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

Customer Behavior Analysis in E-Commerce using Machine Learning Approach

[1]Sapna Kumari, [2]Prof. Anjul Rai, [3]Dr. kriti Jain

[1]Research Scholar, [2]Assistant Professor, [3]Professor
Department of Computer Science Engineering,
School of Research & Technology, People's University, Bhopal, India

Abstract : The development of Internet influenced many of our day-to-day activities. Ecommerce is one of the rapid growth areas in the Internet era. People are eager to buy products from online sites like Amazon, ebay, Flipkart etc. Online sites also provide facility for customers to write review on products they buy. These reviews help consumers and vendors for making decision on marketing strategies, and the improvement of products and services. Consumer behaviour models are typically based on machine learning, data mining of customer data, and each model is designed to answer one question at one point in time. Predicting customer behaviour is an uncertain and difficult task. Thus, developing customer behaviour models requires the right technique and approach. 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 machine learning algorithm. Simulation is done using python spyder platform and simulated result shows the improvement in the accuracy of the prediction model.

IndexTerms **- 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].

Figure 1: Artificial Intelligence & E-commerce

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 presented 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].

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. BACKGROUND

E. Manohar, et al.,[1] presents collective value from various classification methods makes the proposed method more effective. The accuracy metric, precision metric, recall metric and F-Score metric values are calculated to show the effectiveness of this approach. V. Shrirame, et al.,[2] proposes novel approaches (1) to identify unobserved consumer characteristics and preferences by analyzing the target consumers' and other prior reviewers' DFs; (2) to extract product-specific product content dimensions (PCDs) from review text data; (3) to predict individual consumers' sentiment before they make a purchase or write a review; (4) to classify consumers' sentiment toward a specific PCD by using context-based word embedding and deep learning models..

B. Lebichot et al.,[3] discusses the design and implementation of transfer learning approaches for e-commerce credit card fraud detection and their assessment in a real setting. The case study, based on a six-month dataset (more than 200 million e-commerce transactions) provided by the industrial partner, relates to the transfer of detection models developed for a European country to another country. X. Chen et al.,[4] studies the automated control method for regulating air conditioner (AC) loads in incentive-based residential demand response (DR). The critical challenge is that the customer responses to load adjustment are uncertain and unknown in practice.

S. Wu et al.,[5] Oversampling Technique (SMOTE) is applied to the training set to deal with the problems with imbalanced datasets. The 10-fold cross-validation is used to assess the models. Accuracy and F1-score are used for model evaluation. F1-score is considered to be an important metric to measure the models for imbalanced datasets since the premise of churn management is to be able to identify customers who will churn. K. Ali et al.,[6] we propose TagSee, a multi-person tracking system based on monostatic RFID imaging. TagSee is based on the insight that when customers are browsing the items on a shelf, they stand between the tags deployed along the boundaries of the shelf and the reader, which changes the multi-paths that the RFID signals travel along, and both the RSS and phase values of the RFID signals.

E. Umuhoza et al.,[7] focus on African customers and African financial institutions as (i) little has been done so far when it comes to understanding the spending behavior of African credit card holders; and (ii) because we believe that this segmentation will allow boosting credit card usage in Africa, thus allowing Africans to fully benefit from credit cards as other parts of the world do. L. Fan et al.,[8] proposes two new energy load forecasting methods, enhancing the traditional sequence to sequence long short-term memory (S2S-LSTM) model. Method 1 integrates S2S-LSTM with human behavior patterns recognition, implemented and compared by 3 types of algorithms: density based spatial clustering of applications with noise (DBSCAN), K-means and Pearson correlation coefficient (PCC).

Y. Yuan et al.,[9] propose a new metric, the coincident monthly peak contribution (CMPC), that quantifies the contribution of individual customers to system peak demand. Furthermore, a novel multi-state machine learning-based segmentation method is developed that estimates CMPC for customers without smart meters (SMs): first, a clustering technique is used to build a databank containing typical daily load patterns in different seasons using the SM data of observable customers. F. Zheng et al.,[10] the Multi-faceted Telecom Customer Behavior Analysis (MTCBA) framework for anomalous telecom customer behavior detection and clustering analysis is proposed. In this framework, we further design the hierarchical Locality Sensitive Hashing-Local Outlier Factor (hierarchical LSH-LOF) scheme for suspicious customer detection, and the Autoencoders with Factorization Machines (FM-AE) structure for dimension reduction to achieve more efficient clustering.

