SALES PREDICTION OF E-COMMERCE USING MACHINE LEARNING ALGORITHM EXTREME GRADIENT BOOST IN COMPARISON WITH K-NEIGHBOR REGRESSION TO IMPROVE ACCURACY .

Sumanth Mekala[1], S. Ashok Kumar[2*]

Article History: **Received:** 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: The main objective of this study is to predict the Sales of E-commerce using Extreme gradient boosting which is a base model and compared with Linear regression algorithms.
Materials and Methods: Extreme gradient boosting and K-neighbor regression Algorithms are used to predict the Sales of E-commerce. Sample size is calculated using G Power calculator and found to be 25 per group has been taken and a total of 50 samples are used. Where Pretest power is 80% and CI of 95%.
Results: Based on the analysis Extreme gradient boosting has significantly more accuracy (90.50) compared to K-neighbor regression algorithm (80.50). There is a Statistically Significant difference between the two groups with p=0.02 (p<0.05).
Conclusion: According to this study Extreme gradient boosting has better accuracy than the K neighbor regression algorithm to predict the sales prices in E-commerce **.**

Keywords: Novel Extreme Gradient Boosting, k-Neighbor Regression, E-Commerce, Machine learning, Accuracy.

[1]Research Scholar Department of Computer Science and engineering, Saveetha School of engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, TamilNadu, INDIA, Pincode-602105.
[2*]Project Guide, Department of Computer Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, TamilNadu, INDIA, Pincode-602105.

1. Introduction

In the modern world, sales forecasting is essential for product-based businesses to expand their offerings, boost production, and analyse sales, profits, and losses. Data analytics is most often used in this context (Al-Qahtani, Fahad H., and Sven F. Crone. 2013). A novel algorithm for forecasting UK electricity demand is described in "Multivariate K-Nearest Neighbor Regression for Time Series Data." International Joint Conference on Neural Networks 2013, held in 2013 (. As artificial intelligence applications are growing large the usage of machine learning algorithms has become more in e-commerce companies. There are many algorithms that can be used in future esteems.In this work, we employ extreme gradient boosting, which forecasts the sales of e-commerce sites. (2013) O'Meara, Wendy Prudhomme, Andrew Obala, Harsha Thirumurthy, and Barasa Khwa-Otsyula "The Association between Price, Competition, and Demand Factors on Private Sector Anti-Malarial Stocking and Sales in Western Kenya: Considerations for the AMFm Subsidy." Malaria Journal 12 (June): 186. Extreme gradient boost predicts exact accuracy value whereas other algorithms only just classify the output. (Parvatiyar, Atul, and G. Shainesh. 2001). Customer Relationship Management: Emerging Concepts, Tools, and Applications. (Tata McGraw-Hill Education. Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2020.)Algorithms, Worked Examples, and Case Studies: Fundamentals of Machine Learning for Predictive Data Analytics, Second Edition. The MIT Press. Machine learning models to predict e-commerce sales which plays a major role in sales prediction this is compared to k-neighbor regression to improve accuracy. (Basan, Ghillie. 2013). The Moon's Our Nearest Neighbor. Hachette UK. (Taghipour, and Atour. 2020). Planning orders and predicting demand in supply chains and humanitarian logistics. IGI Worldwide Extreme gradient boosting and K-neighbor regression can be applicable to many real life problems to sort out those. These algorithms can predict weather forecasts and sales forecasts and play a vital role in technical problems to sort out and give new ways of solutions.

There are 3000 research articles published based on real estate price prediction based on linear regression and K-nearest neighbor algorithm in science direct and also 325 research articles in google scholar and 22 research articles were published in IEEE xplore for house price prediction. The research gap between existing one and for this is four years. The existing research K-neighbor regression having the value 86.50. This research is developed with minimum gained experience through learning. The aim of the study is to improve accuracy of the proposed algorithm when compared to the existing algorithm K-neighbor regression, which is the algorithm used for existing research with an average accuracy rate. (Xiaoling, Wang, Zhou Aoying, and Ji Wendi. 2018). Time-Aware Conversion Prediction For E-Commerce. World Scientific. (Vandeput, Nicolas. 2021.) Data Science for Supply Chain Forecasting. Walter de Gruyter GmbH & Co KG. Feiura, Milan. n.d. "Forecasting Foreign Exchange Rate Movements with K-Nearest-Neighbor, Ridge Regression and Feed-Forward Neural Networks." SSRN Electronic Journal. Our team has extensive knowledge and research experience that has translated into high quality publications (Mohan et al. 2022; Vivek et al. 2022; Sathish et al. 2022; Kotteeswaran et al. 2022; Yaashikaa et al. 2022; Saravanan et al. 2022; Jayabal et al. 2022; Krishnan et al. 2022; Jayakodi et al. 2022; Mohan et al. 2022)

