

Content is available at: CRDEEP Journals Journal homepage: http://www.crdeepjournal.org/category/journals/ijrem/

ngineering an

SJIF: 4.45

International Journal of Research in Engineering and Management (ISSN: 2456-1029)

A Peer Reviewed UGC Approved Quarterly Journal

Research Paper Linear Regression based Sales Prediction using Web Usage Big Data

Pushpalata Patel and Prabhat Pandey*

Department of Computer Science Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India

ARTICLE DETAILS	ABSTRACT
Corresponding Author: Prabhat Pandey	The exponential growth of web usage has generated huge amounts of user interaction data, which offers valuable insights into consumer behavior and market trends. As online shopping continues to grow, e-commerce strategies have increasingly adopted modern
<i>Key words:</i> Web Usage Big Data, E- Commerce, Review Rating, Customers Review, Linear Regression, Sales Prediction	technologies, such as big data analytics and machine learning. Predictive analysis of customer reviews using big data from web usage is becoming a key element in shaping online marketing tactics. This research presents a linear regression-based approach for predicting sales using big data from web usage. A clear connection exists between customer review ratings and product sales. We extracted the key feature review ratings from the product review sections on e-commerce sites. This article presents a sales prediction model using linear regression to support e-commerce marketing using customer review data from web usage big data sources. The proposed model allows forecasting initial sales and repeat purchases via review analysis. The linear regression model was trained and evaluated on real-world e-commerce review rating data, demonstrating its effectiveness in capturing correlations between sales growth and review ratings

1. Introduction

Customers' purchasing habits have shifted toward e-commerce as a result of recent technological advancements, faster delivery, and better review systems in Internet technology, which is an ideal platform for the socialization of e-commerce applications [1, 2]. Big data from web usage can help competitive e-commerce businesses improve their analytical decision making and intelligence strategies so that effective demand and sales prediction and supply chain management can be designed [3]. The customers review analysis improves the performance of product sales prediction and it is the potential step in the e-commerce business. The review section of ecommerce platforms contains Big Data with a huge volume of customer-contributed comments, including feedback, reviews, experiences, and ratings regarding products that can be analyzed for sales prediction [4, 5]. Large, globally popular e-commerce platforms such as Amazon, eBay, Alibaba, Walmart, Shopify, and others are constantly launching initiatives to increase sales through customer experience, feedback, and reviews [6, 7]. In e-commerce, customers cannot touch, feel, and interact physically with the products, so they have concerns and doubts about the features, functions, and qualities of the products. Therefore, reviews and ratings of products play a potential role in online shopping [8–10]. Nowadays, it is crucial for the e-commerce industry to analyze a large number of credible product reviews in the form of Big Data from web usage so that customers can make purchasing decisions based on review ratings [11].

As an illustrative method to promote the customers' trust in products, the online review, rating, suggestion, and feedback system has become a potential and distinguished department in e-commerce [12, 13]. E-commerce industries want customers to be satisfied with their purchases, to have as few returns as possible, and to repurchase as frequently as possible, so they try to provide online platforms that maximize product certainty for customers while balancing the demand-supply ratio [14]. Product review ratings have appeared as potential information sources that influence prepurchase aspects of the products. The review analysis assists social media and community to improve perceptions of online business [15, 16]. Review rating represents a valuable potential path to success for firms by analyzing customers'

^{*}Author can be contacted at: Department of Computer Science Awadhesh Pratap Singh University, Rewa, Madhya Pradesh, India

Received: 20-04-2025; Sent for Review on: 27-04- 2024; Draft sent to Author for corrections: 10-05-2025; Accepted on: 18-05-2025; Online Available from 30-05-2025

Patel & Pandey

behaviors, attitudes, habits, and experiences with the products they used in real time and planning marketing strategies, manufacturing, distribution, and overall supply chain management accordingly [17].



