

Enhancing sales volume using machine learning algorithms

Amal Elsayed Aboutabl¹, Ola Mahmoud Moawad², Ahmed Mohamed Abd-Elwahab³

¹Department of Computer Science, Faculty of Computer and Artificial Intelligence, Helwan University, Cairo, Egypt

²Department of Business Information Systems, AlGazzera Academy, Mokattam, Cairo, Egypt

³Department of Information Systems, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

Article Info

Article history:

Received Dec 4, 2024

Revised Oct 23, 2025

Accepted Nov 16, 2025

Keywords:

Bayesian ridge

Machine learning algorithms

Random forest regression

Sales forecasting models

Sales volume prediction

Stacking

ABSTRACT

In today's highly competitive business landscape, companies face a significant challenge in making accurate decisions based on vast amounts of historical data. Reliance on human data analysis often leads to biases and errors, hindering the ability to extract effective insights for sales forecasting. To address this challenge, this research presents an advanced model that integrates 14 machine learning (ML) regression algorithms, including XGBRegressor and LGBMRegressor, to provide accurate sales predictions using a comprehensive global store dataset. The results demonstrate that XGBRegressor and LGBMRegressor achieved the highest test accuracy (92%) and the lowest error rates, proving their ability to handle complex prediction tasks efficiently. This high accuracy in sales forecasting enables companies to make more effective strategic decisions, such as optimizing inventory management, allocating resources optimally, and exploring new growth opportunities. Consequently, the use of these advanced algorithms directly contributes to increasing sales volume and achieving a sustainable competitive advantage.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ola Mahmoud Moawad Mahmoud

Department of Business Information Systems, AlGazzera Academy

Mokattam, Cairo, Egypt

Email: ola.mahmoud2043@commerce.helwan.edu.eg

1. INTRODUCTION

Under current developments, supermarkets are facing difficulty in accurately forecasting their annual sales due to a lack of knowledge and resources. Although traditional statistical methods are important for building predictive models, machine learning (ML), a branch of Artificial Intelligence, has now become the new standard for sales forecasting. In the age of big data and immense computing power, ML has proven its ability to process vast amounts of data using complex algorithms to discover hidden patterns and make accurate decisions. This capability makes it an invaluable tool for sales forecasting by analyzing the interaction between historical data, market trends, and consumer behavior. However, to ensure maximum benefit from ML, it is essential to understand the nature of the data and select the appropriate models to apply to real-world problems [1], [2]. This paper aims to optimize sales prediction for global stores. this paper aims to identify the most effective ML algorithms and data pre-processing techniques within the proposed framework. Specifically, the study will seek to determine which algorithms exhibit the highest efficacy in predicting sales and what are the most optimal methods for refining and preparing sales data to improve model accuracy.

Due to increasing competition, shopping malls and large supermarkets are now meticulously recording sales data and various related factors. This data is of great importance for forecasting future demand and managing inventory efficiently. For this reason, the reliance on ML in data science is growing,

as it allows for the fast and highly accurate processing of this data, which in turn provides valuable business insights. Sales forecasting is essential for the success of companies, as it supports vital functions like strategic planning, budgeting, and risk management by providing reliable estimates of future revenues [3]-[5]. Because of its ability to provide accurate sales forecasts, ML significantly enhances corporate profitability, leading to a profit increase of more than 10%.

These forecasts support effective strategic decision-making, improve resource allocation, and help guide marketing campaigns. Prediction accuracy is fundamental in ML, as it is used to evaluate data precision and enhance analysis results across multiple fields. Ultimately, sales forecasting is of paramount importance to companies because it provides them with future insights, boosts investor confidence and market credibility, and also helps improve production cycles to meet demand and reduce waste [6].

ML can significantly boost business profitability by providing accurate sales predictions, which supports strategic decision-making and resource optimization. It has the potential to increase profitability by over 10% and is vital for targeted marketing. Accurate demand forecasting streamlines operations and drives higher profits [7]. Prediction accuracy is a key aspect of ML, used to evaluate data precision and enhance analysis in fields like financial analysis and sales forecasting. Accurate predictions are essential for informed decision-making across various sectors [8]. Ultimately, sales forecasting is vital for businesses, as it offers insights into future trends, boosts investor confidence, and helps production firms optimize cycles to meet demand and reduce waste [9].

2. RELATED WORK

In this section, we will explore the various ML algorithms applied in this context and evaluate their performance based on prediction accuracy and other relevant metrics. Amrutkar and Mahadik [10], aimed to predict sales for an e-fashion store using three years of historical data (2015-2017). The final dataset was refined by removing non-usable and redundant entries, resulting in a smaller, cleaner dataset. The study evaluated various prediction models and found the Gradient Boost algorithm to be the most accurate for future sales forecasting, achieving 98% overall accuracy. It was followed by the Decision Tree algorithm (71%) and the Generalized Linear Model (64%). Nana *et al.* [11] focused on sales forecasting using Big Mart data, with Gradient Boosted Tree achieving the highest accuracy of 95.84%. While the results are promising, further comparison with other algorithms and deeper error analysis could enhance the study's insights.

