Journal of Advanced College of Engineering and Management, Vol. 5, 2019

CLASSIFICATION OF LOAN APPLICATIONS OF GARIMABIKAS BANK LTD USING DECISION TREE CLASSIFICATION METHOD

Subik Shrestha[1], Laxman Paudel[2]

[2]Lecturer, Department of Mechanical, Pulchowk Campus, T.U.

Abstract

There is a possibility in finding hidden patterns that might help find a relationship between the information provided by the Loan Applicants during the Loan Application process and the status of their loan repayment. This paper highlights on finding such patterns by building a Decision Tree with the help of the data provided during the loan application process. Eleven attribute information of Five Hundred sixty four loan applicants were collected from Garima Bikas Bank Ltd. A decision tree model with a depth of 6 has been built by calculating the entropy and information gain at each split and selecting the feature with the highest information gain.

1. Introduction

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.(Peng, Chen, & Zhou)

Decision Trees fall under the category of Data mining which is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. Data mining is about solving problems by analyzing data already present in databases. (Witten, Frank, & Hall, 2011)

2. Rationale

Banking Sectors in Nepal decide the granting of loan applications by analyzing the Income status, Income Source assessments and Collateral analysis. There is a possibility of research in determining a decision tree model that can help identify hidden patterns in the data which has not yet been explored. Loan applications of Garima Bikas Bank Ltd have been divided into two categories, namely safe and risky. Garima Bikas Bank uses Equated Monthly Installment for calculating the fixed amount of payment to be made by the borrower each month during the term of the loan. If the customer has not made the payment for more than three months, then those Loan Applications are categorized as Risky loans and those Loan Applicants who pay back the EMI amount on a regular basis or doesn't delay more than three to six months are classified as Safe loans.

Eleven attribute information of Five Hundred sixty-four Loan Applicants of Garima Bikas Bank till year 2017 was collected. It was found that about 32% of the total loan applications were risky i.e. the loan applicants delayed the loan payment process by more than 3 to 6 months whereas 68% of the total loan applications were safe and paid back the EMI amount duly within 3 months. A model that reliably classifies loan applicants that are risky and not risky is needed. Therefore, a study must be done on a decision tree model that can be built for credit classification to classify the loan applicants so that the creditor can make proper decision based on the predictive model and classify loan applicants as safe or risky where a good applicant is the one who is credible whereas a bad applicant is the one who should be rejected due to the probability of the applicant not paying the loan duly.

3. Literature Review

This section introduces some of the similar works that have been done in this particular field. (Jahromi, 2009)have constructed predictive models for Customer churn in Talia Telecommunications Co. by utilizing Neural Networks algorithm and different algorithms of Decision Tree. They have used eight input features to the tree with 75% of data as training set and by utilizing various decision tree algorithms such as CART, C5.0 and CHAID algorithms along with Neural Networks.

(Buo & Kjellander, 2014)have investigated if customer churn can be predicted at the Swedish CRM-system provider Lundalogik. They found out that old customers with low number of licenses are likely to churn. Customers tend to churn when they are part of corporate fusion. They also found that customers with low average in participation during marketing events are more likely to churn and customers that are involved in marketing events and educational seminars are not very likely to churn.

(Fu & Liu) have compared various classifiers such as Logistic Regression, SVM, Random Forest, Boosting Classification Tree and Neural Networks on the dataset used in Simplifying Decision Trees. They found out that linear models work pretty well. They found out that deep learning methods such as boosting and neural network has no apparent advantages over other methods. One problem that they found with boosting classification is that it easily over fitted the data.

(Qureshi, Rehman, Qamar, & Kamal, 2013)have applied different machine learning algorithms such as linear and logistic regression, Artificial Neural Networks, K-Means clustering, decision trees in order to classify churners and active customers. The data set contained telecommunication traffic data of 106,000 customers along with their usage behavior for 3 months. The results were compared based on the values of precision, recall and F-measure. The best results were obtained with Exhaustive CHAID algorithm, a variant of the standard decision trees algorithm.

(Wu, 2015)introduces multi-criteria decision making into group decision making to propose a multi-criteria group decision making model for credit risk analysis.

