for sales forecasting at Walmart.

C. EDA & Data Preprocessing

In the Exploratory Data Analysis (EDA) and Preprocessing phase, the dataset underwent careful examination and preparation to uncover insights and ensure its suitability for further analysis. The following steps were carried out:

Initially, an overview of the dataset revealed that it included data from all 45 stores, with each store contributing 142 individual data points (Fig. 1). This ensured a balanced representation of data across all stores, eliminating any concerns regarding data imbalance. Among all 45 stores, the 14th store exhibited the highest weekly sales, indicating its exceptional performance in terms of sales. Conversely, the 44th store displayed the lowest weekly sales, suggesting a need for improvement in that particular store (Fig. 2).

A picture containing screenshot, colorfulness, rectangle, line

Description automatically generated

Fig. 1. Store data Visualization

A picture containing diagram, screenshot, plot, line

Description automatically generated

Fig. 2. Weekly sales on all stores

To ensure an unbiased evaluation of the models, a stratified sampling technique called Stratified ShuffleSplit was employed. This technique partitioned the dataset into stratified training and test sets while maintaining the proportion of holiday and non-holiday data in both sets. By doing so, a balanced representation of the holiday flag was preserved within the dataset. However, it is worth noting that the majority of the data is from non-holiday days, while there is a smaller portion of data from holiday days, indicating an imbalance in the dataset (Fig. 3). Nonetheless, this sampling approach ensured fairness and representativeness in the analysis by accounting for the distribution of holiday and non-holiday instances.

A picture containing text, screenshot, rectangle, display

Description automatically generated

Fig. 3. Holiday & Non-holiday day distribution

Categorical features, specifically 'Store' and 'Holiday\_Flag', were transformed using one-hot encoding. This process, facilitated by the pd.get\_dummies() function, created binary columns representing unique values within these features. The utilization of one-hot encoding enabled the inclusion of categorical variables in subsequent analysis and modeling procedures. By converting categorical variables into numerical representations, we can capture the unique characteristics of each store and the presence or absence of holidays, which are important factors influencing weekly sales.

To address differences in scale and enhance analysis and model performance, numerical features were normalized or standardized. The resulting "Scaled Dataframe" showcased the transformed dataframe, df\_scaled, which consisted of standardized numerical features. Standardization ensured consistency in the scales of variables, thereby improving the accuracy of subsequent analyses. This step is particularly crucial when dealing with numerical features that have different measurement units or magnitudes, allowing for fair comparisons and unbiased modeling.

To tackle any class imbalance in the target variable, Weekly\_Sales, an undersampling technique was applied. This method balanced the distribution of instances across different sales categories (low, medium, and high), leading to a fair representation of all sales categories within the dataset. By resampling the data, we ensure that each sales category is adequately represented in the analysis, preventing the model from being biased towards the majority class and improving the overall performance of the predictive models.

Following the preprocessing steps, correlation analysis was performed to explore the relationships between variables and the target variable, Weekly\_Sales (Fig. 4). The correlations between Weekly\_Sales and other key variables were examined. Notably, the correlation between Weekly\_Sales and itself exhibited a perfect 100% matching correlation, indicating consistent patterns within the dataset.

A picture containing text, screenshot, number, square

Description automatically generated

Fig. 4. Correlation matrix

Further exploration revealed interesting correlations between Weekly\_Sales and other variables. Higher temperatures were associated with lower weekly sales, except when the temperature was neither high nor low, which resulted in higher sales. Moreover, medium to low fuel prices corresponded to increased weekly sales, suggesting a potential influence of fuel costs on consumer behavior. Additionally, lower Consumer Price Index (CPI) values were linked to higher weekly sales, while increasing CPI values were associated with lower sales, indicating a negative correlation. Furthermore, low unemployment rates aligned with high weekly sales, while increasing unemployment led to decreased sales. The correlation with 'Holiday\_Flag' was approximately 2.78%, indicating a minimal positive correlation. The correlation with 'CPI' was approximately -8.22%, implying a slight negative correlation. These correlations provide valuable insights into the relationships between various factors and the target variable, allowing for a better understanding of the drivers behind weekly sales.

