Journal of Research in Computer Science and Engineering
Volume 7 Issue 1
Online ISSN- 2457-0818

Data Mining Approach for Credit Analysis in Banking Sector

Ishara Devendra[1], Rasika Vijithasena[2], N.A.U.H. Jayarathna[3], SamindaPremarathne[4]

Lecturer[1, 2], Student[3], Senior Lecturer[4]

Department of Information Technology

University of Moratuwa, Sri Lanka

Corresponding Authors' E-mail: *isharad@uom.lk[1], rasikav@uom.lk[2], 179465N@uom.lk[3], samindap@uom.lk[4]*

Abstract

The banking sector is a significant part of the economy committed to holding financial assets for others. One of the critical services provided by the banking sector is the provision of loans to support individuals and businesses with their financial conditions. A bank must incur many risks due to all of its financial activities, one of which is credit risk. This research focuses on banks' credit analysis, which gives loan facilities to businesses. It empirically evaluated several data mining techniques to classify the possibility of the person or business to fulfil the obligations. The Decision tree and Naive Bayes models outperformed the state-of-the-art by offering the best performance. Finally, this study suggests that a better approach to credit risk analysis is required to maintain bank profitability.

Keywords: *Banking Sector, Bank loans, Credit Analysis, Data mining*

INTRODUCTION

Banks play an important role in the Sri Lankan financial system since they provide measurable benefits to the entire economy while transforming the risk characteristics of assets. Checking accounts, savings accounts, certificates of deposit, and loans are some of the services banks provide to assist individuals and businesses in managing their finances. From globalization and cut throat competition, the banks struggle to gain a competitive edge (Moin and Ahmed, 2012)

The main purpose is usually regarded as maximizing the market value of a company's common or ordinary shares in countries with developed financial markets.

To do so, a bank must be profitable on a constant basis and in terms of inequality. However, conflict arises because the nature of banking is to take risks. The major risks faced by banks include credit, operational, market, and liquidity risk. The credit risk is one of the most important risks to consider. As a result, credit analysis is a crucial task in the banking industry.

The probability/possibility that a borrower would fail to meet his/her obligations to the lender during the contract period is simply defined as credit risk. Since the credit analysis and appraisal procedure is done manually and is reliant on the credit officers' opinion, human errors are most likely to occur while approving loans. As a result, credit analysis is becoming increasingly crucial in the banking sector in order to reduce credit risk. The creditors use credit analysis to determine a corporation's ability to pay back loans.

The credit analysis process involves a thorough review of a business to determine its perceived ability to pay. To do so, business credit managers must analyse financial statements, apply credit analysis ratios, and check trade references to evaluate the information presented in the credit application. Each loan application is considered and evaluated on its own merits in a bank credit analysis. They check each individual or entity's creditworthiness to determine the level of risk that they subject themself by lending to an entity or individual. The objective of credit analysis is to determine a client's risk of default and to assign a risk rating to each client.

The risk rating will determine whether the loan application will be approved (or rejected), as well as the amount of credit that will be granted if approved. In general, banks offer two types of loans. There are two types of loans: personal loans and business loans. Small and medium-scale businesses, as well as large-scale businesses, are eligible for bank loans. In Sri Lanka, Small and Medium Scale businesses play a major role.

The banking industry has realized the importance of Data Mining techniques that can be used to compete in the financial market. Leading banks use data mining techniques for client segmentation and profitability, credit scoring and approval, payment default prediction, marketing, and detecting fraudulent transactions, etc.(Moin and Ahmed, 2012).

In recent years, the amount of data acquired by banks has increased dramatically. The vast volumes of data now available make existing statistical data analysis methodologies difficult to manage. This rapid expansion has necessitated the development of new data analysis techniques and tools in order to uncover the information hidden within the data. (Bhasin, 2006)

Interpreting and transforming massive amounts of data into valuable information and knowledge is one of the major issues in the banking industry. Bank executives can better predict how customers will react to interest rate changes, which customers are more likely to default on a loan, which customers are more likely to accept new product offers, and how to make customer relationships more profitable by analysing patterns and trends. (Shree, 2015)

All the above information clearly indicates the need of studying the best credit analysis techniques in the banking sector in Sri Lanka using data mining approaches.