Tony *et al.* [12] based on the data provided, according to the available data, the random forest (RF) model appears to be the most effective, achieving the highest performance at 89. The Decision Tree model follows with a score of 78, while the performance of the Linear Regression model was the lowest, with a score of 70. These results indicate that the RF model is the best for this predictive task, based on the performance metric used. Ofoegbu [13] evaluated various algorithms for classification tasks, with RF achieving the highest accuracy of 98%. While the results are promising, comparing its performance with other advanced algorithms under different conditions would provide a more comprehensive understanding of its effectiveness and robustness.

This paper differs from previous studies by providing a more comprehensive and systematic evaluation of ML algorithms used in sales forecasting. While studies [10]-[13] offer valuable insights, they are limited by their narrow scope. Specifically, the studies [10], [12] only compare a small number of models (three), making it difficult to draw broad conclusions about which algorithms are truly the most effective. Similarly, the study [11] acknowledges this limitation, noting the need for "further comparison with other algorithms". This paper directly addresses this gap by evaluating 14 different algorithms, offering a much more robust and reliable comparison of model performance across a wide range of techniques. Another key difference is your study's specific identification of the top-performing algorithms. While studies [10], [11] correctly highlight the power of Gradient Boost-based models, this research goes a step further. By testing a broader range of models, you were able to pinpoint XGBRegressor and LGBMRegressor as the most accurate, achieving an impressive 92% accuracy with the lowest error rates. This finding is significant because these are modern, highly optimized versions of gradient boosting, and their superior performance over other models, including those mentioned in studies [12], [13] (which focus on RF), provides a clear, data-driven recommendation for practitioners.

This paper also, fills a crucial gap by focusing specifically on the regression task of sales forecasting. While study [13] shows a high accuracy of 98% for RF, its application was for a classification task, which is a fundamentally different problem. By concentrating on sales forecasting as a regression problem, the results of this research are directly relevant and applicable to the business problem of predicting numerical values. This specificity makes your research highly valuable for companies looking to apply these models to their own sales data. The ultimate contribution of this paper lies in its ability to provide clear, actionable insights for businesses. The high accuracy and low error rates of recommended models (XGBRegressor and LGBMRegressor) enable companies to make better strategic decisions. In contrast to the

limited results of other studies, this research offers a powerful tool for improving inventory management, optimizing resource allocation, and ultimately increasing sales.

Furthermore, this paper offers a more comprehensive and robust approach compared to the referenced research ([10]-[13]) by utilizing 14 algorithms, significantly more than the one or two used previously. While those studies identified either Gradient Boost or RF as the top performers with high accuracy, this paper will provide a more detailed and reliable comparison. A key advantage of my work is the use of a larger dataset of 51,000 records, in contrast to the smaller datasets in previous studies. I also meticulously cleaned the data by removing duplicates, with the original data having only a small percentage of issues. This rigorous methodology, combined with a larger dataset, is expected to yield a more powerful and dependable model than those in the literature shown all studies in Table 1.

Table 1. A comparative analysis of ML algorithms for sales prediction

Refrences	Algorithms (Accuracy)	Strengt and weakness
This Paper	14 Algorithms	In this paper distinguished by its integration of 14 ML algorithms, providing a comprehensive evaluation. The key highlight is that the XGBRegressor and LGBMRegressor algorithms achieved the highest prediction accuracy of 92% and the lowest error rates. This highly accurate sales forecasting model enables companies to make better strategic decisions, such as optimizing inventory management and resource allocation, which directly contributes to increased sales and a sustainable competitive advantage.
Amrutkar and Mahadik [10]	3 Algorithms Gradient Boost Algorithm 98%. Decision Tree Algorithms 71%. Generalized Linear Model 64%.	The study demonstrates strength by using real data from an e-fashion store, which enhances its practical importance. However, its weakness lies in not specifying the actual size of the dataset after cleansing, making it difficult to assess the reliability and statistical significance of the results.
Nana <i>et al.</i> [11]	3 Algorithms GLM 56.03. DT 58.46. GBT 95.84.	While the system effectively identifies the Gradient Boost Algorithm as the best predictive model through a strong analytical approach, Furthermore, the lack of information on data sources and evaluation criteria makes it difficult to fully assess the model's reliability.
Tony <i>et al.</i> [12]	3 Algorithms Random Forest89 . Decision Tree78 . Linear Regression 70.	The study demonstrates an understanding of the importance of data in ML, as it pointed out that increasing data volume enhances the predictive power of models. However, its weakness lies in not specifying the actual size of the data after cleansing and the absence of quantitative values for performance metrics, which reduces the credibility of the results and makes them difficult to verify.
Ofoegbu [13]	4 Algorithms DecisionTreeClassifier%88 DeepLearningANN 67% RandomForestClassifier 98% Naïve Bayes 55%	The research's strength lies in its treatment of a vital business problem using effective ML techniques to support decision-making and increase profits. However, its main weakness is evident in data processing challenges, as some records were discarded with the acknowledgement that the fields and attributes used were not suitable for deeper analysis.

