Machine learning based classification and segmentation techniques for CRM: a customer analytics

Narendra Singh

GL Bajaj Institute of Management and Research,
Knowledge Park-III,
Greater Noida, Uttar Pradesh, 201306, India
Email: narendra.naman09@gmail.com

Pushpa Singh*

Delhi Technical Campus,
Knowledge Park-III,
Greater Noida, Uttar Pradesh, 201306, India
Email: pushpa.gla@gmail.com
*Corresponding author

Krishna Kant Singh

KIET Group of Institutions,
Delhi-NCR, Ghaziabad,
Uttar Pradesh, 201206, India
Email: krishnaiitr2011@gmail.com

Akansha Singh

ASET, Amity University,
Uttar Pradesh, Noida, 201313, India
Email: akanshasing@gmail.com

Abstract: Machine learning and data mining help companies to build a tool that can make and take actions based on customer knowledge and information. Customer information is the base of maintaining long term relationship with customers and also known as customer relationship and management (CRM). Classification and segmentation of customer data set is utilised to maintain efficient relation with customers and subsequently increase the profitability and productivity. In this paper, author proposed customer segmentation based on demographic properties like gender, age and spending score and analysed the data set for interesting fact. The derived attribute data set is investigated for classification. Classification is used to categorise each customer into a number of classes, i.e., 'gold', 'silver', 'elite' and 'occasional'. Comparison of different classification algorithm is simulated by WEKA tool. Multi-layer perceptron (MLP) is found as the best classification algorithm with an accuracy of 98.33% compared to Naïve Bayes, regression and J48.

Keywords: machine learning; classification; segmentation; customer relationship management; CRM; WEKA; multi-layer perceptron; MLP.

Reference to this paper should be made as follows: Singh, N., Singh, P., Singh, K.K. and Singh, A. (2020) 'Machine learning based classification and segmentation techniques for CRM: a customer analytics', *Int. J. Business Forecasting and Marketing Intelligence*, Vol. 6, No. 2, pp.99–117.

Biographical notes: Narendra Singh is an Assistant Professor in the area of IT and Marketing at G.L. Bajaj Institute of Management and Research, Greater Noida. He has 15 years of teaching and research experience. He is pursuing PhD (Management) from Dr. A.P.J. Abdul Kalam Technical University, UP, Lucknow and Master in Business Administration (MBA) from Uttar Pradesh Technical University, Lucknow. His research areas include consumer behaviour, customer relationship management, electronic commerce, management information system and IoT. He has 14 publications of repute in his credit. His research work has appeared in various refereed international journals with publication houses like, Emerald and Inderscience.

Pushpa Singh is working as an Associate Professor in Computer Science and Engineering in IEC College of Engineering, Gr. Noida, India. She is having more than 16+ years exposure to teaching BTech and MCA students. She has acquired MCA, MTech (CSE), and PhD (CSE) from AKTU Lucknow in the area of wireless network. Her current areas of research include performance evaluation of heterogeneous networks, machine learning, and cryptography. She has 30 papers in reputed international journals and conferences. She has published four books and contributed in book chapter. She is also a member of Computer Society of India (CSI).

Krishna Kant Singh is working as an Associate Professor in Electronics and Communication Engineering in KIET Group of Institutions, Delhi-NCR, Ghaziabad, India. He has wide teaching and research experience. He has acquired BTech, MTech and PhD (IIT Roorkee) in the area of image processing and remote sensing. He has authored more than 54 research papers in Scopus and SCIE indexed journals of repute. He has also authored 28 technical books. He has also served several SCIE indexed journal as reviewer. He is also an Associate Editor of *Journal of Intelligent & Fuzzy Systems* (SCIE Indexed), *IEEE ACCESS* (SCIE Indexed) and Guest Editor of *Open Computer Science* (ESCI & Scopus Indexed). He is also member of Editorial board of *Applied Computing and Geosciences* (Elsevier).

Akansha Singh has BTech, MTech and PhD in Computer Science. She received her PhD from IIT Roorkee in the area of image processing and machine learning. Currently, she is working as an Associate Professor in ASET, Amity University, Uttar Pradesh, Noida, India. She has to her credit more than 40 research papers, 20 books and numerous conference papers. She has been the editor for books on emerging topics with publishers like Elsevier, Taylor and Francis, Wiley etc. She has served as reviewer and technical committee member for multiple conferences and journals of High Repute. She is also the Associate Editor for IEEE Access journal which is an SCI journal with impact factor of 4.018. She has also undertaken government funded project as Principal Investigator. Her research areas include image processing, remote sensing, IoT and machine learning.

1 Introduction

The digital revolution and the increasing amount of data generated by firms/organisations in the past few decades has led to a great interest in the field of machine learning (ML) and deep learning. Organisations and technology companies are employing *ML* based *predictive analytics* to gain an edge over their competitors. The objective of ML techniques is to discover 'hidden' information in data, which is almost impossible by traditional means based on human analytical skills. The ML techniques are used for mining data for business intelligence and saleable strategies for the customers like their classification in different categories, making strategies of promotional schemes, and for improving the customer relationship management (CRM) (Cioca et al., 2013; Rahman and Khan, 2017). In present scenario, business processes are becoming more and more customer oriented and placed as a top priority of management. Due to the technological advancement in e-commerce, M-commerce, virtual marketing and digital marketing, every product is just one click away from the customers. This has led to an extremely fierce competition, which is necessary to ensure that the consumers receive the highest possible standards of quality in order to retain them (Singh and Agrawal, 2019a; Adebiyi et al., 2016) and reduce the churning rate. The focus of CRM is to expand the customer-service and support in customer retention.

The customer is vital for each firm and organisation. For the identification and retention of their target customer, it is very important to have data analysis, which is used to explore valuable insights and trends to know customer metrics and traits. It is also very important to observe foremost key factors affecting the buying decisions of customers to buy any product and services (Singh et al., 2018; Yadav et al., 2018). ML is one of the popular data analyses that mechanise analytical model structure, which is valuable to growth in buying behaviour. ML techniques is extensively used in prediction of customer segmentation, customer lifetime value (CLTV), churning, sales etc. Customer segmentation is useful in understanding what demographic and psychographic sub-populations, there are within your customers in a business case (Sgaier et al., 2017) and utilised this information to increase profits, image, values and inventory management. Companies in any businesses already recognise that attaining new customers is not enough for lasting success and efforts need to make in order to identify the customer segmentation towards retention.

