JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1866-1874
https://publishoa.com
ISSN: 1309-3452

Performance Study of Proposed Predictive Data Mining Model for analysing Online Customer Buying Behaviour

Teena Vats

Orchid Id: (0000-0002-3378-3051)

Research Scholar, Department of CSE,

Jagannath University NCR, Haryana, India

tina.summer27@gmail.com

Dr. Kavita Mittal

Orchid Id: (0000-0002-2967-0804)

Associate Professor, Department of CSE,

Jagannath University NCR, Haryana, India

kavitamittal.it@gmail.com

Received 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

ABSTRACT

These days web-based business climate assumes a significant part to trade product information between shoppers regularly with others. Precisely foreseeing client buy designs in the web-based business marketplace is the basic place to use the information mining. To accomplish high benefit in internet business, the connection among client and product are vital. Additionally, numerous web-based business sites increment quickly and immediately and contest has become quite recently a just a click away. For that reason, the significance of remaining in the business, and further developing the benefit needs to precisely anticipate buy conduct and focus on their clients with customized administrations as per their inclinations. In this paper, an information mining model has been planned to improve the exactness of foreseeing and to observe affiliation rules for incessant thing sets. Various algorithms have been used to calculate the customer behaviour like Tree Classifier Boost Classifier, gradient Boosting Classifier, Random Forest Classifier. Planned model has been verified on three city Yangon, Mandalay, Naypyitaw Mall's sales dataset and the results shows that data uses is 94.0%.

Keywords: Decision Tree Classifier, Boost Classifier, Random Forest Classifier, Data mining.

Introduction: The method of inspecting information from an alternate classification is known as information mining. This information contains significant data, likewise in information mining extra information will be removed. Additionally, it is a useful methodology for removing and recognizing designs in colossal informational collections that integrate strategies from AI, measurements, and data set framework. These days corporate association is trying to embrace a computerized showcasing system and selfless business sectors to acquire overall business benefits, online business organizations should initially appreciate their clients' feelings. surveys and seasons according to their items and administrations. Control and handling of this information in different ways to give a model expanded expectation precision take into consideration the extraction of novel information that guides in balanced promoting, personalization, expanded deals, and client maintenance [6]. Network advertising has turned into a critical showcasing method, and as web innovation has progressed many organizations have constructed internet-based stores to give clients buying materials. As a result of the various advantages of web-based business the quantity of individuals who participate in web-based exchange as well as the volume of exchanges has fundamentally extended [7]. Trouble in information mining applications is recognizing considerable, significant and clear data from simple and inadequate information by digging continuous examples for information disclosure [8]. Perhaps the main utilizations of data mining in the internet business area are precisely expecting client buy tendencies in light of the fact that the quantity of internet business sites (both client and product) develops quickly and in a flash and contest is just a mouse click away. To remain in business, suppliers should have the option to dependably device client purchasing conduct and target them with altered administrations in light of

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1866-1874
https://publishoa.com
ISSN: 1309-3452
their inclinations. AI (ML) strategies are perhaps the most strategy that are utilized as information mining methods, additionally utilizing of ML to use temporary the student model in light of past encounters and get new information when the size of information becomes huge. ML has been utilized in many fields for example security, clinical field, web-based business pitch and others. internet business information is mentioned to as "Large Data" subsequently managing this information to separate the information is viewed as a test [11]. All the same size utilizing scientific methodologies and answers for separate examples in secret connections to settle on better choices and get new information makes it more confounded. Moreover, picking appropriate calculation to get the best example and concentrate the information to further develop the exhibition is likewise difficult. The information mining applications have issues in the mining of intermittent examples looking for information expose to recognize large, helpful and reasonable data out of unpolished and scanty information. Applying an information mining model to improve the precision of expectation with regards to internet business and managing huge information to extract the information at a sensible time is a significant job.

