Predicting Customer Loyalty Using Data Mining Techniques

Simret Solomon
University of South Africa (UNISA), Addis Ababa, Ethiopia
simrets2002@yahoo.com

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

This research aimed on prediction of customer loyalty (Non loyal or Loyal) using the application of data mining in microfinance that helps to build a classification model which supports during loan decision making in the organization.

In this study a classification model is built based on the loan data obtained from Joshua Multi Purpose Limited Liability Cooperative (JMPLLC). Experiments using ZeroR, Naïve Bayes, and J48 classifier algorithms of the WEKA 3.6.6 software have been conducted using the pre-processed dataset with selected attributes and parameter settings in order to find the optimal model. The classification model J48 with the best accuracy level of 97.83%) is selected to predict customer loyalty class label (Non Loyal or Loyal) and J48 algorithm was employed to generate rules.

Keywords: Data Mining; Customer Loyalty; Loan; Microfinance; Decision Trees

1. Introduction

Economical development of a nation is directly related with the economical development of an individual. However, third world nations couldn't play a significant role in curbing this problem due to the limited roles of the banking sector. Even though cooperatives and microfinance institutions are there for more than half a century, their contribution to poverty reduction was not that much as it is now.

Data mining techniques perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific research [3].

In most cases, data mining is treated as synonym for knowledge discovery in database (KDD). According to Han and Kamber [3], data mining is viewed as an essential step in the process of knowledge discovery in databases. Data mining is also defined as a process of selection, exploration and modeling of large quantities of data to discover regularities and relations that are at first unknown, with the aim of obtaining clear and useful results for the owner of the database [4].

Bearing this in mind, 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. Examples of these types of business organizations are microfinance and cooperatives.

JMPLLC envisioned being one of the leading, efficient, and dynamic credit and savings societies by offering a competitive financial and social services for the socio-economic empowerment of its members. To achieve its vision, JMPLLC keeps on mobilizing extensively savings and deposits, providing innovative and very competitive financial and social services in a well organized and efficient manner based on ethical values and principles.

2. Related Work

Even though there are not that much research conducted on predicting customer loyalty on microfinance using data mining, many related researches have been conducted using different data mining techniques, algorithms, and tools on various customer profile analysis and customer relation management related to unsafe borrower, predicting loan default and loan risk assessment in financial sectors locally and internationally.

Today, many businesses such as banks, insurance companies, and other service providers realize the importance of Customer Relationship Management (CRM) and its potential to help them acquire new customers, retain existing ones, and maximize their lifetime value. At this point, close relationship with

customers will require a strong coordination between IT and marketing departments to provide a long-term retention of selected customers [5].

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customer, increasing revenue from existing customers, and retaining good customers. By determining characteristics of a good customer (profiling), a company can target prospects with similar characteristics, by profiling customers who have not bought that product (cross-selling). By profiling customer who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition, because it is usually far less expense to retain a customer than acquire a new one [8].

Data mining techniques are used quite well to a variety of critical business decisions in the pharmaceutical industry. It is also used for forecasting production schedules for the manufacturing plants, determining market potential in critical go/no decisions on continuing work on development compounds, or making financial projections for stock holders and investors on Wall Street [10].

Getachew *et al.* [1] has conducted a research on the potential application of data mining techniques for investigating customer loyalty and to predict loan default incidences. The study used mixed research methodology which is qualitative and quantitative. Experiments were conducted with 9551 dataset records using J48 classifier algorithm. The result indicates technical solution to be in place in order to satisfy the information requirement of the company.

In [2], the target was the overall process of exploiting customer data and information, and using it to increase the revenue generated from an existing customer and attract new customers by creating good relationship with them accordingly.

In the research the applicability of clustering and classification data mining techniques to implement CRM in the Ethiopian Electric Power Corporation (EEPCo) have been explored within the approach of CRISP-DM process model. The K-means clustering algorithm was used to segment customer records into clusters with similar behaviors. In the classification sub-phase, J48 decision tree and Naïve Bayes algorithms were employed. Using the final dataset, different clustering models at K values of 4, 5, and 6 with different seed values have been experimented and evaluated against their performances. Consequently, the cluster model at K value of 4 with seed size 1000 has shown a better performance.

Finally, its output is used as an input for decision tree and Naïve Bayes classification models. First the different classification models with decision tree and Naïve Bayes algorithms are experimented with different parameters. Among these, J48 decision tree model that showed a classification accuracy of 99.89% was selected.

Jia *et al*. [6] conducted a research which focuses on applying data mining technology in developing a loan risk assessment system. The paper focused on comparing different data mining methods when applied to loan data.

