SALES PREDICTION USING E- COMMERCE BY USING MACHINE LEARNING ALGORITHMS EXTREME GRADIENT BOOSTING IN COMPARISON WITH THE ARIMA MODEL 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 Arima model algorithm.

Materials and Methods: Extreme gradient boosting and Arima model 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 (83.50) compared to random arima model algorithm (79.50). 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 Linear regression algorithm to predict the sales prices in E-commerce **.**

Keywords: Novel Extreme Gradient Boosting, Arima Model, Accuracy, E-Commerce, Forecast 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

Sales prediction is very much important in this present world to analyze the sales of gods and to get an idea on profits and losses. (Omar, Hani, Van Hai Hoang, and Duen-Ren Liu. 2016). "A Hybrid Neural Network Model for Sales Forecasting Based on ARIMA and Search Popularity of Article Titles." Computational Intelligence and Neuroscience 2016 (May): 9656453.(Brownlee, Jason. 2018). Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery. Due to the increase of population across the world, the usage of the internet has become so frequent. It led to the growth of mobile shopping and a new industry came into the picture 'E-commerce' . In this sector they have to analyze sales frequently to make profits. There are many algorithms that can be used in future esteems. In this study we are using extreme gradient boosting which predicts the e-commerce sites sales prediction. (Mentzer, John T., and Mark A. Moon. 2004). Sales Forecasting Management: A Demand Management Approach. SAGE. (Hazim, Mohamad, Nor Badrul Anuar, Mohd Faizal Ab Razak, and Nor Aniza Abdullah. 2018). ("Detecting Opinion Spams through Supervised Boosting Approach." PloS One 13 (6): e0198884. Chatfield, Chris. 2000. Time-Series Forecasting. CRC Press). (Chatfield, Chris. 2000). Time-Series Forecasting. CRC Press. Arai, Kohei, and Supriya Kapoor. 2019. Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC),(Volume 2. Springer.). Algorithms can be applicable in different sectors in real world problems. Extreme gradient boosting helps in forecasting techniques whereas the arima model helps in multiple ways in forecasting as well as in developing deep learning models and contributing solutions to real world problems.

There are 3000 research articles published based on real estate price prediction based on Extreme gradient boosting and Arima model in science direct and also 325 research articles in google scholar and 22 research articles were published in IEEE xplore for sales prediction. The research gap between existing one and for this is four years. The existing research Arima model has the value of 79.50. 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 Arima model, which is the algorithm used for existing research with an average accuracy rate. Wei, Leyi, Wenjia He, Adeel Malik, Ran Su, Lizhen Cui, and (Balachandran Manavalan. 2021). "Computational Prediction and Interpretation of Cell-Specific Replication Origin Sites from Multiple Eukaryotes by Exploiting Stacking Framework." Briefings in Bioinformatics 22 (4). (Gupta, G. S. 1993). Arima Model for and Forecasts on Tea Production in India. (Wu, Yhao Yang. 2010). Stock Index Prediction Based on Gray Theory, ARIMA Model and Wavelet Method. (Ieee Staff. 2018.) 2018 17th Ieee International Conference on Machine Learning and Applications (ICMLA).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 Arima model 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 Arima model 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 Arima model 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,goods,production… 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/somyaagarwal69/gold-forecasting). (Hazim, Mohamad, Nor Badrul Anuar, Mohd Faizal Ab Razak, and Nor Aniza Abdullah. 2018). 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

The extreme gradient boost algorithm is one of the key algorithms of machine learning algorithms which gives more accurate results. Decision trees, in their simplest form, are easy to visualize and interpret, but building intuition for the next-generation of trees-based algorithms requires a lot more sophistication and expertise.(Perez, Bruno C., Marco C. A. M. Bink, Karen L. Svenson, Gary A. Churchill, and Mario P. L. Calus. 2022. "Prediction Performance of Linear Models and Gradient Boosting Machine on Complex Phenotypes in Outbred Mice." G3 , February.