RELATED WORK

1**.** Moazzam, Y. Farwa and Y. Fu, M. Yang [1,2] The ascent in notoriety of web-based entertainment has introduced a new time for internet business, changing web-based shopping. A few studies have been proposed to improve the exhibition for the expectation in internet business [2]. Additionally, some of them used to anticipate the client assessment in view of the remarks [1]. This area presents different order calculation that has been utilized in the writing for information emulating or for ordering.

2**.** Lu Y. B and S. C. Zheng [3] gives a work three distinct datasets, highlight choice and outfit classifier SVM-AR is utilized for arrangement. The better precision is acquired by the group classifier SVM-AR (Association Rules) and minimum performance time.

3.Khammas et. al [4] Offerings a effort of relative review with various element determination techniques and machine learning strategies. This effort obtains improved outcomes with the mix of Principal Components Analysis (PCA).

The significant commitments of this paper are as per the following:

1) Put on information mining calculations in such a method for giving a model that predicts customers level of purchasing and prescribes it to them.
2) Giving a correlation among various choice tree arrangement calculations to pick the best characterization calculation for an item proposal framework in light of a bunch of thought about boundaries.
3) Boosting the information to improve the runtime and utilizing optimization of random forest affiliation rule calculation to separate the arrangement of things.

Data Description: Dataset is occupied from Cleveland repository of Kaggle and this dataset has various types of characteristics definite, text, short text and numerical. The categorical attributes list is male and female. Some of numeric attribute's list are tax, total, gross salary, cogs, unit price and rating. In Data set researcher found 1000 records 499 for male and 501 for female. Dataset is stable. The dataset covers 13 attributes and 6 classes

Table 1. Exposed Sample of Characteristics from Dataset

No.	Attribute Term	Category	Characteristics Properties
1	City	Text	Yangon, Mandalay, Naypyitaw
2	Customer type	Short Text	Member, Normal
3	Gender	Categorical	Male, Female
4	Product	Text	Food product, home accessories, lifestyle, health and beauty, electronic etc.
5	Unit Price	Numeric	Per item
6	Quantity	Numeric	Total purchased by a customer
7	Tax 5%	Numeric	On total bill
8	Total	Numeric	Bill including tax
9	Payment	Text	Cash, Credit card, Wallet
10	Cogs	Numeric	Bill after discount
11	Grossmargin percentage	Numeric	Total discount
12	Gross income	Numeric	Total saving
13	Rating	Numeric	1-10

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1866-1874
https://publishoa.com
ISSN: 1309-3452

Exploratory Data Analysis: Figure 1 histogram shows the graphic picture of arithmetical data distribution. It is mostly used to characterize data available in a form of groups. Histogram is a type of bar plot where X-axis denotes the basket ranges while Y-axis gives info about frequency**.** Researcher can check the ups and downs of the attributes like unit price, quantity, tax, total, cogs, gross margin percentage, gross income, Rating.

