International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 279-283
Published Online March 2021 in IJEAST (http://www.ijeast.com)

A NOVEL APPROACH FOR USER INTENTION ANALYSIS

Dhaarani. S
Department of CSE
Velalar College of Engineering and Technology
Erode, Tamilnadu, India

Keerthana. B
Department of BCA
Kongu Arts & Science College
Erode, Tamilnadu, India

***Abstract*— E-shopping commonly known as online shopping has become a trend-setter in today's business sector. Due to the medical pandemic, most of the people around the world have changed their lifestyle from physical shopping to online shopping. People surf the products that they are looking for in the shopping site and use their Net-banking (or) Google Pay linked with their bank account to make the payment. But from the vendor's point of view, the job is quite tedious. The vendor has to record the action of every customer opening the site. Based on the database, the vendor has to provide product suggestions for the customers when they visit the site next time. A classification algorithm named Naive Bayes is implemented to provide the product recommendations. Alongside, ensemble techniques were also used to enhance the performance of the Naive Bayes algorithm. The combination of classification algorithm with ensemble techniques shines out with the highest accuracy of about 83.47%. This study is made unique by considering the knowledge of ensemble domain to the dataset.**

Keywords— **Business sector, Online shopping, Internet, Product Recommendation, Classification, Ensemble Techniques.**

I. INTRODUCTION

Online shopping has technically arisen from the word Business to Consumer communication. The use of smartphones and the internet has become a mandatory need in one's life to survive. The increase in use of internet has eventually given way to many online businesses. People either go for browsers or smart phones to surf the products they need and make their payments either with debit card or Net banking services. The visit to online shopping sites has almost raised to about 22 billion in June 2020 from 16.07 billion visits in January 2020 [1]. As of 2020, online shopping sites has the biggest share of online purchases across the world and Amazon stands the leading online retail website [2]. Convenience in using the site, bulk number of products, easy exchange of items if damaged or misplaced, and the reviews given for each product that is purchased stand behind the success of online shopping [12]. How the customer feels while using the site is very much important for increasing online sales. If the customer is not comfortable in accessing the site, then it greatly affects the marketing of that particular online shopping site.

Predicting the purchase intention of the buyers is quite challenging because there will be no individual conversation between the vendor and the buyer [3]. To improve marketing and sales, understanding the customer's behavior and providing relevant suggestions is much more important. These product recommendations can be done efficiently by examining the purchase history of each customer as well as the database containing the action of each customer while using the site.

Fig. 1. Working of Online Shopping

The vendor maintains the database of every customer visiting the site. The database contains the record of the pages that the customer has just made window shopping, the pages where the purchase was made, the pages whose links are copied by the customers to share with others, and the products that the buyer adds to the shopping bag. The Purchase history database includes [11] the product's unique ID, name, quantity, reviews of that particular product, mode and time of payment, shipping charges, and the delivery date. In the field of machine learning and data mining, purchase prediction with the help of empirical data of buyers has become an emerging area of research. In this paper, experimentation is done with the help of a classification algorithm alongside an ensemble approach to boost up prediction accuracy.

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 279-283
Published Online March 2021 in IJEAST (http://www.ijeast.com)

II. LITERATURE REVIEW

Mohammad Hasan Aghdaie et.al., proposed a novel combined outlook for supplier clustering and cluster evaluation and selection with integrating data mining and MADM methods [3]. Suppliers are grouped using a data mining tool called Two-stage cluster analysis. The SWARA method was implied to weigh the features for cluster analysis. The outcome of SWARA was used as weighted inputs for VIKOR which then ranks the clusters from best to worst. A real case history was picked up to convey the performance and application of the model.

Michael Shekasta et.al., proposed a content-based algorithm named Purchase Intent Session-based (PISA) algorithm is employed to weigh up the purchase intention for cold start session-based scenarios [4]. This perspective hires deep learning techniques for modeling the content in addition to purchase intent guessing. Though content-based approaches go wrong while executing uneven data files, this technique handles such situations. Experiments were conducted showing that associating PISA with baseline in a non-cold start framework additionally enhances performance. The inquiry demonstrates that PISA outplays a notable deep learning baseline when new products are launched.

