IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

E-commerce customer spend prediction through Hyper parameter tuned regression models

Vishwath Shankar

Foothill Highschool, Pleasanton California - 94588

1. ABSTRACT

With the emergence of the internet, the last decade has seen a revolution in online shopping, and the number of users around the world has dramatically increased. During the pandemic, even more have started shopping online. E-commerce growth has been enormous this decade and has paved the way for numerous studies on customer purchase behavior and spending predictions. This proposed study is based on customers spending time on web and mobile applications to predict their annual spending. The study on customer purchase behavior can improve business strategies, as vendors are able to handle stocks according to customer purchase behaviors. This proposed work uses machine learning regression algorithms to predict annual spending; specifically, regression models such as Decision Trees and Random Forest. The models use hyperparameters tuned using the Grid Search Cross Validation (GSCV) technique. Experimental results showed that the hyperparameter-tuned Random Forest model has the highest accuracy in e-commerce customer spending prediction.

Key Words: Machine Learning, Regression models, Decision Tree, Random Forest, Grid search cross validation, Hyper parameter tuning.

2. INTRODUCTION

In the evolving world of online shopping, customer spending predictions are increasingly vital, as they increase opportunities for business. The shift from direct shopping to online shopping has induced a large amount of data generated every day. This data has shown the potential to get deeper insights into understanding customer behaviour for spending, as well as challenges. Understanding this data eventually helps e-commerce websites like Amazon and Flipkart to handle their inventory effectively and increase customer satisfaction. Customer spending is the most important factor in the e-commerce business, and we can make use of artificial intelligence, machine learning techniques, and their predictive capability for decision-making and predicting customer spending.

Existing works mainly focus on customers' purchase behaviour and patterns, which helps understand what items customers prefer to purchase. This analysis is done through the past purchasing behaviour of customers and predicts future item sales using various factors including product reviews, and ratings. However, this study has a major challenge in the sparsity of data, as the total number of items in the portal is massive relative to the number of items purchased every day, moreover, all customers may not rate or review the items. Recommendation systems are also widely used in the existing research, which can recommend items or services to customers based on their spending prediction. Similarly, time series studies were also used for forecasting customer purchases based on the history of date-wise purchasing data.

Making use of artificial intelligence increases the accuracy of spending prediction, which in turn helps in inventory management, revenue growth, enduring the competitive market, and improving customer satisfaction. The objective of the proposed study is to implement an e-commerce customer spending prediction application with a machine learning regression algorithm.

The proposed work is customer spending prediction on e-commerce websites using machine learning techniques. The few important attributes considered for this work are session length, time on the website, and time on the application. By analyzing this data, the objective of the study is achieved. The hyperparameter tuning of the machine learning algorithm is executed to improve the performance of the algorithm. The cross-validation helps in identifying the best-fit parameter from the given search space. Grid Search cross validation, one of the popular techniques for cross validation is used. The Decision Tree and Random Forest ML models are being used. Regression models are preferred over Classification, as the target attribute is a continuous value.

The main contributions of this paper are:

- Ecommerce customer spend prediction through hyper parameter tuned machine learning regression models and evaluate its performance
- Implement Grid search cross validation to find the best parameter fit for the regression models.
- Improving the performance of the model by selecting the best fit parameter from the given search space.

This paper is organized as Chapter 2 gives detailed study of literature survey and related work in ecommerce prediction. Chapter 3 discusses proposed regression models and hyper parameter tuning. Chapter 4 gives detailed discussion on results arrived for ecommerce customer spend prediction. Chapter 5 draws conclusions based on this work and further enhancement possibilities.

RELATED WORK

E-commerce has been evolving over the past decade and has become a key area for research. There are existing works available in e-commerce for analyzing product reviews, the sentiment of products on social media, statistical analysis of customer sales data for analyzing sales projects, and purchase patterns. This chapter discusses some of the recent research on this platform.

The e-commerce shopping behavior of users is analyzed in [1], which considers the most influencing parameters such as date, time, the environment used for purchase, session time, clicks, etc. This paper mainly researched mega sale events, the dataset used is a web access log by deep packet inspection, and there are seven types of classifiers used for purchase prediction. This work aimed at early detection of purchases to enhance the business model. Experimental results showed that logistic regression has the highest accuracy of 92.4%.

