

Customer Behavior Analysis in E-Commerce using Decision Tree Machine Learning Approach

¹Pooja Sharma, ²Dr. Sachin Kumar, ³Dr. Lokesh Kumar

¹Research Scholar, ²Assistant Professor, ³Assistant Professor

Department of Computer Science and Engineering

Tula's Institute Dehradun, India

poojasharma1028@gmail.com, sachin.kumar@tulas.edu.in, cse.hod@tulas.edu.in

Abstract—The e-commerce industries are growing rapidly. Many of the Customers are purchasing products through the online platform. The Customer, who don't buy product online but they still search online products before buying. The review rating of any product plays important role for making decision among Customers. If any product gets good rating and reviews then the selling chances is more of this product. Sometimes, the false or wrong rating is also done by the companies or persons. The artificial indigence based machine learning techniques are capable to predict he true and false review or provide the prediction model. This paper presents the Customer behavior analysis in e-commerce using decision tree machine learning algorithm. Simulation is done using python spyder platform and simulated result shows the improvement in the accuracy of the prediction model.

Keywords—Machine learning, E- Commerce, Python, Accuracy, Error rate.

I. INTRODUCTION

Client created content as audits, appraisals, and remarks can be investigated for more noteworthy experiences for big business use. The investigation of such buyer conduct is useful to figure out the customer's prerequisites and foresee their future expectations towards the assistance. Through this mental review, Web based business Associations can follow the use and opinions appended to their items and adopt proper showcasing strategies to give a customized shopping experience to their buyers, consequently expanding their authoritative benefit. [1]. Computer based intelligence is a captivating innovation that will wear the pants on different elements of life so as to come. Computerized reasoning capacitates the machines to reproduce human knowledge. Machine Learning is one of the pivotal subsets of Man-made brainpower. The expression Machine Learning (ML) is plain as day meaning the machines that will learn on their own utilizing their related knowledge. The machines are not imperative to be customized expressly for learning new communications. Today organizations put an incredible time and asset in mining the information of clients. As client's information has covered examples and patterns which are rewarding for the organizations. Organizations execute artificial intelligence procedures onto the client information to group the expected clients for their items and administrations[2].

In the present computerized world, headway in machine learning has had a significant impact on the customary viewpoint towards business investigation. Customary business examiners didn't consider client surveys as practical contribution for investigation in light of the fact that previous

bringing client audits were exorbitant. The rise of the web flipped around the entire world. Presently, the client's feeling examination is the new companion of all business investigators[3]. In an e-commerce setting, the large volume of online reviews may become a source of data to predict the repurchase intention. Repurchase intention is important for a company because it is related to customer loyalty. A machine-learning based methodology is proposed in this paper to perform the prediction of repurchase intention based on online customer reviews, in order to obtain the insights from a large volume of the available data [4].



Figure 1: Artificial Intelligence & E-commerce

3 different bunching calculations (k-Means, Agglomerative, and Meanshift) are been executed to fragment the clients lastly analyze the consequences of groups got from the calculations. A python program has been created and the program is been prepared by applying standard scaler onto a dataset having two highlights of 200 preparation test taken from nearby retail shop. Both the elements are the mean of how much shopping by clients and normal of the client's visit into the shop yearly. By applying bunching, 5 portions of group have been shaped named as Imprudent, Cautious, Standard, Target and Reasonable clients. Nonetheless, two new groups arose on applying mean shift bunching marked as High purchasers and incessant guests and High purchasers and intermittent guests [11].

II. LITERATURE SURVEY

V. Shrirame et al.,[1] present work means to utilize information driven advertising apparatuses, for example, information representation, normal language handling, and machine learning models that assistance in grasping the socioeconomics of an association. We likewise fabricate recommender frameworks through cooperative separating, brain organizations, and feeling investigation.

S. Sharma et al.,[2] presents, creators have executed regulated machine learning calculations for example Support Vector Machine (SVM), Arbitrary Woodland, Calculated Relapse, k-Closest Neighbor on web-based client shopping dataset for ordering regardless of whether the client wound up buying the item. The creators have likewise made a basic examination among the order correctnesses of these ML Calculations. The work uncovers that Arbitrary Backwoods performs better with the arrangement of unmitigated reaction variable.