The following steps to be followed for extreme gradient boosting algorithm to group the data

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

Arima Model

This is an acronym which stands for Auto Regressive Integrated Moving Average, and it actually describes a set of models which uses its own lags and forecast errors to 'explain' a time series and use it to forecast future values. Three terms define an arima model: p, d, and q. The 'Auto Regressive' (AR) term's order is 'p.' It's the number of Y delays that will be used as predictors. The order of the 'Moving Average' (MA) word is 'q.' It relates to how many lagged forecast mistakes should be included in the arima model. A pure Auto Regressive (AR alone) model is one in which Yt is only determined by its own lags. To put it another way, Yt is a function of the 'lags of Yt. where $Yt-1$ is the series' lag1, $beta1$ is the lag1 coefficient predicted by the model, and $alpha$ is the intercept term estimated by the model. where the error terms are the errors of the autoregressive models of the respective lags. The errors Et and E(t-1) are the errors.(Van Calster, Tine, Bart Baesens, and Wilfried Lemahieu. 2017. "ProfARIMA: A Profit-Driven Order Identification Algorithm for ARIMA Models in Sales Forecasting." Applied Soft Computing.

ARIMA model in words: Predicted Yt = constant + linear combination lags of Y (upto p lags) + linear combination of lagged forecast errors (upto q lags)

Algorithm Steps

1. Visualize the Time Series Data
2. Identify if the date is stationary
3. Plot the Correlation and AutoCorrelation Charts
4. Construct the ARIMA Model or Seasonal ARIMA based on the data

Statistical Analysis

The statistical software which is doing for analysis is IBM SPSS version 22(64 bit) which is analysis software which is done by uploading dataset to software which is done by uploading dataset to the software which gives the output as independent variables N,means,Std deviation,std. error means with the accuracy as the output for given models extreme gradient boost and random forest regressor. (Lu, Shaobo. 2021. "Research on GDP Forecast Analysis Combining BP Neural Network and ARIMA Model." Computational Intelligence and Neuroscience 2021 (November): 1026978).

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 shows pseudo code of extreme gradient boost algorithm. Table 2 represents pseudo code of the Arima model algorithm.

Table 3 represents the data set with attributes and explains the group statistics algorithm and accuracy using sample values of 50 for extreme gradient boosting and 20 for arima model, Mean=83.50 for extreme gradient boosting and Mean=79.50 for arima model. Std.Deviation=3.028 is equal for both models Table 4 explains about the independent variables, which defines the equal variances assumed and equality of means with sig.(2-tailed)=0.02 for both assumed and non assumed variances and mean differences=10.000 for both assumed and non assumed variances and 95% of confidential value respectively. Figure 1 gives the comparison of mean accuracy of the proposed and the existing algorithm. The accuracy of the extreme gradient boost algorithm is found to be 83.50% and the arima model algorithm has accuracy of 79.50%.

4. Discussion

The data evolution was done using IBM SPSS software version21.To analyze data for performing independent sample T-test and group statistics be carried out.which represents the comparison of two algorithms with their accuracy percentages 83.50% for Extreme gradient boosting and 79.50% for Arima model. There are many studies which are related to similar studies of proposed research where findings are (Chen, Shijun, Xiaoli Han, Yunbin Shen, and Chong Ye. 2021). "Application of Improved LSTM Algorithm in Macroeconomic Forecasting." Computational Intelligence and Neuroscience 2021 (October): 4471044. Alamrouni, Abdelgader, Fidan Aslanova, Sagiru Mati, Hamza Sabo Maccido, Afaf A. Jibril, A.(G. Usman, and S. I. Abba. 2022). "Multi-Regional Modeling of Cumulative COVID-19 Cases Integrated with Environmental Forest Knowledge Estimation: A Deep Learning Ensemble Approach." International Journal of Environmental Research and Public Health 19 (2). (Huang, Zeying, Haijun Li, and Beixun Huang. 2021). "Regional Distribution of Non-Human H7N9 Avian Influenza Virus Detections in China and Construction of a Predictive Model." Journal of Veterinary Research 65 (3): 253–64. (Hyndman, Rob J., and George Athanasopoulos. 2018). Forecasting: Principles and Practice. OTexts. (Sudarshan, Vidya K., Mikkel Brabrand, Troels Martin Range, and Uffe Kock Wiil. 2021). Performance Evaluation of Emergency Department Patient Arrivals Forecasting Models by Including Meteorological and Calendar Information: A Comparative Study.Computers in Biology and Medicine 135 (August): 104541. 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 extreme gradient boosting has better significance value compared to the arima model. The accuracy of the Extreme gradient boosting Algorithm is 83.50% and the arima model has 79.50%. It proves that extreme gradient boosting is an efficient method compared to the arima model algorithm. Independent sample T-test result is done with confidence interval as 95% and significance level as 0.02 (Extreme gradient boosting appears to perform significantly better than arima model algorithm with the value of $p<0.05$).