User rating is one of the most influential factors for e-commerce sales, as most customers in the set reviewed the product, and it enables users to understand the product quickly. The work [2] addressed the similarity search by the technique of compact binary sketch for identifying similar types of users. The relationship has been built for viewed and bought products to facilitate the ranking prediction. Experimental results showed that MAE error is less for the proposed work ES_USE, predicting the more accurate raking for user purchase.

The work in [3] discovered the process of identifying consumer purchase behavior for Omni channels, though Omni channel increases the customer purchase experience, it is more challenging for businesses to track the purchase behavior of consumers. There are several miner algorithms implemented including fuzzy, Heuristic, inductive, and alpha. These studies identified that the consumers who spend time on social media are the majority of consumers in the Omni channel. Experiment results showed that the Fuzzy Miner has the highest fitness and precision score of about 1.0 for consumer purchase behavior analysis.

E-commerce purchase patterns using classification models are studied in [4], the work addressed feature selection by fitting a linear regression to predict the sales. The most significant features are only taken for the training. Experimental results showed that the consumer sales prediction has been achieved with an accuracy of 84% using the Naïve Bayes algorithm.

User churn prediction on e-commerce portals is another major study [5], which helps in understanding the reason for customers leaving and taking necessary steps to retain customers. The features considered in the dataset include session, purchase, behavior of customer, interactions made, actions, and rating. Clustering algorithms are implemented for churn prediction including K-means, DBscan, and Birch. Experimental results showed that Birch has achieved the highest accuracy of 81.3% for customer churn classification.

E-commerce data as big data analytics is studied in [6], has arrived recommendation system by analyzing the positive and negative effects of big data analytics. The paper analyzed in terms of vendors and customers. According to the vendor, there are positive effects including customer satisfaction, enhanced advertisement, quick decision-making, etc, whereas negative effects are the high cost of analyzing tools, and challenges handling big data. In terms of customers, the positive effects are satisfaction, improved service, and purchase decisions, whereas the negative effects are shopping addiction.

This literature survey helps in understanding the techniques used in the e-commerce market. It is inferred that the e-commerce industry has been an interesting area of research, there are existing works based on machine learning classifier models, clustering models, miner algorithms, etc. There are works represented on customer behavior prediction, purchase pattern, sales prediction [8, 9, 10, 11,13], sales projections [12], churn prediction [7], etc. Though the works are on different perspectives to increase the business of e-commerce, the customer's annual spending prediction is not addressed. The customer annual spend prediction enhances the business portal to understand the revenue of the portal and design the strategies according to prediction to enhance the revenue. Thus an effective model to be addressed for customer annual spend prediction is addressed in this work.

PROPOSED WORK

E-commerce consumer spend analysis has great importance on business growth and plant strategy for business, inventory, increased satisfaction, and more. This proposed predictive analysis is aimed at annual spending prediction for consumers on e-commerce platforms. The proposed work implements regression analysis to predict customer spending annually using hyper-tuned ML models. The Grid Search Cross validation is performed for the given search to identify the best-fit parameter.

DATASET DETAILS

The e-commerce dataset with a total of eight attributes as described in the below table is used for this project. This dataset is publicly available on kaggle.com. The dataset attributes and their descriptions are given in the below table.

Table 1: Ecommerce dataset details

Variable name	Attribute Description
Email	Customer's email id
Address	Customer's address
Avatar	Color selected by the customers on their profile
Average session length	Length in minutes spent by users on average
Time on app	Minutes spent by customer on application
Time on website	Minutes spent by customer on website
Length of membership	Years the customer has been using store
Yearly amount spent	Money spent yearly by customer on store

Exploratory data analysis

Exploratory data analysis (EDA) is visualizing and analyzing the dataset using plots. It helps understand data patterns and the purchasing behavior of customers. The dataset is visualized to understand the customer's time and application and website and how much yearly amount spent in the form of histograms. This plot is used to identify the data statistic behind every attribute in the dataset.

In the below plot, the average session length is shown, the number of users between the session lengths 32 to 34 is high. There are much fewer customers using the small session length 30 to 31 and the high session length users (35 to 36). The below plot shows a histogram for 500 customers, which accounts for 28% of customers using session lengths of 32, and 29% of users using under 34 session length. The average session length of 33 has the highest number of consumers. Overall 32 to 34 sessions cover around 94% of the total number of users in the e-commerce dataset.