Lastly, the 'Date' column was converted to a datetime data type using the pd.to\_datetime() function. This conversion facilitated easier manipulation, analysis, and interpretation of dates in subsequent operations, allowing for more efficient and accurate time-based analysis. By treating the dates as datetime objects, we can easily extract temporal information, identify seasonality patterns, and analyze time-dependent trends, enabling more informed decision-making in sales forecasting and strategy development.

D. Project Workflow

The project workflow, illustrated in Fig 5, outlines the sequential steps followed in the pipeline. It commenced with the collection of data, which was then subjected to a comprehensive preprocessing stage. The preprocessed data was subsequently fed into multiple models for analysis and evaluation of the results.

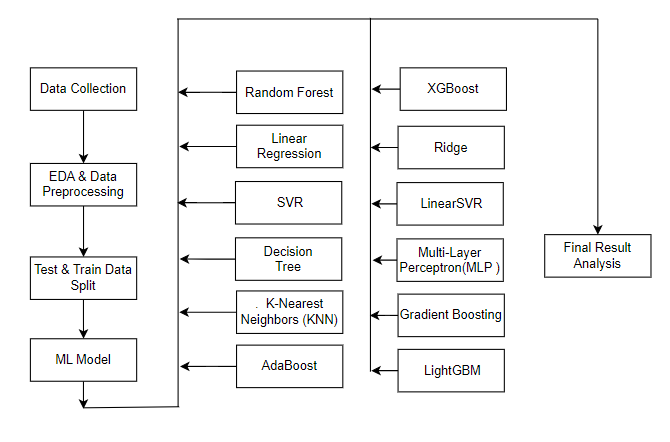


Fig. 5. The workflow diagram of the project.

E. Random Forest

Random Forest is a popular ensemble machine learning model that is highly effective for sales forecasting. Its versatility allows it to handle complex datasets with multiple variables, making it well-suited for analyzing sales data. One of its key advantages is its ability to capture nonlinear relationships between variables, enabling accurate modeling of sales patterns that may not follow linear trends. Additionally, Random Forest provides insights into feature importance, allowing businesses to identify the key drivers that impact sales performance. Overall, Random Forest is a versatile and powerful tool for sales forecasting, providing accurate predictions and valuable insights for businesses. MSE, or mean squared error, is a cost function that is often used to evaluate the performance of a random forest model. The mathematical Equation given below:

MSE = (1/n) \* ∑(y - ŷ)^2 (1)

F. Linear Regression

Linear Regression is a fundamental model for predicting numerical values based on linear relationships between variables. It aims to minimize the difference between actual and predicted values using the best-fit line. However, it may struggle with interactions and outliers, resulting in less accurate predictions. While the mean squared error (MSE) is minimized during training, it may not fully capture the complexities of sales data.

G. SVR

Support Vector Regression (SVR) is a machine learning model used for regression tasks. It is an extension of Support Vector Machines (SVM) and is particularly effective in predicting numerical values. SVR aims to find the best-fit line that maximizes the margin while minimizing the error between the predicted and actual values. It uses kernel functions to handle non-linear relationships between variables, allowing it to capture complex patterns in the data. The cost function used in SVR optimization depends on the chosen kernel and loss function, such as the epsilon-insensitive loss. SVR is widely used in sales forecasting due to its ability to handle non-linear relationships and make accurate predictions.

H. Decision Tree

The decision tree is a popular machine learning algorithm used for classification and regression tasks. It forms a tree-like structure, with leaf nodes representing class labels or predicted values and internal nodes representing attributes or traits. Decision trees mimic human decision-making, offering an intuitive and understandable approach. They are versatile, capable of handling both numerical and categorical data. However, decision trees are prone to overfitting, particularly when the tree becomes complex. Pruning is a technique used to address overfitting. The cost function employed in decision trees varies, with commonly used measures including Gini impurity and entropy. In sales forecasting, decision trees provide valuable insights into the factors that influence sales outcomes in a clear and interpretable manner. The mathematical equation for the Gini impurity and entropy explained below:

Gini Impurity: Gini(p) = 1 - ∑(pi^2) (2)