Y. Zhang et al.,[11] investigate relations between operators and customers, finding that some features of 4G service plans provided by operators indeed affect switching behaviors of 4G customers remarkably. This provides insight into reasonable design of 4G service plans for operators in the future. Our framework uncovers the root causes of intra-operator customer churn and solve the problem well. Experimental results based on real data demonstrate the effectiveness of our framework. E. A. E. Dawood et al.,[12] study aims at using k-mean, improved k-mean, fuzzy c-means and neural networks. The used dataset is labeled and creating a ýnew label as a target for neural network classification is the main aspect of this study, which helps to reduce the clustering execution time and get the best accuracy results.

III. METHODOLOGY

The methodology of the proposed research work is as followings-

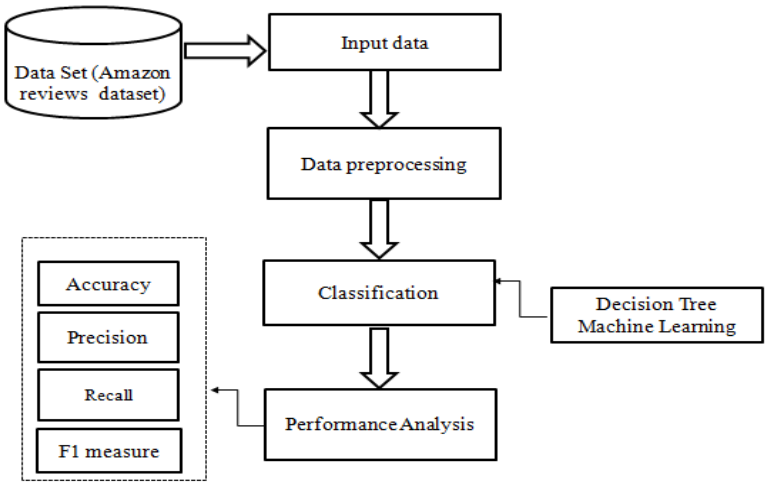

Figure 2: Flow Chart

- **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 naïve bayes and logistic regression algorithm to predict the customer rating on products

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 AND 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 IDE is platform using by the python for the implementation and simulation of the proposed concept.

```
"LOGISTIC REGRESSION"
from sklearn.metrics import confusion_matrix
print()
logreg = linear_model.LogisticRegression(solver='lbfgs' , C=4.5)
logistic = logreg.fit(Y_train_tfidf, train["sentiment"])
y_pred_lr = logistic.predict(Y_test_tfidf)

lr_df = pd.DataFrame(y_pred_lr, columns=['Predicted'])
label_encoder = preprocessing.LabelEncoder()
lr_df['Predicted']= label_encoder.fit_transform(lr_df['Predicted'])

lr_df['Original'] = label_encoder.fit_transform(test["sentiment"])
result_lr = (metrics.accuracy_score(lr_df['Predicted'] , lr_df['Original']))
print(f"The Accuracy of the Logistic Regression {result_lr}")
print("----------------------------------------------")
print("Logistic Regression Result Analysis")
```

Figure 3: Logistic Regression

Figure 3 is presenting logistic regression 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.

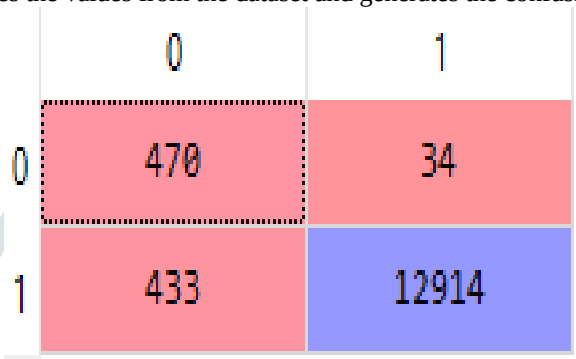

Figure 4: Confusion Matrix

The predicted value from decision tree method is as followings-
True Positive (TP) = 470
False Positive (FP) = 34
False Negative (FN) = 433
True Negative (TN) = 12914