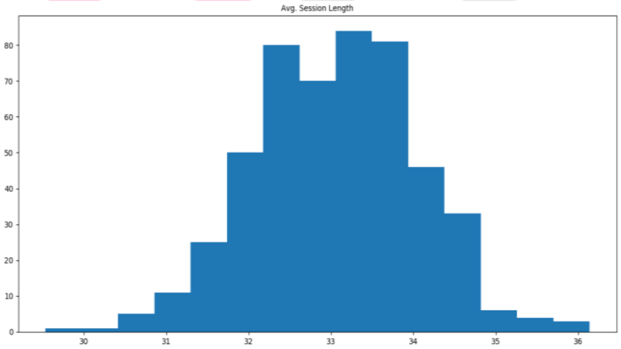

Figure 1: Average session length on ecommerce website

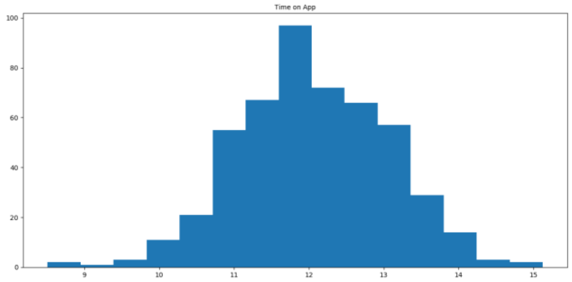

Figure 2: Time on App for ecommerce dataset

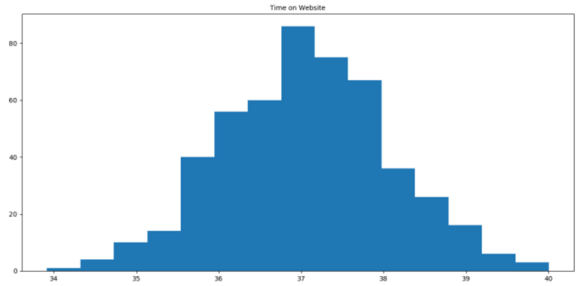

Figure 3: Time on website for ecommerce dataset

The above plot on the left side shows the time on the app, the plot shows that there are a high number of users in 12 minutes and there are an average number of customers who use 11 minutes and 13 minutes. The time on the website is comparatively higher than the time on mobile apps. The high number of users used a session length of 37 minutes, and the average number of users used 36 to 37 minutes. The left side plot time on the app shows that 23% of users used a session length of 11, 40% of users used 12 minutes of session length and 26% of users used a session length of 13. The average time on the app is 12, which has the highest number of users, which accounts for 40% of total customers. Overall from 11 to 13 minutes of time, the app constitutes around 89% of users. Similarly, in the right side plot, 23% of users used time on the website of 36 minutes, 39% of users used it for 37 minutes, and 24% of users used it for 38 minutes. The average time on the website of 37 minutes has the highest number of customers at 39%. Overall 36 to 38 minutes substitutes for 82% of total users.

3. RESULTS

The proposed e-commerce customer spending prediction using regression models DT, and RF is implemented with hyper tuned models. The hyper tuned models are evaluated with 30% of the dataset. Decision tree and random forest models considered five parameters for the hyperparameter tuning experiment. Then cross validation was carried out three times. The algorithm is evaluated with accuracy metric, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Table 3 shows the comparison of DT, RF models and DT-hyper tuned and RF- hyper tuned model's performances.

The below plot shows the e-commerce customer annual spending original data and predicted data by the Random forest regressor model. The plot shows that the predicted values are very slightly deviated from the original value, thus giving the highest accuracy for the annual spending prediction.

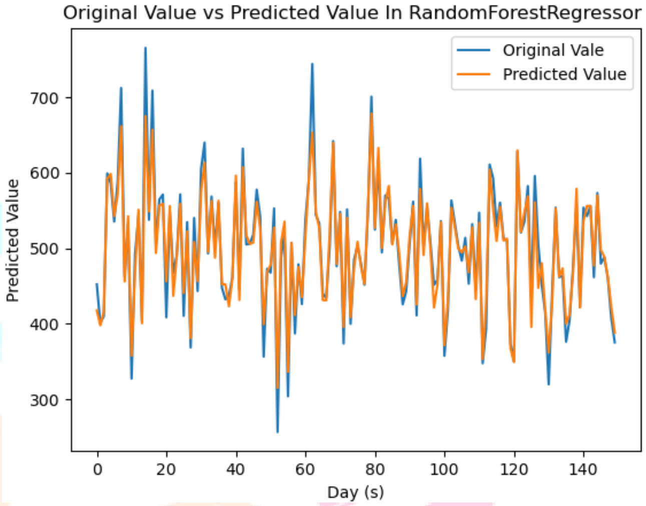

Figure 5: Original Vs Predicted value of Ecommerce customer annual spend by Random forest model

