Machine Learning Approach for Customer Segmentation and Prediction: The Case of Oromia Saving and Credit Shared Company

Kabu Ayele Mersha
Faculty of Engineering and Technology, Rift Valley University, Department of Computer Science
Post Box No: 80734, Addis Ababa, Ethiopia
Email:kabuayele2010@gmail.com

Abstract

Identifying customers which are more likely potential to have product and service offers are an important issue. In customers' identification, the machine learning approach has been used extensively to segment and predict potential customers for a product and service.The aim of this study is to create a model that helps to classify customers for Oromia Credit and Saving Share Company microfinance institution products and services. Since there are not any predefined classes, that describe the purchasers of the institution, the researcher uses clustering techniques that resulted within the appropriate number of clusters. Then, a predictive model was developed to predict the potential of the purchasers. This predictive model achieved an accuracy of 94.1%. For modeling purposes, data was gathered from an establishment head office. Since irrelevant features end in bad model performance, data preprocessing was performed so as to work out the inputs to the model.Thus, various data processing techniques and algorithms were wont to implement each step of the modeling process and alleviate related difficulties. K-means was used as a clustering algorithm to segment customers' records into clusters with similar characters. Different parameters were wont to run the clustering algorithm before reaching a segment that made business sense. The J48 decision tree algorithm was used for prediction purposes. Additionally to those attributes that are believed by the experts to possess a high impact on customer segmentation, attributes value of loan amount features a big influence.

Keywords: Microfinance, machine-learning, predictive analytics, segmentation, data mining
DOI: 10.7176/CEIS/13-2-03
Publication date: March 31st 2022

1. INTRODUCTION

In banking and finance institutions, machine learning algorithms have been applied for various purposes, which include customer segmentation and prediction, profitability, credit scoring, and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management, and forecasting operations, optimizing stock portfolios, and ranking investments mainly for credit risk assessment and customer scaling [1]. Segmentation is the process of developing meaningful customer groups that are similar based on individual explanations characteristics and behaviors [2].

According to [3], segmentation is also viewed as a way to have more targeted communication with the customers; and the process of segmentation describes the characteristics of the customers' groups (called segments or clusters) within the data. Customers vary in terms of behavior, needs, wants, and characteristics and therefore the main goal of clustering techniques is to spot different customer types and segment the customer base into clusters of comparable profiles in order that the method of target marketing is often executed more efficiently. The advantage and drawbacks of the clustering algorithm technique have also been discussed for better clustering efficiency.

2. STATEMENT OF THE PROBLEM

Microfinance like saving and credit shared companies in Ethiopia does not apply modern technology like machine learning techniques, which have registered positive outcomes over the last few years. But the services that they contribute in the area are still low and need higher effort to achieve better results. Some of the weaknesses that characterize the business industry are gap in serving more structured micro-enterprises; low technical capacity; difficulties in accessing funds from donors; lack of product diversification and inadequate management of information system (MIS); lack of research to understand the needs of clients; lack of business development services to clients; weak internal control system, and weak marketing strategy [4] [5]. As the studies of [4], the retention of customers without identifying the best customer segments from bad customers segments creates unhealthy competition. Oromia Saving and Credit Shared Company (OSCSCO) shares the same problems as discussed above in the microfinance industry.

Therefore, this thesis work is will initiates to come up with an applications of machine learning approach that helps to segment and predict profitable customers, so that the institution can make proper decisions in designing strategies in looking for additional clients and opportunities for provision of loan services.

3. RELATED WORK AND LITERATE REVIEW

There are many more researches were conducted on application of machine learning approach on microfinances. As [6] conducted a study to predict credit risk in peer-to-peer lending the data containing default and non-default were extracted from European peer-to-peer from 2009 to 2015. The variables used include credit decisions, new credit customers, age, country, loan duration, loan amount, marital status, country of residence, age, total income, employment status, actual default, credit group, loan purpose, new offer made, and application type. A neural network approach with a back propagation algorithm built-in R programming language was employed. Seventy percent (70%) of data were used for training and thirty (30%) for testing. Overall, the performance was good in classifying defaulters and non-defaulters. As the authors of [7] applied ML methods on the payment histories dataset to estimate the individual creditworthiness. Machine learning methods used were RF, KNN, and bagged KNN. Finally, the study concluded that ML methods can be utilized in credit rating prediction and RF exceeded other ML algorithms in analyzing the creditworthiness of loan applicants. The study of [8] conducted a comparative study for business analytics using random forest trees. The study used German credit data from the University of California Irvine (UCI) repository. The dataset was spliced into two parts that are seventy percent (70%) for training and thirty percent (30%) for testing. Finally, they found that RF algorithms performed well in credit risk prediction. Also, the study concluded that the benefit of using RF algorithms in credit rating analysis lies in simplicity and accuracy. As the author of [9] applied ML algorithms to spot customers' credit risk. The objective of the study was to determine the appropriate variable related to good or bad risk and compare bad customers and good customers. In achieving the defined objective the LR model was applied in analyzing good and bad customers. The info used was German Credit data which incorporates 1000 observations and thirty-one (31) variables. The performance of the model was 74% for Sensitivity and 75% for Specificity. According to the authors [10] used LR with 548 customers' samples to analyze predictors of loans in Ghana's non-traditional banks. They revealed six factors for credit rating which included marital status, dependents, collateral types, duration, kind of loan, and assessment that had high weight for loan default. Likewise,[11] used a binary logistic model within the study of the performance of loan repayment determinants in Ethiopia microfinance. Fourteen variables were used for evaluation and nine of them seemed to be statistically significant and five of them were statistically insignificant.