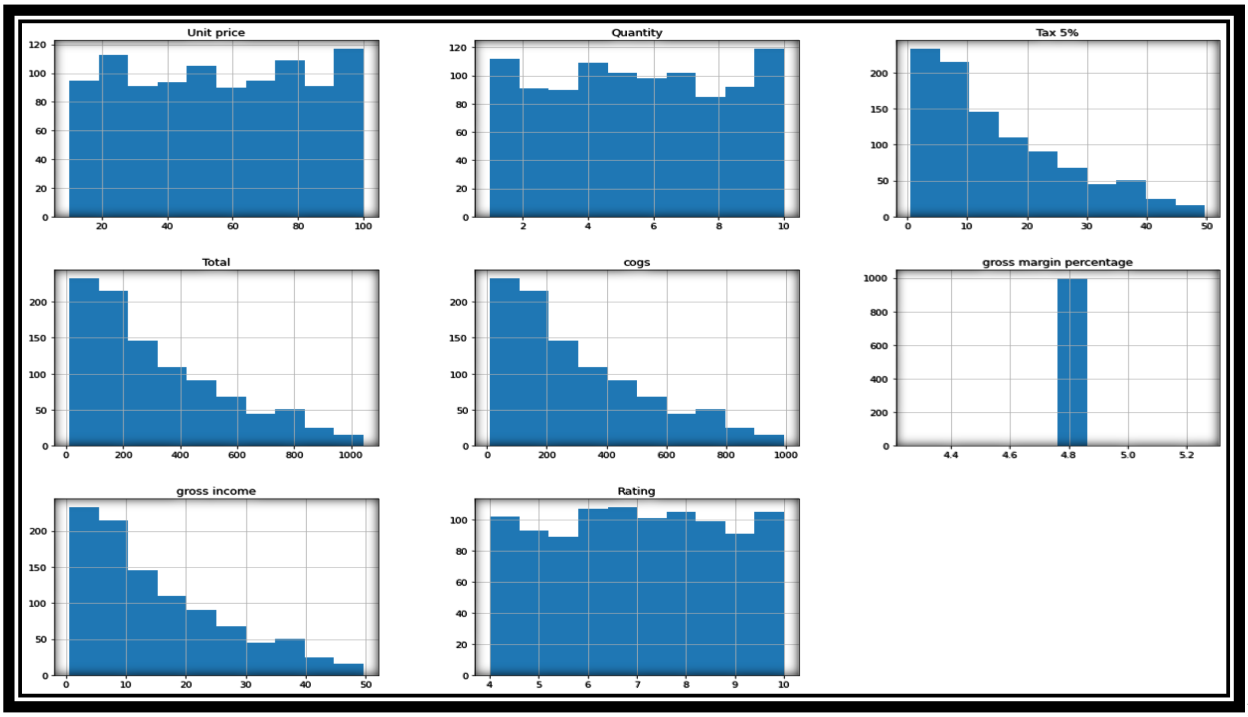

fig .1. Data Anaysis

Figue 2. Shows a graphic depiction in python of data that uses a structure of color-coding to represent different values by using heatmap. Heatmap is used in many forms of analytics but most usually used to show user's behaviour on precise webpages or webpage patterns.