Humphrey Sheil et.al proposed a twin tasking approach of user classification and content ranking in e-commerce done using a non-restrictive dataset. Gradient Boosted Machine (GBM) is a straightforward idea to train as it ceaselessly extends an ensemble of classification and regression trees (CART) to foster judgment on unseen data [5]. New trees are continually added throughout the training to stronger the objective function and for correcting the flaws made by the initial trees. The features are calculated and are saved in LIBSVM (label: value) format which is used by GBM to mould a forest of CART trees. In all the experiments done, GBM performed compatibly well that the outcome attained depends greatly on feature engineering than algorithm improvements.

Dietmar Jannach et.al., came up with a better understanding of features that can make e-commerce suggestions successful in practice. Recommendations including the products that are previously recognized, fast-moving and that are given reduction at present can be very useful [6]. Resting on those factors, an innovative algorithm is designed that associates a neighbourhood-based scheme with a deep neural network to speculate the related items for a prescribed shopping zone. The data file is taken from Zalando, a well-known e-commerce site that has an annual ledger of client interactions. This work not only shows more effective outcomes in offline trials but also a perceptible raise with respect to business metric of online retailing.

Arthur Torth et.al., contemplates real web interactions [7] from a US e-commerce branch of Rakuten to speculate three possible outcomes: purchase, abandoned shopping basket and browsing-only. Clickstream data is handled to investigate with mixtures of high-order Markov Chain Models (MCMs) and mixtures of Recurrent Neural Networks (RNNs) that make use of Long Short-Term Memory (LSTM) architecture. Then each model is compared and contrasted at different lengths and reports on precision, recall and F-measures are made. It is demonstrated that LSTM RNNs generalize better and with less data than high-order Markov chain models.

III. PROPOSED ALGORITHM

The study is designed in a way to analyze the prediction and performance using the classification and ensemble technique. First and foremost, the pre-processing of data must be done to use the dataset. Removing unwanted and missed data, replacing data into standard format is the process of data pre-processing.

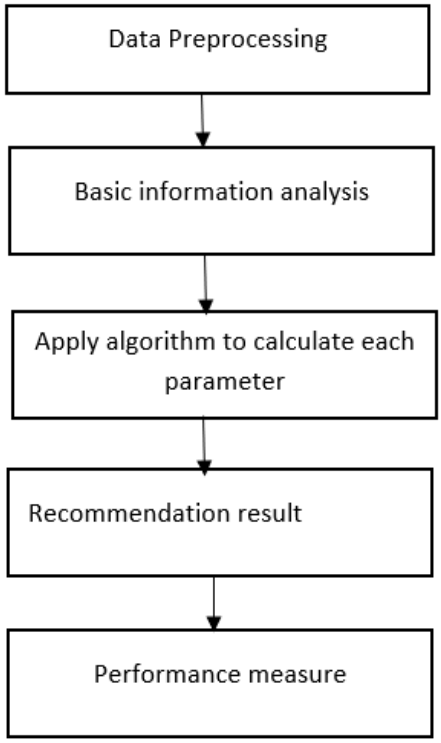


Fig. 2. Study Framework

A. Random Forest –

Random Forest algorithm involves a collection of trees that finally produces a correlated decision tree based on prediction. Each tree in the Random Forest is considered to be an ensemble model [9]. Since all the trees join as a group and produce the final result, the efficiency of Random Forest is

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 279-283
Published Online March 2021 in IJEAST (http://www.ijeast.com)

based on the strength of each tree and the correlation between them [10]. Number of attributes a node has and the total number of trees stands the basic parameters of this algorithm. Overall rating of Random Forest algorithm can be evaluated by considering the error rate and accuracy. Working of Random Forest can be split up as two phase- Creating Random Forest by combining N trees and then calculating predictions for each tree. Working steps in Random Forest algorithm is as follows:

Step 1: From the training set, select the data points and build a decision tree for each data point.

Step 2: Repeat step 1 for all the data points.

Step 3: Find predictions for each tree and update the tree with new data.

B. Naive Bayes –

Bayes theorem stands behind the success of Naive Bayes algorithm. It makes an assumption that every parameter in the dataset is not dependent on other parameters- that is why it is called Naive. It is known for its effective classification and building fast machine learning models which can make quick predictions [8]. Naive Bayes algorithm works well for very small datasets and it needs a well-trained dataset for making classifications and predictions. Formula for the Bayes theorem can be defined as:

$$P(A|B) = (P(B|A).P(A))/(P(B)) \quad (1)$$

Working steps in Naive Bayes algorithm is as follows:

Step 1: Generate frequency table for the given dataset.

Step 2: For each feature in the dataset, evaluate the probability and generate the likelihood table.