In Ethiopia, there is no previous work done which help us for the application of machine learning approach for customer segmentation and prediction in case of Oromia Credit and Saving Share Company. Therefore, this research focuses application of machine learning for customer segmentation and prediction: the case of oromia microfinance for the business industry.

4. METHODOLOGY

a. RESEARCH DESIGN

This research is experimental research because the experimental analysis is conducted on microfinance data collected from Oromia Credit and Saving Share Company. In this research, the researcher aimed to use clustering and classification ML algorithms model to achieve the stated objectives. In clustering algorithms, specifically, the researcher uses k-means for segmentation purposes and in the classification, the researcher uses decision trees for predictive purposes.

b. SOURCE OF POPULATION

All the sources of customer's data set of OSCSCO for this study located in Aje and Arsi Negelle from January 2021 to March 2021.

c. DATA UNDERSTANDING

The initial dataset is collected from OSCSC MFI. This data is organized in the format of Microsoft Office Excel 2010 for further processing. The data has 14 attributes and were total of 14114 records. Based on the objective of the research, from a total of 14114 records, 14089 clean data are selected for modeling. The reason is that the data is full of outliers, noise and contains many missing values. From the selected dataset, missing data are removed and some changes of attribute names are made. Some less important attributes are also removed from the selected and a total of 6 attributes (for clustering purposes) and 14089 records have been taken for analysis.

d. DATA COLLECTION METHOD

The data sets are gathered for the study is through, first data was collected by using interviews with domain experts, and therefore the second data was gathered from different written documents, conference articles, and journal publications. A dataset was collected from the dataset of the microfinance using a secondary data collection method, also referred to as the retrospective method.

e. EVALUATION

In this research work from Internal Index measures the researcher used the Sum of Squared Error measure to judge cluster validity. And the algorithm selected for classification purposes was the J48 decision tree.

5. MODELING

The model-building machine learning algorithms mainly consist of the cluster algorithm modeling and classification algorithm modeling subsections and selecting a model that shows better performance.

5.1. SELECTION OF MODELING TECHNIQUES

A good segmentation model can divide customers into homogeneous groups on the basis of the shared common attribute. In this research, automatic cluster detection (K-means clustering algorithm), and J48 decision tree classifier techniques are the selected data mining modeling techniques for customer segmentation and classification respectively.

6. EXPERIMENTATION AND RESULTS ANALYSIS

In this study, all the sample size or available dataset cleaned has been used for training and testing. In the case of decision tree classification models, different experiments have been done by splitting the dataset into training and testing sets and by adjusting the default parameter values. Finally, the classification model that shows better accuracy performance has been selected.

6.1. CLUSTER MODEL BUILDING

The cluster model building process consists of three activities namely: attribute selection, clustering, and selection of best clustering model. This is followed by building different decision tree classifier models and selecting the one that shows better accuracy are part of the process.

6.2. ASSESSMENT OF CLUSTERING CUSTOMER MODELS

There is no actual defined good clustering output. Hence, assessing the clusters based on certain crucial attributes is reasonable. However, it doesn't mean that other attributes have no importance; rather it expresses the weight given to these attributes by the domain expert in the institutions to cluster customers. To observe different changes in the distribution of the segments, the researcher compares the seven (where K= 3, 4and 5 with randomly selected seed size 10,100, 1000) clustering models. The comparison of the clustering model was done in a way that the attribute value of each cluster in the model is compared to other clustering models with the number of iterations, inter-class similarity error, the objective of institutions, and the domain expert's judgment. Experiment One The first experiment was conducted using simple K-means to build a cluster model. 14089 instances and 6 attributes were used in the experiment. Here to have natural segmentation (correctly clustered instances) of customers, classes to clusters evaluation mode was selected from four cluster modes of WEK's tool. This was important to know the measure of incorrectly clustered instances. In turn, this helps to choose the better model from the others. Besides this, a number of iterations, inter-class similarity error, and the objective of the institution were examined through the experiment. The output of experiment one is depicted in the following Table 1. Table 1 Cluster description based on values of attributes for K=3 and seed size 10.