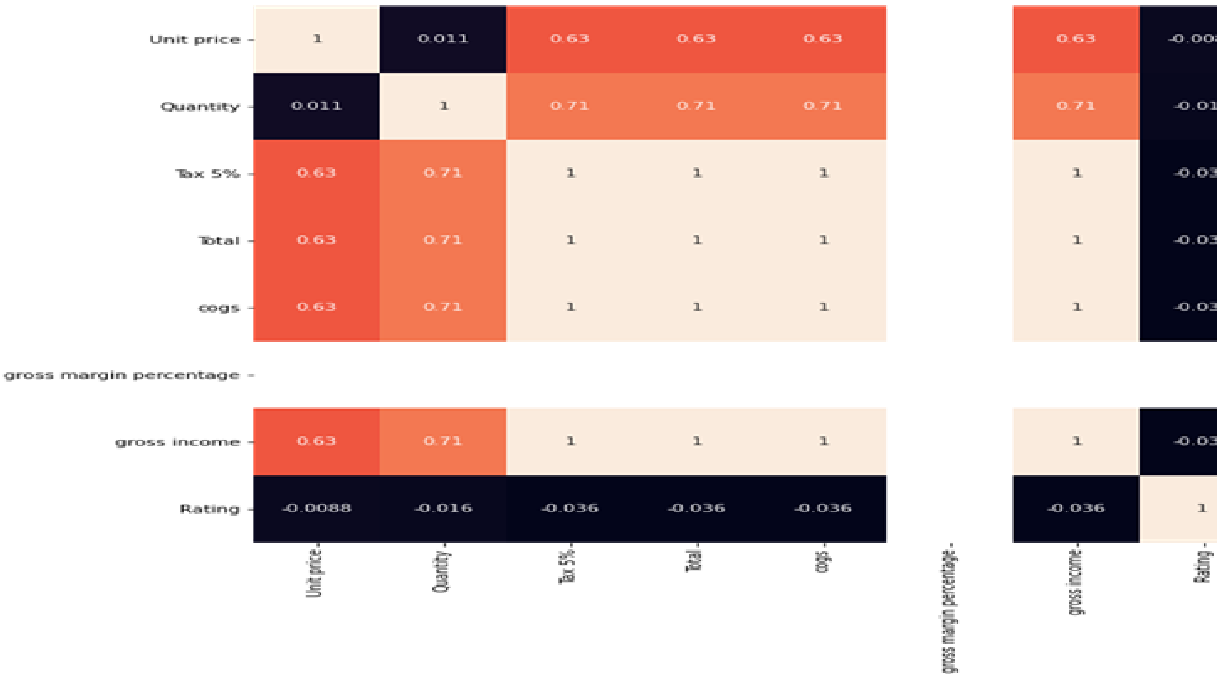

Fig. 2. Confusion matrix of the features

METHODOLOGY

Strategy purposed for the supplementary item goes the complementary advances. The initial step is to parted the quantity into two subsets, first is the Training set and the subsequent one is the Testing set. Whenever information is parting happen then first, researcher apply preparing information on classifiers/learning Models to become familiar with our model then researcher achieve testing by getting information from test dataset. The proportion of parting Training and Testing information subsets are 60% and 40% separately. Researcher prepared the model by giving 70% information as info and in the wake of preparing our model is prepared to predict the name based on past information. Presently researcher can put their test part to check how effective our model is prepared. Models or gaining classifier gets prepared from preparing information. [5,6].

Accuracy: To evaluation the correctness of a test, researcher must calculate the percentage of true positive and true negative in all estimated cases. Exactly this can be stated as: TP-true positive, TN-true negative, FP-false positive false negative.

Accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$

Recall: Recall recognized as a completeness of a classifier. It also known as an unfair classification problem with two classes, recall is planned as the number of true positives divided by the total number of true positives and false negatives.

Recall = $\frac{TP}{TP+FN}$

Precision: Precision is recognized as the correctness of classifiers. it is also recognized as an indicator of a machine learning model's performance. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

Precision = $\frac{TP}{TP+FP}$

F1-Score: F1 score takes the balance between the precision and the recall. F1 score known as harmonic mean of precision and recall. It is also called the F-Measure. F1 Score is the 2*((precision*recall)/ (precision recall)).

F1 – Score = $2 * (precision * recall)/(Precision + Recall)$

Random forest

Random forest is a greedy algorithm to select an optimal split point. Random forest uses bootstrap replicas means subsamples the input data with replacement. The selection of cut points in order to split nodes Random Forest picks the best split. In random forest there are many numbers of trees. Best split method is used to split each subset of features of decision node Bagging is applied in random forest. Random forest is a greedy algorithm so that it used best split method [6,7].

Process of random forest

- Random forest is a group of many decision tree and it is less sensitive to the training data.
- Random forest is a greedy algorithm to select an optimal split point.
- Method used in random forest is bootstrapping (row sampling) and random feature selection that’s why it’s called random process.
- Bootstrapping ensures that we are not using the same data for every tree so that it helps our model to be fewer sensitive to the original training data.
- Random feature selection helps to reduce the correlation between the trees that reduce the variance.
- Problem of high variance in decision tree over come in random forest it changes high variance into low variance.
- Random forest usages bootstrap copies mean subsamples the input data with replacement.
- The selection of cut points in order to split nodes Random Forest chooses the optimum split.
- In random forest there are many numbers of trees. Best split method is used to split each subset of features of decision node.
- In random forest process uses two steps.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1866-1874
https://publishoa.com
ISSN: 1309-3452

- first is perform bootstrapping and 2nd is aggregation this process also called bagging.
- Bootstrapping is applied in random forest. Random forest is a greedy algorithm so that it used best split method.