R. Katarya et al.,[3] In this work, do a relative report between the two techniques. One in which we utilize common boundaries like normal rating, normal audit counts. In second, we utilize a message

audit as a boundary for opinion characterization. The viability of these strategies is assessed, and the ideal model is chosen.

D. Suryadi et al.,[4] In a web based business setting, the huge volume of online surveys might turn into a wellspring of information to anticipate the repurchase expectation. Repurchase expectation is significant for an organization since it is connected with client devotion. A machine-learning based philosophy is proposed in this work to play out the expectation of repurchase aim in view of online client audits, to get the experiences from a huge volume of the accessible information..

S. Ghosh et al.,[5] Forecast examination of client buy conduct is a fascinating and testing task in current life. Our goal is to present the idea of machine getting the hang of involving an irregular woodland calculation inside and out. In this work, a model has been proposed for foreseeing which cloud administrations have been bought on various variables. An irregular woodland model is assembled utilizing various boundaries, for example, commercial snap arrangement, recently bought cloud administrations, and so forth and preparing our model.

Z. Wang et al.,[6] In the field of retail industry and showcasing, recognizing client fragments is quite possibly the main undertaking. A significant division can assist the supervisors with upgrading the nature of items and administrations for the focusing on fragments. A large portion of customary techniques utilized POS information to characterize the client dedication as "weighty" fragment while others are having a place with "light" section.

S. M. A. M. Manchanayake et al.,[7] Upselling is an important procedure for expanding the net revenue of any assistance giving business space. It assumes an essential part in development of an organization. Among those organizations, telecom industry is a noticeable industry where upselling is exceptionally impacted on agitate decrease and settling the client base.

K. Rasmee et al.,[8] Current buyer conduct dealers are utilized to make dynamic devices for business visionaries to deliver items or administrations that address customer issues. Hence, subsequently zeroing in on information examination involving SMO methods for foresee framework for the way of behaving of shopper purchasing individual vehicle choice in view of a sum of 1,110 information got, with a sum of 6 applicable vehicle exchanging data highlights.

I. Lieder et al.,[9] For deals and showcasing associations inside huge undertakings, recognizing and seeing new business sectors, clients and accomplices is a key test. Intel's Deals and Promoting Gathering faces comparable difficulties while filling in new business sectors and areas and advancing its current business. In the present complex innovative and business scene, there is need for insightful computerization supporting a fine-grained comprehension of organizations to assist SMG with filtering through large number of organizations across numerous geologies and dialects and distinguish significant headings.

S. Satpathy et al.,[10] For imparting and communicating insights, an individual can involve the virtual space in the web and Online entertainment is a stage for things like this; where talk about on specific issues, remarks on various realities and contrasting things and others. Then again, customer promoting and brand the board are associated with it and make considerably more intricacy in this worldwide, rapidly developing innovation world. Likewise it impacts purchaser buying conduct and buying demeanor.

T. Kansal et al.,[11] The climate of current time is development, where everybody is entangled into contest to be preferable over others. The present business run based on such development having capacity to enchant the clients with the items, yet with such an enormous pontoon of items leave the clients perplexed, what to purchase and what to not and furthermore the organizations are confused about which part of clients to focus to sell their items.

A. Inoue et al.,[12] The quantity of Portable Virtual Organization Administrators (MVNO) clients is expanding in Japan, however the cell phone market in Japan was essentially shared by three Significant Versatile Transporters (3MMC): docomo, au and SoftBank. The reason for this study is to figure out the inclination for 3MMC versus MVNO considering the ongoing cell phone market.