Fig. 1. Detailed Review and Comment based Product Analysis

The products can be analyzed based on some particular factors (Figure 1) that are based on the detailed reviews and comments by the customers, including:

• Client: Based on the reviews and comments, the target clients can be identified for the target market in ecommerce.

• Size: For proper supply chain management, the products' portability and convenience can be predicted.

• Function: Review analysis can improve a product's function, workings, and performance.

• Cost: In the competitive e-commerce market, cost is a critical factor. Analyzing the cost-related discussion and review is always beneficial for product sales at an optimized price.

• Manufacture: The materials used in product manufacturing are always identified by the customers after utilizing them. The manufacturing materials are associated with the quality of the products.

• Aesthetics: The aesthetics (looks, colors, themes) is importance factor that extremely reviewed by the customers.

• Safety: Safety features and risk factors are the crux of discussion in social media, and reviews are important to improve product quality.

A potential customer deciphers between credible and non credible reviews, and finds out helpful and useful information regarding the products that are to be purchased. The customer created information and opinions about products in the review section of an e-commerce platform are the pathway for other customers, and they are more valuable and beneficial than retailers' or sellers' descriptions or information [18].

The rest of this article is structured as follows: The importance of review ratings given by customers about products on an e-commerce platform is presented in section II. section III discusses the related work about sales prediction using web usage big data analytics of product reviews. The overview of the proposed linear regression-based model and the illustrative computation are given in section IV. section V reports the simulation results of the proposed method, which represent sales prediction. Finally, the conclusion and future directions of this research work are presented in section VI.

2. Importance Of Review Rating In E-Commerce

The most effective way to minimize customers' concerns about the quality of a product is through review ratings that influence the majority of customers in their online shopping [19]. If credibility is considered, then the review rating of the product plays a critical role for e-commerce industries, and the popular brand name of the product does not matter. More than 90% of online customers read the product's review before deciding to purchase, so review-rating is a significant factor in e-commerce [20–22]. Almost all the online e-commerce platforms facilitate a review section for products that permits the customers to provide feedback, share their experiences, and write reviews. There are possibilities of fake review also, false reviews can be a issue for online shoppers, as the growth of digital marketplaces has led to a rise in misleading feedback aimed at attracting or deterring potential buyers [23]. The review ratings help other customers to get clear details about the particular products before purchasing them. They can make their own decision about whether it is worth shopping for or not [24].

There are several reasons for the importance of review ratings in e-commerce businesses; the Figure 2 depicts these major factors as follows:

• Trust Building: The review rating builds trust in the product, and customers first look for reviews before putting it in their online shopping cart. Positive reviews influence customers' purchasing decisions because they place their trust in the product and its brand.

• Better Perception of Product: Customers' reviews for the products help others form perceptions or deep insights about the product's performance, flaws, benefits, and much more.

• Product Issue Rectification: The issues related to products can be rectified using reviews. For a popular branded product, if many customers are identifying and pointing out similar types of issues, then the manufacturer can rectify the defect to resolve that issue. Customers' reviews and feedback assist the manufacturer in detecting flaws in the products and providing improvements.

• Social/Community Proof: The product reviews or testimonials from social networking sites are beneficial to promoting the brand, which can attract many potential customers and grow sales.



Fig. 2. Importance of Review Rating on Products in E-Commerce

• Conversion Boosting: Positive reviews by customers can enhance the product conversion rate effectively. The positive remarks given by the customers encourage many other customers to purchase or repurchase that product.

• Less Possibility of Product Return: Products with positive reviews, feedback, and ratings may face fewer returns because they have a large number of satisfied customers. After reading positive reviews, customers will be confident enough in the product's quality to purchase it.

3. Related Work

With the progression of Internet technologies, the habits of customers have changed. Online marketing and sales are very common today and create a vast amount of data. Various researchers have proposed their methods to predict business strategies using web usage big data analytics.

Wang et al. [25] suggested a big data analytical method for the transformation of customers' record and data from several sources into valuable information. The purchasing patterns of the customers can be predicted using segmented customer records so that business strategies can be developed for appropriate marketing.

