JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Decision Tree Machine Learning Approach for Customer Behavior Analysis on Online Product Review

Siddhant Sharma[1], Akhilesh A. Waoo*[2]
AKS University, SATNA, MP, India
*** Corresponding Author-** Dr. Akhilesh A. Waoo,

Abstract : A significant proportion of the Clients are making their product purchases via the internet platform. The Client, in this case, does not buy the thing online but does research online before making a purchase. Consumers' decisions are heavily influenced by the reviews and ratings that other people have given certain products. If a product receives positive ratings and comments from customers, there is a greater likelihood that it will be purchased. There are situations when businesses or individuals may provide ratings that are either fake or incorrect. The artificial intelligence-based machine learning approaches can create a prediction model or make forecasts about what reviews will be truthful and which will be incorrect. Using a decision tree machine learning technique, this study gives an examination of customer behavior in online retailing. Python Spyder is the platform that is used for simulation, and the results of the simulation demonstrate an increase in the accuracy of the prediction model.

IndexTerms **– Machine learning, E-Commerce, Python, Accuracy, Error rate.**

I. INTRODUCTION

The client is possible to study user-generated material, such as reviews, ratings, and comments, to get more meaningful experiences for usage in large businesses. The analysis of such buyer behavior is beneficial for determining the customer's requirements and estimating the future expectations the consumer will have about the help. Web-based business associations can monitor the uses and opinions attached to their products through this mental review, and they can adjust their marketing strategies accordingly to provide a more personalized shopping experience for their customers. As a result, these associations can increase their authoritative value. [1]. The use of artificial intelligence (AI) in computers is a fascinating development that will revolutionize many different aspects of life in the years to come. The capacity for computerized thinking gives robots the ability to duplicate human knowledge. One of the most important aspects of man-made intelligence is the field of machine learning. It should come as no surprise that "Machine Learning" (ML) refers to computers that can teach themselves new information by drawing on previous experiences and data. The devices don't need to be modified specifically to pick up new communications. Mining the information that customers provide requires a significant investment of time and resources in today's businesses. Seeing as how the client's material has covered patterns and instances that are beneficial to the businesses. Companies use artificial intelligence processes to the information they have about their customers to categorize the customers who are most likely to buy their products and services [2].

The conventional method of thinking about business research has been significantly shifted as a direct result of developments in machine learning, which have become more prevalent in today's hyper-digitized environment. As a result of the high costs previously associated with bringing in clients for audits, conventional company auditors did not see client surveys as a contribution that might be practically useful to investigations. The proliferation of the internet completely reshaped the whole planet. At the moment, the customer satisfaction survey has emerged as the new best friend of all company investigators [3]. When it comes to e-commerce, the enormous number of online reviews may become a source of data that may be used to anticipate a customer's desire to repurchase. Because of its connection to client loyalty, repurchase intention is a vital metric for a firm to track. In this research, a technique that is based on machine learning is provided to conduct the prediction of repurchase intention based on online customer evaluations. The goal of this approach is to extract insights from a vast amount of data that is now accessible [4, 5].

Figure 1: Artificial Intelligence & E-commerce

Three distinct bunching calculations (k-Means, Agglomerative, and Meanshift) are carried out [6]. These calculations are then used to assess the results of the groups that were obtained from the computations. Python code has been written, and this code is now being prepared by using a conventional scaler on a dataset that contains two highlights of 200 preparation tests procured from a local retail store [7]. All of these factors represent the typical amount of shopping done by customers and the typical number of times customers come into the business each year. Bunching was used to divide the customers into five distinct groups, which were then given descriptive names such as Imprudent, Cautious, Standard, Target, and Reasonable. On the other hand, when mean shift bunching was used, two new groups emerged. These new groups were identified as High buyers and persistent visitors, and High purchases and sporadic visitors [8, 9].