III. METHODOLOGY

The methodology of the proposed research work is as followings-

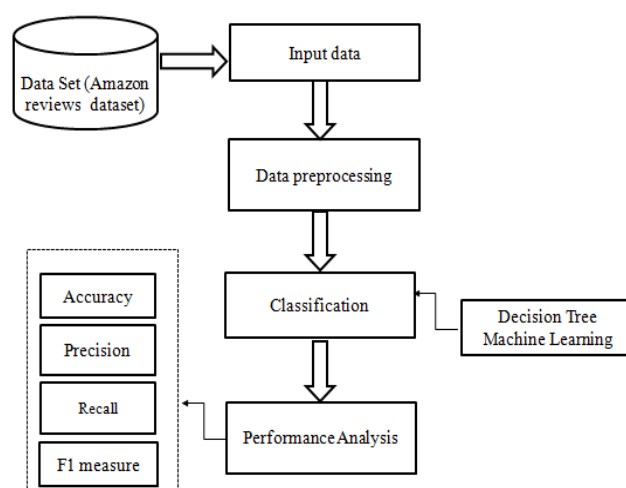


Figure 2: Flow Chart

- **Collect data set**

For implementation the research works, the customer review on online product behavior data set of Amazon website will be taken from kaggle machine learning repository.

This dataset contain 69000 customer reviews of various products.

- **Preprocess of data**

Data pre-processing involves converting any string variable to the numerical one so that it gets easy for evaluation. Also handle missing and null values.

- **Feature Extraction**

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. Here consider product name, rating, user name etc features for extraction.

- **Classification**

We are using decision tree algorithm to predict the customer rating on products

Algorithm

Input: CustomerBehaviour analysis of Reviews of Amazon Products dataset.

Take the initial data features reviews rating, reviews text, reviews title and reviews, username.

Filtering the null value

Classify the text based on sentiments

Output: Optimal Precision, Recall, F-Measure, Accuracy and Error rate

Step: 1. Split train and test dataset Y_train, Y_test, X_train and X_test

2. Feature extractions, features = {} for word in words: features [word] = True

3. Vectorization

Y train counts

Y train transformer

4. Apply the decision tree machine learning classifier.

5. Generate confusion matrix and show value of TP, FP, TN and FN

6. Calculate Accuracy, error rate, precision, recall and f-measure

7. Plot the ROC Curve

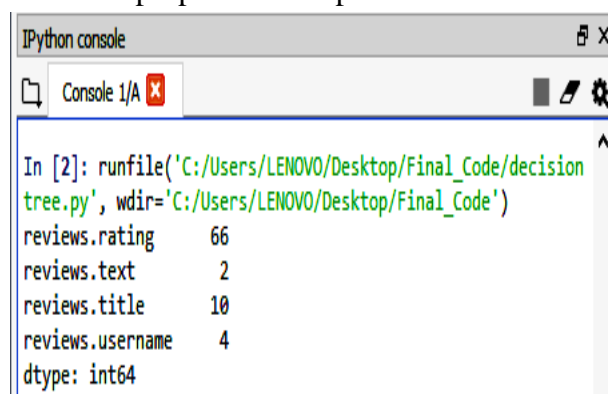
Evaluation

The confusion metrics used to evaluate a classification model are accuracy, precision, and recall.

- Precision = True Positive / (True Positive + False Positive)
- Recall = True Positive / (True Positive + False Negative)
- F1-Score = 2x (Precision x Recall) / (Precision + Recall)
- Accuracy = [TP + TN] / [TP + TN + FP + FN]
- Classification Error = 100 - Accuracy
-

IV. SIMULATION & RESULTS

Simulation is to be done using Python Software. Python is open source software having large library of AI, machine learning etc work. The spyder ISE is platform using by the python for the implementation and simulation of the proposed concept.



```
IPython console
Console 1/A
In [2]: runfile('C:/Users/LENOVO/Desktop/Final_Code/decision
tree.py', wdir='C:/Users/LENOVO/Desktop/Final_Code')
reviews.rating    66
reviews.text      2
reviews.title     10
reviews.username  4
dtype: int64
```