In this paper, a customer segmentation and classification technique for the analysis purpose of business intelligence is proposed. Demographic property of customer is taken as parameter for customer segmentation to know analytics about the customer. That would help in CRM, efficiency and productivity of shopping Mall. Further, machine-learning technique is used to predict grading of the customer.

2 Related work

The basic objective of CRM is to suggest that buyers should identify their most profitable customers and should focus on building a relationship with them (Malmi et al., 2004). Application of data mining techniques in CRM performs by using classification, clustering, and prediction based on customer knowledge (Ngai et al., 2009). An efficient CRM-data mining framework was used for the prediction of customer behaviour to enrich the decision-making processes for the customer retention (Bahari and Elayidom, 2015). Technology is always helpful to understand the customer perception, needs and expectation of the consumer for example IoT in a restaurant (Sudhagar, 2019), Blockchain in consumer relationships (Boukis, 2019), ML in crude price (Mitra and Banga, 2019), customer demographic (Al-Zuabi et al., 2019), etc. The key factors that affect the CRM is necessary to identify in any field (Singh and Gupta, 2020). Information about the customer remains a key strategic point in different application. This brings the need to consider powerful tools available from Big Data technologies, which have already been successfully used in other fields such as Bioinformatics, healthcare, finance or hospitality (George et al., 2014; Talón-Ballestero et al., 2018). IoT and big data are providing voluminous data to be processed (Ahmed et al., 2017). ML concepts with econometric and theory-based methods are required to process the data (Wedel and Kannan, 2016).

ML is a recent data analysis technique which has the capability to learn without explicitly programmed and without the intervention of human (Alpaydin, 2014; Marsland, 2015). A case study on B2B sales was represented to identify new customers by using K-means and PU-learning with a random forests adapter (Norlin and Paulsrud, 2017). Customer classification in different categories are utilised in the decision making process of a firm or organisation. Classification of Customer into 'loyal' or 'not loyal' were discussed in Singh and Agrawal (2018a, 2018b) and utilising the channel allocation problem in wireless network. In fact, in order to survive in the market, telecom operators are devoting more in retaining the valuable customers rather than obtaining new customers (Singh and Agrawal, 2019b). KNN, SVM, Naïve Bayes etc., machine-learning techniques were used for customer classification (Kotsiantis et al., 2007). ML methods are utilised to solve the challenging problem such as a customer churning prediction in the telecommunications industry (Vafeiadis et al., 2015; Ahmad et al., 2019). ML could support in customer assistance to increase the sales of fashion store (Fares et al., 2019), as brand passion became a significant part of consumer behaviour for loyalty, satisfaction and trust (D'lima, 2018). By taking advantage of ML techniques with the CRM database, it is possible to find the valuable customers. The ML technique is used to evaluate dynamic customer segmentation analysis for mobile users (Dullaghan and Rozaki, 2017).

ML technique is a tool to analyse customer or user behaviour in any business seeking to remain significant and gain a competitive age. In this paper, author proposed segmentation and classification of customer who visit the mall for purchasing. Demographic analysis can be used by an organisation to target specific customers and increase their overall profit. To fulfil the aim, author intends following major contribution in this paper:

- customer segmentation on demographic property to analyse customer data
- customer classification through different ML technique in order to perform decision making strategy in CRM.

3 Data set and derived attribute

In our proposed work, we have taken *Mall Customer's* data for analytics and classification purpose from open source (Choudhary, 2018). The data set is available in csv (comma separated value) format. This csv file has columns such as Customer ID (Unique ID assigned to the customer), gender, age, annual income, and spending score.

Spending score is allocated by the Mall based on customer behaviour and spending nature. In this proposed work, we used to perform segmentation on demographic data such as gender and age. The data set contains 200 instances and 4 inputs attributes. The statistics of the data set is given in Table 1.

Table 1 Statistics of data set

	Cust_ID	*Age*	*Income*	*Spend_Score*
Count	200	200	200	200
Mean	100.5	38.85	60.56	50.2
Std.	57.88	13.97	26.26	25.82

Two columns are also derived to categorise the data set. It is important to understand that derived attributes are new variables that are based on the original attributes. The derived attribute is used to grade the customer based on buying behaviour of the customer. The buying behaviour of a customer must be related to the income of the customer. It is significant to discuss that at what percentage of the amount of his income, customer spends in purchasing activities. With this concept, we have taken a new column as attribute updated_score (U_s) as given in (1) and inserted in the data set.

$$U_s = Spend\_Score / Income \tag{1}$$

On the basis of U_s values a new column attributes 'grade' is derived based on following rules:

- if U_s is greater than 3.5 then the customer is 'gold'
- if U_s is less than equal to 3.5 and greater than 1.5 then the customer is graded as 'silver'
- if U_s is less than 1.5 and greater than 0.75 then the customer is graded as 'elite'
- if U_s is less then equal to 0.75 then the customer is graded as 'occasional'.

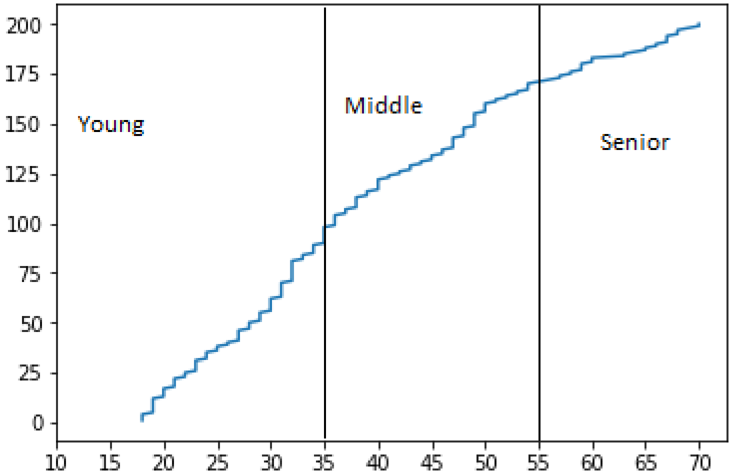
This new derived attribute have a categorical value of the customer used to classify them into different grade. These graded users can be prioritised in the distribution of promotional scheme, discount etc., and more focus on attracting, and retaining of valuable customer.