The existing K-neighbor regression algorithm is poor in finding the better accuracy for E-commerce sale prediction (forecast). So, This paper is about the proposed system Extreme gradient boosting that has better performance and accuracy than the K-neighbor regression algorithm in e-commerce sales prediction. The aim of this paper is to make an intelligent system using an approach based on the novel Extreme gradient boosting algorithm to perform better accuracy in comparison with the K-neighbor regression algorithm.

2. Materials and Methods

This research work is done in the Department of Computer Science and Engineering, Saveetha School of Engineering (SSE), Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai. In this study two sample groups were taken. Group 1 was Linear Regression algorithm and group 2 was Gradient Boosting algorithm. Sample size is calculated using Gpower, consider the pretest power to be 80% and threshold 0.05. Which is mainly dependent on two algorithms,which have the sample sizes of Linear Regression (220) and Gradient Boosting (220) which is 440. The work has been carried out with 3000 records and the dataset consists of different features like sales, production, goods… Which is taken from the kaggle dataset. Accuracy is predicted using two different groups. Here the data is from the kaggle website (https://www.kaggle.com/vinaypratap/knearestneighbor). (Yusuf, Ayomide, and Shadi Alawneh.

2018.) "GPU Implementation of Sales Forecasting with Linear Regression." International Journal of Innovative Research in Computer Science**.**
The model is tested on the setup with Hardware requirements i7 processor, 16GB RAM and 256 SSD by using Hp laptop. The software configuration is Windows 10. The tool which is used to execute the process is jupyter notebook version 6. Algorithm is implemented using the python3 code and accuracy of both groups is determined based on the dataset.

Extreme Gradient Boost
Extreme gradient support calculation is one of the critical calculations in artificial intelligence calculations which give more precision. Choice trees, in their least complex structure, are not difficult-to-envision and genuinely interpretable calculations yet fabricating instinct for the up and coming age of tree-based calculations can be a precarious piece. See beneath for a basic similarity to more readily comprehend the advancement of tree-based calculations.(Brdar, Sanja, Marko Panić, Esther Hogeveen-van Echtelt, Manon Mensink, Željana Grbović, Ernst Woltering, and Aneesh Chauhan. 2021.) "Predicting Sensitivity of Recently Harvested Tomatoes and Tomato Sepals to Future Fungal Infections." Scientific Reports 11 (1): 23109.
The following steps to be followed for grouping the algorithm extreme gradient boosting
Specify the following as input:
Input data N
Number of iterations M
A base-learner h
A loss function l
Initialize l0 to a constant
for t = 1 to M:
compute the negative gradient
fit a new base-learner function hi
Find the best gradient descent step-size p
update the function estimate

K-Neighbor Regression
k-neighbor regressor is a famous artificial intelligence calculation which is broadly utilized in numerous applications. The thought behind k-neighbor regression is that given an example of occurrences in an example space, another case is comparative assuming it has a place with a similar class as the previously existing example. The thought is to initially choose the closest neighbor to the example whose class we need to anticipate. (Panay, Belisario, Nelson Baloian, José A. Pino, Sergio Peñafiel, Horacio Sanson, and Nicolas Bersano. 2020.) "Feature Selection for Health Care Costs Prediction Using Weighted Evidential Regression." Sensors 20 (16).

Algorithm Steps
Step 1: First step is Date pre-processing .
Step 2: Fitting KNN algorithm to the Training set.
Step 3: Predict the test result.
Step 4: Test Accuracy of the result.
Step 5: Visualizing the test result.
The tool used to execute the process in jupyter notebook version 6. Algorithm is implemented using python code and accuracy of both groups is determined based on the dataset.