Figure 3: Dataset loading and preprocessing

Figure 3 is presenting the online Amazon product review dataset in the python environment. Then the preprocessing is start, the followings features are extracted- Reviews rating, reviews text, review title, review username etc.

```

89 #
90 "Decision Tree"
91 from sklearn.tree import DecisionTreeClassifier
92 from sklearn.metrics import confusion_matrix
93 from sklearn import model_selection
94 from sklearn.ensemble import BaggingClassifier
95 seed = 7
96 kfold = model_selection.KFold(n_splits=10, random_state=seed)
97 cart = DecisionTreeClassifier()
98 num_trees = 100
99 model = BaggingClassifier(base_estimator=cart, n_estimators=num_trees,
100 y_predD = model_selection.cross_val_score(model, Y_train_tfidf, train[
101
102 clf = DecisionTreeClassifier()
103
104 clf = clf.fit(Y_train_tfidf,train["sentiment"])
105
106 y_predD = clf.predict(Y_test_tfidf)
107
108 lr_dfd = pd.DataFrame(y_predD, columns=['Predicted'])
109 label_encoder = preprocessing.LabelEncoder()
    
```

Figure 4: Decision tree classifier

Figure 4 is presenting decision tree classifier algorithm in the python editor window. After the data splitting, the classification method is applied. Then this classifier classifies the values from the dataset and generates the confusion matrix or predicted model.

	0	1
0	782	144
1	121	12804

Figure 5: Confusion Matrix (DT)

The predicted value from decision tree method is as followings-

True Positive (TP) = 782

False Positive (FP) = 144

False Negative (FN) = 121

TrueNegative(TN) = 12804

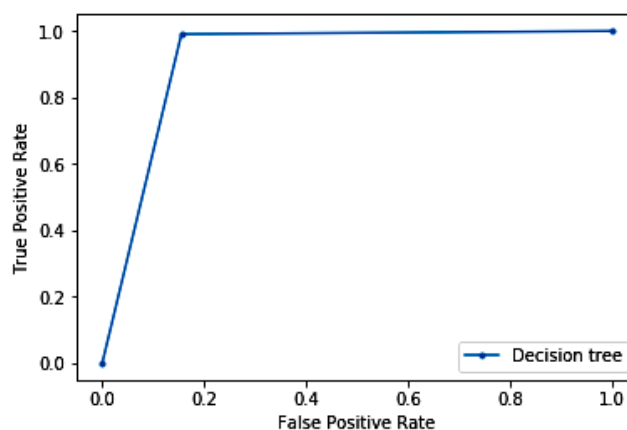


Figure 6: ROC of Decision Tree

Figure 6 is presenting the Receiver Operating Characteristic curve (ROC). The True Positive Rate (TPR) is on the y-axis, and the False Positive Rate (FPR) is on the x-axis.

Table 1: Simulation Result

Sr. No.	Parameters	Decision Tree Method
1	Accuracy	98.08
2	Classification Error	1.91
3	Precision	84.44
4	Recall	86.60
5	F-measure	85.51

Table 1 is showing the simulation results when of the decisiontree machine learning classification algorithm.

Table 2: Result Comparison

Sr. No.	Parameters	Previous work [1]	Proposed Work
1	Method	Naive Bayes	Decision Tree
2	Accuracy(%)	93.41	98.08
3	Classification error(%)	6.59	1.91