Entropy: Entropy(p) = - ∑(pi \* log2(pi)) (3)

I. K-Nearest Neighbors (KNN)

A popular machine learning technique for classification and regression problems, such as sales forecasting, is K-Nearest Neighbors (KNN). KNN groups similar data points into clusters based on a predetermined number of centroids. The algorithm calculates the nearest data points to these centroids using distance metrics like Manhattan or Euclidean distance. KNN is a versatile and user-friendly tool that can handle both categorical and numerical data. However, selecting the appropriate K value and distance metric is crucial for accurate forecasts. Forecasting performance can be evaluated using criteria such as accuracy and mean absolute error. KNN provides valuable insights for sales forecasting by considering neighborhood data and local patterns.

J. AdaBoost

AdaBoost is a popular ensemble learning algorithm that combines multiple weak classifiers to create a strong classifier. It works by iteratively training classifiers on different subsets of the data and assigning weights to each sample based on their classification performance. AdaBoost focuses on misclassified samples, continuously adjusting weights to emphasize difficult instances. The final classifier is an ensemble of weak classifiers weighted by their performance. AdaBoost is robust and effective, particularly for tasks with complex decision boundaries. However, it can be sensitive to noisy data and outliers.

K. XGBoost

Extreme Gradient Boosting, often known as XGBoost, is a potent and popular gradient boosting technique. By utilizing parallel tree boosting, it aims to maximize efficiency and performance. To enhance predictive accuracy, XGBoost combines gradient boosting and regularized regression algorithms. It performs feature selection, automatically handles missing values, and offers sophisticated regularization options. Handling large datasets, speed, and scalability are among the strengths of XGBoost. It has been widely adopted in academia and industry, delivering cutting-edge results in numerous machine learning competitions. The objective function in XGBoost, represented by below Equation:

Objective = Loss(y, ŷ) + Regularization (4)

Minimizes the loss between the actual target values (y) and predicted values (ŷ), incorporating a regularization term to control model complexity and prevent overfitting.

L. Ridge

Ridge regression, also known as Tikhonov regularization, is a linear regression technique that introduces a regularization term to the cost function. It aims to prevent overfitting by shrinking the coefficients towards zero. The regularization term, controlled by a hyperparameter called the regularization parameter (λ), penalizes large coefficient values. Ridge regression is particularly useful when dealing with multicollinearity, where predictor variables are highly correlated. It helps stabilize the regression model and improves its generalization performance. The mathematical equation for the cost function in Ridge regression is as follows:

Cost Function = Least Squares Loss + λ \* (sum of squared regression coefficients) (5)

M. LinearSVR

LinearSVR is a powerful linear support vector regression approach used for solving regression problems. It aims to find the optimal linear hyperplane that fits the training data by minimizing margin violations. This technique assumes a linear relationship between the target variable and features, utilizing the support vector machine (SVM) algorithm for optimization. By adjusting the feature weights, LinearSVR achieves the best possible fit. It is particularly useful in tasks where linear correlations are expected between the features and the target variable.

N. Multi-Layer Perceptron(MLP )

A popular neural network architecture for both classification and regression tasks is the MLP, or multi-layer perceptron. It consists of multiple layers of interconnected nodes, or neurons, that process input data and generate predictions as outputs. During training, MLP utilizes backpropagation to adjust the connection weights in addition to forward propagation for prediction computation. It is renowned for its ability to identify intricate patterns and relationships in data. MLP offers flexibility in terms of architecture and activation functions, enabling it to handle various problem domains. It has achieved remarkable success in applications such as time series forecasting, natural language processing, and image recognition.

O. Gradient Boosting

Gradient Boosting is a powerful machine learning technique that combines weak predictive models to create a stronger ensemble model. It iteratively adds new models to the ensemble, focusing on reducing the errors made by the previous models. The iterative process is guided by the gradient of a loss function, allowing for more accurate predictions with each iteration. Gradient Boosting is widely used for regression and classification tasks, achieving state-of-the-art performance in various domains. It excels in handling complex data patterns and providing robust predictions.