In Table 1 showed the proposed paper offers a more comprehensive and robust methodology for sales prediction, as it compares 14 ML algorithms on a significantly larger dataset of 51,000 records. This rigorous approach, which includes data cleaning and de-duplication, ensures a more reliable selection of the optimal model. As a result, the proposed model is expected to be more powerful and dependable compared to previous studies that used a limited number of algorithms and smaller datasets.

3. METHOD

This paper outlines a comprehensive methodological framework for sales prediction using ML, built upon a structured, five-stage process. The first stage, Data Collection, involves gathering a large dataset (51,000 records) from a global store to ensure a representative sample of sales patterns. Following this, Data Preprocessing is performed to clean and transform the raw data. This step includes managing missing values, handling outliers, encoding categorical features, and scaling numerical data to prepare it for model training. The third stage is feature Selection, where the most impactful variables that influence sales are identified and chosen. This is a critical step to ensure that the models are trained on the most relevant information, which improves their efficiency and helps to prevent overfitting. After feature selection, the study moves to Model Training and Selection, applying fourteen different ML algorithms. This extensive list includes powerful regressors like XGBRegressor, LGBMRegressor, and CatBoostRegressor.

The final stage, Prediction and Evaluation, assesses the performance of each trained model using precise metrics such as prediction accuracy and error rates. This systematic approach not only provides a

clear comparison of each model's effectiveness but also aims to offer an accurate and generalizable analysis that can help companies make informed decisions to optimize their operations and increase profits. Shown proposed framework in Figure 1.

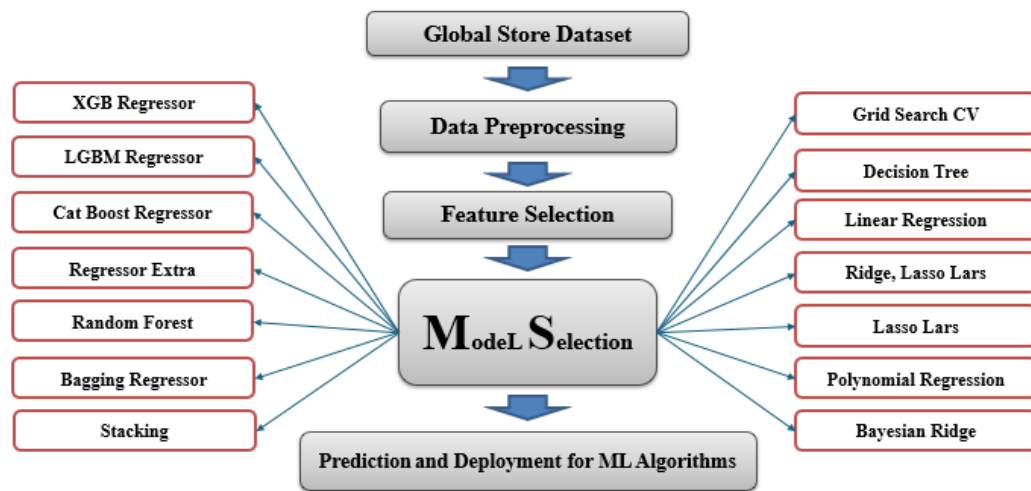


Figure 1. The proposed framework prediction sales for global stores

3.1. Global store dataset

The initial phase, which consists of multiple layers that are prerequisites for training and testing the model, data load, and exploratory analysis: the dataset has been uploaded from the Kaggle global stores. The Global Store Dataset is a valuable tool for sales forecasting and business analysis, as it contains 24 detailed features on sales transactions. It includes crucial information like order details, product categories, and customer data, which are all essential for building accurate predictive models. By analyzing these features, companies can make informed decisions to optimize operations, increase profits, and improve customer satisfaction. The Global Store Dataset is one of the most widely used datasets for analyzing sales, profits, and customer behavior in the retail sector. This dataset contains detailed information on sales transactions, including order details such as order ID, order and shipping dates, as well as product information like product ID, name, category, and subcategory. It also provides comprehensive customer data, including customer ID, name, and segmentation into different categories such as consumers, businesses, or home offices.

3.2. Data preprocessing

Data preprocessing is a crucial step in transforming raw data into a clean, usable format for ML algorithms. This process involves cleaning the data by identifying and handling missing values and errors, such as removing outliers, imputing missing data, or correcting formatting issues. In this context, it was found that the dataset contained 41,296 missing values in the postal code column, and this error was addressed using an imputation technique.

3.3. Feature selection

Feature selection is a key process in ML that involves choosing a subset of the most relevant features from a dataset [14]. This technique is important because it improves model performance and accuracy by reducing noise, while also decreasing computational complexity and speeding up training. In this study, the authors used a correlation matrix to identify the most important features for the Global Store dataset. The features with the highest importance scores were found to be: 'Ship Mode', 'Quantity', 'Delivery Days', 'ProductID', 'OrderPriority', 'Profit', and 'ShippingCost'.] shown in Figure 2.

3.4. Model selection

This section provides an overview of various algorithms and methods used for predictive modeling, including their core functionalities and key characteristics.

