International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

ANALYSING CUSTOMER BUYING BEHAVIOUR IN ONLINE SHOPPING USING RANDOM FOREST CLASSIFIER

B.Pavani[1], A.Siva Nandini[2], B.Thanuja[3], D. Sai Hasitha[4]

[1-4]*Lendi Institute of Engineering & Technology, Jonnada, Vizianagaram, Andhra Pradesh*

Abstract - *Marketing and Customer Analytics is one of the hottest areas for the application of data science in the modern world. Customer buying behavior is identified by people's personality and character. These personality characters vary from person to person. The character includes quality, motivation, occupation and income level, perception, psychological, personality, reference groups and demographic reasons learning, beliefs, attitude, Culture and social forces. Data mining is normally used to investigate the customer activities on shopping by using various algorithms and methods. In order to take advantage of available data, modern businesses need the analytics tools that will provide them with the insight they need to deliver a personalized consumer experience. In our project, we will explore a machine learning algorithm called Random Forest Classification. Classification algorithms such as this one can increase our understanding of the customer and improve our marketing and engagement strategy.*

Key Words: Classification, random forest algorithm, behaviour analysis and data mining.

1. INTRODUCTION

There are a lot of methods and techniques to be available to analyze customer behaviour. Customers who visit online shopping sites leaves important information when they logon on server side. This valuable information is used to determine the business performance. The future customer behaviour is predicted by analysing the previous data of the customer. The profile is created by entering the data by the customer when they visit the sites. Data mining software analyses relationships among patterns based on the customer request. There is a huge amount of data available in the information industry. This data is of no use until it is converted into useful information. It is necessary to analyse this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible values: for example, high credit rating or low credit rating. Multiclass targets have more than two values: for example, low, medium, high, or unknown credit rating. In the model build (training) process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown.

It is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

First, start with the selection of random samples from a given dataset.

Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

In this step, voting will be performed for every predicted result.

At last, select the most voted prediction result as the final prediction result.

It overcomes the problem of over fitting by averaging or combining the results of different decision trees. Random forests work well for a large range of data items than a single decision tree does. Random forest has less variance then single decision tree. Random forests are very flexible and possess very high accuracy. Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling. Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

1.1 OBJECTIVE

Figuring out these problems in the analysis of customer buying behavior in online shopping to improve the efficiency and Analyzing the customer behavior using different learning problems.

2. RELATED WORK

Masud Karim et. al., (2013) had developed algorithms like decision tree and naive bayes for classification and generation of actionable knowledge for direct marketing. The goal of this work is to predict whether a client will subscribe a term deposit. We also made comparative study of performance of those two algorithms. Publicly available UCI data is used to train and test the performance of the algorithms. Besides, we extract actionable knowledge from decision tree that focuses to take interesting and important decision in business area.

Mahendra Pratap et. al., (2012) developed a mining of the customer behavior using web usage in e-commerce. The main purpose of this paper is to study the customer's behavior using the Web mining techniques and its application in e-commerce to mine customer behavior. The concept of Web mining describing the process of Web data mining in detail: source data collection, data pre-processing, pattern discovery, pattern analysis and cluster analysis.

R.Roselin et. al., (2014) developed customer behaviour analysis for credit card proposers based on data mining techniques. This study investigates the shift of consumers towards the use of plastic money, with emphasis on credit cards. A survey of consumers holding one or no credit card was used for data collection. Variables related to demographics such as age, income level and gender have also been taken into consideration.

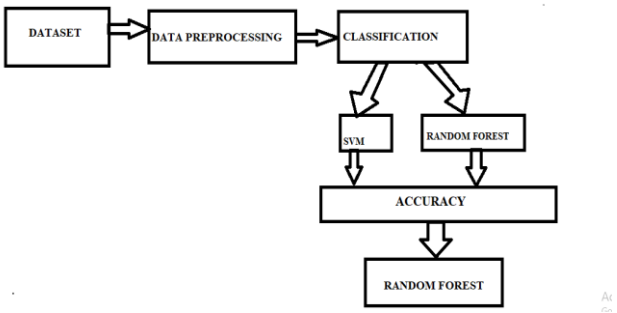
Xiaohua Hu et. al., (2005) had proposed a data mining approach for retailing bank customer attrition analysis. In this paper, we present a data mining approach for analyzing retailing bank customer attrition. We discuss the challenging issues such as highly skewed data, time series data unrolling, leaker field detection etc, and the procedure of a data mining project for the attrition analysis for retailing bank customers.

Neeraj Sharma et. al., (2013) had proposed data mining as a tool to predict the churn behaviour among Indian bank customers. The customer churn is a common measure of lost customers. By minimizing customer churn a company can maximize its profits. Companies have recognized that existing customers are most valuable assets.

Maheswari. K et. al., (2017) had developed predicting customer behavior in online shopping using SVM classifier. In this paper, the dataset is used to analyze and categorize the customer based on their purchase behavior. The classification is performed by SVM algorithm. The inventory data set and sales data set which is available in the internet is used in this work and the performance is evaluated by using the algorithms.

2.1 ARCHITECTURE

System architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships (e.g. the behavior) between them. An architectural design is the design of the entire software system; it gives a high-level overview of the software system, such that the reader can more easily follow the more detailed descriptions in the later sections. It provides information on the decomposition of the system into modules (classes), dependencies between modules, hierarchy and partitioning of the software modules.



