RESEARCH ARTICLE

SAMRIDDHI Volume 14, Issue 1, 2022 Print ISSN: 2229-7111 Online ISSN: 2454-5767

Extreme Gradient Boosting Model-based Forecasting of Big Data Online Sales Record

Gagan Sharma*, Sunil Patil

Department of Computer Science and Engineering, RKDF University, Bhopal, India

Abstract

Nowadays, big data plays a crucial role for many online e-commerce businesses to generate more sales. Big data is a huge collection of data and information which are utilized by many organizations to forecast which products, costs, and advertisements are better to maximize their business profits. This paper aims to apply the extreme gradient boosting (XGBoost) based model to forecast sales growth of online products, specifically books and magazines, from massive datasets present in online shopping. PySpark, as the best suitable and compatible framework, is used for data analysis. The result shows that the proposed model has higher forecasting accuracy with a minimum error rate than other models. A comparative visualization and conclusion are presented in terms of the proposed system's prediction accuracy, error rate, and efficiency.

Keywords: Big Data, E-Commerce, Extreme Gradient Boosting, Forecasting, PySpark.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2022); DOI: 10.18090/samriddhi.v14i01.18

Introduction

With the enhancement of network-based logistics management, internet, and e-commerce has been prospering in recent years. Nowadays, online shopping, business, and e-commerce have become indispensable component of humans' daily life. With the tremendous growth, the volume of e-commerce data has evolved extensively, to form big data. The surroundings in which human lifestyles exist have been more community-based with the changes in social conditions. The inception of community-based business and e-commerce has come; for community products, online business, and sales, the internet as infrastructure is fulfilling the demand of customers for community residents as the target. Big data enriched the models and applications of e-commerce,[1] five basic properties, called the 5 V's: "Volume, Velocity, Variety, Veracity, and Value," are frequently characterized by big data.[2,3]

Now, it is possible to acquire huge amount of information with the huge volume of big data. The primary challenge is to inquire the right query to get reliable information and appropriate processing of data. Big data's possible and suitable applications are supply chain context and business models.[4,5] Unprecedented opportunities and ideas for companies have been created by big data analytics to exploit salient features of datasets for business-to-customer (B2C) and business-to-business (B2B) market commencements. In addition, through assembling, integrating, and utilizing the features of big data, highly-reputed companies such as Facebook, Amazon, Google, and Apple have all given tremendous efforts in the area of commercial and industrial marketing. As an essential element in global business and marketing operations, all of these underline the grandness of big data.[6] The progressive growth in the diversity and quantity of big data contributed towards such datasets that are very larger to manage by the traditional data management frameworks and tools. Advanced methods and techniques of big data with modern applications for predictive analytics have now been developed to control and manage these potentially novel valuable datasets.[7,8]

The vast volume of e-commerce product reviews and feedback has been posted to various forums and online social media in this era of big data.[9] Various repositories collect these vast amounts of review and feedback data for data analysis and visualization for business purposes. The work presented in this article is based on product review and sales growth rate using big data acquired from https://www.kaggle.com hosted by an online book store. The datasets

Corresponding Author: Gagan Sharma, Department of Computer Science and Engineering, RKDF University, Bhopal, India, e-mail: gagansharma.cs@gmail.com

How to cite this article: Sharma, G., & Patil, S. (2022). Extreme Gradient Boosting Model-based Forecasting of Big Data Online Sales Record . *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 14(1), 112-119.

Source of support: Nil

Conflict of interest: None

provided historical sales of books and magazines of one year are in comma-separated value (CSV) format. Around one year's datasets are applied for forecasting modeling, and based on the model, the following three years' growth rate of product sales forecasting reports are visualized. This article presents the forecasting model of sales growth using the "extreme gradient boosting (XGBoost)" algorithm, and using this proposed model various product sales rates can be predicted.

This article is organized as follows; Section II explains the importance of extensive data analysis for business and eCommerce growth. Section III explores the previous works and exploratory surveys related to the forecasting of e-commerce using big data technology. Section IV presents the mathematical calculations of the proposed methodology for evaluation and analysis of result. Section V presents visualization of results and comparison with various algorithms. Finally, Section VI concludes this research with recommended future scopes.