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

Auffarth, Ben. 2021. Machine Learning for Time-Series with Python: Forecast, Predict, and Detect Anomalies with State-of-the-Art Machine Learning Methods. Packt Publishing Ltd.

Chatfield, Chris. 2000. Time-Series Forecasting. CRC Press.

Dudek, Grzegorz, Pawel Pelka, and Slawek Smyl. 2021. "A Hybrid Residual Dilated LSTM and Exponential Smoothing Model for Midterm Electric Load Forecasting." IEEE Transactions on Neural Networks and Learning Systems PP (January).

Dunstan, Jocelyn, Marcela Aguirre, Magdalena Bastías, Claudia Nau, Thomas A. Glass, and Felipe Tobar. 2020. "Predicting Nationwide Obesity from Food Sales Using Machine Learning." Health Informatics Journal 26 (1): 652–63.

Ghafouri-Fard, Soudeh, Hossein Mohammad-Rahimi, Parisa Motie, Mohammad A. S. Minabi, Mohammad Taheri, and Saeedeh Nateghinia. 2021. "Application of Machine Learning in the Prediction of COVID-19 Daily New Cases: A Scoping Review." Heliyon 7 (10): e08143.

Gupta, G. S. 1993. Arima Model for and Forecasts on Tea Production in India.

Hazim, Mohamad, Nor Badrul Anuar, Mohd Faizal Ab Razak, and Nor Aniza Abdullah. 2018. "Detecting Opinion Spams through Supervised Boosting Approach." PloS One 13 (6): e0198884.

IEEE Staff. 2018. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA).

Lu, Yu-Sin, and Kai-Yuan Lai. 2020. "Deep-Learning-Based Power Generation Forecasting of Thermal Energy Conversion." Entropy 22 (10).

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

Sales Forecasting Management: A Demand Management Approach. SAGE. (Hazim, Mohamad, Nor Badrul Anuar, Mohd Faizal Ab Razak, and Nor Aniza Abdullah. 2018)

Sudarshan, Vidya K., Mikkel Brabrand, Troels Martin Range, and Uffe Kock Wiil. 2021. "Performance Evaluation of Emergency Department Patient Arrivals Forecasting Models by Including Meteorological and Calendar Information: A Comparative Study." Computers in Biology and Medicine 135 (August): 104541.

Time-Series Forecasting. CRC Press. Arai, Kohei, and Supriya Kapoor. 2019. Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 2. Springer.

Wu, Yhao Yang. 2010. Stock Index Prediction Based on Gray Theory, ARIMA Model and Wavelet Method.

Jayabal, Ravikumar, Sekar Subramani, Damodharan Dillikannan, Yuvarajan Devarajan, Lakshmanan Thangavelu, Mukilarasan Nedunchezhiyan, Gopal Kaliyaperumal, and Melvin Victor De Poures. 2022. "Multi-Objective Optimization of Performance and Emission Characteristics of a CRDI Diesel Engine Fueled with Sapota Methyl Ester/diesel Blends." Energy. https://doi.org/10.1016/j.energy.2022.123709.

Jayakodi, Santhoshkumar, Rajeshkumar Shanmugam, Bader O. Almutairi, Mikhlid H. Almutairi, Shahid Mahboob, M. R. Kavipriya, Ramesh Gandusekar, Marcello Nicoletti, and Marimuthu Govindarajan. 2022. "Azadirachta Indica-Wrapped Copper Oxide Nanoparticles as a Novel Functional Material in Cardiomyocyte Cells: An Ecotoxicity Assessment on the Embryonic Development of Danio Rerio." Environmental Research 212 (Pt A): 113153.