3.1 Customer behaviour pattern analysis with demographic property

Further, the data set is analysis with respect to customer demographic profile such as age and gender. There are 56% female and 44% male in given data set of a mall. In this paper, gender is used to find the analytics with respect to age, spending score and grade. Age is divided as young, middle and senior as shown in Figure 1 according following rule:

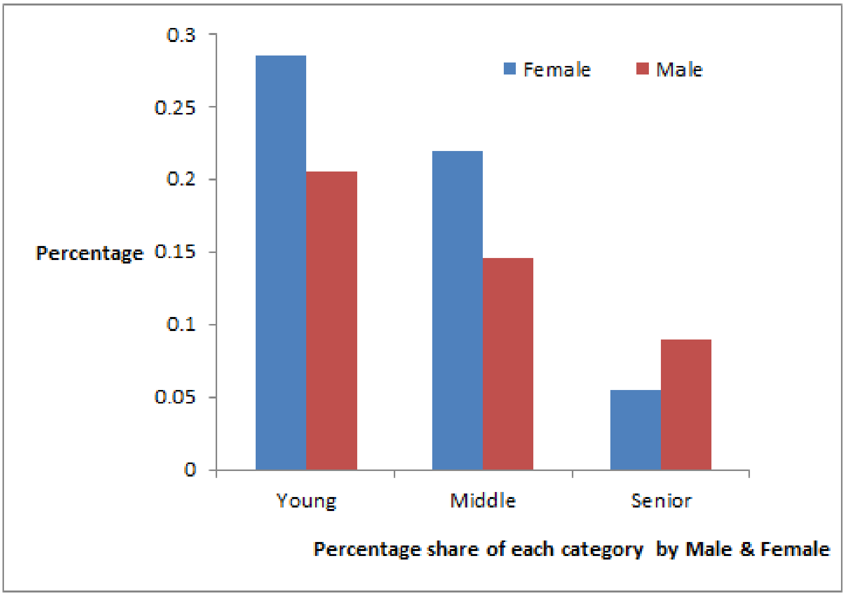
- if age ≤ 35 then age is 'young'
- if age > 35 and < 55 then age is 'middle'
- if age ≥ 55 then age is 'senior'.

Figure 1 Age type of customer (see online version for colours)



'Age' type is applied on given data set and following differences are observed in term of male and female. In given data set young and middle age female share is higher than young and middle age male. However, senior female is less than the senior male as shown in Figure 2.

Figure 2 Percentage share of male and female with different 'age' type (see online version for colours)



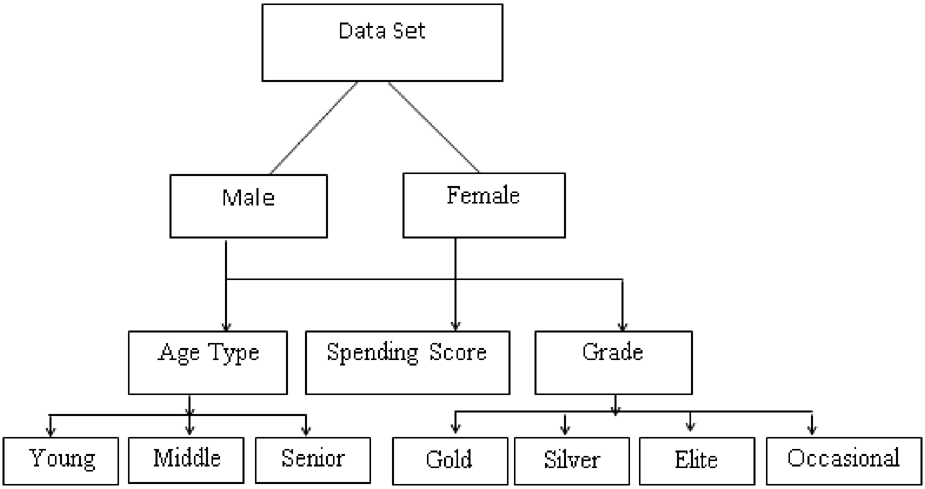
4 Customer segmentation and related analytics

Customer segments enable you to understand the pattern by which differentiation could occur. The pattern may be based on demographic property. Analysing customer segments enable you to do the followings (Sabbeh, 2018):

- identify the most and least profitable customer
- identify the certain customers are more likely to purchase other products based on past purchasing activities
- kinds of incentives apply to each segment to build customer loyalty
- one can develop more effective pricing strategy for the product selected by the segment.

In this paper customer is segmented first on the basis of gender and after that, each segment is further segmented on the basis of 'age' type and derived attributes 'grade' or category. Customer segmentation is represented in Figure 3. Customer analytics are taken age, spend score and grade. The data set is analysed with spend score and type of age that are 'young', 'middle' and senior male and female.

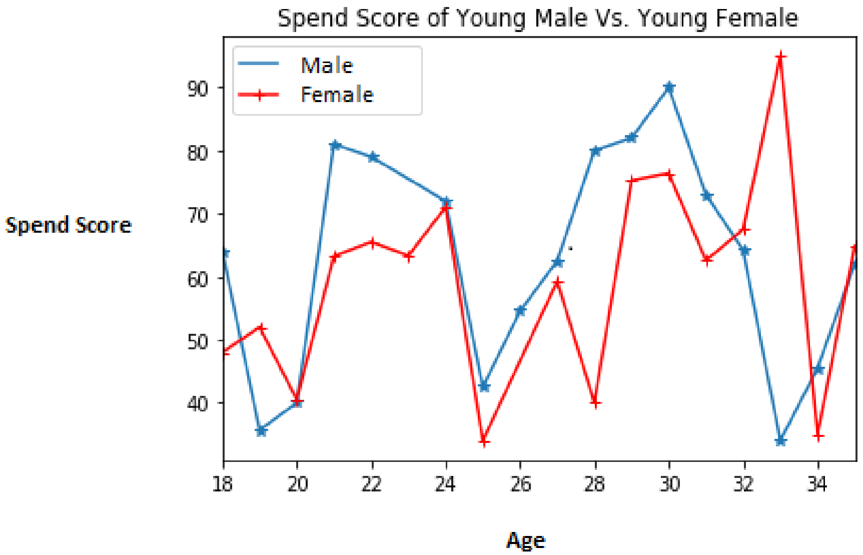
Figure 3 Customer segmentation base



4.1 Spend score of young age male and female

Figure 4, interpreted that young male has more spending score than female. Mean spending score of young male is 62.462745 and mean spending score of young female is 59.611064. However, maximum spends score of young female is 95 of mean age 33, while maximum male spend score in this category is 90 of mean age 30.

