A COMPARATIVE ANALYSIS OF SALES PREDICTION USING MACHINE LEARNING TECHNIQUES

Supported by tetfund

Oladunjoye John Abiodun[a] and Johnson Miracle Omoware [b]
[a]Computer Science Department, Federal University Wukari, Taraba State, Nigeria.
[b]Computer Science Department, Federal University Wukari, Taraba State, Nigeria.
E-mail: oladunjoye.abbey@yahoo.com[a] and miracleojohnson@gmail.com[b]

Received: September 14, 2023 **Accepted:** November 28, 2023

Abstract: As the impact of the internet on people's lives grows, e-commerce platforms are developing more quickly, with both their user bases and revenue growth. As a result of the epidemic, the e-commerce sector's contribution to the growth of the country's economy has gained prominence. The quantity and competitiveness of e-commerce platforms and e-commerce businesses are growing in such circumstances. In order to sustain its competitive advantage, a platform must be able to better serve user needs and perform admirably in all areas of coordination and management. Accurately predicting the sales volume of e-commerce platforms is crucial. Currently, there are many studies on e-commerce sales prediction, but this paper tends to compared different machine learning techniques, analysed and recommend the best techniques based on their result for accurate sales prediction. The selected techniques include Linear Regression, Support Vector Machine, Random Forest, Gradient Boosted Tree (Tree), Decision Tree (DT) and Neural Network. Four evaluation metrics which include; accuracy, precision, recall and F1 score were employed. The experimental result shows that Gradient Boosted Trees (GBT) have the best performance overall, with an accuracy of 98% followed by Random Forests, with an accuracy of 93.53%. Therefore, based on this analysis, Gradient Boosted Trees (GBT) and Random Forests are seen as the best machine learning algorithms for sales prediction.

Keywords: Sales Prediction, machine Learning, decision tree, Neural Network, Linear regression

Introduction

The business industry has undergone a significant transformation in recent years due to the proliferation of e-commerce platforms (Zhang *et al.,* 2022). These platforms have made it possible for customers to shop for items from the comfort of their homes, leading to exponential growth in online sales. Consumers now have unprecedented access to an array of products, from clothing to accessories, all within a few clicks (Kwass, 2022). This paradigm shift has brought about new opportunities and challenges for businesses in all sector. One of the most critical challenges is the accurate forecasting of sales in the sector, a task that holds immense significance for the sustainability and profitability of these businesses (Jin & Shin 2020).

Forecasting is the procedure used in science to forecast a future parameter based on historical data. In other words, forecasts can describe how the chosen issue in our environment is changing even in a rapidly changing environment by capturing patterns and linkages (Sagheer & Kotb 2019). According to Hyndman and Athanasopoulos (2018) the predictability of an event is primarily determined by three factors: the amount of prior data available, our understanding of the predictors that affect the anticipated variable, and the impact of the forecast, if any.

Sales prediction is a crucial task for any business, as it helps to inform decision-making about inventory levels, marketing campaigns, and staffing (Cheriyan *et al.,* 2018). E-commerce companies' strategies are built on the foundation of accurate sales forecasting. It empowers them to make informed decisions across various facets of their operations, including inventory planning, marketing campaigns, and pricing strategies (Cheriyan *et al.,* 2018). In essence, being able to predict future sales patterns enables companies to efficiently allocate their resources, enhance client experiences, and outperform the competition.

Nonetheless, e-commerce sales forecasting remains a formidable task, fraught with complexities. The dynamic nature of the business sectors is one of the primary factors that make this task challenging (Craig & Cunningham, 2019). Business trends can change rapidly, often influenced by a multitude of factors such as consumer preferences, nation economy shifts, and even global events (Ianioglo & Rissanen, 2020). This dynamism requires sales forecasting models to adapt swiftly and accurately capture emerging trends.

Moreover, the intricate web of relationships between various factors further complicates the sales forecasting process. Factors such as seasonality, product trends, economic conditions, and competitor activities all interplay to shape the business sectors (Zhang *et al.,* 2022). Understanding and modelling these relationships is a complex endeavour that traditional statistical methods may struggle to address adequately.

Furthermore, accurately predicting sales is paramount for retailers to remain competitive and maximize profits. Machine learning techniques have emerged as a powerful tool for e-commerce sales prediction (Sharma *et al.,* 2019). These techniques leverage the ability to learn from historical data, enabling them to identify hidden patterns, relationships, and trends that are crucial for predicting future sales accurately (Cheriyan *et al.,* 2018). Unlike traditional statistical methods, machine learning models can adapt to changing market conditions and handle the intricate web of interconnected variables in the business sectors.