This study mainly focuses on two objectives.

- To identify the limitations on the methods and techniques which are currently used for credit analysis and appraisals.
- To assess the most appropriate data mining technique/s which can be used for credit analysis and appraisals in order to get the accurate decisions.

This paper is organized as follows: Section 2 presents the relevant literature regarding the credit risk analysis. Section 3 describes the data collection, data pre-processing, model creation and model evaluation processes. Section 4 presents the predictive analysis and test results while, section 5 discusses the conclusion and future works of this study.

RELATED WORK

In the field of banking and the financial sector, many studies have been conducted using data mining techniques. This section summarizes some of the methodologies used in credit analysis and loan risk management, as well as their findings.

Banks make loans to their customers by confirming various information regarding the loan such as the loan amount, the lending rate, the payback time, the type of property mortgaged, the borrower's demography, income, and credit history. Customers who have been with banks for a long time and who belong to higher income groups are more likely to be approved for loans. Even though banks are cautious when making loans, there is a potential for loan defaults by customers. Data mining algorithms can assist in distinguishing between borrowers who pay back their loans on time and those who don't (Desai and Anita, 2004).

Using data mining techniques, Aboobyda and Tarig created a new classification model for classifying loan risk in the banking sector (Jafar and Ahmed, 2016). Understanding the definition of risk is necessary for risk analysis in bank loans. Furthermore, the number of transactions in the banking sector is increasing, and large data volumes representing consumer behaviour are available and loan risks are growing. Data mining is one of the most exciting and important fields of research, with the goal of extracting knowledge from massive amounts of data.

Fennee Chong investigated the factors that influence loan delinquencies from the perspective of borrower characteristics and loan characteristics. According to Fenee's analysis using 516 loan customers data, approximately 35% of respondents did not pay their obligations on time. The probability of loan repayment delinquency was predicte dusing a logistic regression technique. Gender, education level, total monthly household income, collateral availability, distance from loan providers (in kilometres), and financial planning availability such as a monthly budget were chosen as dependent variables, and the probability of loan delinquency was predicted. The findings show that the borrower-lender distance factor, collateral, level of education, and the availability of a monthly budget have a substantial impact on loan delinquencies, whereas income and gender have no significant impact on repayment behaviour (Chong, 2021).

Ting and Miklos created a prediction algorithm to assist credit card issuers in modelling the likelihood of credit card delinquency. As a dataset 711,397 credit card holders' data from the Brazil bank was collected. Several Machine Learning models were implemented including naïve Bayes, traditional artificial neural network, decision tree, logistic regression, and deep neural network, to evaluate the credit card delinquency risk considering the customer's personal characteristics and their spending behaviours. Among them, the Deep Neural Network model outperformed by providing the highest AUC and F1-score (Ting et al., 2018).

Archana et al. developed a loan credibility prediction system that considers client information such as occupation, family background, financial condition, and other factors to determine whether a loan request would be approved or rejected. They created a number of machine learning models, including the Decision Tree, Random Forest, Support Vector Machine, Linear Regression, Adaptive Boosting Model, and Neural Network, with the state-of-the-art Random Forest model outperforming the others by delivering the best accuracy (Gahlaut, Tushar and Singh, 2017).

Vimala and Sharmili have implemented a loan risk prediction model combining Naïve Bayes and Support Vector Machine techniques, to predict the status of loans. According to their findings, combining these two Machine Learning algorithms improves accuracy and speed (Vimala and Sharmili, 2018).

Research done by Ketaki et al used multiple Decision Tree approaches to analyse credit risk. From this, they explain credit score modelling, which divides loan applicants into two classes: "Good Credit" borrowers who have a financial obligation to repay, and "Bad Credit" borrowers who do not have a financial obligation to repay (Pramila and Chawan, 2012).