Proposed Model

Optimization of random forest algorithms

Researcher start his/her study with decision tree but due to the limitation of structure of decision tree the study move to the random forest tree where different tree combine for generating the result. It based on greedy approaches for splitting the nodes that why it comes in mind to optimise the random forest tree. Optimization of random forest algorithms is a type of collaborative learning technique which aggregates the results of multiple de-correlated choice trees collected in a "forest". To produce the output classification will randomly sample the features at each split point of a decision tree. But by default, it uses the whole input sample. This may increase variance because bootstrapping makes it more expanded. In the Optimization of random forest algorithms sklearn implementation there is an optional parameter that allows users to bootstrap replicas is the selection of cut points in order to split nodes. Optimization of Radom Forest chooses them randomly**.** Optimization of random forest algorithms is a collaborative machine learning algorithm that trusts the predictions from several decision trees. In optimization of random forest algorithms there are many numbers of trees. Random split method is used to split each subset of features of decision node. Boostrapping (Drawing Sampling without replacement) in not applied in optimization tree.

Optimization of Radom Forest algorithm (Process)

Split a node(S)

Input: The local learning subset S corresponding to the node we want to split.

Output: a split [$a<a_{c}$] or nothing.

If **Stop_split(S)** is TRUE then return nothing.

Otherwise select K attributes {a1……..ak} among all non-constant (in S) candidate attribute;

Draw K splits {s1…………s_k},where s_i=**Pick a random split**(S,a_i), $\forall_{i=1\ldots\ldots\ldots K}$;

Return a splits s_k such that Score(s_* .S) = max i=1…….k Score(s_i , S).

Pick a random_split(S,a)

Input : a subset S and an attribute a.

Output: a split

Let a^s_{max} and a^s_{min} denote the maximal and minimal value of a in S;

Draw a random cut-point a_c uniformly in [a^s_{min}, a^s_{max}];

Return the split [$a<a_c$].

Stop-split(S)

Input: a subset S

Output: a boolean

If |S|< n_{min},then return True;

If all attribute are constant in S,,then return True;

If the output is constant in S, then return True;

Otherwise, return False.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1866-1874
https://publishoa.com
ISSN: 1309-3452