Hallikainen et al. [26] used big data analytics to manage business-to-business (B2B) customers' relations and verified the significance of big data analysis on the performance of customers relations and the emergence of sales using industrial datasets. They investigated the impact of analytics culture on firm moderation. They found that big data analysis considerably increased sales growth.

Zheng and Huang [27] examined data from the "Baidu Index" and the "360 Index" from January to October 2015 and discovered a link between fresh energy vehicle sales and the network's search index. Their outcomes present better sales forecasting using network search indexing as compared to without indexing. In a real-world scenario,

Zhu et al. [28] proposed "Gated Recurrent Units neural network with Reinforcement Learning (GRURL)" for improving the accuracy of car sale forecasting. Their proposed model integrates reinforcement learning with "Gated Recurrent Units (GRU)" to optimize the parameters of network update in accordance with dynamic input data. Based on the previous data, they applied a model to predict car sales in Yancheng City, China, and compared it with conventional deep learning techniques.

Koochakpour and Tarokh [29] applied the profit from neural network capabilities to the "adaptive neuro fuzzy inference system (ANFIS)" and "particle swarm optimization (PSO)". This model uses both of these abilities and features to optimize each component of the system so that it generates more accurate outcomes.

Lingxian et al. [30] proposed an integrated framework using a neural network and K-mean for the prediction of online retail sales. They first used K-mean clustering as an unsupervised learning technique for the classification of UK retail data and then used it as an auxiliary attribute to predict the model. Secondly, they used an innovative time sequence model named "Long Short Term Memory" for the future prediction of online retail sales.

Aguilar-Palacios et al. [31] proposed a novel predictive model inspired by the know-how of commercial business analysts and forecasters that would undertake the issues for the prediction of sales on a day-by-day basis. They used k-NN technique, which generates interpretative and easily modifiable predictions. They integrated the feature selection with prediction, accomplished by the cost function, which is reduced with a "non-negative least squares (NNLS)" solver. This technique is supported by online learning, so the forecasting of sales is measured with the current available data, so it avoids retraining the model. Hajek and Abedin [32] proposed "profit function maximizing inventory back-order prediction system" with the help of Big Data analytics. Four basic steps are followed in their proposed approach: first, an improved variant of "clustering-based under-sampling (CBUS)" technique is accomplished, which uses the k-means method for balancing inventory back-order cases. Second, the profit-based classification measurement is determined, which estimates the profits and costs of inventory back-order. Then, the training phase is started using machine learning algorithms using cluster specific classifiers or balanced training datasets to increase data subset accuracy. Finally, a search process based on a genetic algorithm is applied for optimization of the profitability measurement.

Kumari et al. [33] proposed a model for accurately predicting sales of the outlet's items using a two-level statistical method that minimizes the value of the mean absolute error up to 39.17%. The performance of the two-level statistical method is proven to be better as compared to single predictive methods, and it contributes higher predictions to the Big Mart dataset.

4. Proposed Method

The review ratings of a particular product are based on the feedback and opinions of the consumers. Due to the importance of the review rating, many e-commerce industries maintain a review department on their system to permit customer reviews, feedback, ratings, and experiences of the product they purchased online. There is a linear relationship between rating and sales of the products, so review ratings help the customers form a certain opinion of the particular product before purchasing it, and finally, the business model can be developed for the products. Table I represents the review ratings of products by the customers on the e-commerce platform and their impact on business. The review rating of the product determines the significance level, customers' views on the product, business impact, and product recommendation with repurchasing probability. The five-star-rated products are excellent, highly recommended, and demanded products, and their repurchasing probability is greater than 80%. Similarly, the four-star-rated products are also considered awesome, demanded, and highly recommended. Three-star-rated products are considered good, whereas two- and one-star-rated products are of poor quality and have little market impact.