Bandara et al. analysed the influence of credit risk on the banking sector's profitability in Sri Lanka. A Return on Assets (ROA) is used to determine profitability. At the same time, four indicators are used to assess credit risk: the non-performing loan ratio (NPLR), the loan to deposit ratio (LDR), the net charge-off ratio (NCOR), and the capital adequacy ratio (CAR). Panel data regression analysis method was used to analyse the data from thirteen banks over a period of eight years, from 2010 to 2017.

The findings reveal that major characteristics such as credit risk have influenced the banking sector’s profitability in Sri Lanka. In order to maintain Sri Lankan banks' existing profitability, the report proposed that credit risk management be strengthened (Bandara, Jameel and Athambawa, 2021).

Laryea et al. studied the relationship between non-performing and profitability of selected banks in Gahan throughout the period from 2005 to 2010. This study found that non-performing loans negatively and significantly influence the profitability measured in terms of return on equity and return on assets employing the panel data analysis (Laryea et al., 2016).

Some researchers have explored the effect of non-performing and profitability of universal banks. They considered the Gross Domestic Product (GDP), unemployment rate, and lending rate as control variables and employed regression analysis over the quarterly data from 2000 to 2014. The study found that non-performing loans were negatively related to profitability (Mwinlaaru et al., 2016).

Financial decisions can increase the worth of a company or an individual's existence, or they might cause a company to go out of business. As a result, for businesses or people, financial decisions or financial assumptions are critical. The likelihood of losing money as a result of an asset investment is referred to as risk in financial assumptions. Financial assumptions can be used to take precautions against potential future dangers.

The Logistic Regression Analysis (LR), a traditional method, and the machine learning algorithm, Support Vector Machines (SVM) technique, a novel approach, are compared in this study in the loaning process. Its goal is to establish the importance of the comparative approaches, the model's correctness, estimating power, estimation performance, the importance of the independent variables that affect loan non-repayment, and the superiority of the techniques (Kocoglu, Ersoz, 2021).

Researchers have analysed individual and corporate customers who applied for a loan to a bank with data mining. The dependent variable is identified as several days with delayed loan payments, independent variables, on the other hand, are identified as loan amount, number of instalments, payments made to other banks, customer type and number of debtor banks. The aim of the study is the classification of customers as "not eligible for credit" or "eligible for credit" based on historical data. It is thought that the loan requests of the customers could be concluded faster thanks to the classification.

According to the results the dependent variable, the delayed loan instalment, is positively and moderately correlated with the loan amount, negatively and moderately related to the number of instalments, and the payments made to other banks are weakly related, and the customer type is strongly negatively correlated. This determination shows that customers with large loans and payments to other banks are more likely to pay delayed instalments (Ersoz et al., 2016).

METHODOLOGY

1. Data Collection

The dataset containing the information of business loan borrowers was collected from a reputed private bank of Sri Lanka. It included 1500 records including performed and non-performed loans. Dataset includes 916 records of accepted loan borrowers and 584 records of not accepted loan borrowers.

2. Data Pre-processing

Data Pre-processing is a major step in the Machine Learning process. In this step, data is cleaned, transformed and encoded to make an algorithm to easily interpret the features of the dataset. Real-world data is frequently inconsistent, incomplete, erroneous and lacking in specific behaviours or trends. For resolving such problems, pre-processing is used in the Machine Learning Model development process (Pandey, 2019).

Data pre-processing steps that have been used in this research are as follows.

- Missing values were replaced by mean values of each column.
- Discretization was done to the two numerical attributes called customer Id (CID) and amount. CID value was divided into 0-1500 range and amount attribute was divided into 0-12 range.

3. Feature Selection

In this research study, initially there were 17 attributes and among those attributes, most correlated attributes have been selected using correlation matrix by considering the relationship between independent variables and the dependent variable. Correlation matrix determines correlation between all attributes, and it can produce a weights vector based on these correlations (Kumar, 2018). Correlation is a statistical technique that can show whether and how strongly pairs of attributes are related. Correlation Matrix has been used to identify the special parameters which have strong positive relationships towards the label of the data set. According to the results received from the correlation matrix, out of 17 parameters there were 12 special parameters which have a strong positive relationship towards the label. Considering the correlated matrix results, following features were selected for model implementation.