3.4.1. XGB regressor

A powerful gradient boosting algorithm known for creating highly accurate predictive models [15]. XGB Regressor is particularly well-suited for regression tasks because it is able to handle both linear and

nonlinear relationships between input features and the target variable. In Table 2, shown that the XGBRegressor model demonstrated strong performance with a training score of 0.963, indicating a good fit to the training data. The corresponding training error was 4074.603. When evaluated on the unseen test data, the model's performance remained high, with a test score of 0.924 and a test error of 8507.671. The cross-validation score of 0.917 further confirms the model's reliability and its ability to generalize well to new data without significant overfitting.

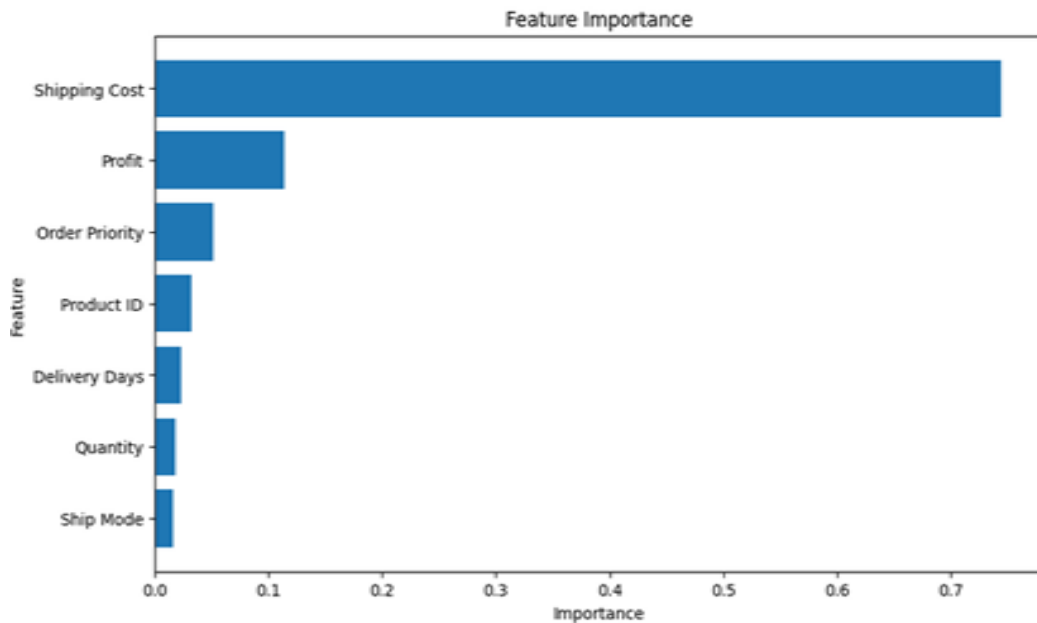


Figure 2. Feature selection

Table 2. XGB regressor

Model	Train score	Train error	Test score	Test error	cross_validation_score
XGB Regressor	0.963	4074.603	0.924	8507.671	0.917

3.4.2. LGBM regressor

A fast and efficient gradient boosting algorithm that utilizes a leaf-wise growth strategy, making it well-suited for large datasets [16]. In Table 3, the provided paragraph describes a light gradient boosting machine (LGBM) Regressor model with a high train score (0.972) and low train error (3081.155), indicating it fits the training data well. However, a noticeable drop in performance on new, unseen data is evident from the lower test score (0.924) and higher test error (8504.054). This suggests a degree of overfitting, where the model has memorized the training data rather than learning general patterns. The cross-validation score (0.917) supports the test results, providing a reliable measure of the model's performance on new data. Despite the overfitting, the model shows good overall generalization and is still considered a suitable choice for the task.

Table 3. LGBM regressor

Model	Train score	Train error	Test score	Test error	cross_validation_score
LGBM Regressor	0.972	3081.155	0.924	8504.054	0.917

3.4.3. CatBoost regressor

A ML algorithm distinguished by its ability to effectively handle categorical features [17]. In Table 4, the results show that the CatBoostRegressor model achieved excellent performance on the training dataset, with a train score of 1.0, which indicates that the model learned the data perfectly. However, the train error was relatively high at 12.169, which may suggest the presence of overfitting. On the test dataset, the performance

declined, as the test score dropped to 0.92, and the test error increased significantly to 9014.459. This discrepancy between the training and testing performance confirms the existence of overfitting.