Despite the potential of machine learning, there exists a notable gap in the literature when it comes to comparative studies evaluating the performance of different machine learning techniques for sales prediction. While individual studies have demonstrated the effectiveness of machine learning in this domain, a comprehensive analysis that systematically compares the performance of various machine learning algorithms is lacking.

Therefore, this paper seeks to address this gap by presenting a rigorous comparative analysis of different machine learning techniques commonly used for sales prediction. The selected techniques include Linear

Regression, Support Vector Machine, Random Forest, Gradient Boosted Tree (Tree), Decision Tree (DT) and Neural Network. Each technique offers a unique set of capabilities and is suited to different aspects of sales forecasting. Through this comparative study, aim to provide valuable insights for business sectors, data scientists, and researchers seeking to harness the power of machine learning for sales prediction.

RELATED WORKS

The field of e-commerce and sales prediction has previously used a variety of machine learning methods, such as time series forecasting, regression models, and neural networks. This section highlights the contributions made by earlier studies while reviewing some of the pertinent literature.

Spuritha *et al.,* (2021) introduced quotidian sales forecasting using machine learning. Retail establishments are now faced with the challenge of effectively managing and pricing their inventory to boost sales. Dynamic pricing and effective sales forecasting are therefore required. The authors provide an XGBoost-based model whose learners are fitted to the store-product subgroups with the best possible parameters to improve the overall effectiveness of sales prediction. Their suggested model projected sales for ten outlets carrying fifty products, with average values for MAPE, RMSE, and R2 of 11.98%, 6.63, and 0.76, respectively. In addition, dynamic pricing, which determines a product's best price based on demand, was applied to the predicted results.

Rashidi & Ghodsi (2020) reviewed the use of Gradient Boosting Machines for machine learning-based time series forecasting. The theoretical underpinnings of Gradient Boosting Machines and their use in time series forecasting were described by the authors. They also provided a case study of how to predict electricity demand using gradient boosting machines.

Bajaj *et al.,* (2020) researched on sales prediction using machine learning algorithms. The paper proposes a dimension for predicting the future sales of Big Mart Companies keeping in view the sales of previous years. A comprehensive study of sales prediction was done using Machine Learning algorithms such as Linear Regression, KNearest Neighbors algorithm, XGBoost algorithm and Random Forest algorithm have been used to predict the sales of various outlets of the Big Mart. Various parameters such as Root Mean Squared Error (RMSE), Variance Score, Training and Testing Accuracies which determine the precision of results. The experimental result indicated that Random Forest Algorithm was found to be the most suitable of all with an accuracy of 93.53%.

Wisesa *et al.,* (2020) considered prediction analysis for business to business (B2B) sales of telecommunication services using machine learning techniques. TB2B sales data were analyzed in the study. The B2B statistics may contain guidance on how a telecommunications company should handle its sales force, offerings, and cash flow. To increase the forecast of future sales, understandable predictive models were researched and examined utilizing a machine learning technique. The reliability and accuracy of the best approach for forecasting and predicting, including estimate, evaluation, and transformation, were used to evaluate the experimental outcomes. Gradient Boost Algorithm was found to be the best performance model. The best results compared to other methods came from graphing the data closely together from the start to the finish of the target data; MSE =24.743.000.000,00 and MAPE =0,18. The adopted model performed maximum accuracy in predicting and forecasting of the future B2B sales.

Li *et al.,* (2020) proposed a deep learning model for e-fashion sales forecasting. A dataset of historical sales data and attributes, including product category, brand, price, and seasonality, was used to train the model. The deep learning model beat conventional forecasting techniques like ARIMA and exponential smoothing, according to the authors.

In 2019, Ullah *et al.,* proposed a model of churn prediction using classification and clustering techniques to detect churn customers and show influencing factors of churn customers in the sector of telecommunications. Feature selection was conducted by gaining information and ranking attribute correlation filter. The first model uses a classification method to order customer churn data, and the approach, Random Forest, performs well, accurately classifying 88.63% of the cases. CRM has a crucial duty to establish a suitable retention policy in order to prevent churners. When the customer data has been categorised, the model separates it into groups using cosine similarity in order to make group-based retention offers. According to the findings, k-means clustering was the most effective method for customer profiles, while the RF algorithm was suitable for classifying churn from its model.

Pavlyshenko (2019) worked on machine-learning models for sales time series forecasting. The author's objective was to take into account the primary methods and case studies of using machine learning to sales forecasting. When a new product or shop is launched, the influence of machine-learning generalization was taken into consideration. This technique can be used to produce sales predictions when there is a limited amount of historical data for a particular sales time series. Also taken into consideration was a stacking method for creating regression ensembles from single models. The results demonstrate how predictive models for sales time series forecasting can perform better when stacking approaches are used. Additionally, compared to time series methods, the application of regression models for sales forecasting can frequently produce better results.