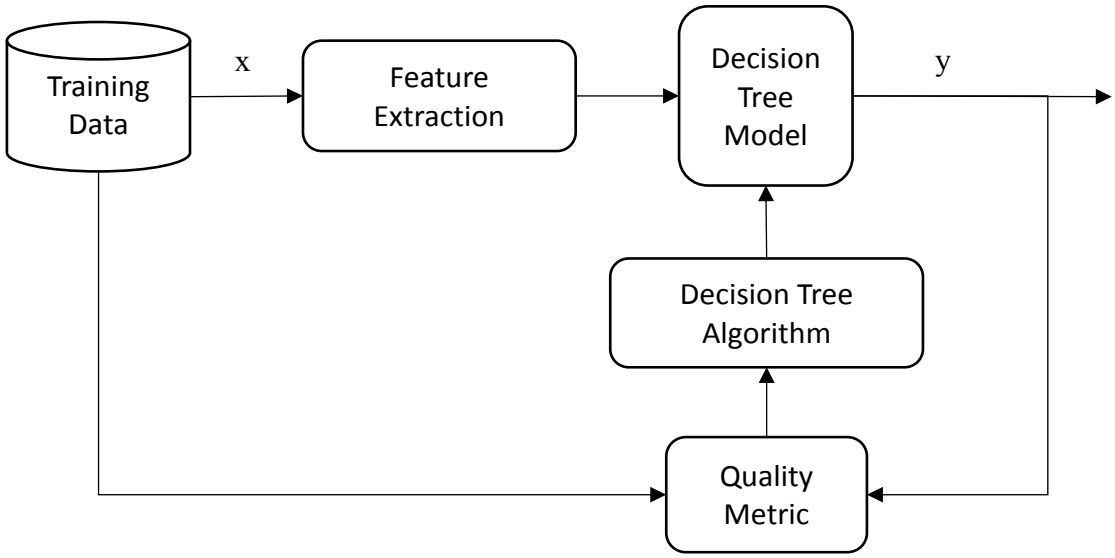
Financial lending institutions continuously look at improving their credit risk models. The study by (Salame, 2011) examined the performance of three estimation methods: logistic regression, decision tree, and neural networks, in terms of their misclassification rates of credit default. The study uses 17.328 loans of grain producers for the period of 2006 – 2010.

(Saardchom, 2011)have proposed detailed steps in developing credit scoring model based on a group analytic hierarchy process to be used in the automated loan approval system. By using expert judgment, they created an information hierarchy. Using the pair-wise comparison, the experts set priorities for each component of the information hierarchy. The credit scoring model by AHP was tested on 7081 applications, 3800 of which were approved and 3281 of which were rejected. They found that only 3.1% of automated decisions were inconsistent with actual decisions made by the experts.

Data mining methods are often implemented at advanced universities today for analyzing available data and extracting information and knowledge to support decision making. The research paper by (Kabakchieva, 2013) presents the initial results from a data mining research project implemented at a Bulgarian university, aimed at revealing the high potential of data mining applications for university management. The research work aimed at finding patterns in the available data that could be useful for predicting student's performance at the university based on their personal and pre-university characteristics. The author has also analyzed the performance of different data mining classification algorithms on the provided dataset.

4. Methodology

Eleven attribute information provided during the loan application process was used as a base for constructing the decision tree model. The data obtained was first separated into two groups namely train set and test set. 75% of the data was separated as the train set and the remaining 25% of the data was separated as the test set. The Decision tree algorithm was then applied to the 75% of the training dataset. The building of the decision tree was started with a root node with all the data in the root node. Then a feature was selected on which to split the data. All of the eleven attributes were used as the feature on which to split the dataset. The entropy and information gain was calculated for each split when splitting on a particular feature. The feature which resulted in the highest information gain was selected as the feature to split the data. Recursive and greedy approach is utilized in constructing the decision tree. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of the current node. Branches can be established based on different values of the attributes and the process above is recursively called on each branch to create other nodes and branches until all the samples in a branch belong to the same category.