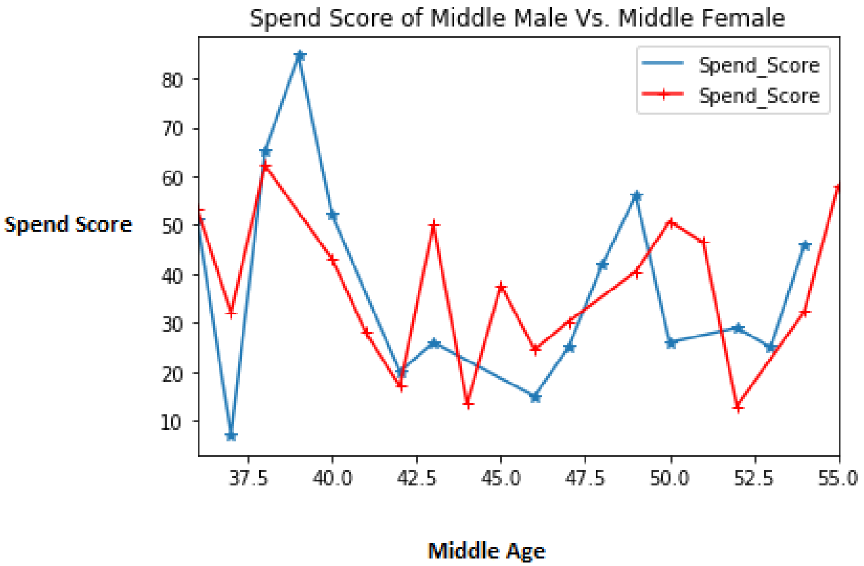
Figure 4 Spend score of young male and young female (see online version for colours)



4.2 Spend score of middle age male and female

Next, interpretation is based on the statistics of spending score of senior male and senior female. The result is represented in Figure 5. The mean spends score of senior male is 37.964444 and senior female 37.205882.

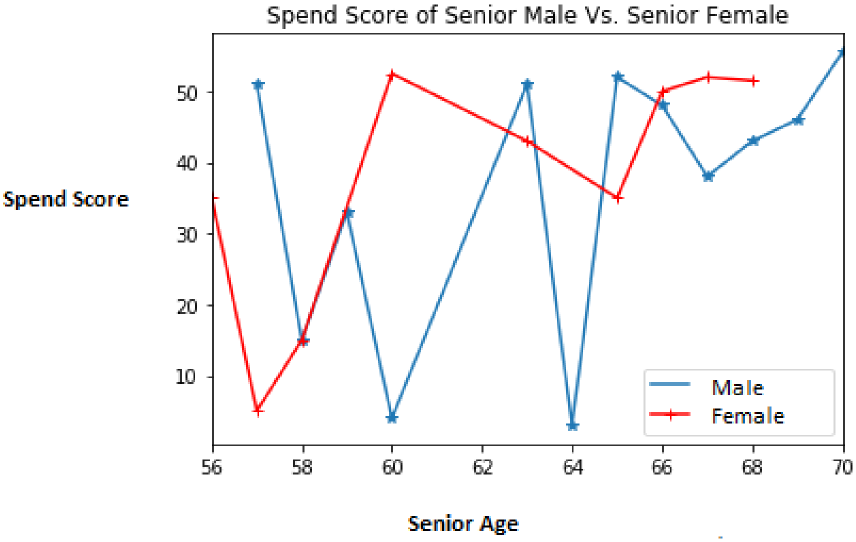
Figure 5 Spend score of middle male and middle female (see online version for colours)



4.3 Spend score of senior male and female

Next, interpretation is based on the statistics of spending score of senior male and senior female. The result is represented in Figure 6. The mean spends score of senior male is 36.625000 and senior female is 37.666667.

Figure 6 Spend score of senior male and senior female (see online version for colours)



According to Figure 6, following interpretation is made to the spends score of senior age male and female.

- the mean spending scores of 'senior female' are greater than 'senior male'
- maximum and minimum spend scores of 'senior female' age are 52.5, 5 respectively
- maximum and minimum spend scores of 'senior male' age are 55 and 3 respectively.

4.4 Interpretation of 'grade' with attributes 'age' and 'gender'

The data set is analysed with 'grade' and type of 'age' that is 'young', 'middle' and 'senior' male and female. The interpretation is represented in Figure 7 and associated data is shown in Table 2.

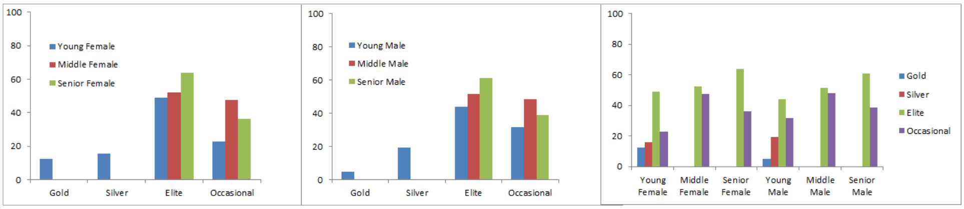
Table 2 Customer analytics with 'grade'

Female	*Young*	*%*	*Middle*	*%*	*Senior*	*%*
Gold	7	12.28	0	0	0	0
Silver	9	15.79	0	0	0	0
Elite	28	49.12	23	52.27	7	63.64
Occasional	13	22.81	21	47.73	4	36.36

Table 2 Customer analytics with 'grade' (continued)

Male	*Young*	*%*	*Middle*	*%*	*Senior*	*%*
Gold	2	4.88	0	0	0	0
Silver	8	19.51	0	0	0	0
Elite	18	43.90	15	51.72	11	61.11
Occasional	13	31.71	14	48.28	7	38.89

Figure 7 Customer Analytics with grade (see online version for colours)



Following important implication have been drawn on the basis of Table 2:

- only young male and young female belong to 'gold' category
- only young male and young female belong to 'silver' category
- all three 'age' type, i.e., young, middle and senior: male and female have 'elite' and 'occasional' grade.