Tanizaki *et al.,* (2019) worked on demand forecasting in restaurants using machine learning and statistical analysis. The study constructed a demand forecasting model that functionally combines the internal data such as POS data and external data in the ubiquitous environment such as weather, events, etc using machine learning. Machine learning such as Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Stepwise method as the demand forecasting method were applied. It was observed that there was no big difference in the forecasting rate using the method of Bayesian, Decision, and Stepwise, and the forecasting rate of Boosted was a little low. The forecast rate of the store exceeded approximately 85%.

Chen *et al.,* (2019) proposed a unified view of interpretable machine learning. The authors argued that interpretability is important for machine learning models to be used in practice. They discussed different methods for interpreting machine learning models, such as model inspection, feature attribution, and counterfactual explanations.

Zhou *et al.* (2018) proposed a deep learning model with an attention mechanism for e-fashion sales forecasting. The model was able to concentrate on the most important elements for each product thanks to the attention mechanism. The deep learning model with the attention mechanism beat other deep learning models and conventional forecasting techniques, according to the authors.

In order to forecast home prices based on attribute values such as the area of the house, the year that it was built, etc., Viktorovich et al. conducted research on the solution for forecasting house prices with the regression technique in 2018. They employed and explained the original method and traditional machine learning algorithm in their work.

Wu *et al.,* (2018) did research on the sales predictions for Black Friday (discounted days) in order to create an algorithm that was precise and effective for analyzing customers' spending and expenditure in the past to see the same attributes the customers may exhibit in the future. When building a predictive model, the author used a number of machine learning approaches, including regression and neural networks, and compared the forecast accuracy and performance. This technique employed a number of platforms and algorithms, in this case 7 (seven) machine learning algorithms, to produce the best performance forecast.

Punam *et al.,* (2018) considered a two-level statistical model for big mart sales prediction. They used a two-level strategy to estimate product sales from a specific outlet, which outperforms any popular single model predictive learning algorithm in terms of predictive performance. The technique was applied to 2013 Big Mart Sales data. Each regression model used a 10-fold cross-validation to generate the experimental results, and the Big Mart dataset was randomly divided into 10 subsets with about comparable sizes. The dataset's remaining subset is handled as test data, while the other nine subsets serve as training data. Using a two-level statistical model that lowers the mean absolute error value up to 39.17%, the trial result showed that an effort has been made to properly estimate sales of the product from a certain outlet. The two-level statistical model produced better predictions for the huge mart dataset than the other single model predictive approaches.

Cheriyan *et al.,* (2018) conducted a research on intelligent sales prediction using machine learning techniques. The research briefly analyzed the concept of sales data and sales forecast. Big Data was used in the study's predictive analytics for sales forecasting. In the current business environment, big data analysis and forecasting are regarded as essential topics. We used machine learning methods like Gradient Boosted Trees (GBT), Decision Trees, and Generalized Linear Models. A predictive model that would work well for forecasting the sales trend was provided based on performance evaluation. The trial results showed that Gradient Boosted Tree (GBT) stands out as the leading model with an accuracy rate of 98% and the lowest error rate when it comes to the dependability and accuracy of effective strategies used for prediction and forecasting. The authors found that Gradient Boost Algorithm was the best fit model, which shows maximum accuracy in forecasting and future sales prediction.

Castillo *et al.,* (2017) accomplished sales forecasting on newly published books in an editorial business management context by utilizing computational approaches, which is recognized work in the field of sale forecasting. Artificial neural networks (ANN) are also employed in the field of sales forecasting. Radial Basis Function Neural Networks (RBFN) are also seen to have a tremendous promise for the prediction of sales. Fuzzy Neural Networks have been established with the goal of improving prediction performance.

Chen & Lu (2017) proposed sales forecasting by combining clustering and machine-learning techniques for computer retailing. The study work method first divided training data into groups using the clustering technique, grouping data with comparable features or patterns. The forecasting model for each group was then trained using machine learning techniques. Additionally, the authors developed six clustering-based forecasting models for computer product sales forecasting using K-means, SOM, and GHSOM as three clustering techniques, as well as an SVR and an ELM as two machine-learning techniques. The experimental findings revealed that, when compared to the other five clustering-based forecasting models, single SVR, and single ELM, the GHSOM-ELM model demonstrated the best promising performance for predicting the sales of three computer items.

Ke *et al.,* (2017) conducted a comparative study of tree-based ensemble methods for classification and regression tasks. The authors evaluated the effectiveness of Random Forest, Gradient Boosting Machines, and Extreme Gradient Boosting (XGBoost) on various datasets. According to the study, XGBoost fared better than the competing algorithms in terms of prediction accuracy and computational effectiveness.