Big Data in E-Commerce

Nowadays, plenty of companies depend on forecasting big data analysis for higher sales and growth. Big data analysis provides an appropriate direction for customer selection to send product offers because customers feel entirely disappointed, irritated, frustrated whenever companies send those irrelevant and incoherent offers.[10–13] The proper target customers (products of their choice) can be selected using proper analysis of big data, so the individual personalized recommendation can be predicted in online shopping at several individual site visits. Big data benefits for e-commerce are presented in Figure 1, including a portfolio of the products, price, online or in-store experiences, marketing or advertising budgets, customer helplines or services, and inventory. These are the most fundamental benefits of using big data in e-commerce for online sales and marketing.[14–16]

Nowadays, customers look for a more convenient and easier way to purchase products. Big Data analysis shows the retailers to interpret requirements and choice of the customers before entering into store or application to provide a superior personal buying experience.[17] Using big data analytics on purchase history, geographic location, travel record, social networks, and search engine data, consumers' product promotions can be targeted or transmitted directly to their smartphones and applications while they shop.[18]

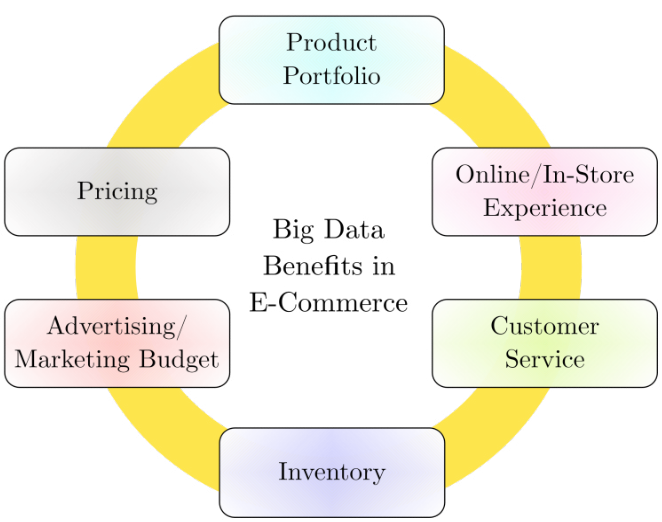

Figure 1: Big Data in E-Commerce

Related Work

E-commerce is completely influenced by data analysis and big data. It has become an indispensable division of modern lifestyle, and this technology has attracted huge attention towards researchers for business and industrial growth.

Many researchers have been concentrating particularly on product sales forecasting for offline stores earlier the emergence of online business e-commerce technologies, such as enterprises and supermarkets. Sales forecasting is beneficial for online business and inventory resources and product management marketing. In both conventional offline commerce domain and online e-commerce industry, inventory resource and product management is a remarkable step for providing better services with higher quality products and fast transactions.[19]

Ragg *et al.*[20] proposed Bayesian learning based on a neural network to forecast sales rates using retailers' big data. The huge volumes of data sets (big data) are not feasible for parametric fitting using cross-validation. Forecasting performance can be improved using "Bayesian learning rules." In the target data sets, noise can be minimized by averaging over data in the selection process.

Ren *et al.*[21] presented a short-term load prediction model based on "extreme gradient boosting (XGBoost)" and load clustering. This model completely evaluates weather data, historical power-load data samples, and calendar data to forecast short-term power loading. Additionally, before the training process, the K-means method is applied for clustering of load data samples so that data having the same features can be clustered to improve this model's accuracy.

Islam and Amin[22] proposed a model to predict the probable backorder products with the help of two machine learning methods. In their method, sales, lead time, the ranges of several inventories, and level of predictions can be easily adjustable. According to the types of business, requirements, and market target, these ranges can be adjusted to forecast profitable backorder products.

Xia *et al.*[23] designed, implemented, and evaluated a sales prediction model, "ForeXGBoost" based on huge datasets that integrated useful information such as vehicle brand name, model number, power, etc. Their comprehensive research shows that "ForeXGBoost" outperforms the criterion methods in forecasting accuracy and overhead.

Chong *et al.*[24] used neural network modeling in big data technology for analyzing the demands of items in online shopping and business. Their outcomes present that the data and variables applied in the analysis are completely effective for predicting online items sales. The outcomes also present that historical reviews of online shopping products are beneficial for the promotional marketing of specific products.

Boone *et al.*[25] proposed a forecasting model for sales of supply-chain using consumers' analytics in big data and associated technologies. They found that data can be applied for acquiring perceptiveness from the consumer's behavior and business prediction.

Sohrabpour *et al.*[26] employed "Genetic programming (GP)" as an artificial intelligence model to forecast the export sales for a "Middle Eastern company" that was confronting variation and irregularities in sales and other applicable prestigious factors. Their "GP-based export sales model" uses four error metrics calculations: "R-square goodness of fit," "correlation coefficient," "mean squared error," and "mean absolute error." Their outcomes indicate greater accuracy and forecasting precision.

Yuan *et al.*[27] designed a novel method to acquire and compute consumers' sentiments towards product quality, usefulness, and price that simultaneously enhance sales prediction. Ultimately, the analytical sentiment distributions with other factors are applied for forecasting sales quantities in the upcoming period.

