Prediction of Consumer Behaviour using Random Forest Algorithm

Harsh Valecha
Department of Computer Science and Engineering,
Jaypee Institute of Information Technology,
Noida, (U.P), India
harshvalecha96@gmail.com

Aparna Varma
Department of Computer Science and Engineering,
Jaypee Institute of Information Technology,
Noida (U.P), India
aparnavarma123@gmail.com

Ishita Khare
Department of Computer Science and Engineering
Jaypee Institute of Information Technology
Noida (U.P), India
khare.ishu18@gmail.com

Aakash Sachdeva
Department of Computer Science and Engineering
Jaypee Institute of Information Technology,
Noida (U.P), India
297aakash@gmail.com

Mukta Goyal*
Department of Computer Science and Engineering
Jaypee Institute of Information Technology,
Noida (U.P), India
*mukta.goyal@jiit.ac.in

***Abstract*- In the ultramodern age of technology, anticipation of market trend is very important to observe consumer behaviour in this competitive world as trends are volatile. Building on developments in machine learning and prior work in the science of behaviour prediction, we construct a model designed to predict the behaviour of Consumer. The aim of this research paper is to examine the relation between consumer behaviour parameters and willingness to buy. First we investigate to find relationship between consumer behaviour to buy products on changing parameters such as environmental factor, organizational factor, individual factor and interpersonal factor . Thus this paper proposes time-evolving random forest classifier that leverages unique feature engineering to predict the behaviour of consumer that affect the choice of purchasing the product significantly. Results of random forest classifier are more accurate than other machine learning algorithm.**

Keywords—random forst algorithm; behaviour; machine learning; customer.

I. INTRODUCTION

Purchasing is an inevitable part of life. Purchasing depends on the choice and utility of the consumers. Therefore, Consumers plays a vital role in purchasing. In purchasing a product to know the consumers mind set is important. Some products appeal to customers while the same one does not appeal to some other. Hence, buying behaviour's usually differs person to person.

To keep up with changes, understanding a consumers requires current analytical methods that refine on data points to reveal the behaviour. Analyzing customer behaviour is a very challenging task so it is very vital to know which strategy should adopt. In the modern technological world, there is need of innovative marketing for analysing consumer behaviour.

So it is very important to understand why, when, how and what other factors that influence buying decision of the consumers. For predicting consumer behaviour intelligent methods required other than finding out what they had purchased earlier [6],[7]. Consumers can classify according to their actual behaviours[8].

This process consists of five sections. Second section explain the literature surveys. It discusses the factors influencing consumer's behaviour, performance of Forest Trees algorithm and comparison of different approaches for intelligent decision. Third section explains the methodology to deduct the buyingu behaviour of consumer using random forest algorithm. Implementation and results discussed in section 4 which find the best configurations for higher prediction accuracy.

II. LITERATURE SURVEY

This paper proposes that there are several factors which affect the buying decision of consumer. To improve and manage multiple sources of enhancing customer relationship it is very important to know once behaviour. There are several factors like external factors such as cultural factor and social factor which depends on nationalities, geographic regions, racial groups, religions, income, profession, and education whereas reference groups, family, role and status. Internal factors such as age, profession, education,, income, personality and life style are depend on what to buy or what not to buy[1]. Sometimes buying behaviour depends on the psychological factors such as perception, motivation, learning from past experience , beliefs and attitude[2].

978-1-5386-5002-8/18/$31.00 ©2018 IEEE

A survey has been done to know the buying behaviour of the customer. This survey was conducted on different states of super market to know about the preferences and choices of customers in different regions. Questions were asked to 200 respondents about importance of price while choosing any product, priority of service; location etc. Rrespondent has to fill the choice accordingly. Survey was conducted by online and offline mode. On the basis of this survey the category of behaviour is assigned to customer.

Following given below are the Survey Questions and responses-

1) How Important is price to you while choosing the product?

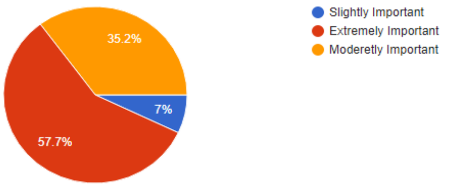


Fig 1. Price Comparison

Fig.1 shows that 57.7% respondent has answered that prices are extremely important to them.