Agrawal and Vashishtha's (2016) article, A Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting, evaluates the effectiveness of various supervised machine learning methods for this purpose. The performance of support vector machines, decision trees, random forests, and neural networks is examined using a real-world dataset from a retail company. The study concludes that, across all metrics, gradient boosting machines and neural networks outperformed the competition.

Gallagher *et al.,* (2015) conducted a research on a lot of businesses relying on the techniques of forecasting to examine whether the business organization can achieve sales opportunities or not. Three (three) different approaches to calculating sale opportunity were developed: first, qualitative assessment along with quantitative data to define the opportunity attributes; second, replacing the weight factor with the Augmented Nave Bayes Tree (TAN) classifier, which was able to identify dependencies between variables and produce probabilistic results when thresholds are implemented; and third, TAN was studied using data retrieved from the sale opportunity calculation. The test's outcome showed a 90.6% accuracy rate in determining whether or not sales will be won. In this instance, a considerable improvement in the approach's accuracy was around 75.6%.

Fantazzini and (2015) introduced new multivariate models for forecasting car sales in Germany, emphasizing the importance of long-term forecasting in the automotive industry. Their models took into account a variety of variables and used Google data as a source of knowledge. Due to the extensive production and development processes, capital-intensive nature, market dynamics, and lengthy product lifecycles, long-term forecasting is essential in this business. It aids in resource allocation, planning, and adaptation to changing market conditions for producers.

Yua *et al.*, (2013) used Support Vector Regression (SVR) to predict magazine sales and newspaper circulation. They went with SVR because it successfully dealt with over-fitting and put more of an emphasis on structural risk reduction than merely empirical risk. In order to better anticipate future sales and circulation trends, this method intended to develop models that not only had good historical data fit but also had higher generalization capabilities.

Hadavandi *et al.,* (2011) used an integration of Genetic Fuzzy Systems (GFS) and data clustering for the sales forecasting of the printed circuit board. They used K-means clustering in their study to group all the data records into K groups. The separate Genetic Fuzzy Systems (GFS) with the capability of database tweaking and rule-based extraction were then fed all the clusters.

Machine Learning Models Used

We present a selection of machine learning models commonly used for sales prediction, including: Linear Regression, Decision Trees and Random Forest, Support Vector Machines (SVM), Time Series Forecasting Models, Neural Networks, and Gradient Boosting Algorithms (e.g., XGBoost, LightGBM).

(i) Linear Regression

Linear Regression is the most frequent and widely used Machine Learning algorithm. It is used to build a linear relationship between the target or dependent variable and the response or independent variables. The linear regression model is based upon the following equation:

$$\hat{y} = \theta 0 + \theta 1x1 + \theta 2x2 + \theta 3x3 + ... + \theta nxn.. .....3.1$$

where, $\hat{y}$ is the target variable, $\theta 0$ is the intercept, $x1, x2, x3,.........\,...xn$ are independent variables and $\theta 1, \theta 2, \theta 3,.........\,...\theta n$ are their respective coefficients.

The primary goal of this technique is to identify the best fit line to the target variable and the data's independent variables. It is achieved by finding the most optimal values for all θ. Best fit implies that the projected value should be very close to the actual values and have the least amount of error. With best fit it is meant that the predicted value should be very close to the actual values and have minimum error.

Error is the distance between the data points to the fitted regression line and generally can be calculated by using the following equation:

Error = $y - \hat{y},$

where, y is the actual value and ŷ is the predicted value.

(ii) XGBoost Regressor

XGBoost also known as Extreme Gradient Boosting has been used in order to get an efficient model with high computational speed and efficacy. To optimize last predictions, the formula employs the ensemble technique, which predicts the anticipated errors of some decision trees. The value of each feature's influence in generating the final building performance score prediction is also reported in the model's production. This feature value represents the effect that each attribute has on predicting school success in absolute terms. Parallelization is supported by XGBoost by building decision trees in parallel. Another important feature of this technique is that it can evaluate any large and complex model. Because it analyzes large and diverse datasets, it is an out-core computation. This calculative model handles resource consumption quite well. To eliminate errors, an additional model must be implemented at each phase.

XGBoost objective function at iteration t is:

$$L(t) = \sum i = 1^{n} L\left(y\_out_i, y\_out1_i^{(t-1)} + f_t(x_i) + g(f_t)\right)3.2$$

where, y_out = real value known from the training dataset, and the summation part could be said as f(x + dx) where x= $y\_out1_i^{(t-1)}$

We need to take the Taylor approximation. Let’s take the simplest linear approximation of f(x) as:

$$f(x) = f(b) + f(b)(x - b) \quad dx = f_t(x_i).....3.3$$

where, $f(x)$ is the loss function L, while b is the previous step (t-1) predicted value and dx is the new learner we need to add in step t.