Table 4. CatBoost regressor

Model	Train score	Train error	Test score	Test error	cross_validation_score
CatBoost Regressor	1.0	12.169	0.92	9014.459	0.908

3.4.4. Extra trees regressor

An algorithm that improves accuracy by creating a large number of Random decision trees and combining their predictions [18]. These results show that the ExtraTreesRegressor model achieved perfect performance on the training data, with a train score of 1.0 and a train error of 0.0. This indicates that the model has fully memorized all the training data, confirming the presence of overfitting. Despite this, the performance on the test data shows a noticeable drop, with the test score at 0.916. The test error also increased significantly to 9386.862, reinforcing the idea of overfitting. Nevertheless, the cross-validation score of 0.907 provides a more realistic evaluation of the model's performance on new, unseen data. show that in Table 5

Table 5. Extra trees regressor

Model	Train score	Train error	Test score	Test error	cross_validation_score
Extra Trees Regressor	1.0	0.0	0.916	9386.862	0.907

3.4.5. Random forest

An algorithm that uses an ensemble of multiple decision trees to make accurate predictions and prevent overfitting [19]. In this Table 6 the results show that the RF model achieved strong performance on the training dataset, with a train score of 0.986 and a train error of 1517.166. While not perfectly ideal, this performance is very good and indicates the model's ability to learn from the data. However, there was a clear discrepancy in performance when tested on new data. The test score dropped to 0.911, and the test error increased significantly to 10054.34.

Table 6. Random forest

Model	Train score	Train error	Test score	Test error	cross_validation_score
Random forest	0.986	1517.166	0.911	10054.34	0.903

3.4.6. Bagging regresso

A ML ensemble technique that enhances accuracy by combining predictions from multiple regression models, each trained on a different subset of the data [20] show that in Table 7.

Table 7. Bagging regressor

Model	Train score	Train error	Test score	Test error	cross_validation_score
Bagging Regressor	0.986	1500.575	0.91	10078.205	0.902

3.4.7. Stacking

An ensemble learning technique that combines multiple ML models by training base models on the same data and using a meta-model to merge their predictions for superior accuracy [21] show that in Table 8.

Table 8. Stacking

Model	Train score	Train error	Test score	Test error	cross_validation_score
Stacking	0.976	2629.98	0.903	10932.631	0.896

3.4.8. GridSearchCV

A hyperparameter tuning technique that systematically explores all possible combinations of hyperparameters to select the best set based on cross-validation performance [22]. These results show that the

GridSearchCV model, after searching for the optimal parameters, achieved balanced performance. The train score was 0.91 and the train error was 9833.712, which indicates a good ability for the model to learn from the data. On the test dataset, the performance was very close to the training performance, with the test score reaching 0.896 and the test error increasing to 11722.636. This relative consistency between training and testing performance suggests that the model did not suffer significantly from overfitting, which is a positive outcome show that in Table 9.

Table 9. GridSearchCV

Model	Train score	Train error	Test score	Test error	cross_validation_score
GridSearchCV	0.91	9833.712	0.896	11722.636	0.884

3.4.9. Decision tree algorithms

Supervised learning methods that create a tree-like structure to either classify or predict data [23]. These results show that the DecisionTree model achieved good performance on the training data, with a train score of 0.968 and a train error of 3452.608. This indicates that the model is capable of learning patterns from the data it was trained on. However, there is a significant drop in performance when tested on new data. The test score decreased to 0.853, and the test error increased substantially to 16543.563. This clear discrepancy between the training and testing performance is a strong indicator of overfitting, as the model appears to have memorized the specifics of the training data rather than learning general, generalizable patterns show that in Table 10.

Table 10. Decision tree

Model	Train score	Train error	Test score	Test error	cross_validation_score
Decision Tree	0.968	3452.608	0.853	16543.563	0.84

3.4.10. Linear regression

A statistical method used to model the linear relationship between a dependent variable and one or more independent variables [24]. In Table 11, the results show that the LinearRegression model achieved balanced and reliable performance. The train score was 0.751 and the train error was 27318.961. This indicates that the model learned well from the data. On the test dataset, the performance was very similar to the training performance, with the test score reaching 0.76 and the test error at 27002.632. This close alignment between training and testing performance is a strong indicator that the model did not suffer from overfitting; rather, it has a good ability to generalize.

Table 11. Linear regression

Model	Train Score	Train Error	Test Score	Test Error	cross_validation_score
Linear Regression	0.751	27318.961	0.76	27002.632	0.754

3.4.11. Ridge regression

A linear regression technique that adds a penalty term to the loss function to prevent overfitting. In Table 12, these results show that the Ridge Regression model achieved balanced and reliable performance. The train score was 0.751 and the train error was 27318.961, which indicates that the model learned well from the data. On the test dataset, the performance was very similar to the training performance, with the test score reaching 0.76 and the test error at 27002.638. This close alignment between training and testing performance is a strong indicator that the model did not suffer from overfitting; rather, it has a good ability to generalize.

Table 12. Ridge regression

Model	Train score	Train error	Test score	Test error	cross_validation_score
Ridge Regression	0.751	27318.961	0.76	27002.638	0.754

3.4.12. LassoLars

A variant of Lasso regression that uses the least angle regression (LARS) algorithm to efficiently perform variable selection and regularization, thereby improving prediction accuracy and interpretability [25]. In Table 13, these results show that the LassoLars model achieved balanced and reliable performance. The train score was 0.751 and the train error was 27258.665. This indicates that the model learned well from the data. On the test dataset, the performance was very similar to the training performance, with the test score reaching 0.76 and the test error at 26923.425. This close alignment between training and testing performance is a strong indicator that the model did not suffer from overfitting; rather, it has a good ability to generalize.