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main balance
1	4966 (35%)	ADDE	AJE	11505.96	10569.09	8316.75
2	8929 (63%)	OBBO	AJE	11693.19	11377.48	9206.89
3	194 (1%)	MSE	ARSI NEGELE	323574.36	68611.07	256831.19

As indicated in the above table 1, cluster one contains attribute values with the frequency of 4966 (35%) records, the names title of the customers ADDE living in AJE, 11505.96 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 10569.09 mounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 8316.75 amounts of birr and their status is active. In the second cluster, with the frequency of 8929(63%) records, the names title of the customers OBBO living in AJE, 11693.1942 amount of money paid out for clients in the form of loan, in 11377.48 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 9206.89 amounts of birr and their status are active. In the third cluster, with the frequency of 194 (1%)records, the names title of the customers MSE living in ARSI NEGELE, 323574.36 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 68611.07 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 256831.19 amounts of birr, the time is taken to build the model is 0.34 seconds and their status are active. In the experiment Sum of squared errors is 5789.37. This measurement was used to compare the goodness of this experiment with others. Besides this, based on the average loan size shown in Table 1, cluster one is categorized into less high preferred customer segmentation, cluster two is categorized into high preferred segmentation and cluster three is categorized into very high preferred segmentation after discussing with domain experiment. According to the expectation of the researcher, the segmentation should contain proportional clusters segmentation like poorly, less, moderately, high, and very high preferred customers category to have a good clustering model. But the result didn't identify as expected. This is because of the default K value and seed number used in the experiment as the starting point. Since the main goal of this clustering experiment is to come up with a good clustering model, the researcher is forced to continue the experiment by increasing the seed number.

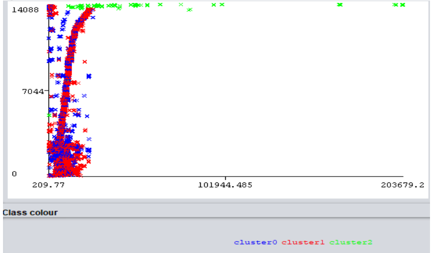

Figure1Visualize cluster assignments

X: axis is install amount and y: axis is instance number

Experiment Two

In the second experiment, the value of K was set to be 2 and the size of seed to 100 but everything the same as the previous experiment. The output of the experiment is depicted in the following table 2. Table 2 Cluster description based on values of attributes for K=3 and seed size 100.

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main balance
1	7139 (51%)	OBBO	ARSI NEGELE	14108.71	13257.3501	11914.32
2	6872 (49%)	ADDE	AJE	11142.87	9118.6014	7299.05
3	78 (1%)	MSE	AJE	602879.83	129218.27	488707.86

As indicated in the above Table 2, cluster one contains attribute values with the frequency of 7139 (51%) records, the names title of the customers OBBO's living in ARSI NEGELE, 14108.7166 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 13257.3501 amounts of birr to the OSCSCO MF, the net balance left on the customer when they start repays the payments are 11914.3232 amounts of birr and their status are active. In the second cluster, with the frequency of 6872 (49%) records, the names title of the customers ADDE living in AJE, 11142.8733 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 9118.6014 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 7299.0514 amounts of birr, and their status is active. In the third cluster, with the frequency of 78 (1%) records, the names title of the customers MSE living in AJE, 602879.8322 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 129218.27 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 488707.8621 amounts of birr, the time is taken to build the model is 0.22 seconds and their status are active. In the experiment Sum of squared errors is 7322.2837. This measurement was used to compare the goodness of this experiment with the previous experiment. Besides this, based on the average loan size shown in Table 2, cluster one is categorized into high preferred customer segmentation, cluster two is categorized into less preferred customer segmentation and cluster three is categorized into very high preferred customer segmentation after discussing with domain experiment. According to the expectation of the researcher, the segmentation should contain proportional clusters segmentation like as said so far in experiment one. But the result didn't identify as expected. Thus, the researcher is forced to continue the experiment by increasing the seed number.

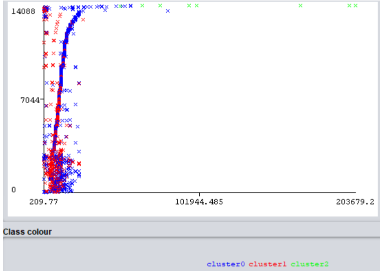

Figure2 Visualize cluster assignments

X: axis is install amount and y: axis is instance number

Experiment Three

The third experiment was conducted by setting the value of K to 2 and seed size to 1000 and it produces the following output as depicted in table 3.