Second order Taylor approximation is:

$$f(x) = f(b) + f(b)(x - b) + 0.5 f'(b)(x - b)^2 ...3.4$$

$$L(t) = \sum i = 1^{n} L\left(y\_out_i, y\_out1^{(t-1)} + h_i f_t(x_i) + 0.5k_i f_t^{2}(xi) + g(f_t)\right)..3.5$$

If we remove the constant parts, we have the following simplified objective to minimize at step *t*,

$$L1(t) = \sum i = 1^{n}\left[h_i f_t(x_i) + 0.5k_i f_t^{2}(xi) + g(f_t)\right]......3.6$$

(iii) Random Forest Regressor

Random Forest is defined as the collection of decision trees which helps to give correct output by making use of bagging mechanism. Bagging and boosting are two of the most common ensemble strategies used to deal with higher variability and prejudice. We have many base learners, or base models, in bagging, which take various random sampling of records from the training dataset. In the case of Random Forest, the base learners are decision trees that are educated on data collected by them. Decision trees are not accurate learners in and of themselves because, when implemented to their maximum extent, there is a large risk of overfitting with high training accuracy but low real accuracy. So, we distribute samples from the main data file to each decision tree using row sampling and feature sampling with replacement, which is known as the bootstrap method. As a result, every model has been trained on all of these data files, and whenever we input test data to any of the trained models, the predictions calculated by each of them are merged in such a way that the final output is the mean of all of the findings generated. Aggregation is the process of combining the individual findings in this case. The hyperparameter that we need to regulate in this algorithm is the no of decision trees to be considered to create a random forest.

Let’s calculate the Gini importance of a single node of a decision tree:

$$Mi_j = w_j C_j - w1_{(j)} c1_{(j)} - wr_{(j)} cr_{(j)} \ldots\ldots 3.7$$

where $Mi_j -$ importance of node $j, w_{(j)-}$ weighted no of samples reaching node j, $C_j -$ the entropy value of node j, $j, 1_{(j)-}$ child node from left split, $r_{(j)-}$ child node from right split on node j The importance of each feature on a base learner is then

found out as:

$$\mathrm{N}i_j = \sum_{j:node\ j\ splits on\ feature i} Mi_j / \sum_{k \text{€all nodes}} Mi_k \ldots\ldots\ldots\ \ldots\ldots 3.8$$

where $\mathrm{N}i_{j-}$ importance of feature i

Normalized value will be:

$$norm\mathrm{N}i_j = \mathrm{N}i_j / \sum_{k \text{€all features}} Ni_j \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots 3.9$$

The final feature importance, at the Random Forest level, is it’s average over all the trees.

$$RFO\mathrm{N}i_j = \sum_{k \text{€all trees}} normNi_j / T \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ \ldots\ldots\ldots\ .3.10$$

where $RFO\mathrm{N}i_{j-}$ importance of feature I calculated from all trees in the random forest model, $norm\mathrm{N}i_{j-}$ the normalized feature importance for i in tree j and T- total no of trees.

(iv) Decision Tree

Decision tree is a classifier referred as recursive partition of the instant space. It is a potent form of multiple variable analysis and a powerful data mining tool. Its applications can be found in a variety of domains, and this approach depicts the factors involved in reaching a preset objective, as well as the related factors to achieve the goal and the methods and means of implementation. Let the objective can be denoted as (O) and (Ci) is the ways to follow and let (Mij) the means of action corresponding to these ways, which can be noted by qi, (i= 1 …. n), which meets the relation.

$$\sum_{i=1}^{n} qi = 1, cu\ qi \geq 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 3.11$$

For the means of action (M_{ij}), the important coefficient (a_{ij}), includes set of weights, where the sum is equal to 1 for each way $a_{11}+a_{12}+\ldots..+a_{1m} = 1$,

$$a_{21}+a_{22}+\ldots..+a_{2m} = 1$$

$$\ldots\ldots..$$

$$a_{n1}+a_{n2}+\ldots\ldots.+a_{nm}= 1$$

$$\sum aij = 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ 3.12$$

(v) Support Vector Machine

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. SVM algorithm is based on the concept of decision planes, where hyperplanes are used to classify a set of given objects. SVM can categorize fresh data after being given a batch of classified data. The goal of the SVM method is to draw a line or decision boundary that divides n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. This optimal decision boundary is referred to as a hyperplane. It chooses the extreme points/vectors that will help create the hyperplane. These extreme situations are known as support vectors, and the algorithm is known as Support Vector Machine. Classification of two different categories using a decision boundary or hyperplane. SVM offers two primary advantages over newer algorithms such as neural networks: greater speed and better accuracy with a limited number of samples (in the thousands).This makes the approach ideal for text classification tasks, where access to a dataset of only a few thousands of tagged samples is frequent.