The equation for the Gradient Boosting model is represented as:

ŷ = f(x) = β₀ + Σᵢ₌₁ᵏ (hₖ(x)) (6)

For regression tasks, the mean squared error (MSE) is commonly used as the loss function. The corresponding cost function can be written as:

Cost = (1/n) \* Σ(y - ŷ)^2 (7)

P. LightGBM

LightGBM is a highly efficient and scalable gradient boosting framework commonly used for regression tasks. It employs a histogram-based approach and a leaf-wise growth strategy, resulting in faster training and lower memory usage. LightGBM utilizes a cost function, such as mean squared error (MSE), to optimize the model. With its accuracy, speed, and ability to handle large-scale datasets, LightGBM is a popular choice in machine learning competitions and real-world regression applications.

The model equation for LightGBM in regression tasks can be represented as:

ŷ = f(x) = β₀ + Σᵢ₌₁ᵏ (hₖ(x)) (8)

## Web Application

A web application created for sales forecasting included serialization of the best-performing model. This online application uses machine learning to forecast whether forthcoming weekly sales are regarded as risky or safe. Fig. 6. shows the web application's user interface. With the help of the serialized model, which contains the learning weights and parameters of our trained models, applications may make precise predictions.

A screenshot of a computer

Description automatically generated with medium confidence

Fig. 6. Screenshot of the web application.

The web application is designed with a front-end and a back-end component. The front end utilizes HTML and CSS to create an intuitive user interface that facilitates text input and displays the classification results.

On the other hand, the back end is implemented using the Flask framework. Flask handles all the preprocessing steps necessary for the classification task.

Fig. 7 showcases the high-level software architecture, outlining the sequence of processes from input to output. This diagram provides a comprehensive overview of how data flows through the system and the different components involved in the classification process.

A picture containing text, diagram, screenshot, plan

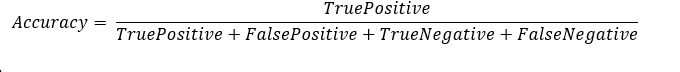
Description automatically generated

Fig. 7. High-level software architecture of the web application.

# IV. RESULTS & DISCUSSION

Various machine learning models were utilized to analyze the results. For the evaluation of machine learning model performance, around 80 percent of the dataset is treated as the training set, and the rest is the testing set. The evaluation metrics and their definitions are given:

Accuracy: Accuracy is the percentage of true positives among all the data points.



Precision: Precision can be defined as the percentage of data points correctly predicted among the positives.

A black text on a white background

Description automatically generated with low confidence

Recall: Recall calculated the percentage of true positives that were successfully identified. It is calculated as per the formula below:

A black text on a white background

Description automatically generated with low confidence

F1 Score: F1 score combines both recall and precision. The recall and precision are used to calculate the harmonic mean.

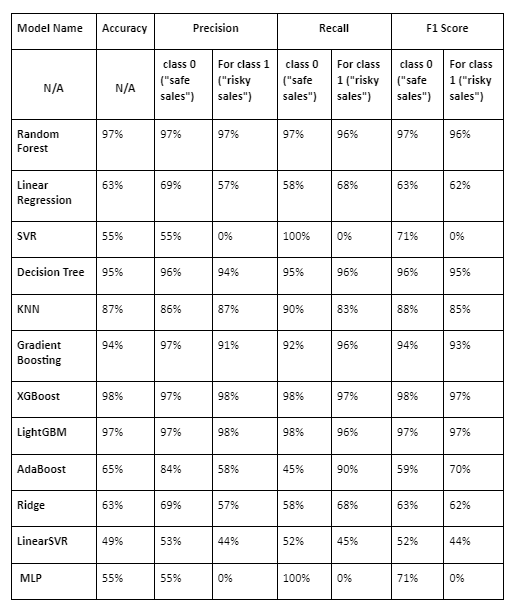
A picture containing font, text, white, diagram

Description automatically generated

Table I shows the comparison of evaluation parameters such as the overall accuracy, precision, recall,f1 score for performance on the Safe & Risky Sale.