Table 3 Cluster description based on values of attributes for K=3 and seed size 1000

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main Balance
1	8310 (59%)	OBBO	AJE	12117.30	11322.50	8623.83
2	90 (1%)	MSE	AJE	539239.32	114600.58	437257.67
3	5689 (40%)	OBBO	ARSI NEGELE	13199.89	11070.87	10954.01

To describe the value of the attribute in table 3, let's start with the first cluster. Cluster one contains attribute values with frequency 8310 (59%) records, the names title of the customers, OBBO living in AJE, 12117.3063 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11322.5048 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 8623.8382 amounts of birr, and their status is active. Cluster two contain attribute values with frequency 90 (1%) records, the names title of the customers MSE living in AJE, 539239.3212 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 114600.5873 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 437257.672 amounts of birr, and their status is active. Cluster three contain attribute values with frequency 5689 (40%) records, the names title of the customers OBBO living in ARSI NEGELE, 13199.899 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11070.8751 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 10954.0099 amounts of birr, the time is taken to build the model is 0.61 seconds and their status are active. In the third experiment Sum of squared errors is 5145.109. These measurements clearly showed that the sum of squared error registered less result than the previous two experiments. Interpretation with domain experts based on the average loan size, as shown in Table 4.4 segment cluster one, into less preferred, cluster two into very highly preferred, and cluster three into high preferred. As shown in the above table and its description, Clusters segmentation did not contain uniquely identified proportional customer segmentation. Therefore, to come up with a good clustering model, the researcher is forced to continue the experiment by increasing the value of K and seed number.

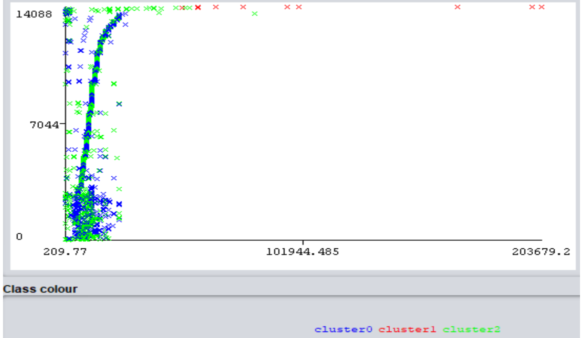

Figure3 Visualize cluster assignments

X: axis is installed amount and y: axis is instance number

Experiment Four

The fourth experiment was conducted by setting the value of K to 3 and seed size to 10 and it produces the following output illustrated in Table 4.

Table 4 Cluster description based on values of attributes for K=4 and seed size 10

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main Balance
1	4907 (35%)	ADDE	AJE	11186.82	10688.58	8334.07
2	5881 (42%)	OBBO	AJE	10735.48	8618.71	7520.23
3	194 (1%)	MSE	ARSI NEGELE	323574.36	68611.07	256831.19
4	3107(22%)	OBBO	AJE	14006.44	16395.28	12355.20

According to table 4 under category of cluster one attribute values with frequency 4907(35%) records, the names title of the customers, ADDE living in AJE, 11186.8231 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 10688.5883 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 8334.073 amounts of birr, and their status is active. Cluster two contain attribute values with frequency 5881 (42%) records, the names title of the customers MSE living in AJE, 10735.484 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 8618.7157 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 7520.2302 amounts of birr, and their status is active. Cluster three contain attribute values with frequency 194 (1%) records, the names title of the customers MSE living in ARSI NEGELE, 323574.3655 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 68611.0785 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 256831.196 amounts of birr, and their status is active. Cluster four contain attribute values with frequency 3107 (22%) records, the names title of the customers OBBO living in AJE, 14006.4469 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 16395.2871 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 12355.2064 amounts of birr, the time is taken to build the model is 0.26 seconds and their status are active.

In this experiment Sum of squared errors is 5786.329. This measurement clearly showed that the sum of squared error registered not better result from all previously conducted experiments. Here also as interpreted with domain experts based on the average loan size as shown in table 4, clusters one and two categorized into less high preferred customers segmentation, three categorized into very high preferred segmentation, cluster four categorized into high preferred customers segmentation. As shown in the above table and its description, Clusters segmentation did not contain uniquely identified proportional customer segmentation. Therefore, to come up with a good clustering model, the researcher is forced to continue the experiment by increasing the value of K and seed number.

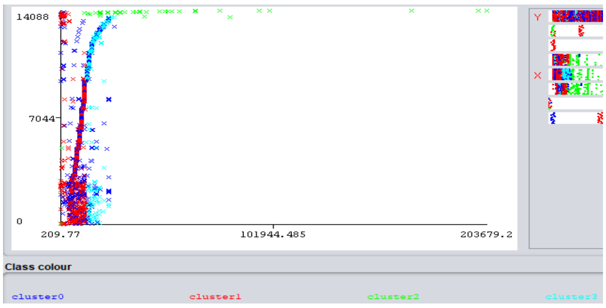

Figure4 Visualize cluster assignments

X: axis is installed amount and y: axis is instance number

Experiment Five

The second experiment was conducted by setting the value of K to 3 and seed size to 100 and it produces the following output illustrated in Table 5.