Statistical Analysis
The statistical programme used for the analysis is IBM SPSS version 22(64 bit), which is analytic software that produces independent variables N, means, standard deviation, and standard error means with accuracy as the output for the specified models of extreme gradient boost and k-neighbor regression. (Kassambara, Alboukadel. 2018.) Machine Learning Essentials: Practical Guide in R. STHDA.

3. Results

Machine Learning Algorithms are used in this study to predict sales of e-commerce, as everything related to the e-commerce market. Here we test the performance of the algorithms and how significantly these algorithms can predict the sales of e-commerce sites. Two algorithms are selected and tested for which algorithm produces the highest rate of accuracy.
Table 1 represents pseudocode of extreme gradient boosting. Table 2 illustrates pseudocode of K-neighbor regression.
Table 3 shows group statistics of algorithms by comparing and getting accuracy using sample values of 50 for extreme gradient boosting and 20 for K-neighbor regression, Mean=90.50 for extreme gradient boosting and Mean=80.50 for K-neighbor regression. Std.Deviation=3.028 for both the algorithms. In Table 4, The results achieved with $p=0.02$ ($p<0.05$), which shows that two groups are statistically significant with 95% of confidential value. Figure1 represents the comparison of mean accuracy of the proposed and the existing algorithm. The accuracy of the extreme gradient boost algorithm is found to be 90.50% and the K-neighbor regression algorithm has accuracy of 80.50%.

4. Discussion

Version 21 of IBM SPSS was used for the data evolution. Data analysis is done in order to perform independent sample T-tests and group statistics. This contrasts two techniques, with the accuracy

percentages of Extreme gradient boosting at 90.50% and K-neighbor regression at 80.50%, respectively. (Biau, Gérard, and Luc Devroye. 2015) There are numerous studies that are connected to related studies of planned study where findings are. The Nearest Neighbor Method lectures. Springer. In 2020, Liu, Cheng-Ju, Tien-Shou Huang, Ping-Tsan Ho, Jui-Chan Huang, and Ching-Tang Hsieh published their research. Model for Predicting Customer Repurchases on an E-Commerce Platform Based on Machine Learning. e0243105 in PloS One 15 (12). Li Hanchao, Jicheng, Chen Hongchang, and Jicheng in 2021. In the context of big data, "Study on a New Method of Link-Based Link Prediction." 2021 December issue of Applied Bionics and Biomechanics: 1654134. Time Series Volatility Forecasting Using Linear Regression and GARCH, Mario Situm, n.d. Electronic Journal on SSRN. (Kevin Turner and Geoff Williams, 2000) Errors in sales forecasts and supply chain. (2018) Yusuf, Ayomide, and Shadi Alawneh "GPU Implementation of Linear Sales Forecasting" Main limitation is the assumption of linearity between the dependent and independent variables. Assumes that there is a straight line relationship between dependent and independent variables which is incorrect many times. Non linearity of prediction relationships. The future scope of this study explains how it is useful for the clients with improved accuracy. Feature Section techniques are used in this algorithm. To simplify the model. To get the best analysis of sales. The feature section algorithm can reduce the computation time and improve the classification across the classifiers.

5. Conclusion

Based on the obtained results the Linear Regression has better significance value compared to the Gradient Boosting algorithm. The accuracy of the Extreme gradient boosting Algorithm is 90.50% and the K-neighbor algorithm has 80.50%. It proves that extreme gradient boosting is an efficient method compared to the K- nearest neighbor algorithm. The results of an independent sample T-test are presented with a 95% confidence interval and a significance level of 0.02 (Extreme gradient boosting looks to outperform K-nearest neighbour with a value of p0.05).

Declaration

Conflicts of Interests

No conflicts of interest in this manuscript.

Author Contribution

Author Mekala was involved in data collection and analysis. Author SAK was involved in the action process, Data verification and validation process.

Acknowledgement

The authors would like to express their gratitude towards Saveetha school of Engineering, Saveetha institute of Medical and Technical sciences (SIMATS) for providing necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support to complete this study.

1. Oracle Tech Solutions Pvt.Ltd.
2. Saveetha Institute of Medical and Technical Sciences (SIMATS).
3. Saveetha University.
4. Saveetha School of Engineering.