- CID - Customer identification number given to each customer
- Amount - Loan amount (values between 500,000 to 100,000,000)
- Facility Type - Loan type (Permanent Over Draft(POD) , Term Loans (TL), Bank Guarantee (BG))

- Legal Status - Status of loan borrower (Individual, Partnership, Proprietorship, Private limited)
- Business Registration - Whether business is registered or not
- Business Age - Duration of business survival
- Business Sector - Retail, Wholesale business
- PPG Rating - Product Process Guideline Rate [$> 80 - 1$, $> 60 - 2$, $< 60 - 3$]
- CRIB - Credit Information Bureau whether Regular or Irregular
- NDB Exposure
- All Bank Exposure
- FOIR / DSCR / Interest Cover - Fixed Obligation Income Ratio (FOIR) and Debt Service Coverage Ratio (DSCR)
- Class variable: Perform/Not Perform - Represent whether loan is approved or not

4. Model Creation

In the model development step, suitable Machine Learning algorithms are selected according to the requirement and the models are developed using Data Mining tools (Nantasenamat, 2020). In this research study, the four Data Mining algorithms are used including Decision Tree, Naive Bayes, Random Forest and K-Nearest Neighbour to develop the prediction model. Rapid Miner Data Mining tool has been used to implement models. Data set was divided into two parts called training data set and testing data set and training dataset was used to train the Machine Learning model.

5. Model Testing

Model evaluation is a core step of machine learning model development. Different evaluation metrics can be used to evaluate the model and the best model is selected according to the evaluation results. In this study, Cross validation, split validation and confusion matrix techniques were used to evaluate the model. Accuracy, recall, precision and F1-Score measures were used to measure the performance of the Machine Learning model.

RESULTS AND DISCUSSION

Evaluation results of each Machine Learning model are shown in the following table. (The accuracy levels of each Machine Learning model have been presented in the table below.

Table 1 - Results of Cross Validation

Algorithm	Accuracy	Precision	Recall	F1-score
Decision Tree	63.22%	64.58%	88.18%	74.55%
KNN	57.11%	63.18%	71.45%	67.06%
Naive Bayes	64.44%	64.74%	91.82%	75.93%
Random Forest	59.44%	63.54%	78.91%	70.39%

Table 2 - Results of Split Validation

Algorithm	Accuracy	Precision	Recall	F1-score
Decision Tree	67.05%	67.42%	89.10%	76.75%
KNN	55.91%	62.75%	68.39%	65.44%
Naive Bayes	66.39%	66.40%	91.01%	76.78%
Random Forest	63.29%	66.74%	79.29%	72.47%

According to the above table 1, Naïve Bays algorithm has provided the highest accuracy 64.44% while decision tree algorithm performed the second highest accuracy as 63.22%. When considering the precision, recall and F1-score measures, Naive Bayes algorithm has achieved the highest results.

Furthermore, as table 2 represents, the Decision Tree algorithm has provided the highest accuracy and precision when evaluating using split validation method.

Naive bayes model has provided highest recall and f1-score measures while the Decision tree model provides second highest values for recall and f1-score.

According to the analysis done using two different validation techniques, it was found that Naive Bays and Decision tree algorithms have been outperformed and selected as best algorithms for predicting credit appraisal decisions.

Political changes in Sri Lanka, pandemic situations, government regulations and policies, inflation, cultural conflicts and other external factors may influence the credit decisions when considering the decision-making in the credit analysis and appraisal process.

CONCLUSION

Banking sector is one of the most important sectors to the economic growth of a country. Giving loans to support the financial conditions of individuals and enterprises is one of the key services provided by the banking sector. Credit risk is a significant risk in the banking industry, and it is simply described as the likelihood that a borrower will fail to meet his or her obligations to the lender during the term of the contract. This paper has analysed loan borrowers' information to identify the hidden patterns among the credit analysis attributes and developed classification models to predict the loan appraisal decision.