Palanimalai[28] discussed several big data analytics strategies, which e-commerce data can importantly exploit to take away greater marketing values and novel business insights. They demonstrated practical experimentation with "customer relationship management (CRM)" data to get over the forecasting solutions that amplify the prediction of the target sales/revenue planning and management. Big data-based analytical solutions alleviate business and marketing to visualize and recognize patterns and trends that can bring perceptive outcomes on the performance of the business.

Kılınç[29] presented a "Spark-based sentiment analysis (SA)" real-time framework that includes four constituents "Spark Machine Learning" and its streaming services, a "Twitter streaming service", a fake account detection system for Twitter, and a reporting with dashboard real-time software solution. They demonstrate the "real-time sentiment analysis (SA)" of Turkish tweets with the help of the machine learning model of the "Spark MLlib" library.

Eapen *et al.*[30] proposed a novel deep learning-based model which integrates multi-pipelines of "Convolutional Neural Network (CNN)" to extract the features from "Bidirectional Long Short Term Memory (LSTM)" data. They discovered that when layers of the CNN are connected to the "Bi-directional LSTM" it presents better performance and prediction than the conventional Support Vector Machine (SVM). Moreover, multi-pipelines of deep layers present outcomes better than a model of the single pipeline.

With the exponential ascending research interest in "Big Data and Predictive Analytics (BDPA)," industrial manufacturing, business operations, and management have continuously been adopting this technology. There is a huge gap in theoretical research on the characterization of "BDPA" in manufacturing performance. Following the call by Oliver,[31] Dubey *et al.*[32] proposed a conceptual model conjugated in institutional theory with the "resource-based view (RBV)" to handle the present shortcomings of the "RBV" and empirical investigation so big data capabilities can significantly assist in accomplishing manufacturing performance.

Proposed Method

The proposed model initiates data collection from various sources such as e-commerce datasets, product review reports, expert discussion data, and consumer forum data sets. These datasets are in several formats, so data analysis requires consistent and uniform re-formatting. The input data must be pre-processed to fulfill the missing value and detect and remove outliers. For clustering analysis, k-means clustering algorithm is applied to the datasets. Finally, XGBoost based proposed system is used to evaluate and analyze results.

For data pre-processing, clustering analysis and implementation of XGBoost proposed model is evaluated in PySpark[33] framework environment. PySpark is an integrated interface for Python with Apache Spark. Big data analysis and processing with machine learning presents an extremely unified analytical engine. The overall process of the proposed model is presented in Figure 2.

Data Collection and Formatting

The data collected from various sources are in several text formats. For uniformity, all datasets are converted into

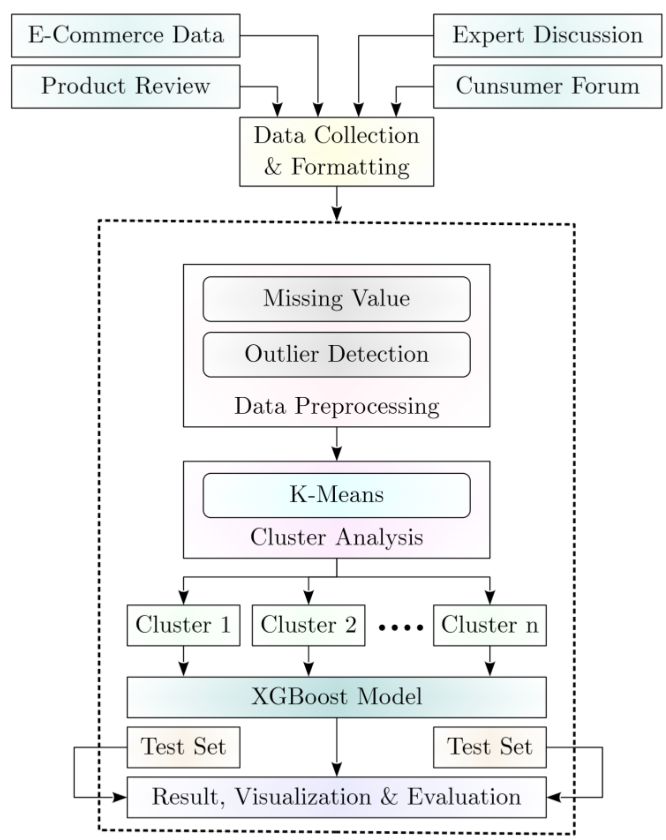

Figure 2: Workflow of Proposed Model

a common "comma-separated value (.csv)" format as an appropriate input for PySpark to analyze datasets.

Data Preprocessing

This research considers the specific input attributes of the product sales from data sets of big data. Some specific attributes are presented in Table 1 are Product ID, Dept. ID, Price, Date, City, etc. In Table 2, the real-life example of sales data from online book shopping e-commerce data is presented with useful parameters.