6. References

Al-Qahtani, Fahad H., and Sven F. Crone. 2013. "Multivariate K-Nearest Neighbor Regression for Time Series Data — A Novel Algorithm for Forecasting UK Electricity Demand." The 2013 International Joint Conference on Neural Networks (IJCNN)..

Brownlee, Jason. 2017. Introduction to Time Series Forecasting With Python: How to Prepare Data and Develop Models to Predict the Future. Machine Learning Mastery.

Hanafi, and Burhanuddin Mohd Aboobaider. 2021. "Word Sequential Using Deep LSTM and Matrix Factorization to Handle Rating Sparse Data for E-Commerce Recommender System." Computational Intelligence and Neuroscience 2021 (December): 8751173.

Kim, Jong-Min, and Hojin Jung. 2018. "Time Series Forecasting Using Functional Partial Least Square Regression with Stochastic Volatility, GARCH, and Exponential Smoothing." Journal of Forecasting.

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2020. Fundamentals of Machine Learning for Predictive Data Analytics, Second Edition: Algorithms, Worked Examples, and Case Studies. MIT Press.

Liu, Weiwen, Yin Zhang, Jianling Wang, Yun He, James Caverlee, Patrick P. K. Chan, Daniel S. Yeung, and Pheng-Ann Heng. 2021. "Item Relationship Graph Neural Networks for E-Commerce." IEEE Transactions on Neural Networks and Learning Systems PP (March).

Mohammed, Mohssen, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. 2016. Machine Learning: Algorithms and Applications. CRC Press.

Mentzer, John T., and Mark A. Moon. 2004. Sales Forecasting Management: A Demand Management Approach. SAGE.

Situm, Mario. n.d. "Time Series Volatility Forecasting Using Linear Regression and GARCH." SSRN Electronic Journal..

Turner, Kevin, and Geoff Williams. 2000. Sales Forecast Errors and the Supply Chain.

Tata McGraw-Hill Education. Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2020. Fundamentals of Machine Learning for Predictive Data Analytics, Second Edition: Algorithms, Worked Examples, and Case Studies.

Xiaoling, Wang, Zhou Aoying, and Ji Wendi. 2018. Time-Aware Conversion Prediction For E-Commerce. World Scientific.

Yusuf, Ayomide, and Shadi Alawneh. 2018. "GPU Implementation of Sales Forecasting with Linear Regression." International Journal of Innovative Research in Computer Science**.**

Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." Energy & Fuels: An American Chemical Society Journal 35 (12): 9930–36.

Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." Journal of Nanomaterials 2021 (July). https://doi.org/10.1155/2021/8115585.

Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." Scientific Reports 10 (1): 18179.

Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." Environmental Toxicology, August. https://doi.org/10.1002/tox.23007.

Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation 29 (1): 65–72.

Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." Environmental Sciences Europe 33 (1). https://doi.org/10.1186/s12302-021-00501-2.

Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." Materials Today Communications 29 (December): 102909.

Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." Environmental Progress & Sustainable Energy 40 (6). https://doi.org/10.1002/ep.13696.

Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." Materials Today: Proceedings 45 (January): 5759–63.

Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." Journal of Circuits Systems and Computers 30 (05): 2130006.

Tables and Figures

Table 1. Pseudo code for Extreme Gradient Boost Algorithm

Input: D={(x_1, y_1), ..., (XN, YN)}, 0,Y
Output: F(x) = 0 F_1(x)
Initialize F_1(x) = arg ming EoL(yi. B) i=0

While (m< M)
di = [OL(y_1, F(x_2))/OF(xi)]F(x)=Fm-1(x1)
9 = {(x_1, di)},i = 1, N
Pm = arg min, E1L(yi, Fm-1(x) + pg(x))
Fm(x) = Fm-1(x) + YPmg(x)

Table 2. Pseudo Code for K-neighbor Regression

Calculate "d(x, x_i)" i =1, 2, ….., n; where d denotes the Euclidean distance between the points.
Arrange the calculated n Euclidean distances in non-decreasing order.
Let k be a +ve integer, take the first k distances from this sorted list.
Find those k-points corresponding to these k-distances.
Let k_i denotes the number of points belonging to the i^{th} class among k points i.e. $k \geq 0$
If $k_i > k_j \; \forall \; i \neq j$ then put x in class i.