Kotteeswaran, C., Indrajit Patra, Regonda Nagaraju, D. Sungeetha, Bapayya Naidu Kommula, Yousef Methkal Abd Algani, S. Murugavalli, and B. Kiran Bala. 2022. "Autonomous Detection of Malevolent Nodes Using Secure Heterogeneous Cluster Protocol." Computers and Electrical Engineering. https://doi.org/10.1016/j.compeleceng.2022.107902.

Krishnan, Anbarasu, Duraisami Dhamodharan, Thanigaivel Sundaram, Vickram Sundaram, and Hun-Soo Byun. 2022. "Computational Discovery of Novel Human LMTK3 Inhibitors by High Throughput Virtual Screening Using NCI Database." Korean Journal of Chemical Engineering. https://doi.org/10.1007/s11814-022-1120-5.

Mohan, Harshavardhan, Sethumathavan Vadivel, Se-Won Lee, Jeong-Muk Lim, Nanh Lovanh, Yool-Jin Park, Taeho Shin, Kamala-Kannan Seralathan, and Byung-Taek Oh. 2022. "Improved Visible-Light-Driven Photocatalytic Removal of Bisphenol A Using V2O5/WO3 Decorated over Zeolite: Degradation Mechanism and Toxicity." Environmental Research. https://doi.org/10.1016/j.envres.2022.113136.

Mohan, Kannan, Abirami Ramu Ganesan, P. N. Ezhilarasi, Kiran Kumar Kondamareddy, Durairaj Karthick Rajan, Palanivel Sathishkumar, Jayakumar Rajarajeswaran, and Lorenza Conterno. 2022. "Green and Eco-Friendly Approaches for the Extraction of Chitin and Chitosan: A Review." Carbohydrate Polymers 287 (July): 119349.

Saravanan, A., P. Senthil Kumar, B. Ramesh, and S. Srinivasan. 2022. "Removal of Toxic Heavy Metals Using Genetically Engineered Microbes: Molecular Tools, Risk Assessment and Management Strategies." Chemosphere 298 (July): 134341.

Sathish, T., R. Saravanan, V. Vijayan, and S. Dinesh Kumar. 2022. "Investigations on Influences of MWCNT Composite Membranes in Oil Refineries Waste Water Treatment with Taguchi Route." Chemosphere 298 (July): 134265.

Vivek, J., T. Maridurai, K. Anton Savio Lewise, R. Pandiyarajan, and K. Chandrasekaran. 2022. "Recast Layer Thickness and Residual Stress Analysis for EDD AA8011/h-BN/B4C Composites Using Cryogenically Treated SiC and CFRP Powder-Added Kerosene." Arabian Journal for Science and Engineering. https://doi.org/10.1007/s13369-022-06636-5.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Algal Biofuels:

Technological Perspective on Cultivation, Fuel Extraction and Engineering Genetic Pathway for Enhancing Productivity." Fuel. https://doi.org/10.1016/j.fuel.2022.123814.

Yaashikaa, P. R., P. Senthil Kumar, and S. Karishma. 2022. "Review on Biopolymers and Composites – Evolving Material as Adsorbents in Removal of Environmental Pollutants." Environmental Research. https://doi.org/10.1016/j.envres.2022.113114.

Tables and Figures

Table 1. Pseudo codes of Extreme Gradient Boosting 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)
{(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 Arima model

procedure FINDOPTIMALARIMA
aic← inf
for p← 0 to 3 do
for d← 0 to 2 do
for q← 0 to 3 do
Model← fit (arima(p,d,q,allow_drift←True, allow_mean← True),X)
aic_crr←compute_Aic(model)
If aic_curr<aic_then
model_apt←model
aic←aic_curr
Return model_opt

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 Arima model Mean= 83.50 for Extreme Gradient boost and 79.50 for X Arima model std.Deviation =3.028 for Extreme Gradient boost and 3.028 for Arima Model Std.Error Mean=.957 for Extreme Gradient boost and .957 Arima model .