Step 3: Apply each value obtained in step 1& step 2 to Bayes Theorem to get the final posterior probability.

IV. EXPERIMENT AND RESULT

A. Dataset and Experimentation –

In our experiment, the dataset is taken from Amazon, the leading online shopping site in recent times. The dataset consists of nearly 4855 records of the users and the summary of the dataset is shown in Fig 3. Each row in the data file describes the session data of one customer. The dataset involves Unique ID, Product name, Manufacturer details, Number of reviews, Average review rating, Product description, Review description, Product category and sub-category in the site. The dataset has to be pre-processed first to replace all the categorical values with numeric values. Data filtering and data tagging is the process involved in data pre-processing stage. Data filtering is done to filter out the cases that are needed to perform the calculation from the entire dataset. In data tagging, unwanted, missed and uneven data are removed and the similar data are sorted out and grouped with a name tag.

Unique ID	Product name	Manufacturer	No. of Reviews	Avg. Review rating	Category & Sub-category	.....
AM0001	Rattle	Hornby	15	4.9 out of 5	Baby& Toddler toys > Rattles	.....
AM0002	Dice	Funky Buys	5	4.5 out of 5	Games > Dice& Dice Games	.....
AM0003	Costume	Generic	8	4.8 out of 5	Fancy Dress > Costumes	.....
AM0004	Blackboard	CCF	17	3.9 out of 5	Arts& Crafts > Blackboard	.....
AM0005	Barbie toys	Kato	1	4.1 out of 5	Characters& Brands > Barbie toys	.....
.	.	.	.	.	.	.....
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
.	.	.	.	.	.	
AM4854	Chain	Fiesta	6	3.5 out of 5	Fancy Dress > Accessories	.....
AM4855	Hand puppets	Golden Bear	7	4.4 out of 5	Puppets& Puppet Theatres > Hand puppets	.....

Fig. 3. Summary of the dataset

Classification algorithms are then applied on the dataset to predict the performance of each algorithm. To train the classification models, 75% of the total data points is taken into account and for calculating the performance of the models, 25% of the data points is taken. The experiments were carried out on a machine that has Intel Core i3 processor with 8GB RAM. For each of the model, evaluation metrics involving accuracy and error rates have been calculated for comparison. Test datafile is the dataset that is used for evaluating the performance. Data in the test datafile should be properly labeled because those labels will be compared with the labels that are predicted from classification.

B. Evaluation metrics –

For a given input data, the binary classifier generates either Yes/ No and 1/0 as the output. Prediction of “Yes” and “1” is considered to be positive while “No” and “0” is considered to be negative. Four outcomes generated as the result of binary classification is used to form the confusion matrix.

1) True Positive (TP): Test cases that are actually positive and are correctly predicted as positive.

2) True Negative (TN): Test cases that are actually negative and are correctly predicted as negative.

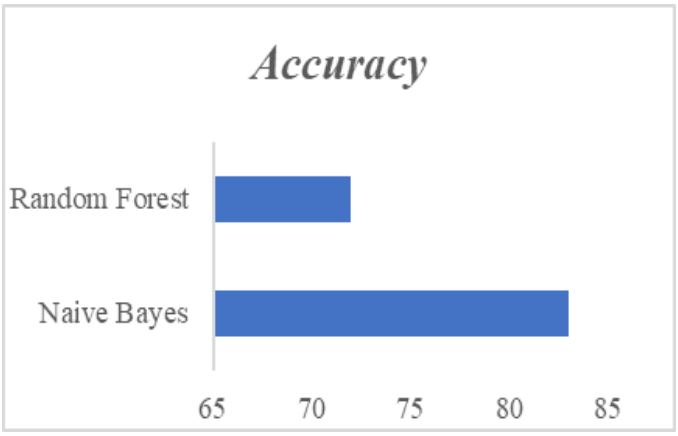
3) False Positive (FP): Test cases that are actually negative and are predicted as positive during classification.

4) False Negative (FN): Test cases that are actually positive and are predicted as negative during classification.

To determine how well the algorithm predicts the data points correctly out of all the data points, accuracy stands one of the best methods. Adding true predictions of the dataset and dividing it by the total dataset gives the accuracy percentage of the classification. 1- ERR i.e., 1- Error Rate is another way of

International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 279-283
Published Online March 2021 in IJEAST (http://www.ijeast.com)

evaluating accuracy. Fig 4 pictures the accuracy rate for both the algorithms.