Table 3. The table explains about the group statistics of the model by comparing the algorithm and accuracy sample values =10 for extreme gradient boost and 10 for k-neighbor regression mean= 90.50 for extreme gradient boost and 78.50 for X k-neighbor regression std.deviation =3.028 for extreme gradient boost and 3.028 for k-neighbor regression Std.error mean=.957 for extreme gradient boost and .957 for k-neighbor regression.

	Group	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Extreme Gradient boost	10	90.50	3.028	.957
	K-neighbor Regression	10	80.50	3.028	.975

Pseudocode

```
from sklearn. ensemble import Gradient Boosting Classifier #For Classification
from sklearn. ensemble import Gradient Boosting Regressor #For Regression
clf = Gradient Boosting Classifier (n_estimators=100, learning _rate=1.0, max_ depth=1)
clf.fit (X_train, y_train)
```

Table 4. The significance value obtained is p=0.02 (p<0.05), which shows that two groups are statistically significant. The graph explains the comparison of the accuracy value with algorithms extreme gradient boost and k-neighbor Regression where the accuracy of extreme gradient boost is 90% and the accuracy value of the k-neighbor regression is 80%.

	Levene's Test for Equality of Variances		**t-test for Equality of Means**					
	F	**sig.**	**t**	**df**	**Sig.**	**Mean**	**Std.Error**	**95% Confidence Interval of the**

					(2-tailed)	Difference	Difference	difference	
								Lower	Upper
Accuracy Equal Variances assumed Equal variances not assumed	.000	0.02	5.170	18.000	.000	10.000	1.354	9.155	10.845
Loss Equal Variances assumed Equal variances not assumed	.000	0.02	5.170	18.000	.000	10.000	1.354	9.155	10.845

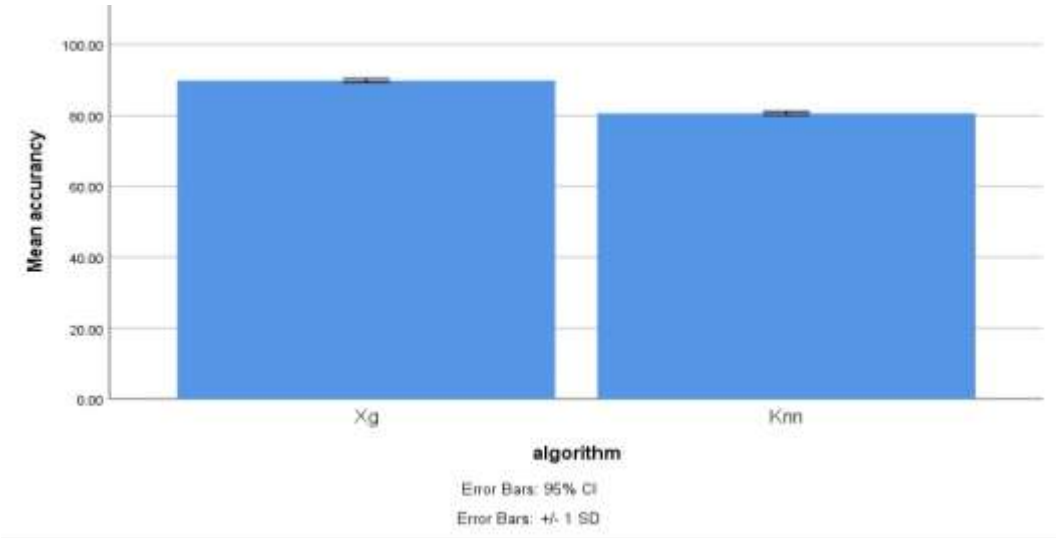

Fig. 1. This bar chart represents the comparison of mean accuracy between extreme gradient boosting and the k-neighbor regression algorithm. The accuracy of the extreme gradient boosting is found to be 90.50% and the k-neighbor regression algorithm is 80.50%.Extreme gradient boosting algorithm gives better results compared to the k-neighbor regression algorithm which has accuracy of 90.50%, The mean accuracy detection is ±1 SD.