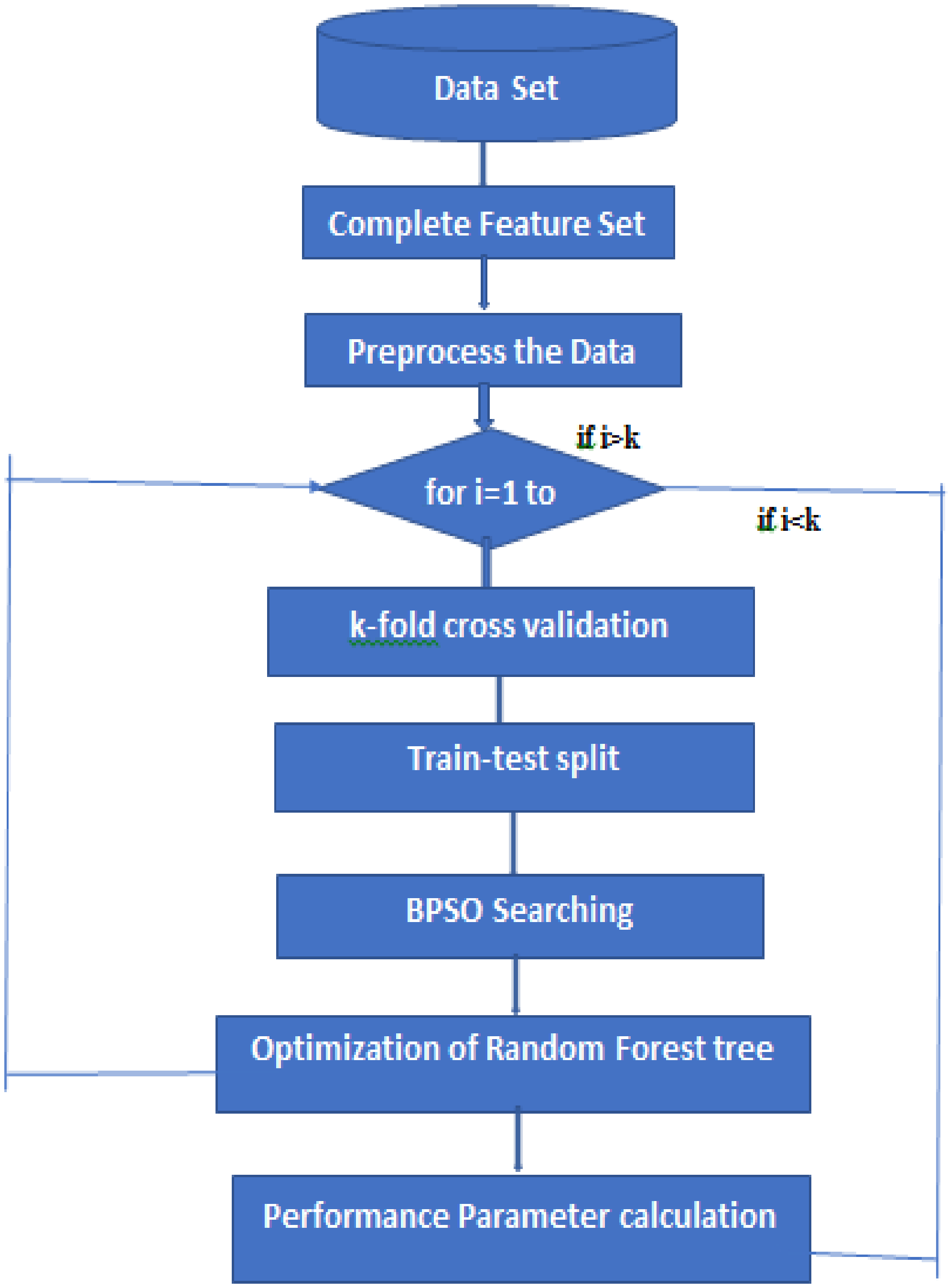

Fig.2. flow chart working of proposed model

RESULTS

This segment covers the details about researches conducted in this research. The first step is the classification for determining the accuracy of classifiers. Trials were carried out with numerous classification models. Accuracy refers to how much our trained algorithm predicts correct classes. Accuracy Results of different classifiers. Researcher's purpose classifier Optimization of random forest gives the better results as compare to random forest, 0.59% that is given more accurate results than other classifiers. The reason behind why Optimization of random forest classifier results are higher is because it's works much faster than the Random tree classifier. Probably it works 3 times faster than Random Tree

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 1866-1874
https://publishoa.com
ISSN: 1309-3452
Classifier. It provides higher presentation. Additional feature of this classifier is it will not over-fit data. It makes additional trees that will help in voting to give best forecasts.

Table 2. Classifier Result of random forest

Class	PRECISION	RECALL	F1-SCORE
0	0.52	0.60	0.56
1	0.56	0.48	0.52
Avg	0.54	0.54	0.54

Optimization of random forest algorithms: it is a type of collaborative knowledge technique which collections the results of many de-correlated decision trees collected in a“forest” to output its classification outcome.

Table 3. Classifier Result of Optimization of random forest

Class	**PRECISION**	**RECALL**	**F1-SCORE**
0	0.52	0.56	0.54
1	0.55	0.51	0.53
Avg	0.54	0.54	0.53

Table 4. Accuracy level of algorithms

Accuracy level	
Random forest	Optimization of random forest
0.51	0.59
Training Score	
91.0	94.0

Table 5. Confusion Matrix of models

Confusion Matrix			
Random forest		**Proposed model**	
142	96	134	104
131	121	124	128

Table 6. Comparison of decision tree, Random Forest and proposed model

Characteristics	**Decision Tree**	**Random Forest**	**Optimize Random Forest**
Number of trees	1	Many	Many
Number of features considered for split each decision node	All Features	Random subset of Features	Random subset of Features
Boostrapping (Drawing Sampling without replacement)	Not applied	No	Yes
Method to Split Node	Best Split	Best Split	Random Split
Throughput	Fast	Fast	Slow
Accuracy	High	Low	High