Table 13. Lasso lars

Model	Train score	Train error	Test score	Test error	cross_validation_score
Lasso Lars	0.751	27258.665	0.76	26923.425	0.754

3.4.13. Polynomial regression

A regression technique that models non-linear relationships between variables using a polynomial equation [26]. In Table 14, these results show that the PolynomialRegression model achieved balanced and reliable performance. The train score was 0.751 and the train error was 27258.665, which indicates that the model learned well from the data. On the test dataset, the performance was very similar to the training performance, with the test score reaching 0.76 and the test error at 26923.425. This close alignment between training and testing performance is a strong indicator that the model did not suffer from significant overfitting; rather, it has a good ability to generalize.

Table 14. Ploynomial regression

Model	Train score	Train error	Test score	Test error	cross_validation_score
Ploynomial Regression	0.751	27258.665	0.76	26923.425	0.754

3.4.14. Bayesian Ridge Regression

A linear regression technique that applies Bayesian principles to regularize the model, estimating parameters by inferring their posterior distribution [27]. In Table 15, the results show that the BayesianRidge model achieved balanced and reliable performance. The train score was 0.751 and the train error was 27258.767, which indicates that the model learned well from the data. On the test dataset, the performance was very similar to the training performance, with the test score reaching 0.76 and the test error at 26923.966. This close alignment between training and testing performance is a strong indicator that the model did not suffer from overfitting; rather, it has a good ability to generalize.

Table 15. BayesianRidge

Model	Train score	Train error	Test score	Test error	cross_validation_score
BayesianRidge	0.751	27258.767	0.76	26923.966	0.754

3.5. Prediction and deployment for ML algorithms

The evaluation metrics are explained as follows:

3.5.1. Accuracy

In this study, we will investigate various ML algorithms and their effectiveness in predicting sales. The focus of the study will be on comparing the performance of different algorithms, as well as on identifying the factors that contribute to the accuracy of sales predictions. To this end, we will use a variety of evaluation metrics and statistical techniques to assess the performance of the different models. Through this study, we aim to provide valuable insights for practitioners and studies interested in using ML for sales prediction Accuracy (R2_score) formula:

$$\text{Accuracy} = (\text{Number of Correct Predictions}) / (\text{Total Number of Predictions}) \quad (1)$$

3.5.2. Training and testing the models

The model training phase is crucial in ML, as it involves feeding the model with data to discover patterns and enhance its predictive ability. In this study, 80% of the data was used to train the model, with the remaining 20% allocated for validation. This split ensures the model learns effectively from the majority of the data and that its performance is tested on an independent set.

3.5.3. Model performance analysis

The analysis provided of the MSE graph highlights the performance of various ML models. The tree-based models on the left, particularly LGBMRegressor and XGBRegressor, stand out as the best performers, exhibiting the lowest testing MSE despite some degree of overfitting (indicated by the gap between low training MSE and higher testing MSE). In contrast, the linear models on the right, such as LinearRegression and Ridge, demonstrate poor performance with very high and similar MSE values for both training and testing data, which is a classic sign of underfitting. The decision tree model is a clear example of severe overfitting, showing a minimal training error but a very high testing error, which means it fails to generalize new data. In conclusion, while many tree-based models show some overfitting, LGBMRegressor and XGBRegressor provide the optimal balance, offering the lowest testing error and thus the highest predictive accuracy shown that in Figure 3.

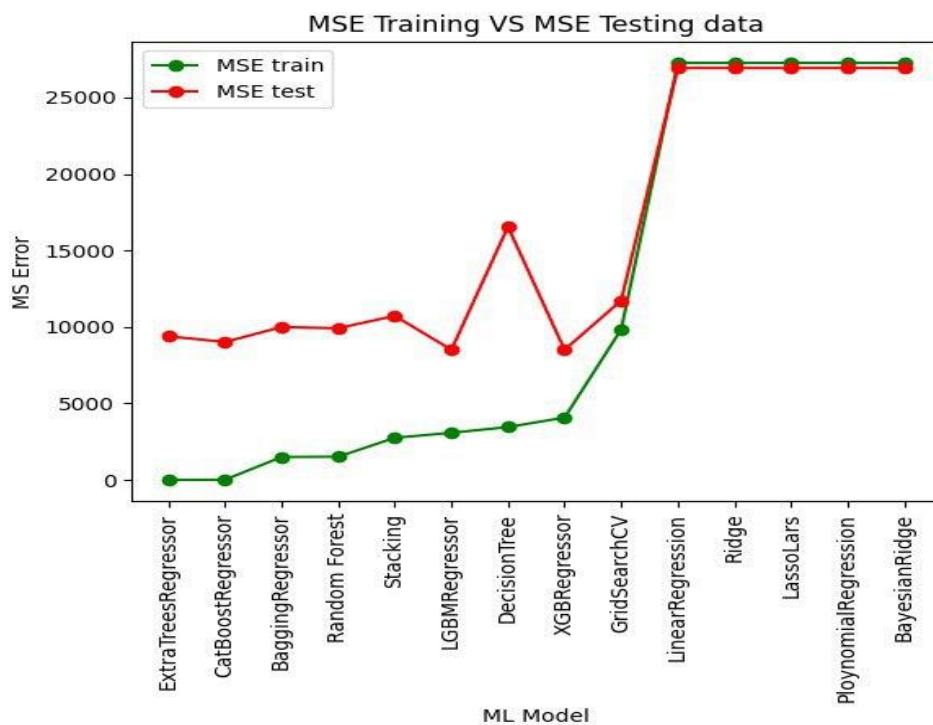


Figure 3. MSE Training vs. Testing: A performance comparison of ML models

Mean squared error (MSE) is a measure of the average squared difference between the predicted and actual values. It is calculated as the sum of the squared differences between the predicted values and the actual values, divided by the number of predictions.