2) What is your priority while choosing the product?

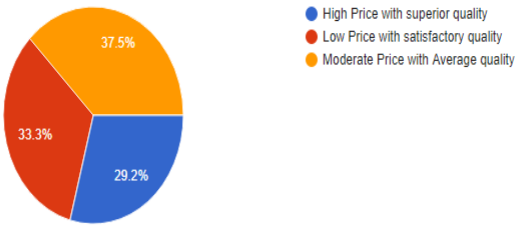


Fig.2. Priority comparison corresponding to Product

There is a negligible difference between the high price with superior quality and moderate price with average quality(fig.2.).

3) What is your main priority?

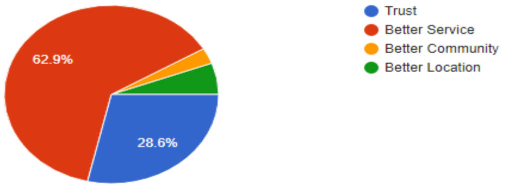


Fig.3. Priority Comparison corresponding to Service

Better services has become an important part of everyone life . Fig.3 shows that 62.9 % agreed on that.

4) For how much social standards affect your buying decision?

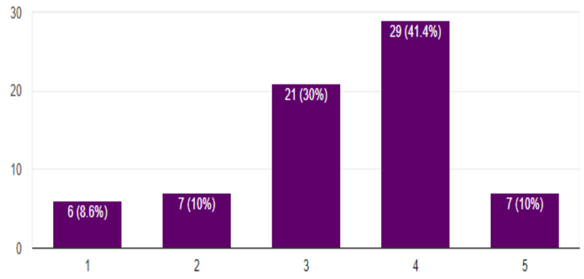


Fig.4. Effect of Social Standard

A likert scale is used to find the effect of social standard on buying behaviour. 29 percent consumer agreed on that social pressure also effect their buying behaviour(fig.4.).

5) Which product fascinates you the most?

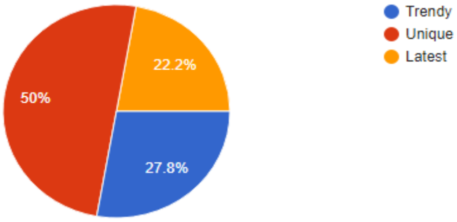


Fig.5. Impact of Product

Fig.5 shows the likes and dislike of the consumer. It seem that most of the consumer do not want that the same product should be consumed by others.

6) What would you do if you go some new place?

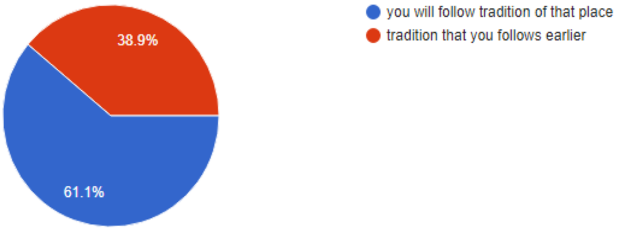


Fig.6. Exploration of new places

Fig. 6. represent that consumer wants to play safe. It follows the other tradition.

7) For which purpose you make frequent purchases?

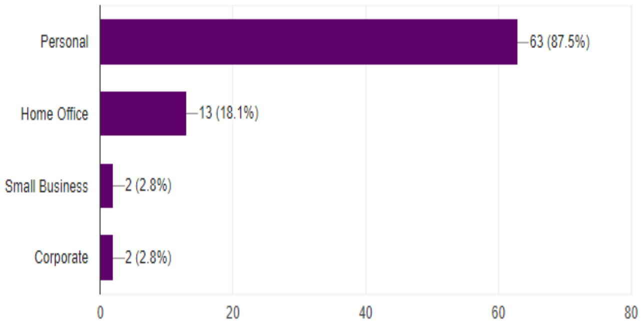


Fig.7. shows purchasing habits

Most of the people buy things use for personal as shown in fig.7.

