Developing Loan Advisor Model for Customer Loyalty using Data Mining

Teshome Bulo
HiLCoE, Computer Science Programme, Ethiopia
World Vision Ethiopia (WVE)
teshebb@gmail.com

Temtim Assefa
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
temtimasefa@yahoo.com

Abstract

Customer relationship management is vital in financial institutions for their successful operations. To develop a good customer relationship, it is required to understand customer characteristics by analysing historical customer data. Financial organizations capture massive data which isn't feasible to do by traditional methods. Wasasa is not exceptional to this problem. Even though the main goal of microfinance is to provide financial service to alleviate poverty, it is also necessary to analyse data and generate new knowledge so as to satisfy its customer needs and consequently its future existence.

The purpose of this paper is to use data mining to analyse customer data and predict customer's behaviour. In this paper, an attempt is made to apply different classification algorithms namely Naïve Bayes, KNN, Rule induction (PART), and Decision Tree. As methodology, we used CRISP-DM methodology and WEKA software to implement the selected algorithms. Different experiments were conducted repeatedly by adjusting different parameters until the required output is reached.

There were twelve experiments conducted using four selected algorithms. According to the result of the experiments, KNN, PART, and J48 achieve an accuracy of 99.96%, 99.93% and 99.91% respectively and ROC area of 1.0 that is equal for all algorithms. The rules are generated using the PART algorithm because it achieved slightly better output than the J48 algorithm.

Based on the rules generated using PART rule induction, a prototype system was developed for domain experts. The system will help Wasasa's employees to understand their customers' need and provide better services. As customers become more satisfied with the services they will remain loyal to Wasasa.

Keywords: Microfinance; Loyalty; Data Mining

1. Introduction

The establishment of microfinance has brought many benefits to low income citizens which have direct relationship on the country's development. Even though microfinance is playing a great role in reducing poverty, recently the explosively growing, widely available, and gigantic body of data becomes an issue in all business firms. Analyzing these data easily through known traditional methods has lots of problems. Thus business firms started searching for powerful and versatile tools that automatically extract valuable information from these data and convert such data into organized and understandable knowledge. Microfinances were one of the business firms which were looking for this solution that support them to analyze their customer information during decision making. The popular and fast-growing technology to solve such problems is data mining or knowledge discovery.

Data mining uses a variety of data analysis tools to discover patterns and relationships in the data which may be used to make valid predictions [1].

Data mining is a component of business intelligence that will extract non-trivial knowledge and hidden patterns from large amount of data which enables a business firm to make the right decision.

The derived knowledge must be new, not obvious, relevant and can be applied in the field where this knowledge has been obtained. It is also the process of extracting useful information from raw data. Data mining is applicable in all areas of business including manufacturing, financial, governmental and health institutions to discover new knowledge. Out of these application areas this research is conducted on financial institutions [2, 4].

Data mining presents an opportunity to increase significantly the rate at which the volume of data can be turned into useful information [5]. The application of data mining is essential in any organization to make important decisions that might be crucial for the well-being of the whole business [3]. Data mining has different algorithms from which a suitable algorithm can be selected to generate meaningful patterns. Hence, data mining is recently becoming a popular tool for analysing big data and reveals relevant information for decision making.

Even though different attributes are in use to define loyalty within the domain area, the definition given to loyalty is not consistent and vary from company to company. The same is true for Wasasa and loyalty is defined from different perspectives in Wasasa; such as based on saving, repayment schedule, gender and others. This shows that within the same company there is no proper definition assigned to loyalty.

Accordingly, Wasasa is facing the same challenges in relation to customer loyalty definition. One of the challenges is that the institution is fully dependent on the credit officer ability and judgments to assign label to its customers which makes the company vulnerable because of the subjective definition assigned to loyalty. Moreover, apart from the operational manual there is no standard set to define loyalty within the company during loan appraisal that all branches follow.

2. Related Work

The work in [6] developed a model on a selected MFI that segments and predicts customers based on historical data. The author follows the CRISP-DM process model and employed k-means clustering to segment the dataset into different clusters and J48 classification algorithm was used to develop the final model. One of the limitations is that clustering is based on the distance between each point and cluster which takes a longer time to calculate when the dataset size becomes large. The other problem is that selecting the initial K is also challenging and the author used only decision tree algorithm to develop the final model with dataset of 13,057 records and 6 attributes for the experiment.