II. METHODOLOGY

The methodology of the proposed research work is as followings-

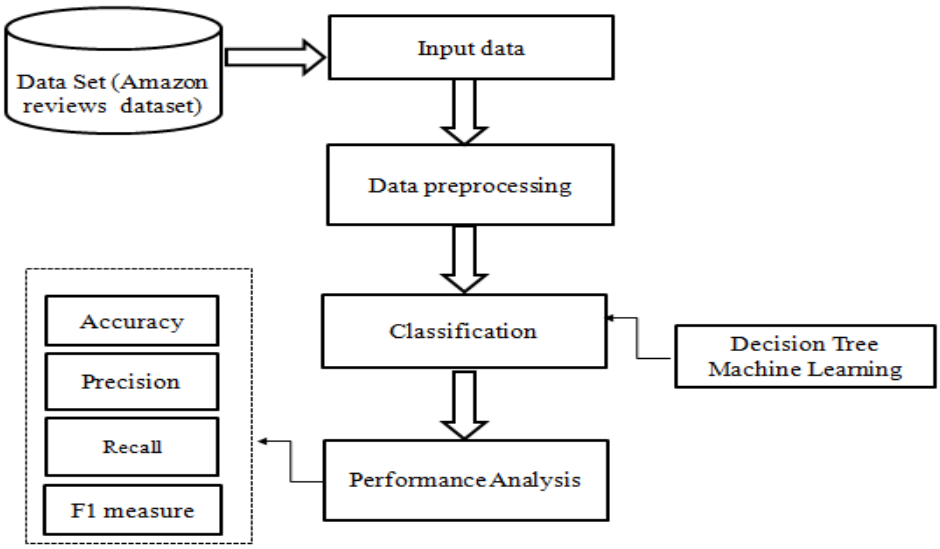

Figure 2: Flow Chart

- **Collect data set**

The customer review on the online product behavior data set of the Amazon website will be collected from the Kaggle machine learning repository to put the research efforts into execution.

This collection of data contains 69000 customer reviews covering a wide range of items.

• Processing of data in advance

To make the evaluation procedure more straightforward, the pre-processing of data comprises transforming any string variables into numerical ones. Take into account missing and null data as well.

• Dimensionality Reduction Feature extraction is a technique of dimensionality reduction in which an initial collection of raw data is reduced to more manageable groups for processing. These groups may then be analyzed. While extracting features, keep in mind things like the product name, ratings, and user names.

• Classification

To forecast how customers would rate various items, we make use of an algorithm called a decision tree.

Algorithm

Input: Customer Behaviour analysis of Reviews of Amazon Products dataset.
Take the initial data features reviews rating, reviews text, review title and reviews, and 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

III. SIMULATION RESULTS

The Python is the software that will be used to carry out the simulation. Python is an open source program that has a wide library of artificial intelligence, machine learning, and other related projects. Python, which will be used to construct and simulate the suggested notion, will use the spyder ISE as its platform [10].

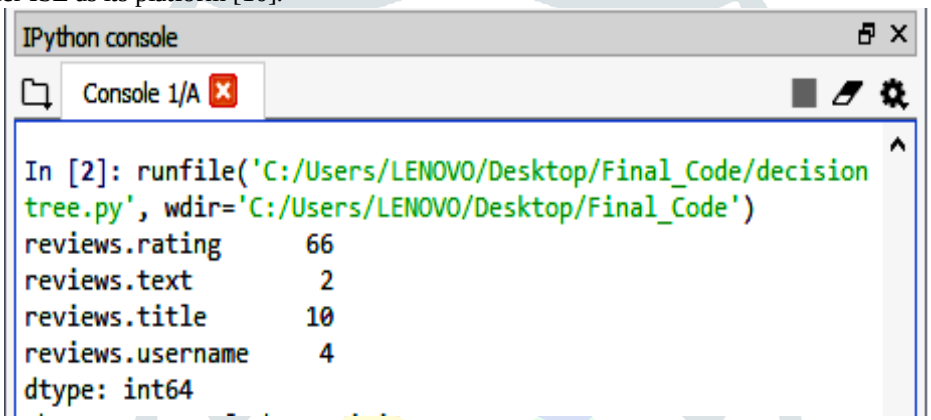

Figure 3: Dataset loading and preprocessing

The online Amazon product review dataset is shown in the python environment in Figure 3, which can be found here. After that, the preprocessing phase begins, during which the following characteristics are extracted: reviews' ratings, reviews' texts, reviews' titles, users' usernames, and so on.