The below plot shows the e-commerce customer annual spending original data and predicted data by the Decision tree regressor model. The plot shows the results for the test dataset predicted by the algorithm for 140 days, which is 30% of the dataset.

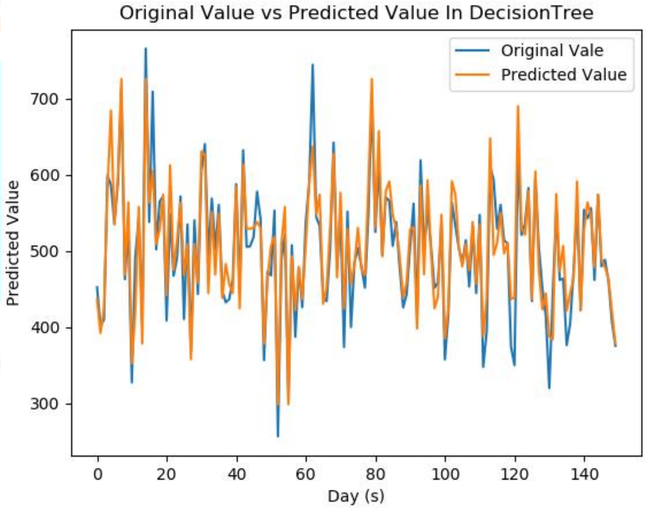

Figure 6: Original Vs Predicted value of Ecommerce customer annual spend by Random forest model

For regression problems, the evaluation metric used are error metrics. Accuracy is computed based on the proportion of correctly classified instances. The formula to calculate accuracy is given in equation 3.

$$\text{Accuracy} = \text{Number of Correctly Classified Instances} / \text{Total Number of Instances}$$

--(3)

The equation (4) shows the computation of mean squared error, where y and yhat is the original and predicted ecommerce customer annual spend.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y_i})^2 \quad \text{--- (4)}$$

The equation (5) shows the computation of mean absolute error, where n is the total number of error, x_i-x is the absolute error.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x| \quad \text{----(5)}$$

Table 3: Algorithm Comparison for Normal and Hyper tuned models

Algorithm	Decision Tree	Hyper tuned Decision Tree	Random forest	Hyper tuned Random Forest
MSE	8.896	9.47	4.409	43.93
MAE	23.20	23.82	15.12	14.86
R-Squared	0.877	0.869	0.939	0.939
RMSE	29.82	30.77	20.99	20.96
Accuracy	86.74	86.95	93.92	93.94

The above table shows that the accuracy of the Random forest model is good compared to the decision tree model. The Decision tree has achieved an accuracy of 86.74%, whereas the Random forest model has achieved 93.92% accuracy. The hyperparameter tuned model has improved the accuracy of both models. The Hyper-tuned DT model has achieved an accuracy of 86.95%, whereas the hyper-tuned RF has achieved an accuracy of 93.94%. Experiments are conducted for cross-validation three times and test set data of 30%.

4. DISCUSSION

Nowadays, online shopping is highly desired by people around the world, as it has many advantages like product reviews, door delivery, return policies, and attractive discounts. The online e-commerce portal is growing at a high rate during this decade, and competition has increased. Purchasing behavior should be carefully studied to maintain the revenue of businesses as well as make strategic plans. The proposed work is the predictive analysis of e-commerce customer annual spending by hyperparameter-tuned machine learning algorithms. Grid search cross validation is implemented for tuning the model. Experimental results showed that the hyperparameter tuning has improved the accuracy of prediction than by the normal ML models. Experiments proved that the Random forest algorithm has achieved the highest accuracy of 93.94% for e-commerce customer spend prediction. As an extension of this work, deep learning models can be studied such as the Convolutional Neural Network model (CNN) and, Deep Neural Network model (DNN). This work can also be further extended by adding more features and statistical analysis of each customer preference and number of visits etc. to personalize the user preference and arrive at more effective strategic planning for the business.

