ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

A STUDY ON LOAN DEFAULT PREDICTION MODELS

AMBOLLA JYOTHI [1] **MBA Student**

MRS. D SUNITHA[2] **ASSISTANT PROFESSOR**

[1,2]MASTER OF BUSINESS ADMINISTRATION

[1,2]Megha Institute of Engineering and Technology for Women, Sy. No. 7, Edulabad Road, Edulabad, Ghatkesar, Telangana.

ABSTRACT

Financial institutions have battled with handling and determining the creditworthiness of their clients in the recent past. The ever-increasing customer base makes it hard for financial institutions, especially banks to follow due process of determining if a customer qualifies for a loan or not depending on one's credit history manually. As a result, there have been delays in processing customer loans, making banks and other financial institutions inefficient. Automation of these tasks has come as a factor of necessity, to improve the speed, cost, and efficiency of processing loans. Thus, an AI web-based application that predicts the probability of borrowers' failure to repay the loans is a handy solution for this time.

The system will auto-collect historical borrowing and repaying data for that particular borrower within the shortest time with high precision whenever an individual uploads the personal data. The prototype will apply cloud cutting-edge AI and machine learning services to analyse the borrowers' creditworthiness and apply the recommendation to achieve the following: Identifying personal information of the proposed borrower, evaluating the prerequisite information for loan approval or decline, determining the credibility, notifying the lender if any loan default history, and recommending approval or disapproval based on the history of the borrower. This application will save financial institutions stress and time, avoid losses in the lending business, reduce the loan process time, decrease the likely risks associated with loans, and save the costs of the admission department.

1 - Introduction

1.1Background Information

In the modern world of business, lending and borrowing money from financial institutions bring new opportunities to financial organisations, yet there is a challenge of incurring losses following the risk of loan defaulters. This business of lending money to people becomes a frequent business activity for financial institutions. Every day people are seeking to borrow money from different financial institutions for different reasons. Conversely, not every person in need of a loan is reliable, and not all people can be given loans. Moreover, a reasonable number of people on yearly basis do not repay the amount advanced from lending institutions, hence these institutions incur huge losses.

In addition to many people seeking loans from different lending financial institutions, bad loans are significantly affecting the financial sector across the globe. Developing a load predictive model can greatly help these financial institutions to deal with the challenge of giving loans to historically loan defaulters and minimising the risk of incurring huge losses from the number of money defaulters end up not paying back. Using historical data of borrowers can be a great tool in developing a better way of predicting the likely behaviour of a loan applicant and being able to classify the person as a defaulter or non-defaulter.

The process of approving or declining an application for a loan is a significant process for any lending institution. In response, the technological advancement of e-commerce and big data technology can be applied to create a predictive model to categorise each borrower as a defaulter or not using these technologies when financial institutions want to give loans. Therefore, this project is based on the concept of artificial intelligence and machine language techniques to use client's data accessed from reliable financial analytics data websites to ascertain relevant information and predict whether a loan application would be able to repay the loan or not, that is, predict whether the loan applicant in question would be a loan defaulter or not.

1.2 Statement of the problem

Loan processing and approval in financial institutions is a major issue and bottleneck problem, yet integrating machine learning and artificial intelligence (AI) technologies can make a significant change in this important business activity. There is an urgency to develop an efficient loan default predicting system that will become a game-changer in loan processing and approval that can be used by all lending institutions to have fair and successful loan approval systems with the least minimal ratio value of loan defaulters by applying emerging technology to perform time-intensive tasks and making problem-solving more efficient (Tariq et al., 2019). In classifying a loan applicant, the borrower's history about finance would be used. It implies that some predictive data variables to predict the targeted variable as defaulter (delayed or failed to repay the loan on time, 1) or non-defaulter (repaid the loans in time, 0) would be used.

1.3 Project Definition and Goals

- This project aims to build a predictive model to categorise a loan application as a loan defaulter or not using the relevant personal data collected from the historical loan default website when lending is processed and given.
- Minimise the default risk of borrowers ending up not repaying the loans using this created model.

1.4 Methodology

The project proposes the utilisation of emerging technologies such as artificial intelligence and machine learning integration to develop a predictive loan default system in loan processing and approval within lending institutions. The predicting model uses data from the loan default database information shared over financial platforms that aim to minimise the credit risk and bad loans in the financial industry. Additionally, other leading individuals in peer-to-peer platforms will find this proposed predictive loan default application relevant in handling the likelihood of losing money through lending historically known loan defaulters.

In data collection to evaluate the loan applicant's historical borrowing records, data will be gathered from reliable and credible databases of renowned institutions to achieve the initial objective of this proposal of minimising the value of money lost through lending to defaulters. Using the data provided by some research websites like the S&P Global Market Intelligence will help in developing this project. This data will be meaningful in testing the functionality of this application in evaluating the creditworthiness of a loan applicant and aiding decision-making as to whether is a defaulter or a non-defaulter. The study employs data analysis algorithms such as Neural networks, decision trees, random forest, and k-nearest neighbours to analyse the data. The techniques help in developing the trends, patterns, and insights about one’s financial status based on history. After analysis, the final results of each algorithm are compared with the others. The algorithm with the highest accuracy, reliability was chosen as the most suitable for the study.

2 - Literature Review

2.1 Literature Review

The credit score is an influential metric in loan origination and is used in loan processing and approval in many financial lending institutions. According to Sengupta and Bhardwaj (2015) credit scoring metric is useful in ascertaining the borrower's creditworthiness in the current loan application. This is the continued use of credit scoring information across lending institutions aimed at minimising the default loss that these financial institutions incur. Further, the credit scoring metric can be meaningful in observing the loan performance because of its ability to determine the likely credit risk the lending institution can presume to incur in the event the borrower defaults the loan approved.