Reliability	High	Low	High
Perform	No Sampling	Sampling with Replacement	Sampling without Replacement
Searching Method	Top-down	Bagging	BPSO
Working	Top to Bottom	Parallel	Sequential

CONCLUSION AND FUTURE WORK

All in altogether, this paper accomplished higher accuracy for predicting in the proposed information mining model. Additionally, the proposed model recommends the it is connected with one another to buy behaviour of belongings. It used to understand the buy behaviour of customer to predict next buy in light of a bunch of chosen boundaries when Optimization of random forest calculation has been utilized. Then again, the proposed model planned to improve the forecast for an incredible information base. The exploratory outcomes show that arbitrary forest calculations produce high exactness estimations contrasted and different calculations. In this paper, Optimization of random forest calculation is applied for a quick and strong principles age in online business client buying field. At long last, the proposed model accomplished 59.0 % precision. In future, new component determination strategies can be utilized for better element choice with new blend of information digging procedures for another classifier. In future researcher can take different data set for same techniques and produce new results.

REFERENCES

1. Moazzam, Y. Farwa, H. Mushtaq, A. Sarwar, A. Idrees, S. Tabassum, BaburHayyat, and K. Ur Rehman, “Customer Opinion Mining by Comments Classification using Machine Learning,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 5, pp. 385-393, 2021.
2. Y. Fu, M. Yang, and D. Han, “Interactive Marketing E-Commerce Recommendation System Driven by Big Data Technology,” Scientific Programming, vol. 2021, 2021.
3. Lu, Y. B., Din, S. C., Zheng, C. F., & Gao, B. J. (2010). “Using multi-feature and classifier ensembles to improve malware detection”. Journal of CCIT, 39(2), 57-72.
4. Khammas, B. M., Monemi, A., Bassi, J. S., Ismail, I., Nor, S. M., & Marsono, M. N. “Feature selection and machine learning classification for malware detection”. Jurnal Teknologi, 77(1) (2015).
5. Orieb Abu Alghanam, Sumaya N. Al-Khatib, Mohammad O. Hiari Al-Ahliyya “Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 2, 2022.
6. P. HarshaLatha, R. Mohanasundaram ,“A New Hybrid Strategy for Malware Detection Classification with Multiple Feature Selection Methods and Ensemble Learning Methods”. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 2, 2022.
7. Baranidharan B, Abhisikta Pal, Preethi Muruganandam “Cardio-Vascular Disease Prediction based on Ensemble technique enhanced using Extra Tree Classifier for Feature Selection”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019.
8. Pierre Geurts · Damien Ernst · Louis Wehenkel “Extremely randomized trees”, Mach Learn (2006) 63: 3–42 DOI 10.1007/s10994-006-6226-1.
9. Ernest Kwame Ampomah *, Zhiguang Qin and Gabriel Nyame, “Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement”, e Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/) 2020.
10. Rana Alaa El-Deen Ahmeda , M.Elemam.Shehaba , Shereen Morsya , “Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining”, 2015 Fifth International Conference on Communication Systems and Network Technologies.
11. Ioannis Markoulidakis *, Ioannis Rallis, Ioannis Georgoulas, “A Machine Learning Based Classification Method for Customer Experience Survey Analysis”, Published in MDPI on 7 December 2020.

12. S. N. Mohanty, E. L. Lydia, M. Elhoseny, M. M. G. Al Otaibi, and K. Shankar, "Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks," Physical Communication, vol. 40, pp. 101097, 2020.
13. Y. Fu, M. Yang, and D. Han, "Interactive Marketing E-Commerce Recommendation System Driven by Big Data Technology," Scientific Programming, vol. 2021, 2021.
14. K. Kang, and J. Michalak, "Enhanced version of AdaBoostM1 with J48 Tree learning method," arXiv preprint arXiv:1802.03522, 2018.
15. N. Kavha, and S. Karthikeyan, "Customer Buying Behavior Analysis: A clustered Closed Frequent Itemsets for Transactional Database," International Journal of Computer Science Issues (IJCSI), vol. 10, no. 3, pp. 113, 2013.