Due to human manipulations or systematic errors, there may be some missing as well as outlier data that is possible. These conditions in the datasets severely interact during the training and forecasting steps of the model and will affect the accuracy and precision of the predicted outcomes. Therefore, it is essential to pre-process the original datasets to ensure the proposed model's smoothness, consistency, and full performance.

Pre-processing of Missing Datasets

Using "Euclidean Distance," the weighted average is computed by Equation 1 for missing values in n number of datasets.

$$d(m) = w_1 d(m_1) + w_2 d(m_2) + \ldots + w_n d(m_n) \quad (1)$$

Where, $d(m)$ represents the missing data, $[d(m_1), d(m_2), \ldots, d(m_n)]$ represents the set of data vectors to $d(m)$, and $[w_1, w_2, \ldots, w_n]$ represents the weight vectors set calculated by the Euclidean distances. The lower values of Euclidean distances have higher weight coefficients.

Table 1: Attributes of Products Sales Datasets

Attributes	*Descriptions*	*Examples*
Product_ID	Product identification number	MATH#537
Dept_ID	Dept. ID Department of the product	Book & Magazine
Brand_ID	Brand name of the product	Pearson India
Price	Price of the product	899.00 ₹
Dim	Dimensions (size, weight etc.) of the product	10x10x15 cm^3, 700 gm
Sale Date	Date of the sale	Mar-21-2021
Trans_ID	Transaction number of the product	T784653
Sale_Qty	Quantities of the sale	350
City	Location of the buyer	Bhopal
Review	Review given as feedback	Good, 4 Stars

Table 2: Examples of Sales Datasets

Prod ID	*Dept ID*	*Brand ID*	*Price (₹)*	*Sale Date*	*Trans ID*	*Sale Qty*	*City*	*Review*
COMP#365	Book	TMH	799.00	Apr-03-2021	T6734	1	Pune	Good 3*
ELEC#177	Book PHI	PHI	569.00	Apr-04-2021	T6898	1	Malda	Good 4 *
COMS#009	Magazine	IEEE	270.00	Apr-06-2021	T6917	1	Nagpur	Bad 1 *
PHYS#445	Book	Wiley	749.00	Apr-06-2021	T6975	1	Amravati	Excellent 5 *
ACCO#423	Book	S.Chand	950.00	Apr-07-2021	T6989	1	Noida	Good 3 *
STAT#776	Book	Pearson	780.00	Apr-10-2021	T7565	1	Ujjain	Good 4 *
ECON#317	Book	S.Chand	650.00	Apr-11-2021	T7578	1	Noida	Good 4 *
GEOM#112	Book	TMH	745.00	Apr-11-2021	T7585	1	Bhopal	Good 4 *
CALC#323	Book	Wiley	870.00	Apr-13-2021	T7593	1	Ajmer	Bad 1 *
INDT#365	Magazine	Living Media	298.00	Apr-14-2021	T7598	2	Gwalior	Good 4 *
FASH#365	Magazine	St.Joseph Comms	310.00	Apr-16-2021	T7611	1	Mohali	Good 4 *
CLOU#969	Book	Cengage	890.00	Apr-17-2021	T7623	2	Bikaner	Good 4 *
AIML#284	Book	O'Reilly	865.00	Apr-19-2021	T7628	1	Surat	Excellent 5 *
PROB#833	Book	Pearson	779.00	Apr-23-2021	T6734	1	Jaipur	Excellent 5 *
REGR#556	Book	Apress	689.00	Apr-24-2021	T7634	1	Mainpuri	Good 4 *
DRON#113	Magazine	RotorDrone	367.00	Apr-25-2021	T7648	2	Faridabad	Excellent 5 *
HADO#909	Book	No Starch	899.00	Apr-27-2021	T7652	1	Bhopal	Good 4 *
NEUR#811	Book	Manning	999.00	Apr-29-2021	T7660 1	1	Ponda	Excellent 5 *

Outlier Detection and Processing

Generally, unexpected conditions, unpredictable events, and systematic errors generate some datasets. The sequence and balancing of the integral datasets are interrupted by these outlier data, resulting in lower forecasting accuracy. Consequently, it is essential to handle these outlier data.