Mathematically, it can be expressed as:

$$MSE = (1/n) * \sum (y - y')^2 \quad (2)$$

4. RESULTS AND DISCUSSION

The analysis of model performance reveals a clear hierarchy among the algorithms tested for sales prediction. Gradient Boosting-based models, including XGBRegressor, LGBMRegressor, and CatBoostRegressor, emerged as the top performers, achieving the highest predictive accuracy with test scores

of 0.92. Notably, LGBMRegressor was the most accurate, recording the lowest test error of 8504.054. Following these were the RF-based models (ExtraTreesRegressor, RF, and BaggingRegressor), which all achieved a commendable test score of 0.91, though their error rates were slightly higher. In contrast, traditional and less complex models such as Linear Regression and its variants performed significantly worse, with an average test score of 0.76. This highlights their limited capacity to capture the complex, nonlinear relationships within the sales data, confirming the superiority of advanced boosting models for this particular task. In Table 16 provides a comprehensive comparison of the performance of various regression models on test data, based on two key metrics: test score and test error.

Table 16. ML models summary

Index	Model	(Accuracy)
1	XGBRegressor	0.92
2	LGBMRegressor	0.92
3	CatBoostRegressor	0.92
4	ExtraTreesRegressor	0.91
5	Random forest	0.91
6	BaggingRegressor	0.91
7	Stacking	0.90
8	GridSearchCV	0.89
9	DecisionTree	0.85
10	LinearRegression	0.76
11	Ridge	0.76
12	LassoLars	0.76
13	PloynomialRegression	0.76
14	Bayesian ridge	0.76

5. CONCLUSION AND FUTURE WORK

This paper presents an advanced model that integrates 14 ML algorithms to address the challenges companies face in making accurate decisions from vast amounts of data. The research demonstrates the clear superiority of the XGBRegressor and LGBMRegressor models, which achieved the highest predictive accuracy (0.924) and the lowest error rate. This exceptional performance makes them vital tools for companies, enabling precise, data-driven decisions. Through accurate forecasting, organizations can enhance operational efficiency by optimizing inventory management and resource allocation, and effectively target customers with customized marketing strategies and promotions. The CatBoostRegressor model also provides a competitive advantage for handling rich categorical data. The significance of this paper lies in its function as a roadmap for companies to leverage ML for sustainable growth and increased profits. To ensure the continuity of this progress, the paper offers clear recommendations for future research, such as further developing boosting models, integrating external factors, and improving techniques for handling categorical data, which will ensure the creation of more robust and accurate models for the future.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Amal Elsayed Aboutabl	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	
Ola Mahmoud Moawad	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ahmed Mohamed Abd-Elwahab	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nspection

R : **R**esources

D : **D**ata Curation

O : Writing - **O**riginal Draft

E : Writing - Review & **E**dit

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.





DATA AVAILABILITY

The current paper relies on the publicly available 'Global Superstore' dataset, which was obtained from the Tableau website and can be accessed via the following link: "http://www.tableau.com/sites/default/files/training/global_superstore.zip".