Simret Solomon [3] conducted a research on predicting customer loyalty using data mining techniques and built a model that predicts customer loyalty with a dataset of 6,447 records and 10 attributes. The researcher used three classification algorithms: Rule Based ZeroR, Decision Tree (J48), and Bayes (Naïve Bayes) and three test option training set, 10 fold and split option to build a model. Finally, the researcher selected J48 decision tree algorithm with 10 fold cross validation among the other algorithms which produces an accuracy of 97.8% for customer loyalty. The researcher concluded that the MFI can use this model during pre-loan disbursement to identify default customers and minimize risks credit management.

Wagari Dibaba [7] conducted a research in the application of data mining in microfinance using a dataset of 9,550 records and 13 attributes. The objective of the research was building customer classification model for better customer relationship management. The author followed CRISP-DM methodology to conduct the research and used a classification decision tree algorithm to develop the final model. There were six experiments conducted using the same algorithm with varying size and other parameter settings. Finally the algorithm that produced an accuracy of 78.54% was selected as best out of the six experiments conducted.

Sara Worku [8] conducted a research in the Addis Credit and Saving Institution. The aim of the research was to develop a predictive model that

investigates credit risk in financial institutions using data mining techniques. The author followed CRISP-DM methodology and uses WEKA tool to develop the model. The data was extracted from four branches of the institution within the same geographical location and contain 4000 records and 7 attributes. There was a class distribution imbalance in dataset so that the author used SMOTE technique to rebalance the minority class distribution. Three classification algorithms were employed to develop the model, namely decision tree, J48, Naïve Bayes and PART rule induction. Cross validation test method was used to validate the model. Finally, the author selected PART rule induction that produces optimal accuracy of 99.83% among the nine experiments conducted for the rebalanced dataset with seven attributes. The author concluded that the result shows that it is possible to extract useful pattern through data mining algorithm that help to make accurate decision for credit risk assessment.

3. The Proposed Solution

The following steps are followed to achieve the objectives stated in this paper.

- Reviewing previous works within the area to understand the domain and attributes selection method, identify gaps and how to employ CRISP-DM methodology that includes business understanding, data understanding, data preparation, evaluation, and deployment.
- Preparing the data in the format that is appropriate for the algorithm selected.
- Importing the prepared dataset to the selected tools and conduct the experiments with 4 algorithms.
- Generating the rule based on the optimal result and develop the prototype
- Present the developed prototype to domain experts and senior management for evaluation.

The main activities are discussed below:

Business understanding: is a process where the core business flows are understood. To understand the business process various documents were reviewed that include loan processing procedure, repayment methods and operational policy.

Data understanding: initially the data were collected from the central database of Wasasa Microfinance. In addition interview and discussion were conducted with selected domain experts. Since the data were extracted for all branches separately, the first task was consolidating these data into one Excel sheet. The data set contains a total of 31,996 records and 44 attributes.

Data preprocessing: is a task that is applied on the dataset to have the dataset a suitable format appropriate for the data mining techniques. There were different steps involved in data preprocessing such as data reduction, attribute selection, data cleaning, data integration, and data transformation [9].

Data Reduction: to enhance the mining process of large data, the reduction of the data size should be performed without jeopardizing the mining results [10]. Attributes that are less useful for the purpose are removed.

Attribute selection: is performed based on the purpose of the research and also the mining tasks applied. Thus, we selected 12 attributes for the final experiments.

Data cleaning: is a task to clean the data by filling missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Data integration: having large amount of redundant data may slow down or confuse the knowledge discovery process [9].

Data Transformation and Discretization: is the last preprocessing step where data is transformed into a suitable form for data mining tasks that increases the mining process efficiency.