On the basis of above interpretation, It has been noted that in 'elite' category, all aged group of female, have more contribution in purchasing frequency in the mall, compare to 'elite' category of male. However, in 'occasional' category, all aged group females are less contributing in purchasing frequency in the mall, compared to 'occasional' category of male. As per marketing strategy middle and senior males and females are a prospect, customer that could be converted from 'occasional' to 'elite' and 'elite' to 'silver' and so on by applying effective pricing and marketing strategies. 'Young' age male and female is the most prospective customer in the shopping in the mall. Their percentage share in data set could help in maintain number of products.

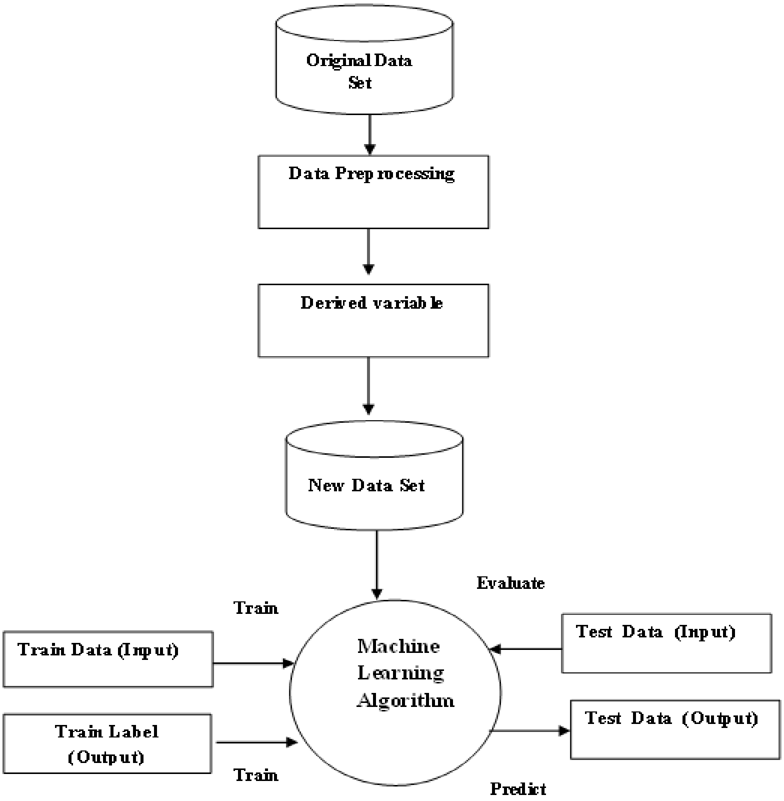
5 Classification techniques

ML techniques are useful in classification of customer data to sustain in a competitive environment by improving CRM and increasing customer retention rate (Tsiptsis and Chorianopoulos, 2011). For the classification, ML classifiers are measured. Classification methods are based on supervised learning techniques. In supervised learning target class label is already known. In this data set we have derived a new attribute 'grade' on the basis of income and spend score. This attribute 'grade' is used to classify the customers of data set in different classes as 'gold', 'silver', 'elite' and 'occasional'. This model can use to retain their valuable customer by providing them better services and pricing strategies. This model can be utilised to offer some promotional scheme to their valuable customer. Business executives can predict the customer churning and may find the measure to reduce it.

For the classification we have used WEKA tool (Sharma and Jain, 2013). Weka is an open source data mining tool. It comprises various ML techniques for classification and clustering. It offers the ability to classify data set through several classifier algorithms. Four candidate classifiers have taken into consideration such as Naïve Bayes, multi-layer perceptron (MLP), regression and J48 (Farid et al., 2014). These classifiers are based on following basic steps and represented by Figure 8.

- collection of customer behavioural data and preparation of the data set
- data preprocessing and the formation of derived variables
- selection of ML model (algorithm)
- training of model
- testing of model for evaluation and prediction
- comparison of models for selection of suitable algorithm for classification.

Figure 8 Basic steps in supervised ML algorithm



5.1 Naive Bayes

Naive Bayes classifier is a probabilistic ML technique that is used for classification job. It is based on Bayes theorem as represented in (2) with the independence assumptions between the predictors.

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{2}$$

5.2 Multi-layer perceptron

MLP is a classification technique based on a feed forward artificial neural network (ANN). MLP classifier used back propagation to learn a multilayer perceptron to classify the instances. A multilayer (feed forward) network contains of at least three layers. First layer is the input layer, then one or more hidden layer and last is output layer. Neurons in input layer use linear activation, neurons in hidden and output layer uses a nonlinear activation function.

5.3 Classification via regression

Regression approach is also applied for classification. A single regression model is constructed for every single value of the class.

5.4 J48

J48 is decision tree based classification algorithm. The decision trees generated by C4.5 can be used for classification. A decision tree is similar to tree structure having diverse nodes. Top most node called as root node, intermediate nodes and leaf node. Node in decision tree consists a decision, to splitting criterion. Splitting criterion in the decision tree is used to find which attribute is the best to split. C4.5 is based on information gain for splitting decision.

6 Performance measures of classification techniques

The proposed classification technique is applied to classify 'grade' of the customers based on the income and spend score. Each classification algorithm is applied and the result is compared to the accuracy. The higher accuracy classifier model is selected for classification and prediction the grade of the customer in the future data sample. For the experimental set up, 70 % split is taken of the data set (Mall_Data.csv). It means out of 200 records, 140 records are used for training of the model and 60 records are used for the testing of the model. Cross validation folds is set 10. Four experiments were accompanied with a given data set: the first one is used analysis of the Naïve Bayes algorithm; the second one is to analysis the performance of the MLP classifier, the third algorithm is used to measure the performance of the Regression and fourth algorithm is used to measure the performance of J48. Kappa statistics measure observed accuracy with an expected accuracy (random chance). If kappa's > 0.75 then it's considered as excellent when interpreting the kappa statistic. A confusion matrix is used to evaluate the performance measures of a classification algorithm. The result obtained from the learned classifiers is analysed through Precision, Recall, F-Measure and ROC area as the performance analysis measure.