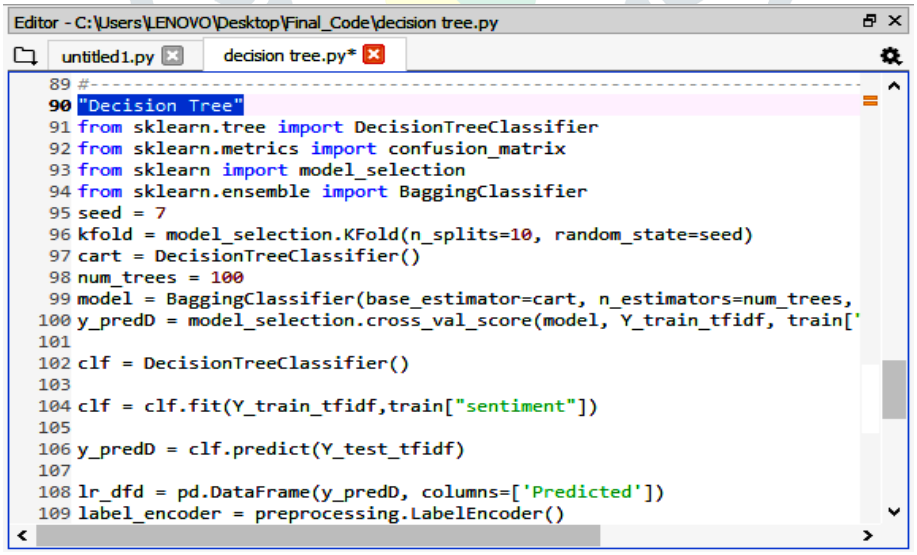

Figure 4: Decision tree classifier

The decision tree classification method is shown in the python editor window in Figure 4. After the partitioning of the data, the classification technique is carried out. After that, this classifier assigns categories to each of the values in the dataset and either produces a confusion matrix or a projected model.

	0	1
0	782	144
1	121	12804

Figure 5: Confusion Matrix (DT)

The predicted value from the decision tree method is as followings-
True Positive (TP) = 782
False Positive (FP) = 144
False Negative (FN) = 121
True Negative (TN) = 12804

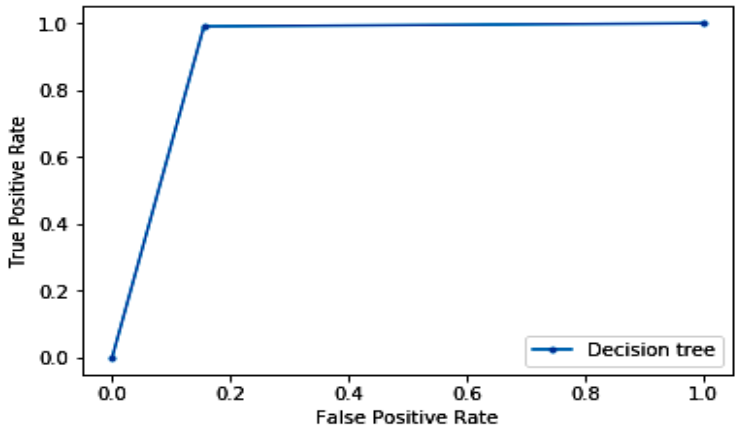

Figure 6: ROC of Decision Tree

The Receiver Operating Characteristic curve is shown in Figure 6, which may be found here (ROC). The True Positive Rate, also known as TPR, may be found on the y-axis, while the False Positive Rate, also known as FPR, can be found on the x-axis.