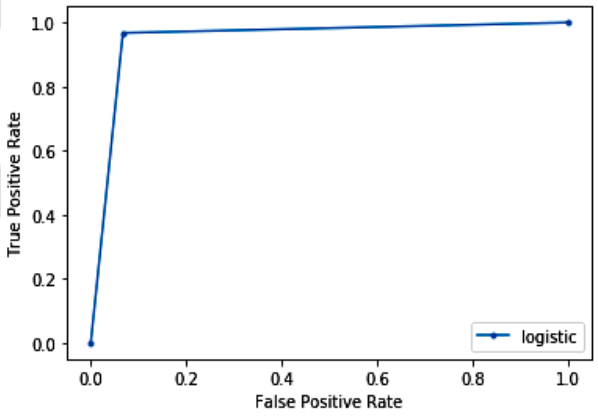

Figure 5: ROC of logistic regression

Figure 5 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 of logistic regression

Sr. No.	Parameters	Values
1	Accuracy	97
2	Classification Error	3
3	Precision	98
4	Recall	97
5	F-measure	97

Table 1 is showing the simulation results of the logistic regression machine learning algorithm.

Table 2: Result Comparison

Sr. No.	Parameters	Previous work [1]	Proposed Work
1	Method	Naive Bayes	Logistic regression
2	Accuracy (%)	93.41	97
3	Classification error (%)	6.59	3

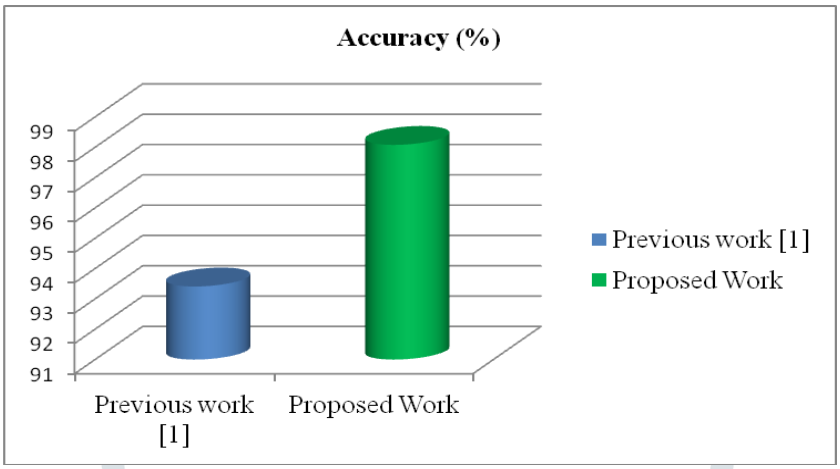

Figure 6: Accuracy Comparison

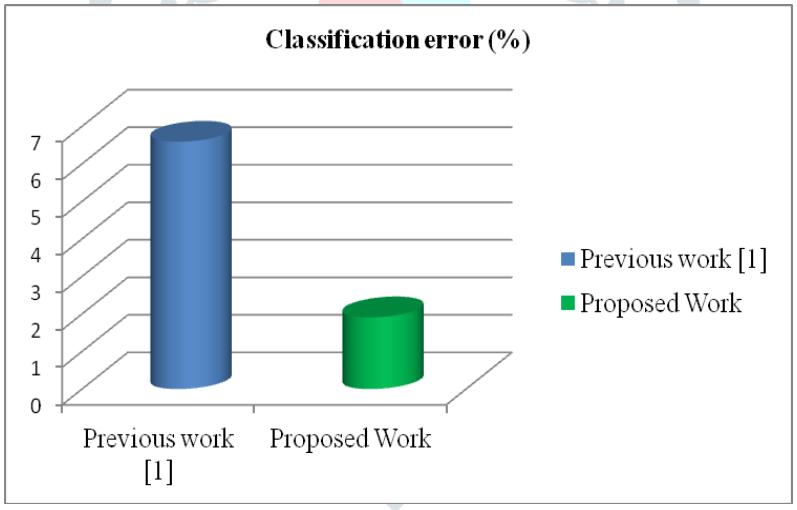

Figure 7: Classification Error Comparison

Figure 6 and 7 is presenting the graphical representation of the performance parameters comparison in terms of the accuracy and error rate.