8) Why do you buy products?

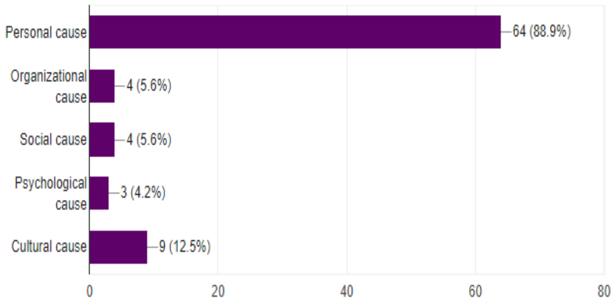


Fig. 8. Shows cause for buying

This question is more similar to the previous one where it also verify that most of the consumer consume items for their personal cause.(fig.8). Based on previous studies and based on the feedback of these questionnaire this paper divide the consumer behaviour on four classes. These behaviours are interpersonal, Individual, organisational and environmental. Table 1 represent the characteristics of these four classes.

Various intelligent techniques such as K-means, decision tree. Hierarchical algorithm are used for clustering. Clustering is done in such a way the objects within the same group are very like/similar to each other but they are very unlike/dissimilar to the objects in some other group. This algorithm creates a forest consisting of a number of trees, as the number of trees in the forest increases more potent and powerful the forest becomes. Similarly, in the random forest classifier, the higher the number of trees in the forest the higher the accuracy results are [4][5]. Table 2 shows the comparison between different techniques.

TABLE 1: CLASSIFICATION OF CONSUMER BEHAVIOUR

Factors	Traits	Description
Individual	Age, Education, Profession, Income,Personality and lifestyle	Refers demographic factors, sex, race, age etc
Environmental	Family, Role and status,	Depend on cultural difference. our surrounding depends mixture of norms, convictions, attitudes, financial values, moral values and habits and including
Interpersonal	Reference Groups	Consumers take reference group in consideration while making buying decisions and these are usually cumulative.
Organizational	Occupation, Learning, motivation, perception, beliefs and attitudes	Every global organization is recognized by its objectives and goals, an efficient organizational structure and a well acknowledged system and producer for buying

TABLE 2: COMPARISON OF CLUSTERING TECHNIQUES

Technique	Feasibility for huge data set	Dependence on iteration	High speed	Input Specified by user	Applicability for all data based
Agglomerative(Hierarchical)	×	✓	×	✓	×
Density based	×	✓	✓	×	×
EM	✓	✓	✓	✓	×
K-Mean	✓	✓	✓	✓	×
Random Forest	✓	×	×	×	✓

III METHODOLOGY

Firstly data set has been collected from kaggle.com on consumer behaviour. The size of the data set was 600, 2000, and 3900. This data set has been divided on the basis of priority, region, customer segment , product category, and product subcategory. An algorithm is applied to select the important feature of the data Seven features has been extracted from the data set . these features are shown in the table 3. They are the Home office, corporation, small business, consumer, furniture, office supply and technology. If the customer shows inetrest to buy home office and furniture then he/she has considered as interpersonal. If the customer shows her interest to buy small business, consumer and office supply then she is considered that she is buying furniture for herself. Likewise rules hase been devised to assign the behaviour of the customer based on the features and sub feature. After that random forest algorithm has been used to used to predict the behaviour of customers . Fig.9 and fig. 10 shows the architecture of classification customer behaviour. Different trees have been generated which has been classified into targeted value.

TABLE 3: SOME OF THE RULES TO PREDICT THE CUSTOMER BYING BEHAVIOUR.