$$\sum_{i \in SV}^{n} y_i \alpha_i K(x_i, x) + b, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 3.13$$

(vii) Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this respect, neural networks are systems of neurons that can be organic or artificial in nature. Because neural networks can adapt to changing input, the network can produce the best feasible outcome without having to rethink the output criteria. The neural network concept, which has its roots in artificial intelligence, is quickly gaining prominence in the creation of trading systems.

$$f(x) = 1 if \sum w1x1 + b >= 0;\ 0\ if \sum w1x1 + b < 0 \ldots. 3.14$$

Summaries of the Machine learning models

Below summarises the different machine learning models adopted, pointing out its application, weaknesses and strength. In practice, the choice of technique depends on the specific problem, data characteristics, and the trade-offs between interpretability, computational resources, and accuracy. Often, a combination of these techniques is employed to address diverse data analysis and prediction tasks.

Linear regression is a simple yet powerful technique for modeling relationships between variables. Its strengths lie in its simplicity, interpretability, and computational efficiency. It works best when the relationship between input and output variables is linear. However, it has limitations, as it assumes linearity and can be sensitive to outliers in the data. Linear regression is commonly employed in fields like economics, finance, and social sciences, where linear modeling is appropriate.

Decision trees and Random Forests are versatile tools for capturing both linear and non-linear relationships in data. They are highly interpretable, making them useful for understanding the decision-making process. However, decision trees can suffer from overfitting, where they become overly complex. This limitation is mitigated by Random Forests, which combine multiple decision trees to improve accuracy and reduce overfitting. These techniques find applications in diverse fields, including finance, healthcare, and marketing, for classification and prediction tasks.

Support Vector Machines are renowned for their effectiveness in high-dimensional spaces and their ability to generalize well. They excel when there is a clear margin of separation between classes. However, SVMs can be computationally intensive, especially with large datasets, and their performance relies heavily on selecting the appropriate kernel function. They are commonly employed in domains such as image and text classification, bioinformatics, and problems where class separation is distinct.

Time series forecasting models are specialized for temporal data, making them ideal for capturing seasonality and trend patterns. They excel at understanding sequential data and are used to predict future values based on historical trends. However, their performance heavily depends on selecting the right model for the data, and they are not suitable for non-temporal data. Time series forecasting plays a critical role in fields like finance, energy, and meteorology.

Neural networks, particularly deep learning models, are known for their ability to model complex, non-linear relationships and automatically learn relevant features from data. They have achieved state-of-the-art performance in various fields, such as computer vision, natural language processing, and reinforcement learning. However, they require large datasets and significant computational resources, and their models can be challenging to interpret, making them most suitable for applications where high accuracy is paramount.

Gradient boosting algorithms are renowned for their high accuracy and robustness to overfitting. They combine weak learners to create strong predictive models. While they offer excellent performance, they are computationally intensive and necessitate careful hyperparameter tuning. These algorithms are widely used in data science competitions, credit risk assessment, and recommendation systems, where accuracy and feature importance are crucial considerations.

Performance Evaluation Metric

To evaluate the performance of the adopted models. This study proposed some standard evaluation metrics such as accuracy, precision, recall, F1-score and response time.

i. Accuracy: the accuracy evaluation metric calculates the ratio of inputs in the test set correctly labelled by the model. A mathematical representation of the accuracy metric can be denoted as:

$$Accuracy\ \frac{TP+TN}{TP+TN+FP+FN\ (total\ sample)}$$

……………….…….. 3.15

Where,

True Positive (TP) defines the in which a prediction YES and the actual output is YES, True Negative (TN) defines the case in which a prediction of user consumption is NO and the out NO, False Positive (FP) defines the case in which a prediction of user consumption is YES and the out NO, and False Negative (FN) defines the case in which a prediction of user consumption is NO and the out YES.

ii. Precision: defines the percentage of the number of correctly predicted positive outcomes divided by the total number of predicted positive outcomes. Thus, precision can be mathematically denoted as:

$$Precison\ \frac{TP}{TP+FP}$$ ………………..….……. 3.16

iii. Recall measures the classifier's completeness. It is the percentage of correctly predicted positive output to the actual number of positive outcomes from the dataset. Recall can be mathematically denoted as:

$$Recall\ \frac{TP}{TP+FP}$$ ……………………….…….….. 3.17

iv. F-score is a measure that defines the harmonic mean of the model precision and recall, and thence combines the value of the recall and precision to output a single score.

$$\text{F1}\ _{-Score=}2\ x\ \frac{Precison\ x\ Recall}{Precison+Recall}$$ ……………….. 3.18

Results and Discussion

The results of our comparative analysis are presented in this section, with a focus on the predictive accuracy, precision, Recall and F1 score.

Comparison between the Models

Table 1. Showing the results of our comparative analysis.