Table 1. Review Rating Of Products By The Customers In E-Commerce

Review Rating	Custormers' View on Product	Significant Level	Business Impact	Recommondation/ Repuchasing Probability
****	Excellent Product	Highly Demanded	Highly Recommended Product	> 0.8
★★★★☆	Awesome Product	Demanded	Highly Recommended Product	0.5 - 0.8
★★★☆☆	Good Product	Demanded	Average Market Impact	0.2 - 0.5
★★☆☆☆	Poor Quality Product	Least Demanded	Poor Market Impact	0.05 < 0.2
$\bigstar \diamond \diamond \diamond \diamond$	Worst Quality Product	Ignored	Poor Market Impact	< 0.05

For analyzing the online business model, the customers' review analysis is an essential step. There is a linear relationship between the customer's product rating and the product's repurchasing probability. The system model of this relation can be expressed in mathematical equations based on Linear Regression. For the total number of reviews n, the mean value for customer review rating of the product Rrev and product sales Sp can be calculated as

$$\overline{R_{\text{rev}}} = \frac{1}{5} \sum_{i=1}^{5} R_{\text{rev}}$$
(1)

$$\overline{S_{\rm p}} = \frac{1}{n} \sum_{i=1}^{n} S_{\rm p} \tag{2}$$

For the best fit line to predict sales, the slop m can be calculated as

$$m = \frac{\sum_{i=1}^{n} \left(R_{\text{Rev}} - \overline{R_{\text{Rev}}} \right) \left(S_{\text{p}} - \overline{S_{\text{p}}} \right)}{\sum_{i=1}^{n} \left(R_{\text{rev}} - \overline{R_{\text{rev}}} \right)^{2}}$$
(3)

Finally the prediction model as presented in Equation 4, where sales are predicted based on the review rating of products by the customers as

$$\overline{S_p} = m(\overline{R_{rev}}) + c$$
 (4)

where c is the constant that represents vertical intercept in the linear relationship.

In this proposed model, initially the customers' reviews and their corresponding sales records are given as the training datasets to the Linear Regression model because product sales highly depend on the review ratings of products given by the customers. The proposed sales prediction model is shown in Figure 3. The previous values (testing datasets) of review rating and sales of products are provided to the Linear Regression model, which then creates a prediction model. Using this model, sales can be predicted based on their corresponding review ratings, as higher-rated products are best suited

for resale in an online e-commerce system. The higherrated products have a higher resale probability as compared to the lower-rated products.



Fig. 3. Proposed Model for Sales Prediction

5. Result Analysis

The simulations were carried out using the Python programming language with PySpark (an effective framework for Big Data) in the Linux operating system. The Big Data required for Linear Regression in this article is based on review ratings of products and its corresponding sales is acquired from Kaggle (https://www.kaggle.com) containing datasets of e-commerce customers review and rating. The Big Data required for Linear Regression in this article is based on review ratings of products and their corresponding sales and was acquired from Kaggle (https://www.kaggle.com), which contains datasets of e-commerce customers' reviews and ratings. Figure 4 shows a heatmap of review-rating-based product sales, as well as a confusion matrix visualization of review rating and product sales. The review ratings were taken from 1 to 5, with an interval of 0.5.



Fig. 4. Heatmap of Review-Rating based Sales

The Figure 5 represents the result of a linear regression model for sales prediction, where the red line represents the best fit regression line (the least error line). Based on the review rating of the product, sales predictions can be estimated using this model.