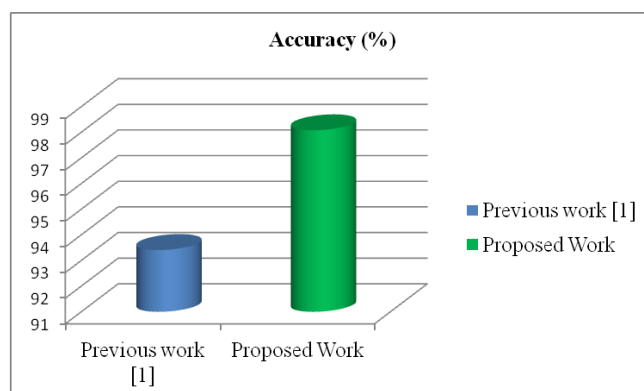


Figure 7: Accuracy Comparison

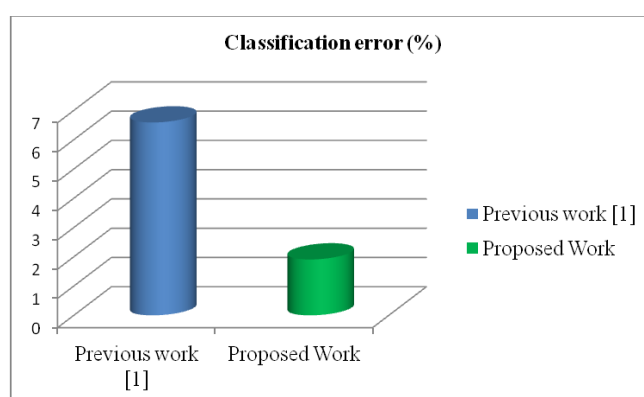


Figure 8: Classification Error Comparison

Figure 7 and 8 is presenting the graphical representation of the performance parameters comparison in terms of the accuracy and error rate.

V. CONCLUSION

Customer behaviour on the online product review after purchasing the product is major key point to make decision by other Customers. This paper presents the Customer behavior analysis in E-commerce using decision tree machine learning algorithm. It is clear from simulated results that proposed approach gives 98.08% accuracy while in previous there is 93.41% accuracy. The classification error is 1.91% in proposed while 6.59% in previous approach. Therefore the proposed approach gives significant better results than previous approach.

REFERENCES

1. V. Shrirame, J. Sabade, H. Soneta and M. Vijayalakshmi, "Customer Behavior Analytics using Machine Learning Algorithms," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198562.
2. S. Sharma and H. Kumar Soni, "Discernment of Potential Buyers Based on Purchasing Behaviour Via Machine Learning Techniques," 2020 IEEE International Conference on Advances and

- Developments in Electrical and Electronics Engineering (ICADEE), 2020, pp. 1-5, doi: 10.1109/ICADEE51157.2020.9368935.
3. R. Katarya, A. Gautam, S. P. Bandgar and D. Koli, "Analyzing Customer Sentiments Using Machine Learning Techniques to Improve Business Performance," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 182-186, doi: 10.1109/ICACCCN51052.2020.9362895.
 4. D. Suryadi, "Predicting Repurchase Intention Using Textual Features of Online Customer Reviews," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-6, doi: 10.1109/ICDABI51230.2020.9325646.
 5. S. Ghosh and C. Banerjee, "A Predictive Analysis Model of Customer Purchase Behavior using Modified Random Forest Algorithm in Cloud Environment," 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), 2020, pp. 239-244, doi: 10.1109/ICCE50343.2020.9290700.
 6. Z. Wang, Y. Zuo, T. Li, C. L. Philip Chen and K. Yada, "Analysis of Customer Segmentation Based on Broad Learning System," 2019 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2019, pp. 75-80, doi: 10.1109/SPAC49953.2019.237870.
 7. S. M. A. M. Manchanayake et al., "Potential Upselling Customer Prediction Through User Behavior Analysis Based on CDR Data," 2019 14th Conference on Industrial and Information Systems (ICIIS), 2019, pp. 46-51, doi: 10.1109/ICIIS47346.2019.9063278.
 8. K. Rusmee and N. Chumuang, "Predicting System for the Behavior of Customer Buying Personal Car Decision by Using SMO," 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2019, pp. 1-6, doi: 10.1109/iSAI-NLP48611.2019.9045571.
 9. I. Lieder, M. Segal, E. Avidan, A. Cohen and T. Hope, "Learning a Faceted Customer Segmentation for Discovering new Business Opportunities at Intel," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 6136-6138, doi: 10.1109/BigData47090.2019.9006589.
 10. S. Satpathy and S. Patnaik, "Role of Social Media Marketing on Customer Purchase Behaviour: A Critical Analysis," 2019 International Conference on Applied Machine Learning (ICAML), 2019, pp. 92-97, doi: 10.1109/ICAML48257.2019.00026.
 11. T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.
 12. A. Inoue, A. Satoh, K. Kitahara and M. Iwashita, "Mobile-Carrier Choice Behavior Analysis Using Supervised Learning Models," 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI), 2018, pp. 829-834, doi: 10.1109/IIAI-AAI.2018.00169.