Behaviour Assigned	Shopping Pattern						
	Home office	Corporation	Small Business	Consumer	Furniture	Office supply	Technology
Interpersonal	✓	✗	✗	✗	✓	✗	✗
Individual	✗	✗	✓	✓	✗	✓	✗
Environmental	✗	✓	✗	✗	✗	✗	✓
Organizational	✗	✓	✗	✗	✗	✓	✗

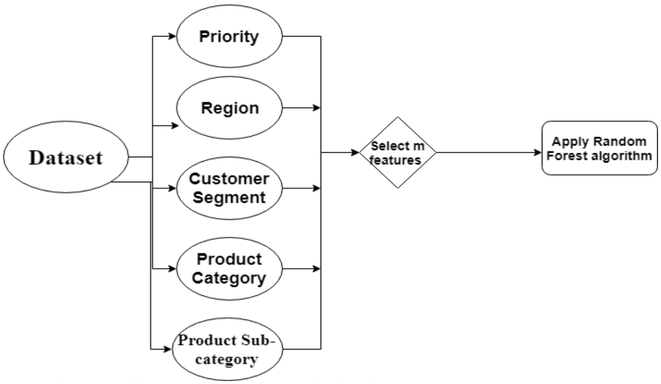


Fig.9. Architecture of customer behaviour

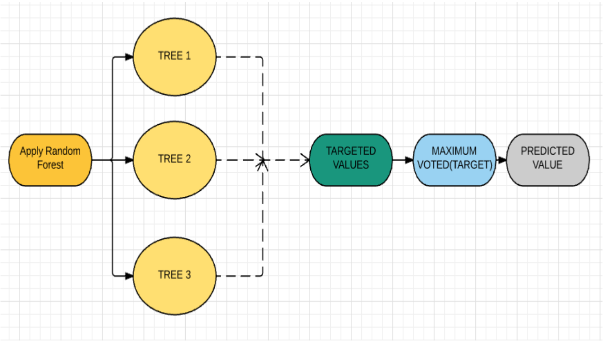


Fig.10. Classifying customer behaviour using random forest algorithm.

```
Algorithm:
  For D= 1 to n where n represent the number of respondent
          for B = 1 to m  where m represent the features
          Trees are the total no. of trees produced
                  for  i=1 to B
                      Choose sample $D_i$ form D
                      Construct tree $T_i$ using D:
      At each node , choose subset of features
      and only consider splitting on these features
      save the output of each tree in Trees
          Take majority voted value(Trees)
              Return voted value
```

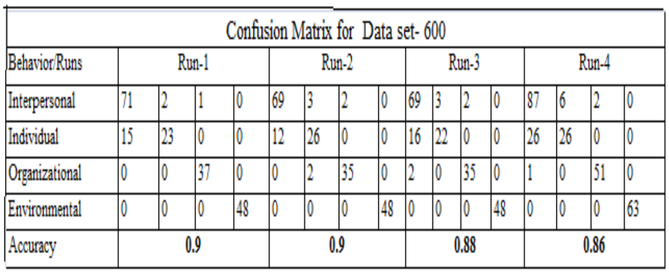
Fig.11. Random forest algorithm

IV IMPLEMENTATION AND RESULTS

The classification of consumer buying behaviour is implemented in python language. The data set has been collected the kaggle repository. This data set has pre-defined categories or segments for customer classification. Data is divided into different categories such as customer segment, product sub category. The customer segment is divided in to different sub category such as Home,office, Corporate, Small Business and Consumer. Product segment is divided into categories such as furniture, office supplies, and technologies. The product subcategories are again divided into book cases, tables, chairs etc.

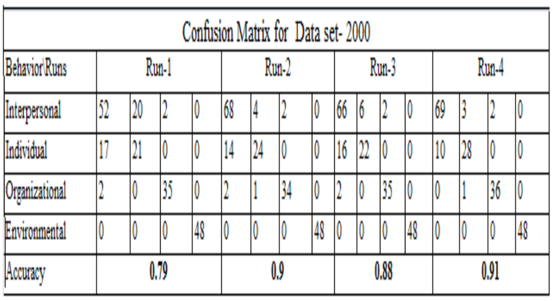
In literature survey, this paper discuss about the preferences and choices of the customer in different region through questionnaire. There were 200 respondents to fill this survey. This survey was conducted by online and offline mode. The result of the survey has been used to assign the behaviour of the customer. The classification of behaviour is divided into four part. The behaviour of customer are Interpersonal, individual, environmental, and organizational. Table 3 shows some of the rules used to assign the buying behaviour of the customer.