After preprocessing, 26,517 records were made ready for the experimentation where 96.28% are

loyal, 2.99% non-loyal, and 0.73% spurious loyalty. But the dataset contains imbalance class distribution within the consolidated data that is 3% of the two classes (spurious and non-loyal) while loyal class is 97% in the dataset which needs rebalancing. A dataset is said to be imbalanced if one class (called the majority or negative class) vastly outnumbers the other (called the minority or positive class). This is a common problem in classification data mining [11]. To rebalance this class distribution there are various techniques available such as Random under Sampling (RUS), Random over Sampling (ROS), and Synthetic Minority Over-sampling Technique (SMOTE). For the purpose of this paper, Synthetic Minority over-sampling Technique (SMOTE) was selected and finally the rebalanced dataset is used for experimentation.

Once the dataset becomes ready and converted into the format supported by the selected tool, that is WEKA 3.8.1, and importing and checking the data is error free, the experimentation follows. There are different algorithms available to develop a predictive model. Among those four algorithms are selected to achieve the objective of this paper. The model developed is evaluated using different metrics to select the optimum result out of the twelve experiments. The metrics used were the one that is available and provided within WEKA tool, and different interview questions designed and conducted with the domain experts. Based on these finally the optimum result is selected and rules generated with the algorithm used to develop the prototype that is easily understandable and used by the business user developed using vb.net programming language.

4. Discussion

To develop the predictive model using data mining, classification algorithms such as Bayes Naïve net, K-Nearest Neighbor (KNN-IBK), PART Rule Induction and Decision Tree (J48) are used. The final cleaned and rebalanced dataset with a total of 75,523 records was given to WEKA tool version 3.8.1 and twelve experiments were conducted using these four selected algorithms. Those experiments are summarized and presented in Table 1.

Table 1: Summary of Experimental Results

No.	*Algorithm*	*Experiments*	*Attributes*	*Evaluative*	*Accuracy*	*ROC*
1.	Naïve Bayes	Experiment 1	12	training set	97.21	0.999
		Experiment 2	12	10 folds	97.19	0.999
		Experiment 3	12	70/30	96.79	0.998
2.	K-nearest Neighbor	Experiment 1	12	training set	99.98	1.000
		Experiment 2	12	10 folds	99.79	1.000
		Experiment 3	12	70/30	99.67	1.000
3	PART	Experiment 1	12	training set	99.97	1.000
		Experiment 2	12	10 folds	99.96	1.000
		Experiment 3	12	70/30	99.94	1.000
4	J48	Experiment 1	12	training set	99.97	1.000
		Experiment 2	12	10 folds	99.95	1.000
		Experiment 2	12	70/30	99.95	1.000

Among the twelve experiments conducted, the model built with algorithm KNN-IBK with training set test method produced the highest accuracy which is 99.98% and ROC area 1.0. Even though the model achieved the highest result, it doesn't have an option to generate rule so that the next model that produced

the highest results was selected to generate the rule that was PART rule induction having accuracy of 99.97% and ROC area of 1.0. The rule generated using this algorithm (PART) is used to develop the prototype shown in Figure 1.