2.2 METHODOLOGY

Step 1: The data is taken in the form of csv file. (dataset.csv)
Step 2: After the input dataset is given, the data will be preprocessed by
Removing Null values from a data frame and replace NaN values with default values.
Sometimes our data will be qualitative form that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.
Then it fit the model to the data, then transform the data according to the fitted model.
Step 3: After the preprocessing, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers. Then using s_to_super function the first column of row (t) is shifted to last column of row (t-1) and concatenated. This act transforms a normal preprocessed dataset to recurrent dataset.
Step 4: Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e. our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import **test_train_split** from **model_selection** library of scikit.
Step 5: Now to build our training and test sets, we will create 4 sets— X_train (training part of the matrix of features), X_test (test part of the matrix of features), Y_train (training part of the dependent variables associated with the X train sets, and therefore also the same indices), Y_test (test part of the dependent variables associated with the X test sets, and therefore also the same indices). We will assign to them the test_train_split, which takes the parameters — arrays (X and Y), test_size. **Step 6:** Now, we need to apply a classification technique. Here the classifier we used is SVM.
Step 7: A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.
Step 8: After applying the support vector machine (SVM) classifier to the customer dataset we are now able to calculate the performance analysis of support vector machine (SVM) on customer dataset.

```
[[1561   18]
 [   0  248]]
Accuracy_score: 0.9901477832512315
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      1579
           1       0.93      1.00      0.96       248

avg / total       0.99      0.99      0.99      1827

Cross validation Train_score 1.0
Cross validation Test_score 0.9849456596544354
```

Fig 2.1 SVM performance

Step 9: Now we are going to apply a semi supervised classifier technique. The technique we used here is Random Forest Classifier.

Step10: Random forests are an ensemble_learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

```
[[1557    0]
 [   4  266]]
Accuracy score: 0.9978106185002736
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1557
           1       1.00      0.99      0.99       270

avg / total       1.00      1.00      1.00      1827

Cross validation train_score 1.0
Cross validation test_score 0.9946626891050581
```

Fig 2.2 Random Forest Performance

Step 11: The prediction class is given to the model with the input data instances. With the help of those

input dataset the classifier analyses our required output here the input dataset is the customer dataset.

Step 12: To classify the data we took RandomForestClassifier ().
Step 13: And we now compare the accuracy scores of both Support Vector Machine (SVM) and Random Forest classifiers from the above results. Hence we find that the accuracy score of random forest classifier is higher than SVM. Therefore, we analyze the customer buying behavior in online shopping using random forest classifier.

3. CONCLUSION

This project details the great potential that classification algorithm like Random Forest Classification. To conquer most of the applications of data science, random forest classification vector is used as one of the data mining classification technique for customer predictive analysis in this modern world. From the experimental results, we will be able to know the customers behavior and increase the marketing. This analysis demonstrates the great potential that lies in the analyzing of customers buying behavior through theirs people's personalities and characters. The results of this approach are analyzed with other classification methods as a future work.

REFERENCES

[1] Masud Karim, Rashedur M. Rahman J," Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for DirectMarketing",Journal of Software Engineering and Applications, 2013, 6, 196.

[2] Yadav, Mahendra Pratap, Mhd Feeroz, and Vinod KumarYadav. "Mining the customer behavior using web usagemining in e-commerce." ICCCNT, 2012 ThirdInternational Conference on, pp. 1-5. IEEE, 2012.

[3] Ranveet Kaur, Sarbjeet Singh,"A survey of data miningand social network analysis based anamoly detectionTechniques", Egyptian Informatics Journal,productionand hosting by Elsevier, 2016,17,199-216.

[4] "Customer Behaviour Analysis for Credit Card Proposers Based on Data Mining Techniques" IJIRAE ISSN: 2349-2163 Volume 1 Issue 11 (November 2014) R.ROSELIN Assistant Professor of Computer Science, Sri Sarada College for Women(Autonomous), Salem – 16 C.HANUPRIYA Assistant Professor of Computer Science, Sri Ganesh College of Arts and Science, Salem– 16.

[5] Xiaohua Hu, (2005) A Data Mining Approach for Retailing Bank Customer Attrition Analysis. AppliedIntelligence. Vol. 22, pp. 47–60. [2] E.W.T. Ngai, Li Xiu. D.C.K. Chau, (2009) Application of data miningtechniques in customer relationship management: Aliterature review and classification. Expert Systems withApplications. Vol. 36, pp. 2592–2602.

[6] ManjitKaur, Dr. Kawaljeet Singh and Dr. Neeraj Sharma,"Data Mining as a tool to Predict the Churn Behaviouramong Indian bank customers", International Journal onRecent and Innovation Trends in Computing andCommunication, September 2013 ISSN: 2321-8169,Volume: 1, Issue: 9, pp:720 – 725.

[7] Rana Alaa El-Deen Ahmeda, M.ElemamShehaba, Shereen Morsya and NermeenMekawiea, "PerformanceStudy of Classification Algorithms for Consumer OnlineShopping Attitudes and Behavior Using Data Mining",(CSNT), 2015 Fifth International IEEE Conference on 4-6 April 2015, Electronic ISBN: 978-1-4799 1797-6,Printon Demand(PoD) ISBN: 978-1-4799-1798-3.

[8] Maheswari, K., and P. PackiaAmuthaPriya. "Predicting customer behavior in online shopping using SVM classifier." 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE, 2017.

[9] He, Benlan, et al. "Prediction of customer attrition of commercial banks based on SVM model." ProcediaComputerScience 31 (2014): 423-430.

[10] Kim, Gitae, Bongsug Kevin Chae, and David L. Olson. "A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models." ServiceBusiness 7.1 (2013): 167-182.