6.1 Experiments result of Naïve Bayes classifier

During the testing of the model by using the Naïve Bayes classifier, 55 instances were correctly classified and model takes negligible time to execute. The summary of experimental result is shown in Table 4a, detailed accuracy is represented in Table 4b and result of confusion matrix is represented in Table 4c.

Table 4a Summary of experiment result of Naïve Bayes classifier

Correctly classified instances	55	91.6667%
Incorrectly classified instances	5	8.3333%
Kappa statistic		0.8634
Mean absolute error		0.0499
Root mean squared error		0.1822
Relative absolute error		16.3174%
Root relative squared error		46.1749%
Total number of instances		60

Table 4b Detailed accuracy by Naïve Bayes classifier

TP rate	*FP rate*	*Precision*	*Recall*	*F-measure*	*MCC*	*ROC area*	*Class*
1.000	0.056	0.667	1.000	0.800	0.793	1.000	Silver
0.250	0.000	1.000	0.250	0.400	0.487	1.000	Gold
0.895	0.000	1.000	0.895	0.944	0.924	1.000	Occasional
1.000	0.069	0.939	1.000	0.969	0.935	1.000	Elite
Weighted avg.							
0.917	0.041	0.935	0.917	0.906	0.888	1.000	

Table 4c Result of confusion matrix by Naïve Bayes classifier

A	B	C	D	←	Classified as
6	0	0	0	|a=	Silver
3	1	0	0	|b=	Gold
0	0	17	2	|c=	Occasional
0	0	0	31	|d=	Elite

6.2 Experiments result of MLP classifier

During the testing of the model by using MLP classifier, 59 instances were correctly classified and model takes 0.26s time to execute. The summary of experimental result is shown in Table 5a, detailed accuracy is represented in Table 5b and result of confusion matrix is represented in Table 5c.

Table 5a Summary of experiment result of MLP

Correctly classified instances	59	98.3333%
Incorrectly classified instances	1	1.6667%
Kappa statistic		0.973
Mean absolute error		0.0193
Root mean squared error		0.0829
Relative absolute error		6.3228%
Root relative squared error		20.9963%
Total number of instances		60

Table 5b Detailed accuracy by MLP

TP rate	*FP rate*	*Precision*	*Recall*	*F-measure*	*MCC*	*ROC area*	*Class*
1.000	0.019	0.857	1.000	0.923	0.917	1.000	Silver
0.750	0.000	1.000	0.750	0.857	0.858	1.000	Gold
1.000	0.000	1.000	1.000	1.000	1.000	1.000	Occasional
1.000	0.000	1.000	1.000	1.000	1.000	1.000	Elite
Weighted avg.							
0.983	0.002	0.986	0.983	0.983	0.982	1.000	

Table 5c Result of confusion matrix by MLP

A	B	C	D	←	Classified as
6	0	0	0	|a=	Silver
1	3	0	0	|b=	Gold
0	0	19	0	|c=	Occasional
0	0	0	31	|d=	Elite

6.3 Experiments result of classification via regression

During the testing of the model by using regression classifier, 54 instances were correctly classified and model takes 0.13s time to execute. The summary of experimental result is shown in Table 6a, detailed accuracy is represented in Table 6b and result of confusion matrix is represented in Table 6c.

Table 6a Summary of experiment result of regression

Correctly classified instances	54	90%
Incorrectly classified instances	6	10%
Kappa statistic		0.8373
Mean absolute error		0.1054
Root mean squared error		0.2047
Relative absolute error		34.4445%
Root relative squared error		51.8735%
Total number of instances		60

Table 6b Detailed accuracy by regression

TP rate	*FP rate*	*Precision*	*Recall*	*F-measure*	*MCC*	*ROC area*	*Class*
0.667	0.056	0.571	0.667	0.615	0.571	0.951	Silver
0.250	0.036	0.333	0.250	0.286	0.245	0.906	Gold
0.947	0.000	1.000	0.947	0.973	0.962	0.961	Occasional
1.000	0.034	0.969	1.000	0.984	0.967	1.000	Elite
Weighted avg.							
0.900	0.026	0.897	0.900	0.897	0.878	0.977	

Table 6c Result of confusion matrix by regression

A	B	C	D	←	Classified as
4	2	0	0	|a=	Silver
3	1	0	0	|b=	Gold
0	0	18	1	|c=	Occasional
0	0	0	31	|d=	Elite

6.4 Experiments result of J48

During the testing of the model by using J48 classifier, 56 instances were correctly classified and model takes 0.02s time to execute. The summary of experimental result is shown in Table 7a, detailed accuracy is represented in Table 7b and result of confusion matrix is represented in Table 7c.

Table 7a Summary of experiment result of J48

Correctly classified instances	56	93.3333%
Incorrectly classified instances	4	6.6667%
Kappa statistic		0.8917
Mean absolute error		0.0333
Root mean squared error		0.1826
Relative absolute error		10.8967%
Root relative squared error		46.2652%
Total number of instances		60

Table 7b Detailed accuracy by J48

TP rate	*FP rate*	*Precision*	*Recall*	*F-measure*	*MCC*	*ROC area*	*Class*
0.667	0.019	0.800	0.667	0.727	0.704	0.824	Silver
0.750	0.036	0.600	0.750	0.667	0.645	0.857	Gold
0.947	0.000	1.000	0.947	0.973	0.962	0.974	Occasional
1.000	0.034	0.969	1.000	0.984	0.967	0.983	Elite
Weighted avg.							
0.933	0.022	0.937	0.933	0.934	0.918	0.956	

Table 7c Result of confusion matrix by J48

A	B	C	D	←	Classified as
4	2	0	0	|a=	Silver
1	3	0	0	|b=	Gold
0	0	18	1	|c=	Occasional
0	0	0	31	|d=	Elite

7 Result and discussion

The accuracy of a classification model is compared and represented in Figure 9. Data associated with Figure 9 are shown in Table 8. From Figure 9, it is observed that the highest accuracy is 98.33% of MLP classifier and the lowest accuracy is 90% of regression classifier.