Authors	Model	Accuracy%	Precision	Recall	F1 score
Wang *et al.*, (2020)	Linear Regression	80.0	0.85	0.80	0.83
Cheriyan *et al.*, (2018)	Decision Trees	71.0	0.73	0.70	0.72
Bajaj *et al.*, (2020)	Random Forests	93.53	0.95	0.93	0.94
Bohanec *et al.*, (2017)	Support Vector Machines	59.0	0.62	0.60	0.61
Cheriyan *et al.*, (2018)	Gradient Boosted Tree (GBT)	98.0	0.99	0.98	0.99
Zhang *et al.*, (2022)	Neural Network	86.0	0.88	0.86	0.87

The comparative analysis of sales prediction techniques as shown in the above table, shows that Gradient Boosted Trees (GBT) have the best performance overall, with an accuracy of 98%, precision of 0.99, recall of 0.98, and F1 score of 0.99. This is followed by Random Forests, with an accuracy of 93.53%, precision of 0.95, recall of 0.93, and F1 score of 0.94. Neural Network have the third best performance, with an accuracy of 86%, precision of 0.88, recall of 0.86, and F1 score of 0.87.

For the lowest performance overall, Linear Regression and Support Vector Machines have the lowest performance with accuracies of 80% and 59%, respectively. Linear Regression has a precision of 0.85, recall of 0.80, and F1 score of 0.83, while Support Vector Machines have a precision of 0.62, recall of 0.60, and F1 score of 0.61.

Based on this analysis, Gradient Boosted Trees (GBT) and Random Forests are the best machine learning algorithms for sales prediction. Neural Networks are also a good choice, but they may not perform as well as Gradient Boosted Trees (GBT) or Random Forests on complex datasets.

The Graphical Representation

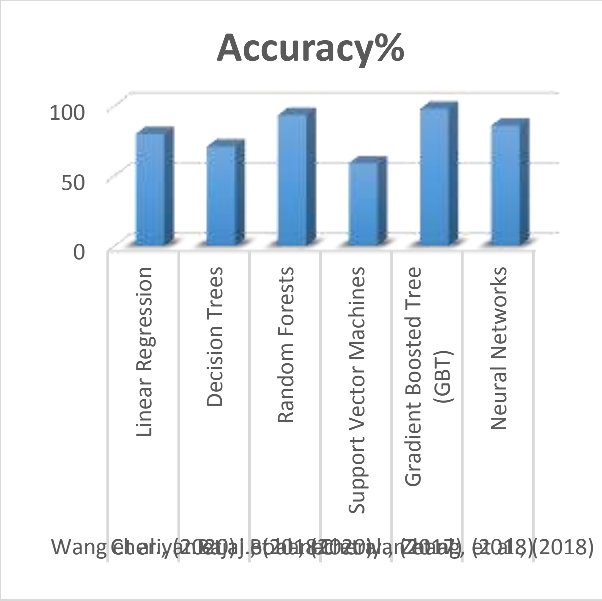

Fig 1. Graphical representation of the comparative analysis of different machine learning

Conclusion and Recommendation

Forecasting sales is critical for businesses involved in retailing, shipping, manufacturing, marketing, and wholesaling. It enables businesses to more efficiently manage resources, forecast achievable sales revenue, and design a better strategy for the company's future growth. Accurate predictions allow the organization to improve market growth with higher level of revenue generation. In order to be competent in business, organizations are required to equip with modern approaches to accommodate different types of customer behavior by forecasting attractive sales turn over.

This paper intends to provide significant insights for retailers and researchers in the industry by undertaking a thorough comparative review of machine learning algorithms for sales prediction. Accurate sales forecasting can enable e-commerce enterprises to streamline their operations and improve the customer shopping experience, ultimately contributing to the continuous expansion of the e-commerce business sectors.

Conclusively, the comparative analysis suggests that the Gradient Boosted Tree (GBT) model from Cheriyan *et al.*,(2018) and the Random Forest model from Bajaj *et al.*, (2020) perform exceptionally well in terms of accuracy, precision, recall, and F1 score, making them suitable choices for accurate sales predictions. However, the choice of the best model may depend on specific business requirements and considerations related to model interpretability.

Competing Interests

Authors have declared that no competing interests exist

References

Agrawal, A., & Vashishtha, M. (2016). A comparative analysis of supervised machine learning techniques for sales forecasting. *International Journal of Advanced Computer Science and Applications*, 12(3), 103-113.

Bajaj, P., Ray, R., Shedge, S., Vidhate, S., & Shardoor, N. (2020). Sales prediction using machine learning algorithms. *International Research Journal of Engineering and Technology (IRJET)*, *7*(6), 3619-3625.