	Group	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Extreme Gradient boost	10	83.50	3.028	.957
	Arima Model	10	79.50	3.028	.957

Pseudocode

```
from sklearn.ensemble import GradientBoostingClassifier #For Classification
from sklearn.ensemble import GradientBoostingRegressor #For Regression
clf = GradientBoostingClassifier(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 Arima model where the accuracy of extreme gradient boost is 83.50% and the accuracy value of the Arima model is 79.50%.

	Levene's Test for Equality of Variances		**t-test for Equality of Means**						
	F	**sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence Interval of the difference**	
								Lower	**Upper**
Accuracy Equal Variances assumed	.000	0.02	2.954	18.000	.008	4.000	1.354	1.155	6.845
Equal variances not assumed									
Loss Equal Variances assumed	.000	0.02	2.954	18.000	.008	4.000	1.354	1.155	6.845
Equal variances not assumed									

GGraph

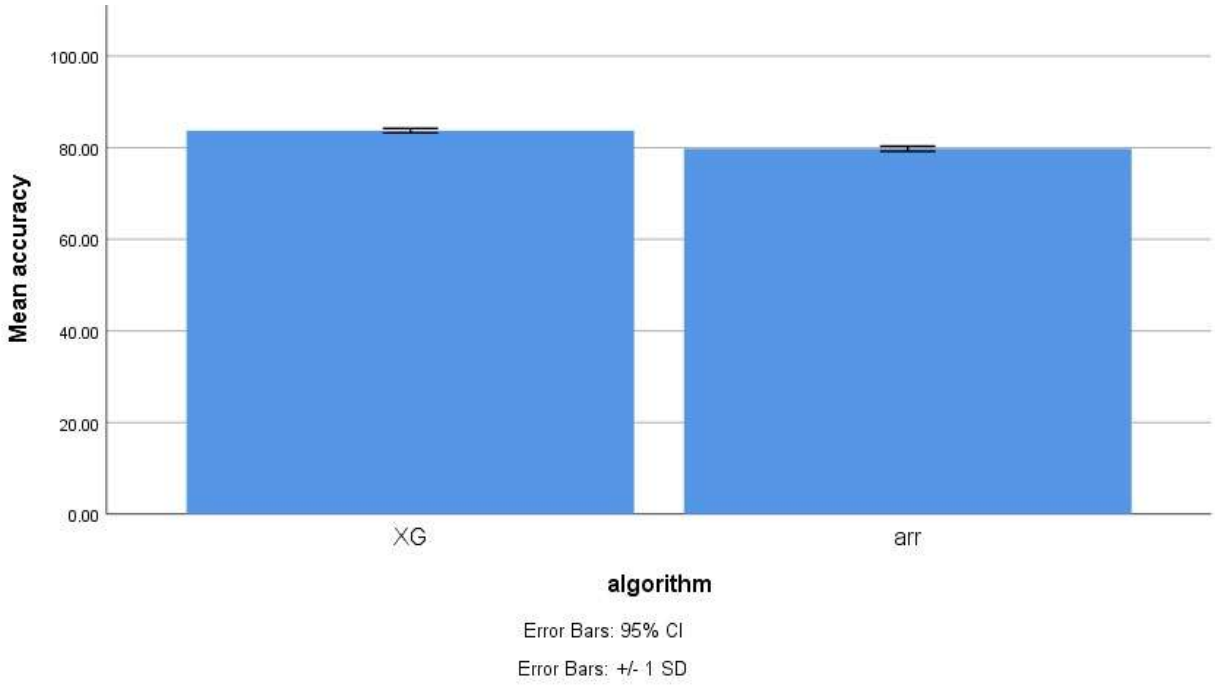

Fig. 1. This bar chart represents the comparison of mean accuracy between extreme gradient boosting and the arima model algorithm. The accuracy of the extreme gradient boosting is found to be 83.50% and the linear regression algorithm is 79.50%. Extreme gradient boosting algorithm gives better results compared to the Arima model algorithm which has accuracy of 83.50%, The mean accuracy detection is ±1 SD.