Figure 9 Comparison of classification algorithm (see online version for colours)

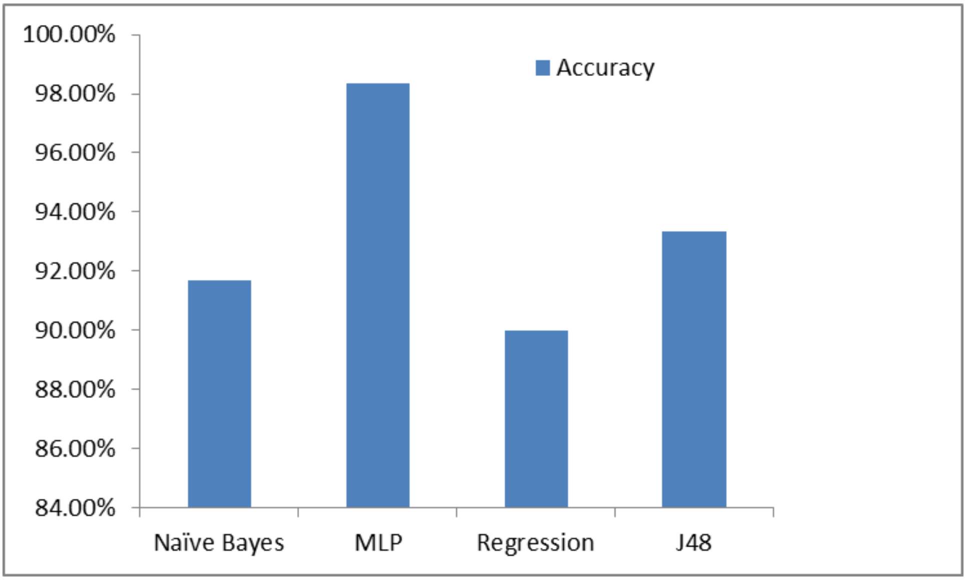


Table 8 Simulation result of each classifier

Algorithm	*No. of record correctly classified in testing (value)*	*No. of record correctly classified in testing (%)*	*Time taken to run (s)*
Naïve Bayes	55	91.67%	0
MLP	59	98.33%	0.26
Regression	54	90%	0.13
J48	56	93.33%	0.02

As indicated in Figure 9, MLP algorithm fares well compared to Naïve Bayes, regression and J48 to classify grade of the customer. MLP classifier correctly classified 59 records out of 60 records in 0.26s. Kappa statistics, mean absolute error (MAE), and root mean squared error (RMSE) of each algorithm has analysed and tabulated in Table 4a, Table 5a, Table 6a and Table 7a which obtained for mentioned classification techniques. MLP classifier has least MAE (0.0193), least RMSE (0.0829) and highest kappa statistics (0.973) when compared to other techniques. Naïve Bayes attained MAE (0.0499), RMSE (0.1822) and kappa statistics (0.8634), regression attained MAE (0.1054), RMSE (0.2047) and kappa statistics (0.8373) whereas J48 attained MAE (0.0333), RMSE (0.1826) and kappa statistics (0.8917). Result indicated that MLP classifier performs well over Naïve Bayes, regression and J48 classifier.

8 Conclusions

The customers are segmented as 'young', 'middle', and 'senior' for male and female. According to data analytics 'young' males spend more than 'young' females. While there is no significant difference between middle male and female in spending score. Further, same data analytics are drawn with respect to derived attribute 'grade'. The analytics suggest that only young male and female lie in 'gold' and 'silver' category. A significant number of middle age male and female is available in 'elite' grade. Middle and senior males and females are prospect, customer that could be converted from 'occasional' to 'elite' and 'elite' to 'Silver' and so on by applying effective pricing and marketing strategies. The ML technique is used to evaluate and investigate four selected classification algorithm based on Weka. The best algorithm is MLP with accuracy 98.33% and total time taken as 0.26s with the lowest error compared to Naïve Bayes, Regression and J48.

The managerial implication of this study offers marketing strategies focusing on specific segments of customers to identify their valuable customer. Furthermore, the study can help to maintain number of products in inventory and offer various purchasing offers to sustain their customer and increase their profitability.

This research work is limited to small data set that contains very specific attributes. Size of data set could be extended in the future for higher accuracy and different perspective of analysis. In future, we would apply clustering algorithms such as K-Mean for segmentation and other ensemble algorithms for classification to improve the accuracy. Expand the customer related input features like product preference, brand, etc., to explore its suitability in the CRM system in more efficient and robust manner.

References

Adebiyi, S.O., Oyatoye, E.O. ad Amole, B.B. (2016) 'Improved customer churn and retention decision management using operations research approach', *EMAJ: Emerging Markets Journal*, Vol. 6, No. 2, pp.12–21.

Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019) 'Customer churn prediction in telecom using machine learning in big data platform', *Journal of Big Data*, Vol. 6, No. 1, p.28.

Ahmed, E., Yaqoob, I., Hashem, I.A.T., Khan, I., Ahmed, A.I.A., Imran, M. and Vasilakos, A.V. (2017) 'The role of big data analytics in Internet of Things', *Computer Networks*, Vol. 129, pp.459-471, DOI: https://doi.org/10.1016/j.comnet.2017.06.013.

Alpaydin, E. (2014) *Introduction to Machine Learning*, MIT Press, USA.

Al-Zuabi, I.M., Jafar, A. and Aljoumaa, K. (2019) 'Predicting customer's gender and age depending on mobile phone data', *Journal of Big Data*, Vol. 6, No. 1, p.18.

Bahari, T.F. and Elayidom, M.S. (2015) 'An efficient CRM-data mining framework for the prediction of customer behavior', *Procedia Computer Science*, Vol. 46, pp.725–731, DOI: 10.1016/j.procs.2015.02.136.

Boukis, A. (2019) 'Exploring the implications of blockchain technology for brand–consumer relationships: a future research agenda', *Journal of Product & Brand Management*, Vol. 29, pp.307–320, DOI: 10.1108/JPBM-03-2018-1780.

Choudhary, V. (2018) 'Mall_Customers.csv', [online] https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python (accessed 06 February 2019).

Cioca, M., Ghete, A.I., Cioca, L.I. and Gifu, D. (2013) 'Machine learning and creative methods used to classify customers in a CRM systems', in *Applied Mechanics and Materials*, Vol. 371, pp.769–773, DOI: https://doi.org/10.4028/www.scientific.net/AMM.371.769.