In order to produce the result, different algorithms like Decision tree induction algorithm, clustering and Naïve Bayes and the data mining tool called WEKA are used. WEKA is a collection of algorithms for solving real-world data mining problems. Using WEKA, different data mining methods are able to be applied to extract data patterns. Finally, the paper introduces a case study of applying different data mining technologies in developing a loan risk assessment system for a sub-prime lender. The challenges and application background are stated. The potential approaches for the problems are examined. Experiments on two datasets are carried out. The results from the different algorithms are analyzed and based on the discussion, decision trees appear to be the most appropriate data mining technology for developing a loan risk assessment system for sub-prime lenders.

Salame [7] conducted a research to provide an additional tool that helps in reducing the proportion of unsafe borrowers which will have a positive effect on the financial institution. Due to the significance of credit risk analysis, this study was done to add additional information to the agricultural loan decision-making process, potentially decrease the cost and time of appraisal of loan applications, and

decrease the level of uncertainty for loan officers by providing knowledge extracted from previous loans.

The results show that the variable representing the number of loans per customer has been significant in the estimated models for two data sets one that excludes these variables that had missing values for the loans refinanced and matured in 2006. The data set that represents the loans refinanced and matured (or defaulted) during the period of 2006 – 2010. It represents customers who have more than one loan (2006 – 2010). This illustrates the presence of contamination on defaulting. The contamination effect means that if a borrower has more than one loan and the borrower defaults on one loan, there is a chance that the borrower will default on the other loan(s).

In conclusion, the research examines the performance of three different methods: logistic regression, decision tree, and neural networks in estimating the probability of default. A comparative examination of these methods was conducted based on the misclassification rate of loan default of the portfolio of a large agricultural financial lending institution. The results show the presence of slight differences between the misclassification rates of the different methods. It was not possible to conclude that one method outperformed the others.

The reviewed researches have shown an encouraging result in integrating data mining with CPA/CRM. Moreover, application of data mining techniques in the finical sector will have a huge impact on effective risk management and strategic decision making.

As a result of exploring and reviewing various data mining tasks and related works, we were able to decide that classification model building will be applied for the research at hand since it is directly related to the data and also it provides methods for predicting the value of a (categorical) attribute (the class) based on the values of other attributes (the predicting attributes). In addition, the researchers after applying different data mining tasks in developing a loan risk assessment system, decision trees appear to be the most appropriate data mining techniques for developing a loan risk assessment system.

3. Methods

Among the various data mining tasks, classification model building will be applied for the mining process, since the main objective of this research is predicting customer loyalty and hence classification is the process of finding a set of models (or functions) that describe and distinguish data classes and concepts for the purpose of being able to use the model to predict the class of objects whose class label is unknown [9].

The real world data can be characterized by missing values, inconsistency, and extreme values (outliers). For these reasons, the data must be preprocessed before it is experimented. Preprocessing is important in order to meet the final objective effectively. The data has no problem with consistency; rather with incompleteness like missing values, outliers, and data must be aggregated.

Data Pre-processing

The activities during this phase include data cleaning, attribute selection, data transformation and aggregation and data formatting.

Data cleaning: The data was cleaned, by removing the records that had inconsistent (outliers) and filling the missing values with most probable value.

Attribute selection: Attributes that are only related to the final objective are selected. Out of 25 attributes, 10 were selected, including target variable.

Data Transformation and Aggregation: A number of fields are categorized to minimized the variation of attribute values

Data formatting: The format is converted from CSV to AREFF

4. Experimentation and Results

To predict customer loyalty using data mining techniques, the classification algorithms such as rules (ZeroR), Decision trees (J48), and Bayes (Naïve Bayes) are used as these classifiers are recommended and commonly used on different literate review to compare and select the best accuracy.

The cleaned data was given to WEKA (version 3.6.6) and the experiments are conducted based on classification models.

Experiment One: In this experiment, classification model building was done using ZeroR classifier with all dataset and attributes. The experiment resulted in an accuracy of 55.34%. Its ROC area is also above 0.5, which is the minimum possible acceptable value for ROC curve.

Experiment Two: In this experiment, classification model building was done using Naïve Bayes classifier with all dataset and attributes. The test model is 10 fold cross validation. It consists of partitioning a dataset into 10 subsets. The experiment resulted with an accuracy of 97.32%. Its ROC area is also above 0.5, which is 0.992.

Experiment Three: In this experiment, classification model building was done using Naïve Bayes classifier with all dataset and attributes by using percentage split option. Out of the total data set, 70% of the records were employed for model building and the remaining 30% records were used for validation set or testing. The experiment resulted in an accuracy of 96.61%. Its ROC area is also above 0.5, which is 0.991.