Table 5 Cluster description based on values of attributes for K=4 and seed size 100

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main Balance
1	4553 (32%)	OBBO	ARSI NEGELE	14475.70	11730.32	12131.14
2	4239 (30%)	ADDE	AJE	11154.55	9950.84	7546.08
3	78 (1%)	MSE	AJE	02879.83	602879.83	602879.83
4	5219 (37%)	OBBO	AJE	12282.78	11825.52	9196.10

According to table 5 under category of cluster one attribute values with frequency 4553 (32%) records, the names title of the customers, OBBO living in ARSI NEGELE, 14475.70 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11730.32 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 12131.14 amounts of birr, and their status is active. Cluster two contain attribute values with frequency 4239 (30%) records, the names title of the customers ADDE living in AJE, 11154.55 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 9950.84 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 7546.08 amounts of birr, and their status is active. Cluster three contain attribute values with frequency 78 (1%) records, the names title of the customers MSE living in AJE, 02879.83 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 602879.83 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 602879.83 amounts of birr, and their status is active. Cluster four contain attribute values with frequency 5219 (37%) records, the names title of the customers OBBO living in AJE, 12282.78 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11825.52 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 9196.10 amounts of birr, the time is taken to build the model is 0.15 seconds and their status are active.

In this experiment Sum of squared errors is 2109.538. This measurement clearly showed that the sum of squared error registered better results than the previous experiments. And also as interpreted with domain experts based on the average loan size in table 5, cluster one is categorized into high preferred customer segmentation, cluster two is categorized into moderate preferred, cluster three is categorized into very high preferred customer's segmentation, and cluster four categorized into less preferred customer's segmentation. This experiment gives better output than all previously conducted.

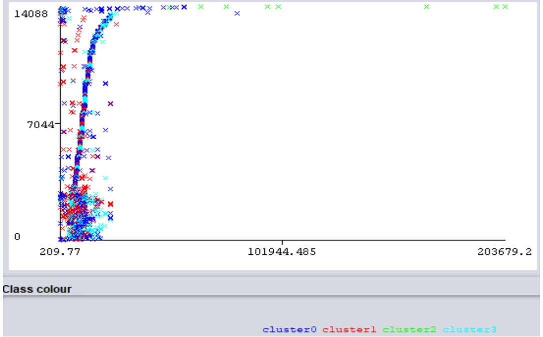

Figure5 Visualize cluster assignments

X: axis is installed amount and y: axis is instance number

Experiment Six

In the six experiments value of K is set to 3 and the size of the seed to 1000 but everything is the same as the previous experiment. The output of the experiment is depicted in the following table 6.

Table 6 Cluster description based on values of attributes for K=4 and seed size 1000

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main Balance
1	4553 (32%)	OBBO	AJE	12282.78	11825.52	9196.10
2	3932 (28%)	ADDE	AJE	11902.66	11183.97	8444.69
3	182 (1%)	MSE	ARSI NEGELE	344137.51	72601.67	273257.65
4	4756 (34%)	OBBO	ARSI NEGELE	10677.53	10193.43	8915.62

As observed from table 6, attribute values with frequency 4553 (32%) records, the names title of the customers, OBBO living in AJE, 12282.78 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11825.52 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 9196.10 amounts of birr, and their status is active. Cluster two contain attribute values with frequency 3932 (28%) records, the names title of the customers ADDE living in AJE, 11902.66 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11183.97 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 8444.69 amounts of birr, and their status is active. Cluster three contain attribute values with frequency 182(1%) records, the names title of the customers MSE living in ARSI NEGELE, 344137.51 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 72601.67 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 273257.65 amounts of birr, and their status is active. Cluster four contain attribute values with frequency 4756 (34%) records, the names title of the customers OBBO living in ARSI NEGELE, 10677.53 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 10193.43 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 8915.62 amounts of birr, the time is taken to build the model is 0.31 seconds and their status are active.

As I observed from the experiment Sum of squared errors is 2075.19. In this measurement, the result of sum squared error decrease than the previous experiment (exp. 5). Here also based on the average loan size in table 6, as interpreted with domain experts, clusters one and three-segmented into high preferred, cluster two-segmented into less preferred and cluster four segmented into very high preferred customer's segmentation.

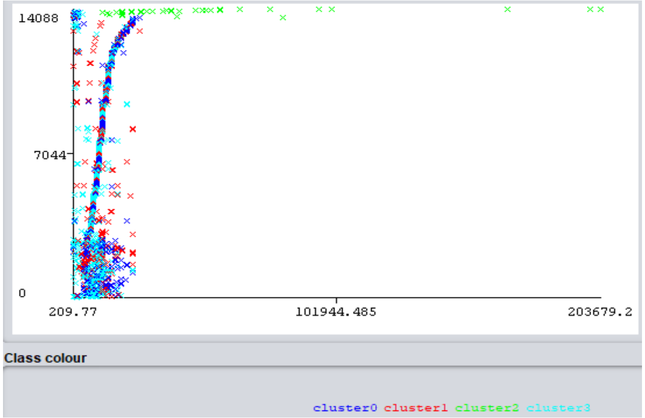

Figure 6 Visualize cluster assignments

X: axis is installed amount and y: axis is instance number

Experiment Seven

The fourth experiment was conducted by setting the value of K to 4 and seed size to 10 and it produces the following output as depicted in table 7.