V. CONCLUSION

There are various types of consumer reviews available in the internet that increasingly affects businesses and customers. Hence it is important to detect and eliminate such fake reviews from online websites. Machine learning techniques are suitable to predict and analysis of various problems. This paper presents the Customer behavior analysis in E-commerce using logistic regression machine learning algorithm. It is clear from simulated results that proposed approach gives 97% accuracy while in previous there is 93.41% accuracy. The classification error is 3% in proposed while 6.59% in previous approach. Therefore the proposed approach gives significant better results than previous approach.

REFERENCES

1. E. Manohar, P. Jenifer, M. S. Nisha and B. Benita, "A Collective Data Mining Approach to Predict Customer Behaviour," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 1310-1316, doi: 10.1109/ICICV50876.2021.9388558.

2. V. Shrirame, J. Sabade, H. Soneta and M. Vijayalakshmi, "Consumer Behavior Analytics using Machine Learning Algorithms," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2020, pp. 1-6, doi: 10.1109/CONECCT50063.2020.9198562.
3. B. Lebichot, T. Verhelst, Y. -A. Le Borgne, L. He-Guelton, F. Oblé and G. Bontempi, "Transfer Learning Strategies for Credit Card Fraud Detection," in IEEE Access, vol. 9, pp. 114754-114766, 2021, doi: 10.1109/ACCESS.2021.3104472.
4. X. Chen, Y. Li, J. Shimada and N. Li, "Online Learning and Distributed Control for Residential Demand Response," in IEEE Transactions on Smart Grid, vol. 12, no. 6, pp. 4843-4853, Nov. 2021, doi: 10.1109/TSG.2021.3090039.
5. S. Wu, W. -C. Yau, T. -S. Ong and S. -C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," in IEEE Access, vol. 9, pp. 62118-62136, 2021, doi: 10.1109/ACCESS.2021.3073776.
6. K. Ali and A. X. Liu, "Monitoring Browsing Behavior of Customers in Retail Stores via RFID Imaging," in IEEE Transactions on Mobile Computing, doi: 10.1109/TMC.2020.3019652.
7. E. Umuhoza, D. Ntirushwamaboko, J. Awuah and B. Birir, "Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa," in SAIEE Africa Research Journal, vol. 111, no. 3, pp. 95-101, Sept. 2020, doi: 10.23919/SAIEE.2020.9142602.
8. L. Fan, J. Li and X. -P. Zhang, "Load prediction methods using machine learning for home energy management systems based on human behavior patterns recognition," in CSEE Journal of Power and Energy Systems, vol. 6, no. 3, pp. 563-571, Sept. 2020, doi: 10.17775/CSEEJPES.2018.01130.
9. Y. Yuan, K. Dehghanpour, F. Bu and Z. Wang, "A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand," in IEEE Transactions on Power Systems, vol. 35, no. 5, pp. 4026-4035, Sept. 2020, doi: 10.1109/TPWRS.2020.2979943.
10. F. Zheng and Q. Liu, "Anomalous Telecom Customer Behavior Detection and Clustering Analysis Based on ISP's Operating Data," in IEEE Access, vol. 8, pp. 42734-42748, 2020, doi: 10.1109/ACCESS.2020.2976898.
11. Y. Zhang, S. He, S. Li and J. Chen, "Intra-Operator Customer Churn in Telecommunications: A Systematic Perspective," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 948-957, Jan. 2020, doi: 10.1109/TVT.2019.2953605.
12. E. A. E. Dawood, E. Elfakhrany and F. A. Maghraby, "Improve Profiling Bank Customer's Behavior Using Machine Learning," in IEEE Access, vol. 7, pp. 109320-109327, 2019, doi: 10.1109/ACCESS.2019.2934644.
13. I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
14. Q. Lin, H. Zhang, X. Wang, Y. Xue, H. Liu and C. Gong, "A Novel Parallel Biclustering Approach and Its Application to Identify and Segment Highly Profitable Telecom Customers," in IEEE Access, vol. 7, pp. 28696-28711, 2019, doi: 10.1109/ACCESS.2019.2898644.