$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN)=(TP+TN)/(P+N) \quad (2)$$

Fig. 4. Accuracy rate

Adding false predictions of the dataset and dividing it by the total dataset gives the error rate of the classification. Fig 5 depicts the error rate for both the algorithms.

$$\text{Error-rate}=(FP+FN)/(TP+TN+FP+FN)=(FP+FN)/(P+N) \quad (3)$$

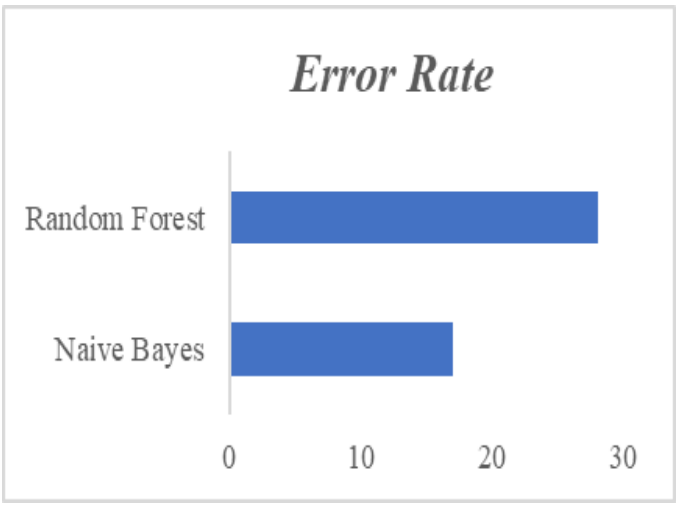


Fig. 5. Error rate

C. Results –

Performance indicators has been calculated for the two classification algorithms which is pictured in figure 3 & 4. Fig 3 depicts the values for accuracy and Fig 4 pictures the error rate. Among the two algorithms, Naive Bayes performs best and stands with an accuracy of about 83.47% and error rate of 16.53%.

V. CONCLUSION

In this paper, two different classification algorithms have been analyzed to identify the performance by using the session data of the customers. The main focus of the work is to predict the purchase intention of the clients visiting the site and evaluating their accuracy. From the experiments carried out so far, Naive Bayes stands the best in predicting the intension and also with the accuracy level of 83.47%. The next plan of the work will be using a very large dataset and also finding an efficient algorithm to get combined with Naive Bayes to increase the accuracy level than that is achieved now.

VI. REFERENCE

[1] Statistics and facts about global e-commerce. Retrieved from https://www.statista.com/topics/871/online-shopping, 2017.

[2] Statistics and facts about e-commerce in India. https://www.statista.com/topics/2454/e-commerce-in-india, 2017.

[3] Mohammad Hasan Aghdaie , Sarfaraz Hashemkhani Zolfani, and Edmundas Kazimieras Zavadskas (2014) "Synergies of data mining and multiple attribute decision making." *Procedia-Social and Behavioral Sciences* 110: 767-776.

[4] Michael Shekasta, et al., (2019) "New Item Consumption Prediction Using Deep Learning." arXiv preprint arXiv:1905.01686.

[5] Humphrey Sheil and Omer Rana (2017) "Classifying and Recommending Using Gradient Boosted Machines and Vector Space Models". *In Advances in Computational Intelligence Systems.*

[6] Dietmar Jannach, Malte Ludewig, and Lukas Lerche (2017) "Session-based item recommendation in e-commerce: on short-term intents, reminders, trends and discounts." *User Modeling and User-Adapted Interaction* 27, 3-5, 351–392.

[7] Arthur Toth, Louis Tan, Giuseppe Di Fabbrizio, and Ankur Datta (2017) "Predicting Shopping Behavior with Mixture of RNNs", *In ACM SIGIR Forum.*

[8] Kohavi, Ron. (1996) "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." Kdd. Vol. 96.

[9] Breiman, Leo (1999) "Random Forests." *UC Berkeley TR567.*

[10] Tin Kam Ho (1995) "Random decision forest." *Proceedings of 3rd international conference on document analysis and recognition.* Vol. 1. IEEE.

[11] Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen. (2002) "Data mining techniques for customer relationship management." *Technology in society* 483-502.

[12] Anil Kumar and Manoj Kumar Dash. (2014) "Factor exploration and multi-criteria assessment method (AHP) of multi-generational consumer in electronic commerce." *International Journal of Business Excellence* 767-776.