A random forest algorithm is applied to classify the behaviour of the customer. The algorithm is applied for different data set for different run. The size of datas et is 600, 2000 and 3900. The data set is divided in to 60-40, 70-30 and 75-25 training and testing data set. Each data set is run for four runs. Fig. 12 shows the confusion matrix for the size of data set 600 for 70-3- training and testing data set.. The algorithm execute for four runs. It is seen that each run has accuracy of above 85% percent The mean accuracy is 89 percent.



Confusion Matrix for Data set- 600																
Behavior/Runs	Run-1				Run-2				Run-3				Run-4			
Interpersonal	71	2	1	0	69	3	2	0	69	3	2	0	87	6	2	0
Individual	15	23	0	0	12	26	0	0	16	22	0	0	26	26	0	0
Organizational	0	0	37	0	0	2	35	0	2	0	35	0	1	0	51	0
Environmental	0	0	0	48	0	0	0	48	0	0	0	48	0	0	0	63
Accuracy	0.9				0.9				0.88				0.86			

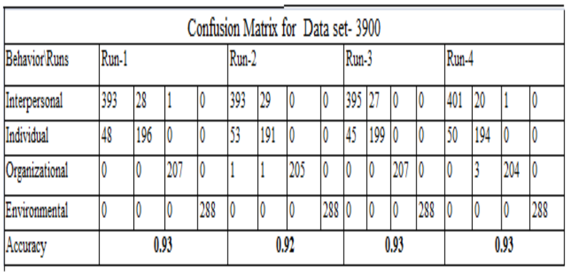
Fig. 12. Confusion matrix for data set 600



Confusion Matrix for Data set- 2000

Behavior\Runs	Run-1				Run-2				Run-3				Run-4			
Interpersonal	52	20	2	0	68	4	2	0	66	6	2	0	69	3	2	0
Individual	17	21	0	0	14	24	0	0	16	22	0	0	10	28	0	0
Organizational	2	0	35	0	2	1	34	0	2	0	35	0	0	1	36	0
Environmental	0	0	0	48	0	0	0	48	0	0	0	48	0	0	0	48
Accuracy	**0.79**				**0.9**				**0.88**				**0.91**			

Fig.13. Confusion matrix for data set 2000.



Confusion Matrix for Data set- 3900

Behavior\Runs	Run-1				Run-2				Run-3				Run-4			
Interpersonal	393	28	1	0	393	29	0	0	395	27	0	0	401	20	1	0
Individual	48	196	0	0	53	191	0	0	45	199	0	0	50	194	0	0
Organizational	0	0	207	0	1	1	205	0	0	0	207	0	0	3	204	0
Environmental	0	0	0	288	0	0	0	288	0	0	0	288	0	0	0	288
Accuracy	**0.93**				**0.92**				**0.93**				**0.93**			

Fig.14 Confusion matrix for data set 3900.

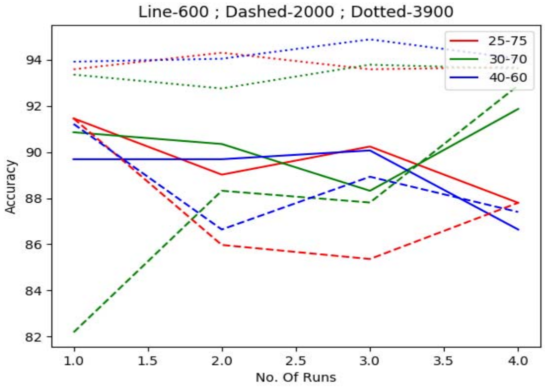


Fig. 15 Accuracy with different training –testing pair for different runs.

Like wise fig. 13. and fig.14. represent the confusion matrix for data set 2000 and 3900 for training and testing pair. The mean accuracy of data set 2000 is 87 percent whereas the mean accuracy for data set 3900 is 94 percent.

Fig. 15 shows the mean accuracy for different testing pairs for different runs. It is clearly seen that increasing the data size, accuracy is improved up to 94% percent. . The data set also analyzed the behavior of customer region wise. Figure 16, 17 ,18 represent the different bar graph for different region for training testing pair 70-30.