Table 7 Cluster description based on values of attributes for K=5 and seed size 10

Cluster number	Frequency of records	Name Title	Residence	Disbursed Amount	Install Amount	Main Balance
1	4182 (30%)	ADDE	AJE	10773.90	10083.08	7556.03
2	3230 (23%)	OBBO	AJE	11264.59	8369.70	7072.64
3	90 (1%)	MSE	AJE	39239.32	114600.58	437257.67
4	2040 (14%)	OBBO	AJE	14537.84	17016.87	12399.34
5	4547 (32%)	OBBO	ARSI NEGELE	14227.11	11690.38	11929.35

To describe the value of the attribute in table 7, let's start with the first cluster. Cluster one contains attribute values with frequency 4182 (30%) records, the names title of the customers, ADDE living in AJE, 10773.90 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 10083.08 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 7556.03 amounts of birr, and their status is active. Cluster two contain attribute values with frequency 3230 (23%) records, the names title of the customers OBBO living in AJE, 11902.66 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 8369.70 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 7072.64 amounts of birr, and their status is active. Cluster three contain attribute values with frequency 90 (1%) records, the names title of the customers MSE living in AJE, 39239.32 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 114600.58 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 437257.67 amounts of birr, and their status is active. Cluster four contain attribute values with frequency 2040 (14%) records, the names title of the customers OBBO living in AJE, 14537.84 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 17016.87 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 12399.34 amounts of birr and their status are active. Cluster five contain attribute values with frequency 4547 (32%) records, the names title of the customers OBBO living in ARSI NEGELE,14227.11 amount of money paid out for clients in the form of loan, in this cluster the amount of money that the customer returns in each scheduled intervals are 11690.38 amounts of birr to the OCSSCO MF, the net balance left on the customer when they start repays the payments are 11929.35 amounts of birr, the time is taken to build the model is 0.09 seconds and their status are active. The seventh experiment Sum of squared errors is 2107.53. As clearly observed from the seventh experiment, other experiments conducted after the six experiments with an increment of K values do not yield better results as the previous one. It registered results beyond the minimum value registered in experiment six. As explained previously, the main goal of the clustering experiment is to come up with a good clustering model. After a good clustering model resulted in six experiments, the researcher did not need to continue the experiment.

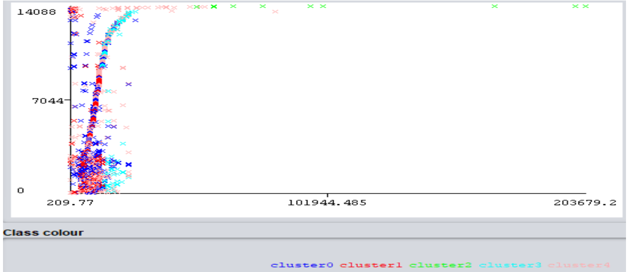

Figure 7 Visualize cluster assignments

X: axis is installed amount and y: axis is instance number

Generally, the overall result of this experiment (the six experiments) looks satisfactory because of the fact that it satisfies the criteria of a good segmentation model used in the research; it is the clarity of the segments to be explained by the domain experiments. The result shows a different group of customers segments and most of the drawbacks indicated in the previous experiments are solved. As clearly indicated, some of the clusters in the previous experiments are suffering from having patterns that are difficult to interpret. In addition to this, the clustering algorithm put customers' segmentations with similar patterns in different clusters.

5.1.CHOOSING THE BEST CLUSTERING MODEL

In finding the best clustering model, seven experiments were conducted. This is to come up with the appropriate clustering model. Seven of them are presented and discussed and a summary of all the experiments is illustrated in bellow table 8. Finally, based on Sum of squared errors, the best clustering model that satisfies the criteria was selected. To validate the obtained output of the experiments, Sum of squared errors is taken as measurement criteria. To measure the goodness of a clustering structure, Sum of Squared Error (SSE) is one of the measures as explained previously. It measures how objects are closely related in a cluster and how distinct or well-separated a cluster is from other clusters.

Besides the above criteria, the comparison of clustering result validity has been done in relation to the values of the key attributes (loan amount) in each cluster, business objectives of the institutions (maximizing profit), and finally, the domain expert's judgment based on the business objective of the institutions. The main objective of every business institution is to maximize its profit. In OCSSCO MFI, this is done by increasing the amount of money lent to customers and by retaining customers for a long period of time. That is why the values of loan amount play a significant role in validating the obtained clustering results. Accordingly, the customer segmentation which has a high loan size has a high probability to be the preferred customer of the institution. Whereas, customer segmentation which has a low loan size, will have less probability to be the preferred customer of the institution based on the objective of the institutions. In the attempt to improve the distribution of instances in different segmentations, different seed values (10, 100, and 1000) with different values of K (3, 4, and 5) have been tried and after a number of experiments conducted, the best cluster had been obtained at experiment 6. The seed value at 1000 and the value of K=4 gives the best distribution of instances in the segments. This is because of the minimum measurement values registered through the experiment compared to others. Comparison of clustering models with different values is shown in the following table 8.