The best algorithms are selected by comparing the evaluation results of the algorithms in order to obtain the accurate decisions in credit analysis and appraisals to minimize the human errors. According to the findings, the Decision Tree model and Naive Bayes models performed the best in forecasting loan performing decisions. Finally, this research suggests that in order to retain bank profitability, a better credit risk analysis method is required.

Many financial institutions, most notably the banking sector, offer credit facilities. This study is limited to credit facilities supplied by the banking sector alone, with a special focus on SMEs in the business loan category, to minimize over complication. Hence, further research can be done in the banking industry using personal loans or large-scale business loans.

REFERENCES

1. A., G., Tushar and Singh, P. K. (2017) "Prediction analysis of risky credit using Data mining classification models," in. IEEE.

2. Bandara, H. M. K. S., Jameel, A. L. M. and Athambawa, H. (2021) "Credit risk and profitability of Banking Sector in Sri Lanka," Journal of Economics, Finance and Accounting Studies, 3(1), pp. 65–71. doi: 10.32996/jefas.2021.3.1.6.

3. Bhasin, M. L. (2006) Data Mining: A Competitive Tool in the Banking and Retail Industries. The Chartered Accountant.

4. Chawan, P. M. (2012) "A Study of Classification Based Credit Risk Analysis Algorithm."

5. Chong, F. (2021) Loan Delinquency: Some Determining Factors. Journal of Risk and Financial Management 14.

6. Desai, B. and Desai, A. (2004) “The Role of Data Mining in Banking Sector",” IBA Bulletin.

7. E., K. and F, E. (2021) “Data Mining Application For Financial Decision Optimization At Risk,” Int. J. of 3D Printing Tech. Dig. Ind, 5(2), pp. 195–209

8. . Ersöz, T., Ersöz, F. and Özbilge, S. (2016) “’ ’Determination of the Bank’s Customer Risk Profile,” Data Mining Applications’’, 10(ue 6), pp. 2199–2203.

9. Hamid, A. J. and Ahmed, T. M. (2016) “Developing prediction model of loan risk in banks using data mining,” Mach. Learn. Appl. Int. J, 3(1), pp. 1–9,.

10. Kumar, S. and Chong, I. (2018) “Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states,” International journal of environmental research and public health, 15(12). doi: 10.3390/ijerph15122907.

11. Laryea, E., Ntow-Gyamfi, E. and Alu, A. A. (2016) “Non-performing loans and bank profitability: evidence from an emerging market,” African Journal of Economic and Management Studies, 7(4), pp. 452–481.

12. Moin, K. I. and B, D. Q. (2012) “Use of Data Mining in Banking”,” Ahmed ,International Journal of Engineering Research and Applications, 2(ue 2,Mar-Apr), pp. 738–742,.

13. MrunalSurve, Pooja Thitme, Priya Shinde, Swati Sonawane, Sandip Pandit (2016) , “DATA MINING TECHNIQUES TO ANALYSES RISK GIVING LOAN(BANK)”, . IJARIIEISSN(O)-2395-4396 Vol-2 Issue-1.

14. Mwinlaaru, P. Y. et al. (2016) Non-Performing Loans and Universal Bank’s Profitability”. University of Cape Coast.

15. Nantasenamat, C. (2020) How to build a machine learning model, Towards Data Science. Available at:

https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1

16. Pandey, P. (2019) Data Pre-processing: Concepts, Towards Data Science. Available at: https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

17. Serwadda, I. (2018) "Impact of credit risk management systems on the financial performance of commercial banks in Uganda," Acta Universitatis Agriculturae et SilviculturaeMendelianaeBrunensis, 66(6), pp. 1627–1635. doi: 10.11118/actaun201866061627.

18. Shree, D. (2015) Data Mining Tools and Techniques used in Banking Sector.

19. Vasarhalyi, T. S. A. M. (2018) "An application of deep neural networks", Predicting credit card delinquencies,. Intelligent Systems in Accounting, Finance and Management.

20. Vimala S., S. K. C. (2018) "Prediction of Loan Risk using Naive Bayes and Support Vector Machine", in International Conference on Advancements in Computing Technologies - ICACT.