5. METHODS

The dataset is split into 70% as training data and 30% as test data, the learning attributes are average session length, time on app, time on website, and length of membership, the target attribute is the yearly amount spent by consumers. The average amount spent is a continuous value, thus a regression model is applied. The model is given a hyperparameter tuned best fit parameter for training and validation. To improve the performance of the proposed predictive analysis, hyper parameter tuning is performed. This enhances the model to get the best fit parameter and helps in effective learning and prediction of e-commerce customer spending.

The below figure represents the proposed system architecture for annual e-commerce spend by customers based on the website and mobile application time they use. The regression models of Decision tree and random forest are applied for the prediction and the model is hyper parameter tuned for improving the model performance. This diagram represents the function flow of the system. The ML regression models are hyper tuned with parameters, the technique used for hyper parameter selection is Grid Search Cross Validation (GSCV) and the best parameter is chosen for ML models. The hyper tuned ML models are evaluated with performance measures such as accuracy and loss metrics.

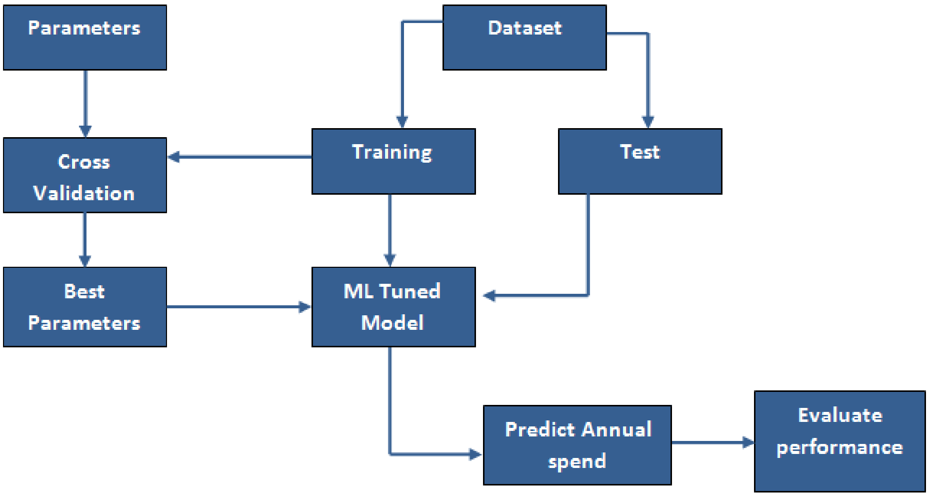

Figure 4: System architecture of Ecommerce annual spend prediction

Hyper parameter Tuning

Hyperparameter Tuning (HPT) can signify the performance of the regression model, this technique searches for the best parameters, which signifies the best prediction performance on the e-commerce consumer spend prediction. For the cross validation, the popular model GridSearch Cross Validation (GSCV) for hyperparameter tuning is used. The search space is defined for DT and RF algorithms are shown in Table 2. Grid search HPT takes a finite set of values and computes Cartesian products. There are a few hyper parameters tuned including the maximum depth of the tree, number of features, etc. GSCV takes the hyperparameter vector and trains the regression ML models DT, and RF, the parameters are given in Table 2 considered as search space parameters and validated, cross-validation computes the mean and Standard deviation (STD) error values are computed and parameters chosen according to low error. The mean is calculated using the equation (1).

$$m = \frac{\sum_{i=1}^{b} err_i}{b} \quad \text{----- (1)}$$

$$se = \sqrt{\frac{var(err)}{b}} \quad \text{------- (2)}$$

In the above equations 'b' is the number of folds.

'm' is the mean error and 'se' is the standard error.

Standard deviation error is computed using equation (2) for cross-validation, the error is calculated and sorted to get the least value, and the least value error parameter is chosen as the best fit parameter for building the ML regression model.