Table 1: Simulation Result of DT

Sr. No.	Parameters	Value (%)
1	Accuracy	98.11
2	Classification Error	1.89
3	Precision	84.45
4	Recall	86.59
5	F-measure	85.45

Table 1 is showing the simulation results when of the decision tree 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	98.11
3	Classification error (%)	6.59	1.89

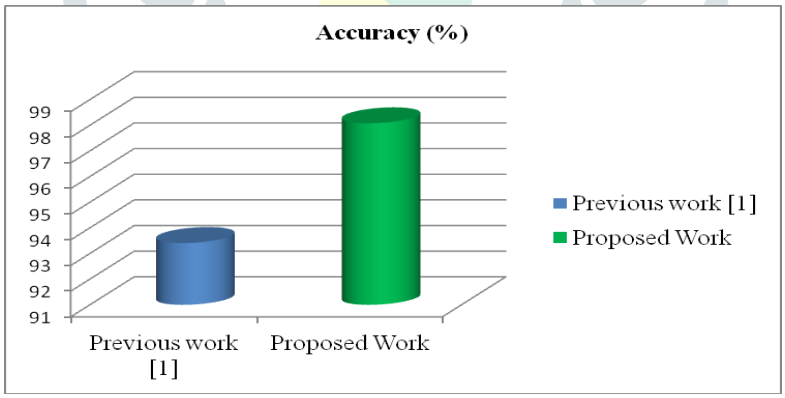

Figure 7: Accuracy Comparison

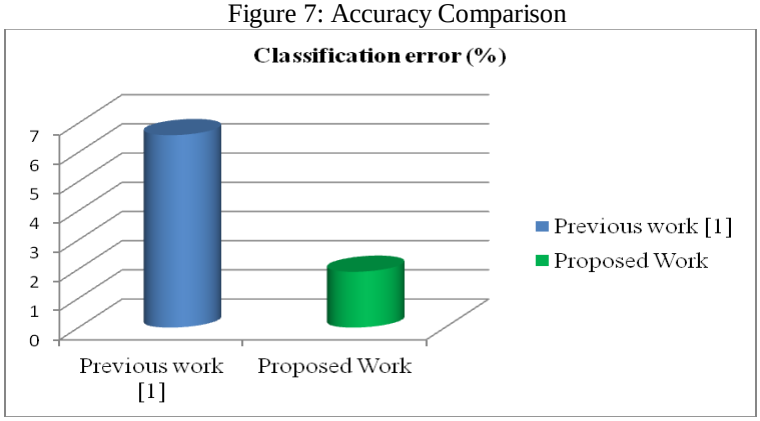

Figure 8: Classification Error Comparison

Figures 7 and 8 give a graphical depiction of the comparison of the performance characteristics in terms of accuracy and error rate, respectively.

IV. CONCLUSION

The behavior of a customer on an online product review after they have purchased the goods is a crucial essential aspect that other customers consider when making a choice. In this research, we offer a study of customer behavior in online retailing using a decision tree-based machine learning method. The results of the simulations make it abundantly evident that the suggested method achieves an accuracy of 98.08%, in contrast to the old method, which achieved an accuracy of 93.41%. The suggested method has a classification error of 1.91%, while the prior method had a classification error of 6.59%. Thus, the technique that was presented yields much better outcomes than the one that was used before.

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. Chauhan, Dipti, Jay Kumar Jain, and Sanjay Sharma. "An end-to-end header compression for multihop IPv6 tunnels with varying bandwidth." 2016 Fifth international conference on eco-friendly computing and communication systems (ICECCS). IEEE, 2016.
6. Jain, Jay Kumar, and Akhilesh A. Waoo. "An Artificial Neural Network Technique for Prediction of Cyber-Attack using Intrusion Detection System." Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799-1172 3.02 (2023): 33-42.
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. Jain, Jay Kumar, Akhilesh A. Waoo, and Dipti Chauhan. "A Literature Review on Machine Learning for Cyber Security Issues." (2022).
9. 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.
10. 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.