Firstly, mean value $M(m)$ and squared error $S(m)$ of the n datasets are calculated as

$$M(m) = \frac{1}{N}\sum_{n=1}^{N} d(m_n) \tag{2}$$

$$E(m) = \sigma_m^2 \tag{3}$$

$$= \frac{1}{N}\sum_{n=1}^{N}[d(m_n) - M(m_n)]^2 \tag{4}$$

Secondly, the rate of deviation is calculated as

$$\rho(m) = \frac{|d(m) - M(m)|}{\sigma_m} \tag{5}$$

The corrected datasets $\hat{d}(m_n)$ are calculated using the outlier correction process on datasets as follows

$$\hat{d}(m_n) = \frac{d(m_{n-1}) + d(m_{n+1})}{2} \tag{6}$$

Normalization of Datasets

Most of the datasets contain various features that vary greatly in terms of attributes, such as ranges and dimensions.

Using machine learning algorithms, the non-standardized data as the input can generate unpredictable wrong output. Therefore, normalization can remove the effect of dimension and range deviation between variables. For standardization of the original data, the normalized data X' is calculated as

$$X' = \frac{X - \bar{X}}{\sigma} \tag{7}$$

where $\bar{X}$ represents the "mean" value of the original datasets, and σ is the "standard deviation" of the original datasets.

Clustering Analysis of Datasets

For clustering of datasets, k-means[34] algorithms are applied to the e-commerce data because it is fast, scalable, and effective for processing big data. Based on the clustering, the distance between data samples is calculated. The "Euclidean Distance" between data samples is calculated using considerable differences in dimensions of the variables in normalized datasets.

$$\mathrm{Dist}(m_i, m_j) = \sqrt{\sum_{n=1}^{N}(x_{i,n} - x_{j,n})^2} \tag{8}$$

Where $\mathrm{Dist}(m_i, m_j)$ represents a distance between i^{th} and j^{th} data samples, whereas, $x_{i,n}$, $x_{j,n}$ represents n variables of i^{th} and j^{th} data samples.

Extreme Gradient Boosting Model

As compared to a conventional boosting algorithm, the "extreme gradient boosting (XGBoost)" algorithm presents high accuracy in machine learning. XGBoost algorithm can be applied to big data using the PySpark framework because it can efficiently handle distributed parallel computation and sparse data. XGBoost also controls model complexity and minimizes model variance by adding regular items into the loss function.

The forecasting accuracy can also be improved using XGBoost at a definite speed.

In the datasets $D = \{(x_i, y_i)\}(x_i \in \mathbb{R}^m, y_i \in \mathbb{R})$ contains n data samples, m number of features, and XGBoost model contains K number of decision trees. The forecasting value is calculated as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{9}$$

where, $\hat{y}_i$ represents the forecasting value for i^{th} target variable, the input data variable is represented by $x_i \in \mathbb{R}^m$ proportionate to the $\hat{y}_i$, whereas, f_k represents forecasting function for k^{th} decision tree, and it is calculated as

$$f_k(x_i) = w_{q(x_i)}, q: \mathbb{R}^m \to T, w \in \mathbb{R}^T \tag{10}$$

For k^{th} decision tree, the structure-function is represented by $q(x_i)$, which maps x_i to the leaf node of the tree, w represents the quantization weight factor for the corresponding leaf node, and T represents a total number of leaf nodes in a tree.

The proposed XGBoost model appends conditional quantities to its loss function by analyzing the complexity and accuracy of the system. By decreasing the loss function L, as presented in Equation (11), the model learns to forecast as

$$L = \sum_{i=1}^{n} d(\hat{y}_i, y_i) + \sum_{k=1}^{K} w(f_k) \tag{11}$$

where, $\hat{y}_i$ represents regular term to assist the model for overfitting prevention. The model complexity function $w(f)$ can be calculated as

$$w(f) = \gamma T + \lambda\frac{\|w\|^2}{2} \tag{12}$$

where, γ represents complex parameter, λ represents predefined fixed coefficient, and the total number of leaf nodes is represented by T.

To analyze the performance of the forecasting model, the error evaluation criterion is applied. The "mean absolute percent error (MAPE)," "mean absolute error (MAE)," and "relative error (RE)" is calculated as

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{13}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{14}$$

$$\mathrm{RE} = \frac{y_i - \hat{y}_i}{y_i} \times 100\% \tag{15}$$

These are the highly popular measures to evaluate the accuracy of forecasting in statistics and data analysis.

Result Analysis

For result analysis and evaluation, PySpark (Python API for Apache Spark) is used with matplotlib, NumPy, and pandas libraries. The Anaconda Python distribution contains the best suitable packages, modules, and libraries for data science, and it is used in Linux (Ubuntu 20.04) operating system.