Table 2: Parameters Details for Hyperparameter tuning

Algorithm	Variable name	Attribute Description
Decision Tree	Splitter	Best, random
	Max_depth	None, 1,3,5,7
	Min_sample_leaf	1,2,3,4,5
	Max_features	None, auto, log2, sqrt
	Max_leaf_node	None,10,20,30,40, 50
Random Forest	Bootstrap	True, False
	Max_depth	10,20,30,40,None
	Max_features	Auto,sqrt
	Min_sample_leaf	2,5,10

	N_estimators	200,400,600,800

GridSearchCV takes input as parameters, estimator, cross validation, the parameters listed in the above table are cross-validated one by one. The estimators are the Decision Tree regressor and Random forest regressor, and the cross validation is an integer value, 3.

The splitter parameter is given as 'best' or 'random'. Based on the information gained the features will be selected for constructing the tree. The max_depth and max_leaf nodes are the parameters, which control the depth of the tree and diminish the overfitting problem.

6. ACKNOWLEDGEMENTS

I have been lucky to receive support and guidance from various individuals who have made this journey both informative and fulfilling. I would like to express my deepest appreciation to my high school teachers and mentors who have provided valuable knowledge and encouraged my curiosity. Their guidance has been key in shaping my understanding of research methods. Special thanks are due to Mr. Sos, my AP Computer Science teacher, whose feedback has been very helpful in refining the methodology of this research.

7. REFERENCES

[1] M. Guan, M. Cha, Y. Wang, Y. Li and J. Sun, "From Anticipation to Action: Data Reveal Mobile Shopping Patterns During a Yearly Mega Sale Event in China," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 4, pp. 1775-1787, 1 April 2022, doi: 10.1109/TKDE.2020.3001558.

[2] R. Yang and B. Li, "ES_Use: An Efficient Rating Prediction Method," in IEEE Access, vol. 9, pp. 110658-110669, 2021, doi: 10.1109/ACCESS.2021.3102961.

[3] F. A. Tridalestari, Mustafid and F. Jie, "Consumer Behavior Analysis on Sales Process Model Using Process Discovery Algorithm for the Omnichannel Distribution System," in IEEE Access, vol. 11, pp. 42619-42630, 2023, doi: 10.1109/ACCESS.2023.3271394.

[4] R. Valero-Fernandez, D. J. Collins, K. P. Lam, C. Rigby and J. Bailey, "Towards Accurate Predictions of Customer Purchasing Patterns," 2017 IEEE International Conference on Computer and Information Technology (CIT), Helsinki, Finland, 2017, pp. 157-161, doi: 10.1109/CIT.2017.58.

[5] P. Berger and M. Kompan, "User Modeling for Churn Prediction in E-Commerce," in IEEE Intelligent Systems, vol. 34, no. 2, pp. 44-52, March-April 2019, doi: 10.1109/MIS.2019.2895788.

[6] S. S. Alrumiah and M. Hadwan, "Implementing Big Data Analytics in E-Commerce: Vendor and Customer View," in IEEE Access, vol. 9, pp. 37281-37286, 2021, doi: 10.1109/ACCESS.2021.3063615.

[7] M. Li, "Research on the prediction of e-commerce platform user churn based on Random Forest model," 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2022, pp. 34-39, doi: 10.1109/ICCSMT58129.2022.00014.

[8] S. K. Punjabi, V. Shetty, S. Pranav and A. Yadav, "Sales Prediction using Online Sentiment with Regression Model," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 209-212, doi: 10.1109/ICICCS48265.2020.9120936.

[9] L. Huang, Z. Dou, Y. Hu and R. Huang, "Online Sales Prediction: An Analysis With Dependency SCOR-Topic Sentiment Model," in IEEE Access, vol. 7, pp. 79791-79797, 2019, doi: 10.1109/ACCESS.2019.2919734.

[10] G. T., R. Choudhary and S. Prasad, "Prediction of Sales Value in Online shopping using Linear Regression," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777620.

[11] H. Yuan, W. Xu and M. Wang, "Can online user behavior improve the performance of sales prediction in E-commerce?," 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 2014, pp. 2347-2352, doi: 10.1109/SMC.2014.6974277.

[12] C. Zhan, J. Li, W. Jiang, W. Sha and Y. Guo, "E-commerce Sales Forecast Based on Ensemble Learning," 2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN), Chongqing, China, 2020, pp. 1-5, doi: 10.1109/ISPCE-CN51288.2020.9321858.

[13] Z. Huo, "Sales Prediction based on Machine Learning," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), Hangzhou, China, 2021, pp. 410-415, doi: 10.1109/ECIT52743.2021.00093.