REFERENCES

- [1] B. Mahesh, "Machine learning algorithms - a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, doi: 10.21275/art20203995.
- [2] H. Al-Sahaf *et al.*, "A survey on evolutionary machine learning," *Journal of the Royal Society of New Zealand*, vol. 49, no. 2, pp. 205–228, Apr. 2019, doi: 10.1080/03036758.2019.1609052.
- [3] V. Kumar, H. Garg, A. Gandhi, and B. Gupta, "Big mart sales prediction using machine learning," in *Artificial Intelligence: Theory and Applications*, Springer Nature Singapore, 2024, pp. 431–443, doi: 10.1007/978-981-99-8476-3_35.
- [4] P. Ranathunga, *Machine learning based sales forecasting system*. Doctoral dissertation, Robert Gordon University, Aberdeen, UK, 2022.
- [5] V. Sohrabpour, P. Oghazi, R. Toorajipour, and A. Nazarpour, "Export sales forecasting using artificial intelligence," *Technological Forecasting and Social Change*, vol. 163, p. 120480, Feb. 2021, doi: 10.1016/j.techfore.2020.120480.
- [6] B. M. Pavlyshenko, "Machine-learning models for sales time series forecasting," *Data*, vol. 4, no. 1, p. 15, Jan. 2019, doi: 10.3390/data4010015.
- [7] D. Reddy, K. S. Reddy, and S. N. R. B. S. Sahithi, "Prediction and forecasting of sales using machine learning approach," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 377, 2021.
- [8] J. E. F. Caroline, P. Parmar, S. Tiwari, A. Dixit, and A. Gupta, "Accuracy prediction using analysis methods and f-measures," in *Journal of Physics: Conference Series*, vol. 1362, no. 1, p. 012040, 2019, doi: 10.1088/1742-6596/1362/1/012040.
- [9] S. N. Gunjal, D. B. Kshirsagar, B. J. Dange, and H. E. Khodke, "Fusing clustering and machine learning techniques for big-mart sales predication," in *2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, Sep. 2022, pp. 1–6, doi: 10.1109/icbds53701.2022.9935906.
- [10] P. Amrutkar and S. Mahadik, "Sales prediction using machine learning techniques," *Journal homepage: www.ijrpr.com*, p. 7421, 2022.
- [11] D. G. S. Nana, D. D. B. Kshirsagar, D. B. J. Dange, D. H. E. Khodke, and D. C. S. Kulkarni, "Machine learning approach for big-mart sales prediction framework," *International Journal of Innovative Technology and Exploring Engineering*, vol. 11, no. 6, pp. 69–75, May 2022, doi: 10.35940/ijitee.f9916.0511622.
- [12] A. Tony, P. Kumar, and S. Rohith Jefferson, "A study of demand and sales forecasting model using machine learning algorithm," *Psychology and Education Journal*, vol. 58, pp. 10182–10194, 2021.
- [13] K. Ofoegbu, "A comparison of four machine learning algorithms to predict product sales in a retail store," *Diss. Dublin Business School*, 2021.
- [14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, pp. 1157–1182, 2003.
- [15] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [16] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.
- [17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.
- [18] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [19] L. Breiman and A. Cutler, "Random forests: classification and regression," *Journal of Machine Learning Research*, 2021.
- [20] S. Raschka and V. Mirjalili, "Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2," *Packt publishing ltd*, 2019.
- [21] Z.-H. Zhou, "Ensemble methods: foundations and algorithms," *CRC press*, 2025, doi: 10.1201/9781003587774.
- [22] L. Liu and Q. Zhang, "An overview of grid search for hyperparameter tuning in machine learning," *IEEE Access*, pp. 24310–24321, 2020.
- [23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017, doi: 10.1201/9781315139470.
- [24] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: with applications in R*. Springer US, 2021.
- [25] T. Hastie, R. Tibshirani, and J. Friedman., "The elements of statistical learning: Data mining, inference, and prediction," *New York, NY: Springer, 2001*, 533 + xvi pp., \$79.95., vol. 68, no. 4, pp. 611–612, Dec. 2003, doi: 10.1007/bf02295616.
- [26] C. Liu and X. Li, "Polynomial regression and its application in predicting the behavior of complex systems," *Journal of Computational and Graphical Statistics*, vol. 30, no. 3, pp. 653–672, 2021.
- [27] Y. Liu and S. Wang, "Bayesian ridge regression for high-dimensional data: theory and applications," *Journal of Statistical Computation and Simulation*, vol. 91, no. 5, pp. 980–994, 2021.

BIOGRAPHIES OF AUTHORS

Amal Elsayed Aboutabl     Professor of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt. Amal Elsayed Aboutabl is currently an Assistant Professor at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from the American University in Cairo and both of her M.Sc. and Ph.D. in Computer Science from Cairo University. She worked for IBM and ICL in Egypt for seven years. She was also a Fulbright Scholar at the Department of Computer Science, University of Virginia, USA. Her current research interests include parallel computing, performance evaluation and image processing. She can be contacted at email: amal.aboutabl@fci.helwan.edu.eg.



Ola Mahmoud Moawad     Teaching Assistant in Business Information Systems Department, AlGazera Academy, Mokattam, Cairo, Egypt. She holds an Academic Diploma in Computer Science from the Faculty of Computers and Information at Menoufia University. She earned her Master's degree in Information Systems from the Sadat Academy for Management Sciences. Furthermore, she has successfully completed the Doctoral Qualifying Courses in the Business Information Systems (BIS) Department at Helwan University. She can be contacted at email: ola.mahmoud2043@commerce.helwan.edu.eg; ola.mahmoud.pbis@commerce.helwan.edu.eg; drola2023@gmail.com.



Ahmed Mohamed Abd-Elwahab     Assistant Professor of Information Systems, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt. Ahmed Abd-Elwahab is an Egyptian academic and researcher in the field of Information Systems (IS), currently holding several prestigious academic and administrative leadership positions at Helwan University in Cairo, Egypt. He serves as an Assistant Professor in the Information Systems Department at the Faculty of Commerce and Business Administration (FCBA), the Deputy Editor-in-Chief for the SJRBS Journal-FCBA-HU, the Academic Registrar for the Financial Markets and Institutions (FMI) program, the IT Unit Manager for the FCBA, and a Consultant for the University's Technology Development Center. He can be contacted at: ahmed.mohamed.abdelwahab@commerce.helwan.edu.eg.