Table 8 Comparison of clustering models.

Experiment No:	K-values	Seed size	Number of iteration	Sum of squared errors
1	3	10	4	5789.38
2	3	100	5	7322.28
3	3	1000	9	5145.11
4	4	10	5	5786.33
5	4	100	4	2109.53
6	4	1000	11	2075.19

5.1. Classification Modeling

The algorithm selected for classification purposes was the J48 decision tree. The researcher tested the algorithm with different parameters and record numbers to improve the classification accuracy. Finally, models are compared and the best model selected. Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. The true purpose of the decision tree is to classify the data into distinct groups or branches that generate the strongest separation in the values of the dependent variable, being superior to predict segments with a desired individual behavior such as response or activation, thus providing an easily interpretable solution [100]. Below, in table 9, the input dataset and the result of decision tree output at different experiments are illustrated.

5.2. Table 9 Output of decision tree with different test modes.

Experiments	No. of records	No.of attributes	No. of leaves	Size of trees	Test modes	Time taken to build model (Sec)	Accuracy (%)
1	14089	6	344	627	All training dataset	0.04	95.0742 %
2	14089	6	344	627	66% percentage split	0.34	93.0063 %
3	**14089**	**6**	**344**	**627**	**10-fold cross-validation**	**0.35**	**94.0663 %**

As observed from Table 10, different experiments were conducted at different test modes in the decision tree algorithm. Results showed that at different test modes there is no impact on the number of leaves and size of the tree while there is a difference in the accuracy of the model built. As explained under Section in the decision tree classification models, different experiments have been done by splitting the dataset into training and testing sets and by adjusting the parameter values into seed split number 4 to have better accuracy. In addition, all training datasets and 10 fold cross-validation test modes are used in experiments to compare models built with each other.

From all experiments conducted, using all training dataset test modes resulted in better classification accuracy (95.1%). However, using all datasets for training has its own limitation. There may be a bias of classification. Therefore, the researcher decides to validate the developed model with the test modes that used some datasets for training and some dataset for testing. From the experiments carried out, 10-fold cross-validation test mode, with all record amounts (14089), the number of attributes 6, numbers of leaves 344 and size of trees 627 scored better accuracy (94.1%) and selected as a better decision tree model than others. The confusion matrix of the selected model is shown in table 10 below.

Table 10 confusion matrix for the experiment

Exper-iments	Test modes			Predicted			Accuracy Rate
				AJE	ARSI NEGELLE	Total	
1	All training dataset	Actual	AJE	8280	98	8378	98.83
			ARSI NEGELLE	596	5115	5711	89.56
		Total		8876	5213	14089	95.07
2	10-fold cross-validation			Predicted			
				AJE	ARSI NEGELLE	Total	
		Actual	AJE	8268	110	8378	98.69
			ARSI NEGELLE	715	4996	5711	87.48
		Total		8983	5106	4089	94.14
3	Percentage split 66%			Predicted			
				AJE	ARSI NEGELLE	Total	
		Actual	AJE	2735	60	2795	97.85
			ARSI NEGELLE	243	1752	1995	87.77
		Total		2978	1812	4790	93.67

As shown in table 10, for the all user training set, 10-Fold cross-validation and percentage split of 66% model evaluation techniques using decision tree (J48) classifier for target class AJE, and ARSI NEGELE. The confusion matrix for the all training dataset depicts that out of total numbers records (14089) supplied, 13395 (95.07%) instances were classified correctly, while the remaining 694 (4.92%) instances were classified incorrectly. From the confusion matrix for the percentage split of 66%, had that out of total numbers records (4790) supplied, 4487 (93.67%) instances classified correctly, while the remaining 4487 (6.32%) of instances were classified incorrectly. In 10-fold cross-validation out of total numbers records (14089) supplied, 13264 (94.14%) instances were classified correctly, while the remaining 825 (5.85 %) instances were classified incorrectly.

6. EVALUATIONS

Evaluation of the segmentation output is based on the dataset of the institution. These are different demographic and financial information of customers. The obtained clustering models have been evaluated using the classes to clusters evaluation approach/cluster mode and SSE. In the case of classification, the obtained classification tree model has been evaluated using the 10 fold cross-validation approach. The dataset is divided into 14 subsets, ensuring that each class is represented with approximately equal opportunities subsets. Then each subset was used for testing and the remaining 6 for training purposes. A total of 14089 (94.1%) instances were correctly classified. The analysis which was closely undertaken with domain experts revealed that the 6th segmentation experiment indeed discovered patterns that are really interesting. As clearly indicated, the clustering model brought customers into different clusters according to their individual behaviors. In addition to this, the decision tree model provides a very good description of the segments and it clearly shows a number of rules that have valuable help to assign potential customers to one of the clusters.

Generally, results identified individual customer behavior in each customer segment that can be used to improve the quality of the institution’s products and services to have profitable and potential customers. Hence, appropriate customer relationship management strategies and programs can be designed and implemented based on market segmentation into meaningful groups according to the institution's needs. These in turn lead institutions to achieve their business objectives.