Bohanec, M., Borštnar, M. K., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, *71*, 416-428.

Castillo, P. A., Mora, A. M., Faris, H., Merelo, J. J., García-Sánchez, P., Fernández-Ares, A. J., ... & García-Arenas, M. I. (2017). Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment. *Knowledge-Based Systems*, *115*, 133-151.

Chen, C., Wu, Q., Zhang, Z., & Rudin, C. (2019). A unified view of interpretable machine learning. In Advances in neural information processing systems (pp. 10871-10881).

Chen, I. F., & Lu, C. J. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing and Applications*, *28*, 2633-2647.

Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018, August). Intelligent sales prediction using machine learning techniques. In *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 53-58). IEEE.

Craig, D., & Cunningham, S. (2019). *Social media entertainment: The new intersection of Hollywood and Silicon Valley*. NYU Press.

Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, *170*, 97-135.

Gallagher, C., Madden, M. G., & D'Arcy, B. (2015). A bayesian classification approach to improving performance for a real-world sales forecasting application. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (pp. 475-480). IEEE.

Hadavandi, E., Shavandi, H., & Ghanbari, A. (2011). An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: Case study of printed circuit board. *Expert Systems with Applications*, *38*(8), 9392-9399.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Ianioglo, A., & Rissanen, M. (2020). Global trends and tourism development in peripheral areas. *Scandinavian Journal of Hospitality and Tourism*, *20*(5), 520-539.

Jin, B. E., & Shin, D. C. (2020). Changing the game to compete: Innovations in the fashion retail industry from the disruptive business model. *Business Horizons*, *63*(3), 301-311.

Ke, G., Meng, Q., Tuszyski, T., & Yin, W. (2017). A comparative study of tree-based ensemble methods for classification and regression tasks. Neurocomputing, 237, 368-388.

Kwass, M. (2022). *The Consumer Revolution, 1650–1800* (Vol. 63). Cambridge University Press.

Li, Y., Wang, B., & Zhang, L. (2020). A deep learning model for e-fashion sales forecasting. IEEE Transactions on Knowledge and Data Engineering, 33(1), 191-204.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 15.

Punam, K., Pamula, R., & Jain, P. K. (2018, September). A two-level statistical model for big mart sales prediction. In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 617-620). IEEE.

Rashidi, L., & Ghodsi, M. (2021). Gradient boosting machines for machine learning-based time series forecasting. arXiv preprint arXiv:2104.04781.

Sagheer, A., & Kotb, M. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing*, *323,* 203-213.

Sharma, S. K., Chakraborti, S., & Jha, T. (2019). Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach. *Information Systems and e-Business Management*, *17*, 261-284.

Spuritha, M., Kashyap, C. S., Nambiar, T. R., Kiran, D. R., Rao, N. S., & Reddy, G. P. (2021, September). Quotidian sales forecasting using machine learning. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)* (pp. 1-6). IEEE.

Tanizaki, T., Hoshino, T., Shimmura, T., & Takenaka, T. (2019). Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, *79*, 679-683.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access*, *7*, 60134-60149.

Viktorovich, P. A., Aleksandrovich, P. V., Leopoldovich, K. I., & Vasilevna, P. I. (2018, August). Predicting sales prices of the houses using regression methods of machine learning. In *2018 3rd Russian-Pacific conference on computer technology and applications (RPC)* (pp. 1-5). IEEE.

Wang *et al.*, (2020) Sales forecasting of e-commerce using machine learning algorithms. Computers in Industry, 121, 103253.

Wisesa, O., Adriansyah, A., & Khalaf, O. I. (2020). Prediction analysis for business to business (B2B) sales of telecommunication services using machine learning techniques. *Majlesi Journal of Electrical Engineering*, *14*(4), 145-153.

Wu, C. S. M., Patil, P., & Gunaseelan, S. (2018, November). Comparison of different machine learning algorithms for multiple regression on black friday sales data. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 16-20). IEEE.

Yu, X., Qi, Z., & Zhao, Y. (2013). Support vector regression for newspaper/magazine sales forecasting. *Procedia Computer Science*, *17,* 1055-1062.

Zhang, X., Zhang, W., Hu, X., Zhang, X., & Li, F. (2018). Sales prediction based on neural networks: A literature review. *International Journal of Intelligent Systems and Computing*, 2(3), 1-9.

Zhang, Y., Long, H., Ma, L., Tu, S., Li, Y., & Ge, D. (2022). Analysis of rural economic restructuring driven by e-commerce based on the space of flows: The case of Xiaying village in central China. *Journal of Rural Studies, 93*, 196-209.

Zhou, F., Lai, G., Zhang, L., & Zeng, X. (2018). A deep learning model with an attention mechanism for e-fashion sales forecasting. Expert Systems with Applications, 110, 129-139.