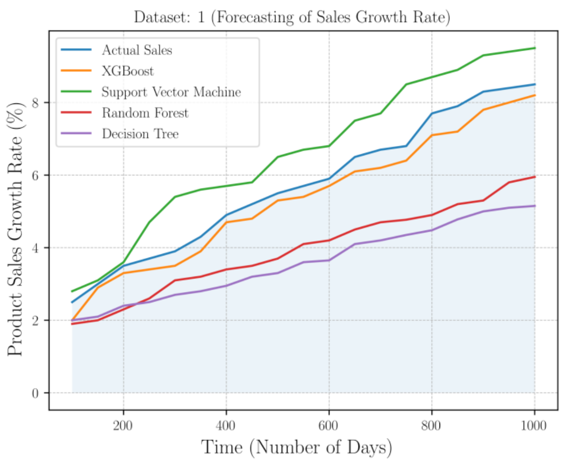

Figure 3: Forecasting Analysis (Dataset 1)

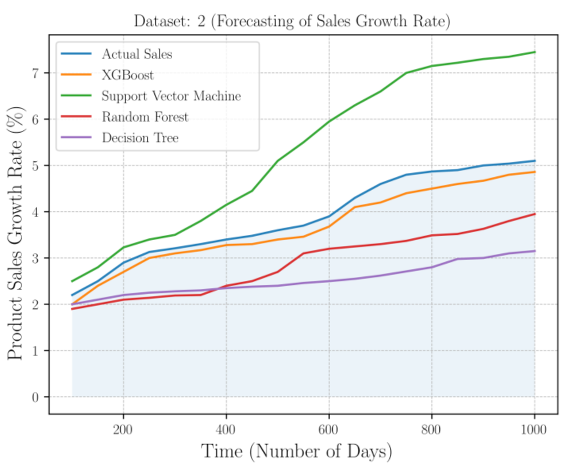

Figure 4: Forecasting Analysis (Dataset 2)

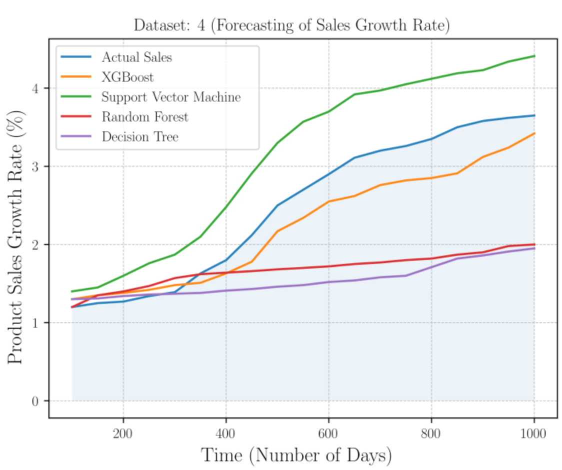

Figure 6: Forecasting Analysis (Dataset 4)

The performance of the proposed "extreme gradient boosting (XGBoost)" model is compared with "support vector machine (SVM)," "random forest," and "decision tree" algorithms. For simulation analysis, four different e-commerce datasets are selected from big data.

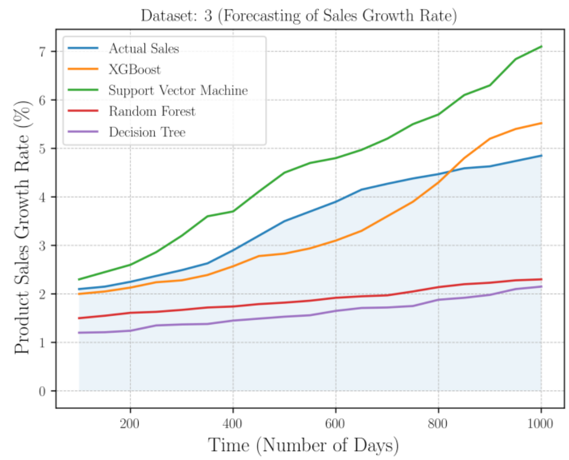

Figure 5: Forecasting Analysis (Dataset 3)

Table 3: Forecasting Error

Model	*MAPE*	*MAE*	*RE*
Decision Tree	17.85%	68.64	38.11%
Random Forest	13.32%	57.38	29.56%
SVM	9.67%	35.66	18.34%
XGBoost	6.89%	28.47	11.87%

Figures 3 to 6 represent the simulation results to evaluate the forecasting of sales growth rate during 1,000 days' time period. The blue, orange, green, red, and purple lines represent the growth rate of actual sales, XGBoost, support vector machine (SVM), random forest, and decision tree-based forecasting.

By observing Figure 3 to 6, it is obvious that the forecasting rate of the XGBoost model is closer to the rate of actual sales as compared to support vector machine (SVM), random forest, and decision tree algorithms. The proposed XGBoost model presents higher forecasting accuracy, which is nearer to the actual sales rate.