7.DEPLOYMENT OF THE RESULT

Deployment of the result means using the machine learning algorithms results investigated through this thesis work i.e. result of unsupervised learning algorithms and supervised learning algorithms (clustering and classification). The new knowledge or pattern discovered should be organized and presented in a way that the organization can understand and use for effective customer relationship management. For the application of the result, different resources, technology, and business process in the institution should be integrated.

At the very beginning, OCSSCO MFI had been established to maximize its profit by providing loan service. As discussed with the officials and domain experts, in the institution, based on the business objective, customers are identified as good and services delivered based on trust. Customers trust each other grouped together in lack of collateral (as in banking) to get loan provision given priority to start the service. But, as a business institution, to be more profitable and to handle customers according to their individual needs in the market, it is better to look at and predict customers before providing loans.

8. CONCLUSION AND RECOMMENDATIONS

CONCLUSION

To uncover the hidden knowledge within the dataset of the Institutions, preprocessing, of the dataset was performed using the Weka tool. The data were analyzed and interpreted using the WEKA 3.8.4 version software. Clustering and then classification models were built to categorize and predict customers. To cluster instances into similar groups, simple K-means algorithms were employed. Thus, using the Simple K-means algorithm, different experiments were conducted with different K-values and seed sizes. Segmentation at K=4 and seed size 1000 with 6 clusters selected as best customers segment model in the institution. To classify instances into the same groups based on results obtained through clustering, the Decision TreeJ48 algorithm was employed. Using the Decision Trees J48 algorithm also different experiments were conducted. Model built by 10-fold cross-validation test mode which registered high accuracy (94.1%), selected as the best model for prediction purpose.

RECOMMENDATION

The researcher believes that the findings of the study will encourage the institution, to work on the application of machine learning techniques for successful achievement of the institutional goal. Because the finding showed the importance of customers profile, and how new knowledge can be generated from those data, to improve their service in a new and modern method than the traditional one. These in turn help them to identify their market status.

Based on the findings discussed above, the following recommendations are forwarded:

- Through using ML approaches loan will be approved to members who are likely to be worth and abandon the problem of defaulting which ruin many OCSSCO.
- OCSSCO must adopt ML usages so that loans will be processed much quicker compared to using human expertise.
- Based on the result of the study, institutions should consider and develop important customer relationship management strategies that could be applied, to expand their service by predicting their potential customers, to retain their customers for a prolonged time by giving optimal satisfaction, and to gain competitive advantage in the industry
- Customer loan repayment prediction helps to increase the speed and consistency of the loan application

process and allows the automation of the lending process
•We used k-means and J48 for segmentation and prediction purposes, further research using ANN, KNN, SVM, for the OSCSCO microfinance

REFERENCE

[1] K. I. Moin and D. Q. B. Ahmed, "Use of Data Mining in Banking," *Int. J. Eng. Res.*, vol. 2, no. 2, p. 5, 2012.
[2] Trappey, V .Charles, Amy J.C. Trappey, Ai-Che Chang, and Ashley Y.L. Huang, "The analysis of customer service choices and promotion preferences using hierarchical clustering.," *Ournal Chin. Inst. Ind. Eng. 5367-376*, vol. 2009.
[3] Bounsaythip, Catherine and Esa Rinta-Runsala, "Overview of data mining for customer behavior modeling.," *VTT Inf. Technol. 18*, pp. 1–53, 2001.
[4] Meklit, "MicroFinance Institution, Progynist and AliseiNGO (Organizers) (2004).".
[5] Befekadu B. and Kereta, "Outreach and Financial Performance Analysis of Microfinance Institutions in Ethiopia.," *Afr. Econ. Conf. U. N. Conf. Cent. UNCC*, 2007.
[6] Byanjankar, A., Heikkila, M., & Mezei, J., "Predicting credit risk in peer-to-peer lending: A neural network approach.," *Proc. - 2015 IEEE Symp. Ser. Comput. Intell. SSCI 2015*, pp. 719-725., 2015.
[7] Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A., "Consumer credit risk: Individual probability estimates using machine learning.," vol. 40, no. 13, pp. 5148–5159, 2013.
[8] Ghatasheh, N., "Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study," *Int. J. Adv. Sci. Technol.*, pp. 19–30, 2014.
[9] Hargreaves, C. A., "Machine learning application to identify good credit customers Machine learning application to identify good credit customers.," *Nternational J. Adv. Eng. Technol. Int.*, pp. 31–35, 2019.
[10] Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K., "Logistic Regression Analysis Of Predictors Of Loan Defaults By Customers Of Non- Traditional Banks In Ghana," *Eur. Sci. J.*, p. 175, 2016.
[11] Pasha, S. A. M., & Negese, T., "Performance of Loan Repayment Determinants in Ethiopian Micro Finance-An Analysis.," *Eurasian J. Bus. Econ.*, pp. 29–49, 2014.