Experiment Four: In this experiment, classification model building was done using J48 classifier with all dataset and attributes by using percentage split option. Out of the total data set, 70% of the records were employed for model building and the remaining 30% records were used for validation set or testing. The experiment resulted in an accuracy of 97.56%. Its ROC area is also above 0.5, which is 0.986.

Experiment Five: In this experiment, classification model building was done using J48 classifier with all dataset and attributes. The test model is 10 fold cross validation. It consists of partitioning a dataset into 10 subsets. The experiment resulted in an accuracy of 97.83%. Its ROC area is also above 0.5, which is 0.987.

As a result, there is a relatively better model prediction in the case of J48 with 10 fold cross validation. The ROC area of the J48 with 10 fold cross validation gives 0.987 with no significance difference with the ROC area in Naïve Bayes which is 0.992 and also much better than the ZeroR ROC area which gives 0.5. The overall model accuracy of J48 with 10 fold cross validation is 97.83% which shows that it has better predication than the accuracy of level of Zero (55.34%), Naïve Bayes (97.31%), Naïve Bayes (96.60%), and J48 with 70/30 percentage split testing mode (97.56%).

For the given data under this study, J48 with 10 fold cross validation has shown better accuracy and the rules generated by this model are used for interpretation and easy understanding among all the experiments.

5. Significance and Contribution

The outcome of the study is highly useful for the microfinance institutes in developing or revising existing loan disbursement and collection policies that all customers who lag behind in their repayments for more than ninety days must not be categorized as a non loyal customer and decided to sell their collateral and blacklist them. But rather negotiate to collect back the money from those customers in less than six months even if the number of months the customer repayment schedule lapsed for more than 270-540 days. Therefore, if the customer pays as per the negotiation, the customer will be counted as loyal and as a result this customer will have an opportunity to get another loan for the future but if the costumer didn't manage to pay as per the negotiation, then the customer will be counted as non loyal and as a result will not get another loan in the future.

6. Conclusion

The characterization of loyalty of customers was done based on the decision tree obtained from selected experiments. Hence, most important rules generated from the tree that characterize customers fall under "Non loyal" and "Loyal". The patterns will help the cooperative to predict the loyalty of a customer before loan disbursement.

From the results of the experiment, it can be concluded that the data mining tools and techniques, especially classification techniques, can be effectively applied on the microfinance and financial institutions data in order to generate predictive models with an acceptable level of accuracy.

Moreover, it will help to revise an existing or develop a new loan disbursement and collection policy for handling new and existing customers for MFIs. It could serve also as a basis for further research on customer profile analysis and customer relation management and an academic exercise as well helping researchers to acquire knowledge on

how to apply data mining tools and techniques in the real world.

References

[1] Getachew Hailemariam, Hill Shawndra, and Sintayehu Demissie, "Exploring Data Mining Techniques and Algorithms for Predicting Customer Loyalty and Loan Default Risk Scenarios at Wisdom Microfinance", October 28-31, Addis Ababa, Ethiopia, 2012.

[2] Hailemariam Abebe, "Application of Data Mining Techniques to Customer Profile Analysis in the Ethiopian Electric Power Corporation", Unpublished Master's Project, Addis Ababa University, 2001.

[3] Han and Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[4] Giudici, P., "Applied Data Mining, John Wiley & Sons Inc, 2003.

[5] Onut Semih, and Erdem Ibrahim, "Customer Relationship Management in Banking Sector and a Model Design for Banking Performance Enhancement, Yildiz Technical University, Istanbul, Turkey, 2002, www.necsi.org/events/iccs/2002/ nap12 onutcrmiccs2002-fixed.pdf.

[6] Wu Jia, Vadera Sunil, and Dayson Karl, "A Comparison of Data Mining Methods in Microfinance", University of Salford.

[7] Salame Emile, "Applying Data Mining Techniques to Evaluate Applications for Agricultural Loans", University of Nebraska-Lincoln, http://digitalcommons. unl.edu/agecondiss/10, last accessed on July 2, 2012.

[8] Two Crows Corporation, "Introduction of Data Mining and Knowledge Discovery", 3rd ed., http://www.twocrows.com, last accessed on September 6, 2012.

[9] Han and Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2004.

[10] Cohen, J. J., Olivia, C., and Rud, P., "Data Mining of Market Knowledge in the Pharmaceutical Industry", Proceeding of the 13th Annual Conference of North-East SAS Users Group Inc., September 24-26 2000.