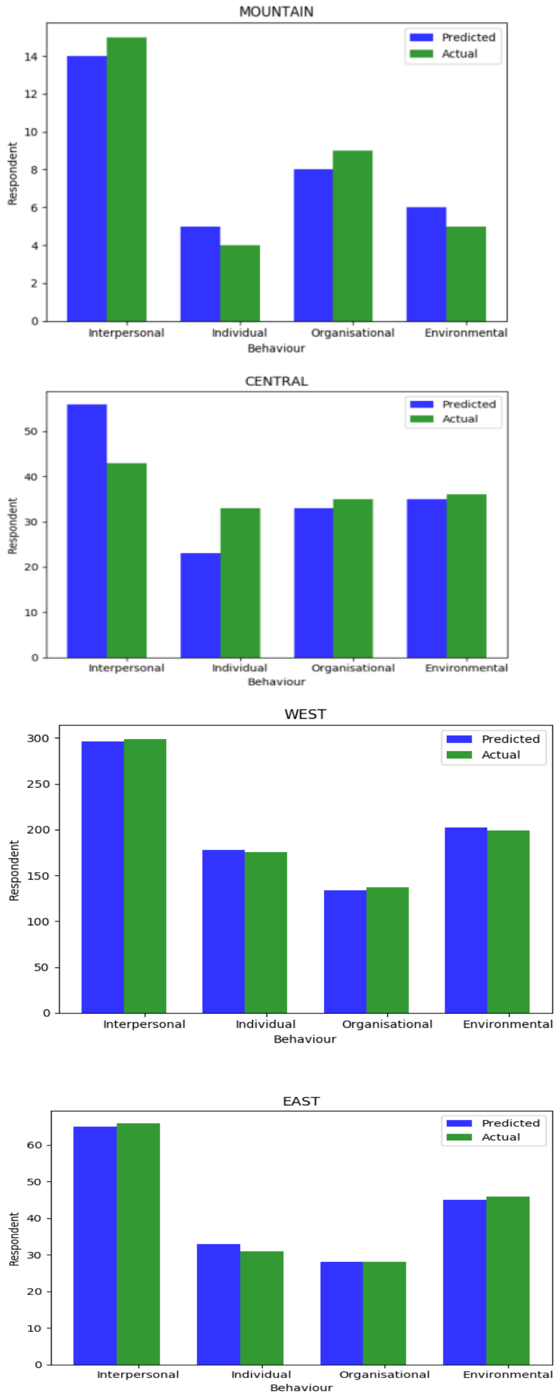
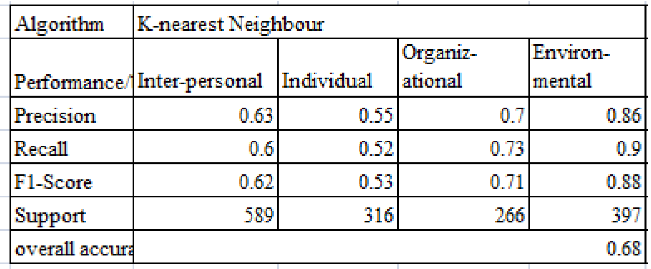


Fig.16. Comparison of different region with different buying behaviour

Fig.16 shows the comparison between the predicted and actual behaviors of particular regions. Behavior is depicted on the X-axis whereas Y-axis represent the number of respondent. Bar graph with blue color shows predicted behavior i.e. output from random forest algorithm whereas green color represent the actual behavior which is taken from data set. These four regions are central, west , east and mountain.

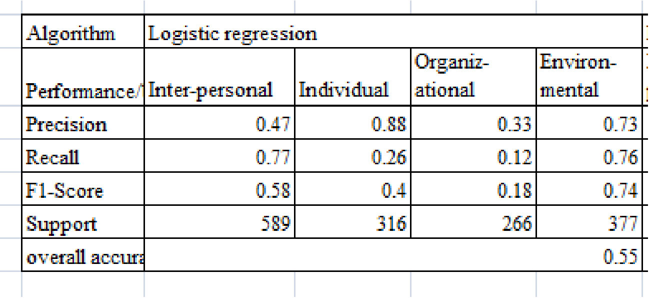
A comparison of performance evaluation with other algorithm as shown in table 3,4,5 are measured on the basis or precision, recall, f1-score and support on 1600 respondent.