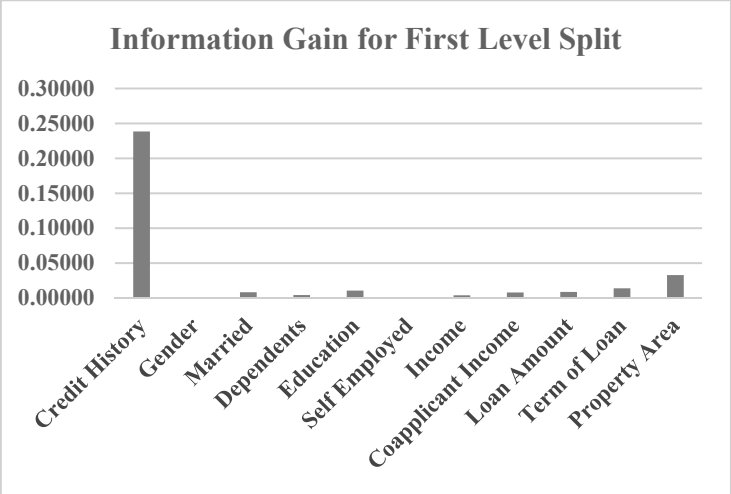


5. Data Analysis

Python was used as the programming language for data analysis. The Decision tree algorithm was implemented using Python programming language and Jupyter Notebook was used as a web based interface for running the python code. The data obtained from the Bank was stored in a csv file. The csv file was then converted to a list in python which was then used as an argument to be passed to the different functions for calculating the entropy, information gain.

6. Findings

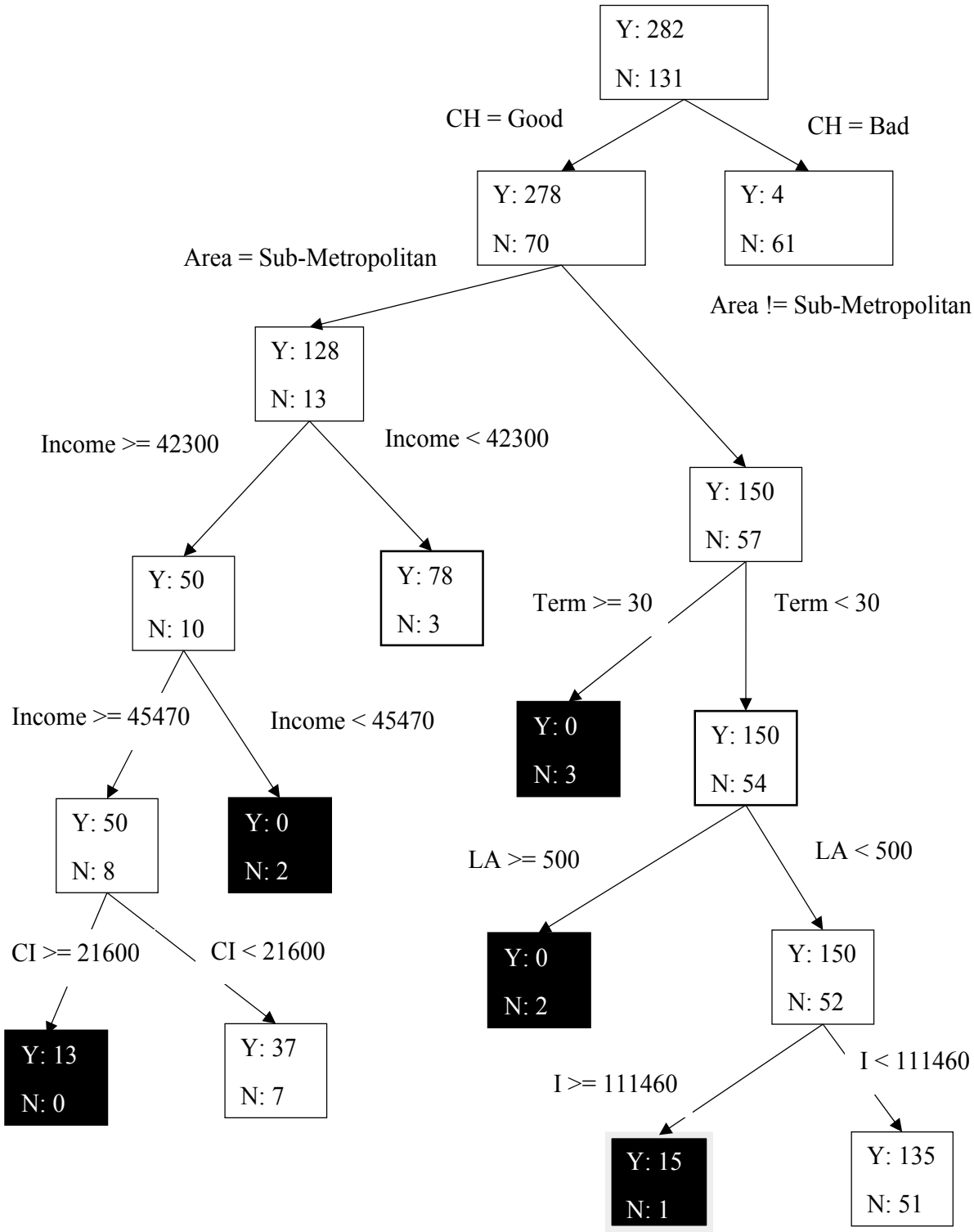
Among all the eleven attributes, the attribute Credit History of the applicant resulted in the highest Information gain, therefore the Credit History was taken as the first feature on which to begin the split and construct the decision tree model. Decision tree with depth of 5 has been constructed and various features are selected at different depths based on the information gain obtained, the feature that resulted in the highest information gain being the one chosen at a particular depth.