Fig. 5. Linear Regression Model for Sales Prediction

6. Conclusion and Future Direction

Many markets and businesses are based on online shopping (ecommerce), and customer satisfaction is the key to business success. Customers who are satisfied leave positive reviews, which can help increase product sales. This article presents a Linear Regression analysis-based model to predict product sales using its review ratings. The other factors, such as

customer sentiments, discussion in social media, and fake reviews, also impact e-commercebased product sales. Deep learning and natural language processing techniques can also be used to extract deep insights from Big Data. R

References

1. Ahn, H.-I., & Spangler, W. S. (2014). Sales prediction with social media analysis. *2014 Annual SRII Global Conference*, 213–222. <u>https://doi.org/10.1109/SRII.2014.37</u>

2. Basalla, M., Schneider, J., Luksik, M., Jaakonmäki, R., & vom Brocke, J. (2021). On latency of e-commerce platforms. *Journal of Organizational Computing and Electronic Commerce, 31*(1), 1–17. https://doi.org/10.1080/10919392.2021.1882240

3. Benlian, A., Titah, R., & Hess, T. (2012). Differential effects of provider recommendations and consumer reviews in ecommerce transactions: An experimental study. *Journal of Management Information Systems*, 29(1), 237–272. https://doi.org/10.2753/MIS0742-1222290107

4. Bosman, D. J., Boshoff, C., & van Rooyen, G.-J. (2013). The review credibility of electronic word-of-mouth communication on e-commerce platforms. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, 22(3), 29–44. <u>https://journals.co.za/doi/abs/10.10520/EJC142270</u>

5. Changchit, C., & Klaus, T. (2020). Determinants and impact of online reviews on product satisfaction. *Journal of Internet Commerce, 19*(1), 82–102. <u>https://doi.org/10.1080/15332861.2019.1672135</u>

6. Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10), 1407–1424. <u>https://doi.org/10.1287/mnsc.49.10.1407.17308</u>

7. Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23–45. <u>https://doi.org/10.1002/dir.20087</u>

8. Hajek, P., & Abedin, M. Z. (2020). A profit function-maximizing inventory backorder prediction system using big data analytics. *IEEE Access*, *8*, 58982–58994. <u>https://doi.org/10.1109/ACCESS.2020.2983118</u>

9. Hajli, M. N. (2014). A study of the impact of social media on consumers. *International Journal of Market Research*, 56(3), 387–404. <u>https://doi.org/10.2501/IJMR-2014-025</u>

10. Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management, 86*, 90–98. <u>https://doi.org/10.1016/j.indmarman.2019.12.005</u>

11. Huang, S., Shen, D., Feng, W., Zhang, Y., & Baudin, C. (2009). Discovering clues for review quality from author's behaviors on e-commerce sites. *Proceedings of the 11th International Conference on Electronic Commerce (ICEC '09)*, 133–141. <u>https://doi.org/10.1145/1593254.1593274</u>

12. Jensen, M. L., Averbeck, J. M., Zhang, Z., & Wright, K. B. (2013). Credibility of anonymous online product reviews: A language expectancy perspective. *Journal of Management Information Systems, 30*(1), 293–324. https://doi.org/10.2753/MIS0742-1222300109

13. Kadry, S., & Hami, A. E. (2016). Innovations in E-Systems for Business and Commerce (1st ed.). Apple Academic Press.

14. Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on amazon.in. *Journal of Retailing and Consumer Services*, 45, 21–32. https://doi.org/10.1016/j.jretconser.2018.08.002

15. Koochakpour, K., & Tarokh, M. (2017). Designing sales budget forecasting and revision system by using optimisation methods. *International Journal of Knowledge Engineering and Data Mining*, 4(2), 93–113. https://doi.org/10.1504/IJKEDM.2017.086225

16. Laudon, K. C., & Traver, C. G. (2022). E-commerce: Business, Technology and Society (17th ed.). Pearson.

17. Lau, R. Y. K., Zhang, W., & Xu, W. (2018). Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management*, *27*(10), 1775–1794. <u>https://doi.org/10.1111/poms.12737</u>

18. Lingxian, Y., Jiaqing, K., & Shihuai, W. (2019). Online retail sales prediction with integrated framework of k-mean and neural network. *Proceedings of the 2019 10th International Conference on E-Business, Management and Economics (ICEME 2019)*, 115–118. <u>https://doi.org/10.1145/3345035.3345048</u>