TABLE I

COMPARISON OF EVALUATION PARAMETERS



However, the machine learning model for sales forecasting effectively categorizes future sales as either safe or risky. In this context, safe sales are defined as projections expected to yield good sales, while risky sales are those anticipated to result in lower sales. The model utilizes a data-driven approach to determine a threshold point, referred to as the midpoint, based on historical sales data. By analyzing the sales data, the model identifies the average sales value from the overall sales data. Consequently, any projected sale value that exceeds the average sales indicates a higher likelihood of profitability. Conversely, sales with a projected value below the average sales value are classified as risky, implying lower sales. The developed machine learning model effectively distinguishes future sales as safe or risky, providing valuable insights into potential sales

forecasts. By incorporating a data-driven average sales value and leveraging historical sales patterns, the model offers a practical and customized approach to sales categorization, ultimately enhancing the accuracy and applicability of sales forecasting.

For better comparative visualization Fig 8 shows the model accuracy.

A picture containing text, screenshot, font, number

Description automatically generated

Fig.8. Accuracy for applied models

In terms of accuracy, the XGBoost model demonstrates the highest performance in sales forecasting, achieving an accuracy of 98%. It exhibits a precision of 97% for class 0 and 98% for class 1. To facilitate better comparative analysis, Figures 9, 10, 11, 12, 13, and 14 depict the evaluation metrics of different machine learning models, showcasing their performance in sales forecasting.

A picture containing text, screenshot, font, number

Description automatically generated

Fig. 9. Precision For class 0 ("safe sales")

A picture containing text, screenshot, font, number

Description automatically generated

Fig. 10. Precision For class 1 ("risky sales")

A picture containing text, screenshot, font, number

Description automatically generated

Fig. 11. Recall For class 0 ("safe sales")

A picture containing text, screenshot, font, number

Description automatically generated

Fig. 12. Recall For class 1 ("risky sales")

A picture containing text, screenshot, font, number

Description automatically generated

Fig. 13. F1 - score for class 0 ("safe sales")

Fig. 14. F1 - score for class 1 ("risky sales")

Regarding precision (class 1), the XGBoost, Random Forest, Decision Tree, and AdaBoost models exhibit the highest precision scores. These models excel at accurately predicting positive sales (class 1) out of all the predicted positive sales. Concerning recall (class 1), the XGBoost, Random Forest, Decision Tree, and Gradient Boosting models demonstrate the highest recall scores. These models effectively capture all actual positive sales (class 1) out of all the positive sales. Lower values of Mean Squared Error (MSE) and Mean Absolute Error (MAE) indicate better performance. Based on MSE and MAE alone, the LightGBM model achieves the lowest values, indicating superior performance in minimizing prediction errors. Figures 15, 16, 17, and 18 visually depict the MSE and MAE values. The XGBoost model attains the highest values in other performance metrics, such as recall and F1-Score, for both class 0 and class 1. Based on the provided metrics, the XGBoost model outperforms others. It exhibits high accuracy, low MSE and MAE, and commendable precision and recall scores for both classes.

A picture containing text, diagram, line, plot

Description automatically generated

Figure 15: Comparative Mean Error Regression Model (MSE)

A picture containing text, screenshot, font, number

Description automatically generatedA picture containing text, screenshot, software, display

Description automatically generated

Figure 16: Bar Plot of comparative Mean Squared Error (MSE)

A picture containing text, line, diagram, parallel

Description automatically generated

Figure 17: Comparative Mean Absolute Error (MAE) of Regression Model

A picture containing text, screenshot, number, software

Description automatically generated

Figure 18: Bar Plot of comparative Mean Absolute Error (MAE)

Figures 19, 20, and 21 present the confusion matrix for the Random Forest, LightGBM, and XGBoost models among all the applied models. The confusion matrix for XGBoost exhibits a significant number of true positives and true negatives, with only a few occurrences of false positives and false negatives.

A picture containing text, screenshot, diagram, design

Description automatically generated

Figure 19: Confusion Matrix for Random Forest.

A blue squares with white text

Description automatically generated with low confidence

A picture containing text, screenshot, diagram, number

Description automatically generatedFigure 20: Confusion Matrix for LightGBM.

Figure 21: Confusion Matrix for XGBoost.