TABLE 3: PERFORMANCE MATRIX FOR K-NEAREST NEIGHBOR



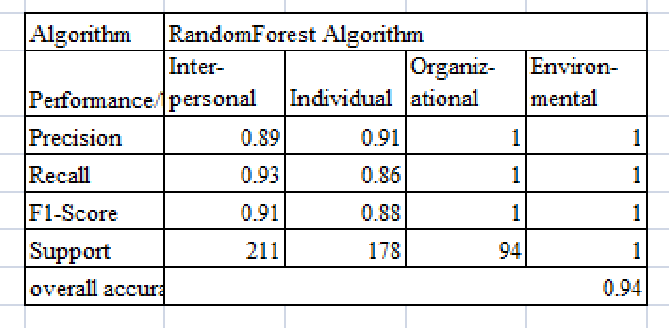
Algorithm	K-nearest Neighbour			
Performance/	Inter-personal	Individual	Organiz-ational	Environ-mental
Precision	0.63	0.55	0.7	0.86
Recall	0.6	0.52	0.73	0.9
F1-Score	0.62	0.53	0.71	0.88
Support	589	316	266	397
overall accura				0.68

TABLE 4: PERFORMANCE MATRIX FOR LOGISTIC REGRESSION



Algorithm	Logistic regression			
Performance/	Inter-personal	Individual	Organiz-ational	Environ-mental
Precision	0.47	0.88	0.33	0.73
Recall	0.77	0.26	0.12	0.76
F1-Score	0.58	0.4	0.18	0.74
Support	589	316	266	377
overall accura				0.55

TABLE5: PERFORMANCE MATRIX FOR RANDOM FOREST ALGORITHM



Algorithm	RandomForest Algorithm			
Performance/	Inter-personal	Individual	Organiz-ational	Environ-mental
Precision	0.89	0.91	1	1
Recall	0.93	0.86	1	1
F1-Score	0.91	0.88	1	1
Support	211	178	94	1
overall accura				0.94

It can be easily seen that the overall accuracy is of random forest algorithm is better than K-nearest neighbor and Logistic Regression.

CONCLUSION

This paper shows the customer behavior in this competitive world is volatile. The behavior of the customer is depend on so many factors such as environmental factor, organizational factor, individual factor and interpersonal factor . Moreover according to their use the behavior of the customer is changed time to time. According to this data set it is seen that most of the furniture is used in corporate. Very few people buy for home office. It is also seen that buying behavior of the customer depend on interpersonal and environment in different region. Machine learning techniques are used to predict the behavior of the customer which give good accuracy. This paper used random forest algorithm, has given 94% accuracy.

REFERENCES

[1] M. Khaniwal,“ Consumer buying behaviour” , International Journal of Innovation and Scientific Research, vol. 14, no. 2, pp. 278–286, April 2015.

[2] A.H.Kumar, F.S.John, & S. Senith, “A Study on factors influencing consumer buying behavior in cosmetic Products”, International Journal of Scientific and Research Publications, vol. 4, no.9, pp.1-6, 2015.

[3] A.O. Abbas, “Comparisons Between Data Clustering Algorithms” ,International Arab Journal of Information Technology (IAJIT), vol 5 no 3,pp 320-325, July 2008.

[4] N. Ghatasheh, “ Business analytics using random forest trees for credit risk prediction: A comparison study”, International Journal of Advanced Science and Technology, vol 72, pp.19-30, 2014.

[5] R.Joshi, R.Gupte, & P.Saravanan,, “A Random Forest Approach for Predicting Online Buying Behavior of Indian Customers”, Theoretical Economics Letters, vol 8, no.3, pp 448, Feb 2018.

[6] I. Stulec, & K.Petljak, “The research on buying behaviour among group buyers: the case of Croatia”, International Journal of Knowledge-Based Development, vol 4 no.4, pp 382-401, 2013

[7] M.Solomon, R.Russell-Bennett,&J.Previte, “ Consumer behaviour.”, Pearson Higher Education AU, 2012.