The overall forecasting errors are presented in Table 3 using "mean absolute percent error (MAPE)," "mean absolute error (MAE)," and "relative error (RE)." Decision tree algorithm presents 17.85%, random forest algorithm presents 13.32%, support vector machine (SVM) presents 9.67%, and XGBoost algorithm presents 6.9% mean absolute percent error (MAPE). The mean absolute error is higher for the decision tree whereas, XGBoost presents lower as 28.47%.

Also, the relative error is minimum for XGBoost as compared to a decision tree, random forest, and SVM. The XGBoost presents minimum error with higher forecasting accuracy in e-commerce big data analysis.

Conclusion

The research helped understand and solidify several machine learning models for predictive data analysis. Various techniques and models come with their own set of benefits and demerits. No perfect model provides highly efficient data analysis. Accordingly, after the comparative analysis and running of various algorithms such as decision tree, random forest, and support vector machines, better results are obtained with XGBoost based model as presented in the result. However, the proposed model based on XGBoost algorithm has some error also. Further, various factors directly affect the forecasting of the e-commerce products sales and recommendations, so, in the future, these factors such as over-fitting of data can be considered in big data analysis. The correlated features must be robust for over-fitting prevention to obtain better results.

References

[1] H. Xia, S. Tang, S. Li, and X. Yu, "Application research of big data e-commerce in closed community," in *Proceedings of the 2018 International Conference on Internet and E-Business*, ser. ICIEB'18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 43–46. doi: https://doi.org/10.1145/3230348.3230425

[2] M. Vanauer, C. B¨ohle, and B. Hellingrath, "Guiding the introduction of big data in organizations: A methodology with business- and data-driven ideation and enterprise architecture management-based implementation," in 2015, pp. 908–917. doi: https://doi.org/10.1109/HICSS.2015.113

[3] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015. doi: https://doi.org/10.1016/j.ijinfomgt.2014.10.007

[4] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management," Journal of Business Logistics, vol. 34, no. 2, pp. 77–84, 2013. doi: https://doi.org/10.1111/jbl.12010

[5] N. K. Gimenez Isasi, E. Morosini Frazzon, and M. Uriona, "Big data and business analytics in the supply chain: A review of the literature," IEEE Latin America Transactions, vol. 13, no. 10, pp. 3382–3391, 2015. doi: https://doi.org/10.1109/TLA.2015.7387245

[6] J. C. Miguel and M. A'. Casado, GAFAnomy (Google, Amazon, Facebook and Apple): The Big Four and the b-Ecosystem. Cham: Springer International Publishing, 2016, pp. 127–148. doi: https://doi.org/10.1007/978-3-319-31147-0 4

[7] S. Mazumder, Big Data Tools and Platforms. Cham: Springer International Publishing, 2016, pp. 29–128. doi: https://doi.org/10.1007/978-3-319-27763-9_2

[8] Y. Arfat, S. Usman, R. Mehmood, and I. Katib, Big Data Tools, Technologies, and Applications: A Survey. Cham: Springer International Publishing, 2020, pp. 453–490. doi: https://doi.org/10.1007/978-3-030-13705-2_19

[9] F. Zhu and X. M. Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," Journal of Marketing, vol. 74, no. 2, pp. 133–148, 2010. doi: https://doi.org/10.1509/jm.74.2.133

[10] B. Schmarzo, Big Data: Understanding How Data Powers Big Business. Wiley, 2013.

[11] J. Liebowitz, Big Data and Business Analytics. Auerbach Publications, 2013.

[12] P. Simon, Too Big to Ignore: The Business Case for Big Data, 1st ed. Wiley, 2013.

[13] E. Stubbs, Big Data, Big Innovation: Enabling Competitive Differentiation through Business Analytics. Wiley, 2014.

[14] S. Sakr, Z. Maamar, A. Awad, B. Benatallah, and W. M. P. Van Der Aalst, "Business process analytics and big data systems: A roadmap to bridge the gap," IEEE Access, vol. 6, pp. 77 308–77 320, 2018. doi: https://doi.org/10.1109/ACCESS.2018.2881759

[15] A. Vera-Baquero, R. Colomo-Palacios, and O. Molloy, "Business process analytics using a big data approach," IT Professional, vol. 15, no. 6, pp. 29–35, 2013. doi:https://doi.org/10.1109/MITP.2013.60

[16] S. Williams, Business Intelligence Strategy and Big Data Analytics. A General Management Perspective, 1st ed. Morgan Kaufmann, 2016.

[17] S. S. Alrumiah and M. Hadwan, "Implementing big data analytics in e-commerce: Vendor and customer view," IEEE Access, vol. 9, pp. 37 281–37 286, 2021. doi: https://doi.org/10.1109/ACCESS.2021.3063615

[18] J. Dean, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. Wiley, 2014.