19. Liu, C., Wang, S., & Jia, G. (2020). Exploring e-commerce big data and customer-perceived value: An empirical study on Chinese online customers. *Sustainability*, *12*(20), 8649. <u>https://doi.org/10.3390/su12208649</u>

20. Mewada, A., Dewang, R. K., Goldar, P., & Maurya, S. K. (2023). Sentibert: A novel approach for fake review detection incorporating sentiment features with contextual features. *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing (IC3-2023)*, 230–235. <u>https://doi.org/10.1145/3607947.3607991</u>

21. Park, D.-H., Lee, J., & Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, *11*(4), 125–148. https://doi.org/10.2753/JEC1086-4415110405

22. Punam, K., Pamula, R., & Jain, P. K. (2018). A two-level statistical model for big mart sales prediction. *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 617–620. https://doi.org/10.1109/GUCON.2018.8675060

23.Qin, Z., Shuai, Q., Wang, G., Zhang, P., Cao, M., & Chen, M. (2022). *E-Commerce: Concepts, Principles, and Application* (Advanced Studies in E-Commerce). Springer. <u>https://doi.org/10.1007/978-981-19-6438-1</u>

24. Qu, Z., Zhang, H., & Li, H. (2008). Determinants of online merchant rating: Content analysis of consumer comments about Yahoo merchants. *Decision Support Systems*, *46*(1), 440–449. <u>https://doi.org/10.1016/j.dss.2008.08.004</u>

25. Ramanathan, U., Subramanian, N., & Parrott, G. (2017). Role of social media in retail network operations and marketing to enhance customer satisfaction. *International Journal of Operations & Production Management, 37*(1), 105–123. https://doi.org/10.1108/IJOPM-03-2015-0153 26. Roy, G., Datta, B., & Basu, R. (2017). Trends and future directions in online marketing research. *Journal of Internet Commerce*, *16*(1), 1–31. <u>https://doi.org/10.1080/15332861.2016.1258929</u>

27. Ullal, M. S., Spulbar, C., Hawaldar, I. T., Popescu, V., & Birau, R. (2021). The impact of online reviews on e-commerce sales in India: A case study. *Economic Research-Ekonomska Istraživanja*, 34(1), 2408–2422. https://doi.org/10.1080/1331677X.2020.1865179

28. Wang, S.-C., Tsai, Y.-T., & Ciou, Y.-S. (2020). A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network. *Journal of Industrial Information Integration*, *20*, 100177. https://doi.org/10.1016/j.jii.2020.100177

29.Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, 24(4), 720–734.

30.Zhang, R., & Tran, T. T. (2009). Helping e-commerce consumers make good purchase decisions: A user reviews-based approach. In Babin, G., Kropf, P., & Weiss, M. (Eds.), *E-Technologies: Innovation in an Open World* (pp. 1–11). Springer. <u>https://doi.org/10.1007/978-3-642-01187-0 1</u>

31. Zheng, S., & Huang, J.-H. (2018). New energy vehicles sales prediction method and empirical research under the environment of big data. In Li, X., & Xu, X. (Eds.), *Proceedings of the Fifth International Forum on Decision Sciences* (pp. 295–306). Springer. <u>https://doi.org/10.1007/978-981-10-7817-0_26</u>

32. Zhu, B., Dong, H., & Zhang, J. (2019). Car sales prediction using gated recurrent units neural networks with reinforcement learning. In Cui, Z., Pan, J., Zhang, S., Xiao, L., & Yang, J. (Eds.), *Intelligence Science and Big Data Engineering: Big Data and Machine Learning* (pp. 312–324). Springer. <u>https://doi.org/10.1007/978-3-030-36204-1_26</u>

33. Aguilar-Palacios, C., Munoz-Romero, S., & Rojo-Alvarez, J. L. (2019). Forecasting promotional sales within the neighbourhood. *IEEE Access*, *7*, 74759–74775. <u>https://doi.org/10.1109/ACCESS.2019.2920380</u>