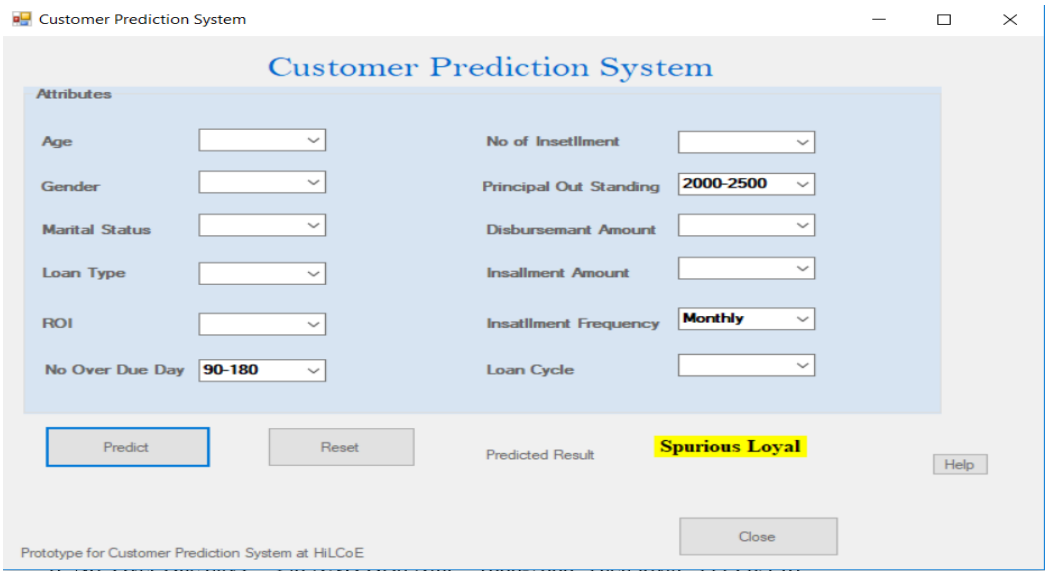

Figure 1: Customer Prediction System Prototype

The model produced better result when compared to a similar previous work [3]. The model developed in this paper is better than that in [3] because: it achieved highest accuracy that approaches to perfection, the model categorized customers into three classes, and there was no model developed previously to allow MFIs to check their customers after disbursement.

5. Conclusion and Future Work

Customer loyalty is vital to any business firm especially in financial institutions like Wasasa. From the dataset prepared for experimentation spurious and non-loyal customers accounted for 3%. Though it looks less when compared with other researches conducted, it has significant impacts on financial institutions business. Analyzing and extracting new knowledge through manual methods from customer records are time consuming and difficult as the size of data keeps increasing from day to day. Not only analysing but also discovering a pattern and relationship will become difficult. Use of the state of the art data analysis technology such as data mining increases organizations' learning capability from their data and solve their problems in a novel way.

From a total of 31,996 records with 44 attributes, 26,517 (82.87%) records with complete data were validated and preprocessed. Among all attributes, 12 relevant attributes were selected to build the model.

The CRISP-DM standard process model and WEKA tool were adopted to achieve the research objective. The classification algorithms Bayes Naïve net, KNN, PART and J48 were employed to develop the model. From experiments conducted on four algorithms, KNN and PART with accuracy of 99.97% and 99.91% respectively were selected to customer loyalty model development. Even though KNN achieved optimum result, it couldn't generate rule. Hence, PART algorithm with better accuracy result next to KNN was used to generate the rules.

Finally, the predictive model that enables the MFI for decision making and loan appraisal is developed.

Future research includes the following:

- Conducting study on developing a prototype integrating to a live database and generating different reports that will support decision making.
- To study further by incorporating additional attributes like income and saving products to widening the study area.

References

[1] Two Crows Corporation, "Introduction to Data Mining and Knowledge Discovery", 3rd edition, 1999.

[2] A. J. Hamid and T. M. Ahmed, "Developing Prediction Model of Loan Risk in Banks using Data Mining", University Khartoum, March 2016.

[3] Simret Solomon, "Predicting Customer Loyalty Using Data Mining Techniques," Unpublished Masters Thesis, HiLCoE School of Computer Science and Technology, Addis Ababa, January 2012.

[4] Anteneh Fentahun, "Mining Road Traffic Accident Data for Predicting Accident Severity to Improve Public Health – Role of Driver and Road Factors: The Case of Addis Ababa", Addis Ababa, 2011.

[5] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining", Daimler Chrysler Research & Technology, 1999.

[6] Belachew Reganie, "Application of Data Mining Techniques for Customer Segmentation and Prediction: The Case of Busaa Gonofa MFI," Unpublished Masters Thesis, Addis Ababa University, 2013.

[7] Wakgari Dibaba, "Application of Data Mining Techniques for Effective Customer Relationship Management of Microfinance: The Case of Wisdom Microfinance," Unpublished Masters Thesis, Addis Ababa University, 2010.

[8] Sara Worku, "Application of Data Mining Technology for Credit Risk Assessment in Addis Credit and Saving Institution", Unpublished Masters Thesis, Addis Ababa University, June 2016.

[9] Women in Microfinance and its Social Effects on the Community: A Case Study in Hue, Vietnam, 2016.

[10] V. K. Tripathi, "Microfinance - Evolution, and Microfinance-Growth, of India," International Journal of Development Research, Vol. 4, No. 5, pp. 1133-1153, May 2014.

[11] T. R. Hoens and N. V. Chawla, "Imbalanced Datasets: From Sampling to Classifiers", John Wiley & Sons, Inc., USA, 2013.