This indicates that XGBoost accurately classifies sales as either safe or risky. Based on these metrics and performance measures, XGBoost consistently exhibits superior accuracy, precision, recall, F1-score, and error metrics in comparison to the other models. Therefore, it can be regarded as the best model among the listed ones for sales forecasting in the Walmart dataset.

# V. CONCLUSION & FUTURE WORK

This study aimed to achieve accurate sales forecasting using Walmart sales forecasting data by implementing various machine learning models. The evaluation of model performance revealed significant variations among the tested models. The XGBoost model emerged as the top performer, demonstrating superior accuracy, precision, recall, and F1 score compared to the other models. Conversely, the LinearSVR model exhibited the lowest performance, highlighting the need for further enhancements.These findings underscore the importance of selecting the appropriate machine learning algorithm for sales forecasting tasks. The strong performance of the XGBoost model suggests its potential for real-world applications in retail stores like Walmart. Future efforts should focus on improving the performance of the LinearSVR model, potentially increasing its efficacy in similar tasks. Moving forward, several areas warrant exploration. Firstly, employing feature engineering techniques can help extract more informative features from Walmart's previous historical data, potentially enhancing model performance. Furthermore, integrating advanced deep learning techniques, such as Convolutional Neural Networks (CNNs) or Artificial Neural Networks (ANNs), may enable capturing complex patterns and temporal dependencies within the data. Collecting and preprocessing more diverse and comprehensive datasets to evaluate model performance in real-world scenarios is also crucial. Overall, this study provides a foundational framework for future research in sales forecasting using machine learning. By delving into advanced algorithms, enhancing feature engineering, and leveraging more extensive and diverse datasets, future work can contribute to developing more accurate and reliable models for sales forecasting. Such advancements will benefit various applications, including retail stores and companies in general.

REFERENCES

1. H. Jiang, "Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast," in Proceedings of the IEEE International Conference on [2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)], 2021.
2. B. Pavlyshenko, “Machine-Learning Models for Sales Time Series Forecasting,” *Data*, vol. 4, no. 1, p. 15, Jan. 2019, doi: 10.3390/data4010015.
3. D. Xie, "Machine Learning Model for Sales Forecasting by Using XGBoost," in Proceedings of the IEEE International Conference on [2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)], 2021.
4. Z. Qiao, "Walmart Sale Forecasting Model Based on LightGBM," in Proceedings of the IEEE International Conference on [2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)], 2020.
5. Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," in Proceedings of the IEEE International Conference on [2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)], 2020.
6. C. Catal, K. ECE, B. Arslan, and A. Akbulut, "Benchmarking of Regression Algorithms and Time Series Analysis Techniques for Sales Forecasting," *Balkan Journal of Electrical and Computer Engineering*, vol. 7, pp. 20-26, Year.
7. J. Chen, "Sales Forecasting Using Deep Neural Network and SHAP techniques," in Proceedings of the IEEE International Conference on [2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)], 2021.
8. Y. Qi, "A Deep Neural Framework for Sales Forecasting in E-Commerce," in Proceedings of the ACM [CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge Management], 2019.
9. H. Weytjens, E. Lohmann, and M. Kleinsteuber, "Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet," *Electronic Commerce Research*, 2019.
10. E. Zunic, K. Korjenic, K. Hodzic, and D. Conko, "Application of Facebook's Prophet algorithm for successful sales forecasting based on real-world data," *International Journal of Computer Science and Information Technology*, vol. 12, no. 2, pp. 23-36, 2020.
11. Gumus M, Kiran M S. Crude oil price forecasting using XGBoost[C]//2017 International Conference on Computer Science and Engineering (UBMK). IEEE, 2017: 1100-1103
12. Y. Hamed, "Walmart Dataset," Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/datasets/yasserh/walmart-dataset>.
13. You, L., Kou J., & Wang S. (2019). "Online Retail Sales Prediction with Integrated Framework of K-mean and Neural Network." In *Proceedings of the 2019 10th International Conference on E-business, Management and Economics (ICEME 2019)*. Association for Computing Machinery, New York, NY, USA, 115-118.