[19] M. S. Daskin, C. R. Coullard, and Z.-J. M. Shen, "An inventory-location model: Formulation, solution algorithm and computational results," Annals of Operations Research, vol. 110, no. 1, pp. 83–106, Feb 2002. doi: https://doi.org/10.1023/A:1020763400324

[20] T. Ragg, W. Menzel, W. Baum, and M. Wigbers, "Bayesian learning for sales rate prediction for thousands of retailers," Neurocomputing, vol. 43, no. 1, pp. 127–144, 2002, selected engineering applications of neural networks. doi: https://doi.org/10.1016/S0925-2312(01)00624-5

[21] L. Ren, L. Zhang, H. Wang, and Q. Guo, "An extreme gradient boosting algorithm for short-term load forecasting using power grid big data," in Proceedings of 2018 Chinese Intelligent Systems Conference, Y. Jia, J. Du, and W. Zhang, Eds. Singapore: Springer Singapore, 2019, pp. 479–490. doi:https://doi.org/10.1007/978-981-13-2288-4_46

[22] S. Islam and S. H. Amin, "Prediction of probable backorder scenarios in the supply chain using distributed random forest and gradient boosting machine learning techniques," Journal of Big Data, vol. 7, no. 1, p. 65, Aug 2020. doi: https://doi.org/10.1186/s40537-020-00345-2

[23] Z. Xia, S. Xue, L. Wu, J. Sun, Y. Chen, and R. Zhang, "Forexgboost: passenger car sales prediction based on xgboost," Distributed and Parallel Databases, vol. 38, no. 3, pp. 713–738, Sep 2020. doi: https://doi.org/10.1007/s10619-020-07294-y

[24] A. Y. L. Chong, E. Ch'ng, M. J. Liu, and B. Li, "Predicting consumer product demands via big data: the roles of online promotional marketing and online reviews," International Journal of Production Research, vol. 55, no. 17, pp. 5142–5156, 2017. doi: https://doi.org/10.1080/00207543.2015.1066519

[25] T. Boone, R. Ganeshan, A. Jain, and N. R. Sanders, "Forecasting sales in the supply chain: Consumer analytics in the big data era," International Journal of Forecasting, vol. 35, no. 1, pp. 170–180, 2019, special Section: Supply Chain Forecasting. doi: https://doi.org/10.1016/j.ijforecast.2018.09.003

[26] V. Sohrabpour, P. Oghazi, R. Toorajipour, and A. Nazarpour, "Export sales forecasting using artificial intelligence," Technological Forecasting and Social Change, vol. 163, p. 120480, 2021. doi: https://doi.org/10.1016/j.techfore.2020.120480

[27] H. Yuan, W. Xu, Q. Li, and R. Lau, "Topic sentiment mining for sales performance prediction in e-commerce," Annals of Operations Research, vol. 270, no. 1, pp. 553–576, Nov 2018. doi: https://doi.org/10.1007/s10479-017-2421-7

[28] S. Palanimalai and I. Paramasivam, "Big data analytics bring new insights and higher business value - an experiment carried out

to divulge sales forecasting solutions," International Journal of Advanced Intelligence Paradigms, vol. 8, no. 2, pp. 207–218, 2016. doi: https://doi.org/10.1504/IJAIP.2016.075728

[29] D. Kılınc¸, "A spark-based big data analysis framework for real-time sentiment prediction on streaming data," Software: Practice and Experience, vol. 49, no. 9, pp. 1352–1364, 2019. doi: https://doi.org/10.1002/spe.2724

[30] J. Eapen, D. Bein, and A. Verma, "Novel deep learning model with cnn and bi-directional lstm for improved stock market index prediction," in 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0264–0270. doi: https://doi.org/10.1109/CCWC.2019.8666592

[31] C. Oliver, "Sustainable competitive advantage: combining institutional and resource-based views," Strategic Management Journal, vol. 18, no. 9, pp. 697–713, 1997. doi: https://doi.org/10.1002/(SICI)1097-0266(199710)18:9<697::AID-SMJ909>3.0.CO;2-C

[32] R. Dubey, A. Gunasekaran, S. J. Childe, C. Blome, and T. Papadopoulos, "Big data and predictive analytics and manufacturing performance: Integrating institutional theory, resource-based view and big data culture," British Journal of Management, vol. 30, no. 2, pp. 341–361, 2019. doi: https://doi.org/10.1111/1467-8551.12355

[33] S. A. Ramcharan Kakarla, Sundar Krishnan, Applied Data Science Using PySpark: Learn the End-to-End Predictive Model- Building Cycle, 1st ed. Apress, 2021.

[34] X. Jin and J. Han, K-Means Clustering. Boston, MA: Springer US, 2010, pp. 563–564. doi: https://doi.org/10.1007/978-0-387-30164-8 425