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.

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, five basic properties, called the 5 V’s: "Volume, Velocity, Variety, Veracity, and Value," are frequently characterized by big data.

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. 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. 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.

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. 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...
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. 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.

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. 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.

**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.

Ragg et al. 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. 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 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. 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. 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.
Extreme Gradient Boosting Model-based Forecasting of Big Data Online Sales Record

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.

Kılınc\[29\] 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

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

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.

\[ d(m) = w_1 d(m_1) + w_2 d(m_2) + ... + w_n d(m_n) \] (1)

Where, \( d(m) \) represents the missing data, \( [d(m_1), d(m_2), ..., d(m_n)] \) represents the set of data vectors to \( d(m) \), and \( [w_1, w_2, ..., w_n] \) 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

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Descriptions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product_ID</td>
<td>Product identification number</td>
<td>MATH#537</td>
</tr>
<tr>
<td>Dept_ID</td>
<td>Dept. ID Department of the product</td>
<td>Book &amp; Magazine</td>
</tr>
<tr>
<td>Brand_ID</td>
<td>Brand name of the product</td>
<td>Pearson India</td>
</tr>
<tr>
<td>Price</td>
<td>Price of the product</td>
<td>899.00 ₹</td>
</tr>
<tr>
<td>Dim</td>
<td>Dimensions (size, weight etc.) of the product</td>
<td>10x10x15 cm³, 700 gm</td>
</tr>
<tr>
<td>Sale Date</td>
<td>Date of the sale</td>
<td>Mar-21-2021</td>
</tr>
<tr>
<td>Trans_ID</td>
<td>Transaction number of the product</td>
<td>T784653</td>
</tr>
<tr>
<td>Sale_Qty</td>
<td>Quantities of the sale</td>
<td>350</td>
</tr>
<tr>
<td>City</td>
<td>Location of the buyer</td>
<td>Bhopal</td>
</tr>
<tr>
<td>Review</td>
<td>Review given as feedback</td>
<td>Good, 4 Stars</td>
</tr>
</tbody>
</table>

Table 2: Examples of Sales Datasets

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

$$\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$$

$$S(m) = \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} d(m_i)$$

Secondly, the rate of deviation is calculated as

$$\rho(m) = \frac{|d(m) - \bar{m}|}{\sigma}$$

The corrected datasets $d'(m_i)$ are calculated using the outlier correction process on datasets as follows

$$d'(m_i) = \frac{d(m_i) - \bar{m}}{\sigma}$$

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

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

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

Clustering Analysis of Datasets

For clustering of datasets, k-means 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.

$$\text{Dist}(m_i, m_j) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2}$$

Where Dist($m_i, m_j$) represents a distance between $i^{th}$ and $j^{th}$ data samples, whereas, $x_{i,k}, x_{j,k}$ 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

$$\hat{y}_i = \sum_{k=1}^{K} \hat{f}_k(x_i)$$

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, $\hat{f}_k$ represents forecasting function for $k^{th}$ decision tree, and it is calculated as

$$\hat{f}_k(x_i) = w_{i,j} \cdot q: \mathbb{R}^m \rightarrow \mathbb{R}$$

For $k^{th}$ decision tree, the structure-function is represented by $q(x)$, which maps $x$ 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

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

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

$$w(f) = yT + \frac{||w||^2}{2}$$

where, $y$ represents complex parameter, $\lambda$ 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

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\text{RE} = \frac{y_i - \hat{y}_i}{y_i} \times 100\%$$

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.
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.

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.

**Table 3: Forecasting Error**

<table>
<thead>
<tr>
<th>Model</th>
<th>MAPE</th>
<th>MAE</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>17.85%</td>
<td>68.64</td>
<td>38.11%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>13.32%</td>
<td>57.38</td>
<td>29.56%</td>
</tr>
<tr>
<td>SVM</td>
<td>9.67%</td>
<td>35.66</td>
<td>18.34%</td>
</tr>
<tr>
<td>XGBoost</td>
<td>6.89%</td>
<td>28.47</td>
<td>11.87%</td>
</tr>
</tbody>
</table>

**Conclusion**

The research helped understand and solidify several machine learning models for predictive data analysis. Various
Extreme Gradient Boosting Model-based Forecasting of Big Data Online Sales Record

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


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