D'lima, C. (2018) 'Brand passion and its implication on consumer behaviour', *International Journal of Business Forecasting and Marketing Intelligence*, Vol. 4, No. 1, pp.30–42.

Dullaghan, C. and Rozaki, E. (2017) 'Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers', *International Journal of Data Mining & Knowledge Management Process*, Vol. 7, No. 1, pp.14–24.

Fares, N., Lebbar, M. and Sbihi, N. (2019) 'A customer profiling' machine learning approach', for in-store sales in fast fashion', in Ezziyyani, M. (Eds.): *Advanced Intelligent Systems for Sustainable Development (AI2SD'2018). AI2SD 2018. Advances in Intelligent Systems and Computing*, Vol. 915, pp.586–591, Springer.

Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M.A. and Strachan, R. (2014) 'Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks', *Expert Systems with Applications*, Vol. 41, No. 4, pp.1937–1946.

George, G., Haas, M.R. and Pentland, A. (2014) 'Big data and management', *Academy of Management Journal*, Vol. 57, No. 2, pp.321–326.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P. (2007) 'Supervised machine learning: a review of classification techniques', *Emerging Artificial Intelligence Applications in Computer Engineering*, Vol. 160, No. 3, pp.249–268.

Malmi, T., Raulas, M., Gudergan, S. and Sehm, J. (2004) 'An empirical study on customer profitability accounting, customer orientation, and business unit performance', in *The EAA 2004 Conference in Practice, and the Research Seminars*, pp.1–30.

Marsland, S. (2015) *Machine Learning: An Algorithmic Perspective*, Chapman and Hall/CRC Press, Boca, Raton, London, New York.

Mitra, P.K. and Banga, C. (2019) 'Predicting Indian basket crude prices through machine learning models-a comparative approach', *International Journal of Business Forecasting and Marketing Intelligence*, Vol. 5, No. 3, pp.249–266.

Ngai, E.W., Xiu, L. and Chau, D.C. (2009), 'Application of data mining techniques in customer relationship management: a literature review and classification', *Expert Systems with Applications*, Vol. 36, No. 2, pp.2592–2602.

Norlin, P. and Paulsrud, V. (2017) *Identifying New Customers Using Machine Learning: A Case Study on B2B-Sales in the Swedish IT-Consulting Sector*, Stockholm, Sweden.

Rahman, A. and Khan, M.N.A. (2017). 'An assessment of data mining based CRM techniques for enhancing profitability', *International Journal of Education and Management Engineering*, Vol. 7, No. 2, p.30.

Sabbeh, S.F. (2018) 'Machine-learning techniques for customer retention: a comparative study', *International Journal of Advanced Computer Science and Applications*, Vol. 9, No. 2, pp.273–281.

Sgaier, S.K., Eletskaya, M., Engl, E., Mugurungi, O., Tambatamba, B., Ncube, G. and Gumede-Moyo, S. (2017) 'A case study for a psychographic-behavioral segmentation approach for targeted demand generation in voluntary medical male circumcision', *eLife*, Vol. 6, pp.1–21, doi: 10.7554/eLife.25923.

Sharma, T.C. and Jain, M. (2013) 'WEKA approach for comparative study of classification algorithm', *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 4, pp.1925–1931.

Singh, N. and Gupta, M. (2020) 'Key factors affecting customer relationship management in real estate sector: a case study of National Capital Region', *International Journal of Indian Culture and Business Management*, Vol. 20, No. 2, pp.194–209.

Singh, N., Gupta, M. and Dash, S.K. (2018) 'A study on impact of key factors affecting buying behaviour of residential apartments: a case study of Noida and Greater Noida', *International Journal of Indian Culture and Business Management*, Vol. 17, No. 4, pp.403–416.

Singh, P. and Agrawal, R. (2018a) 'A customer centric best connected channel model for heterogeneous and IoT networks', *Journal of Organizational and End User Computing (JOEUC)*, Vol. 30, No. 4, pp.32–50.

Singh, P. and Agrawal, R. (2018b) 'Prospects of open source software for maximizing the user expectations in heterogeneous network', *International Journal of Open Source Software and Processes (IJOSSP)*, Vol. 9, No. 3, pp.1–14.

Singh, P. and Agrawal V. (2019a) 'A collaborative model for customer retention on user service experience, advances in computer communication and computational sciences', *Advances in Intelligent Systems and Computing*, Vol. 924, pp.55–64, Springer, Singapore, DOI: https://doi.org/10.1007/978-981-13-6861-5_5.

Singh, P. and Agrawal, R. (2019b) 'A game-theoretic approach to maximise payoff and customer retention for differentiated services in a heterogeneous network environment', *International Journal of Wireless and Mobile Computing*, Vol. 16, No. 2, pp.146–159.

Sudhagar, D.P. (2019) 'IoT in restaurants: an exploratory understanding of customer perception and preferences of IoT in the Indian context', *International Journal of Business Forecasting and Marketing Intelligence*, Vol. 5, No. 4, pp.401–411.

Talón-Ballestero, P., González-Serrano, L., Soguero-Ruiz, C., Muñoz-Romero, S. and Rojo-Álvarez, J.L. (2018) 'Using big data from customer relationship management information systems to determine the client profile in the hotel sector', *Tourism Management*, Vol. 68, pp.187–197, DOI: https://doi.org/10.1016/j.tourman.2018.03.017.

Tsiptsis, K.K. and Chorianopoulos, A. (2011) *Data Mining Techniques in CRM: Inside Customer Segmentation*, John Wiley & Sons, USA.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.C. (2015) 'A comparison of machine learning techniques for customer churn prediction', *Simulation Modeling Practice and Theory*, Vol. 55, pp.1–9, DOI: https://doi.org/10.1016/j.simpat.2015.03.003.

Wedel, M. and Kannan, P.K. (2016) 'Marketing analytics for data-rich environment', *Journal of Marketing*, Vol. 80, No. 6, pp.97–121.

Yadav, N.S., Gupta, M. and Singh, P. (2018). Factors affecting buying behavior & CRM in real estate sector: a literature survey', *Asian Journal of Research in Business Economics and Management*, Vol. 8, No. 6, pp.32–39.