The following shows the information gain obtained at each depth while building the tree.

S.N.	Tree Depth(n)	n^{th} level Splitting Attribute	$(n-1)^{th}$ level Splitting Attribute	Information Gain
1	1	Credit History	N/A	**0.23857**
2	2	Loan Amount	Bad Credit History	**0.0752905**
3	2	Home Loan Area	Good Credit History	**0.0393782**
4	3	Income	Home Loan Area	**0.035873**
5	3	Term	Home Loan Area	**0.027373**
6	4	Income	Income	**0.090521**
7	4	Loan Amount	Term	**0.018995**
8	5	Co-applicant Income	Income	**0.049415**
9	5	Income	Loan Amount	**0.015825**

Decision Tree with Depth of 5



7. Conclusion

From the decision tree with depth of 5 constructed, we can conclude the following;

Loan Applicants with Good Credit History, with monthly income less than 45 Thousand, who applied for home loans in Sub-metropolitan area were found to be risky.

Loan Applicants with Good Credit History, with both the applicant's and co-applicant's monthly income greater than 21 Thousand who applied for Home loan in Sub-metropolitan area were safe.

Loan Applicants with Good Credit History, applying for Loans on Metropolitan or Municipalities with term of the loan greater than 25 years were found to be risky.

Loan Applicants with Good Credit History, applying for loans in Metropolitan or Municipalities with term of the loan less than 25 years and the amount of the loan greater than 5 Lakh were found to be risky.

Loan Applicants with Good Credit History, applying for loans in Metropolitan or Municipalities with term of the loan less than 30 years, loan amount less than 5 Lakh and income greater than 1 Lakh 11 thousand were found to be safe.

References

1. Buo, D., & Kjellander, M. (2014). "*Predicting Customer Churn at a Swedish CRM-system Company"*.
2. Fu, Z., & Liu, Z. (n.d.). "*Classifier Comparision On Credit Approval Prediction"*.
3. Jahromi, A. T. (2009). "*Predicting Customer Churn in Telecommunications Service Providers"*.
4. Kabakchieva, D. (2013). *"Predicting Student Performance by Using Data Mining Methods for Classification*". Cybernetics and Information Technologies.
5. Peng, W., Chen, J., & Zhou, H. (n.d.). "*An Implementation of ID3 - Decision Tree"*.
6. Qureshi, S. A., Rehman, A. S., Qamar, A. M., & Kamal, A. (2013). "*Telecommunication Subscriber's Churn Prediction Model Using Machine Learning"*.
7. Saardchom, N. (2011). *"Credit Scoring Model by Analytic Hierarchy Process"*.
8. Salame, E. (2011). "*Applying Data Mining Techniques to Evaluate Applications for Agricultural Loans"*.
9. Witten, I., Frank, E., & Hall, M. (2011). "*Data Mining Practical Machine Learning Tools and Techniques"*.
10. Wu, W. (2015). *"A Multi Criteria Group Decision Making Model for Credit Risk Analysis*". International Journal of Investment Management and Financial Innovations .