Lending in finance is increasing and as one way of getting monetary support to meet personal needs without the old credit or bank union (Zhu et al., 2020). So, developing a loan default predictive system is becoming a necessity in evaluating the type of borrowers to give and not give a loan because of the bad loan resulting in huge financial losses to lending institutions. Having a good credit score for a loan applicant is vital for one to get a loan approval or else get declined. Different criteria are used to minimise the risk of a financial institution losing its resources when a loan applicant fails to repay the loan given. Majorly, the lending firms use the historical data of a loan applicant to ascertain whether an individual qualifies to get a loan or not.

Further, individuals and investors seek loans from financial institutions unlike those obtained from a bank. Another type of platform that readily lends money to borrowers includes individual to individual (P2P) in need of cash. Currently, there are online platforms for different financial institutions that offer this lending service to new applicants because of reducing lenders' risk of losing the monies to loan defaulters. Tariq et al. (2020) assert that using technological advancement and information sharing across lending institutions and individuals is taking a new perspective in the lending decision-making process. This proves how this proposed loan default predictive system application will be beneficial to lenders.

Several mobile-based systems have been also developed to help microfinance institutions predict the creditworthiness of their clients. By utilising spatial data, such as travel and expenditure behaviour, these organisations can analyse, determine and classify any customer to a credit level. The system then automatically recommends the amount of credit a customer qualifies for.

However, this project aims to develop a web-based solution for large financial institutions like banks, dealing with a large volume of customers and data.

Alomari & Zakaria. (2017) used machine learning classifiers to predict loan default based on 188,124 loan records from lending club. Random forest classifiers yielded the best performance (71.75%) followed by Naïve bayes classifier (61.44%).The worst was 1R with 59.9%. In a similar study (Xu et al, 2021), they used random forest (RF), extreme gradient boosting tree (XGBT), gradient boosting model (GBM), and neural network (NN) to predict loan default. Data from Renrendai.com was used. Random forest was found to be more superior than the rest of the

3 - Project Description

In this project, we will source the best available dataset for the related problem statement of predicting loan defaults within customers. This would involve using multiple steps to proceed with the problem statement which is termed as the CRISP-DM method of solving a data analytics or data science problem. It involves multiple steps of solution like business understanding, data collection, data understanding, data cleaning and preparation, modelling and then providing the insights and recommendations to the stakeholders. The below figure summarises the steps that would be involved within the course of this project which we plan to perform one by one.

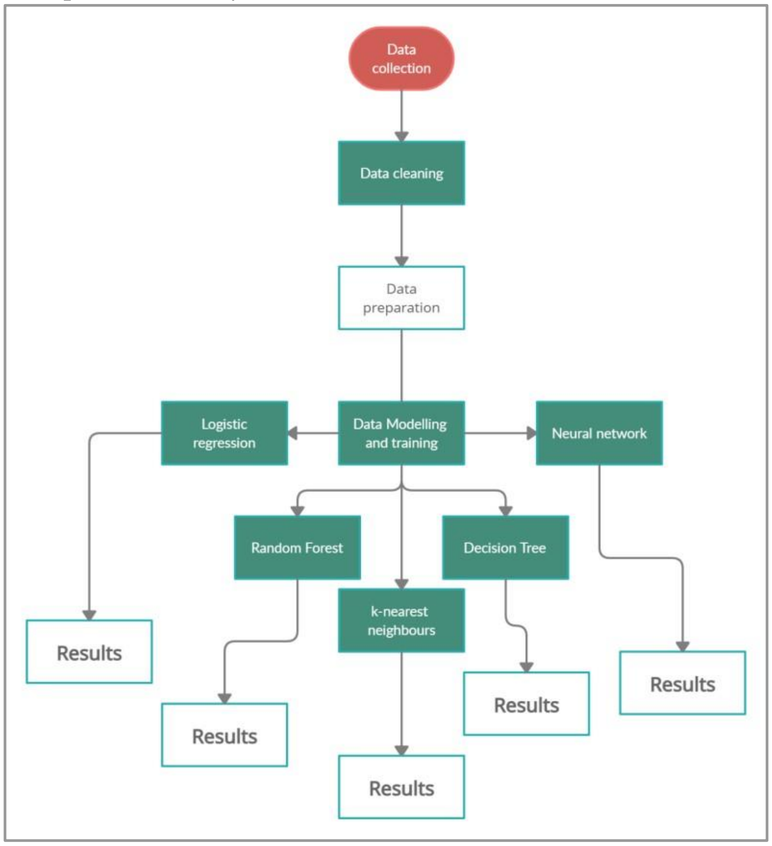

Figure 1: system flow chart

To summarise the steps that would be involved in the course of the project, we outline the major steps below. This would help our readers understand the core steps that involve solving an analytics project in a framework manner.

Business Understanding

It is important that the domain of the report is understood before moving forward with the solution approach. The financial industry along with the loan business has to be properly understood

Data Understanding

Here, the dataset which is obtained from an online repository has been explored well by using data analysis and various statistics to understand the data in depth. Different data description needs to be identified like averages, standard deviations as well as other skewness of the variables in the data

Data Preparation

The obtained data is then prepared by using data cleaning methods to treat missing value and other inconsistencies. This would help us obtain a standardized data for the modeling approach as well as any data analysis

Data Modelling

The cleaned data is then used for the modeling purpose wherein the data is fed into the machine learning models with a split of train and test to the ratio 4:1 for train to test. This helps us to validate the prediction results at the end

Evaluation

In the final step of the entire pipeline, we want to validate and compare the different models that have been obtained. Various evaluation methods are used to identify the best fit model for the solution approach

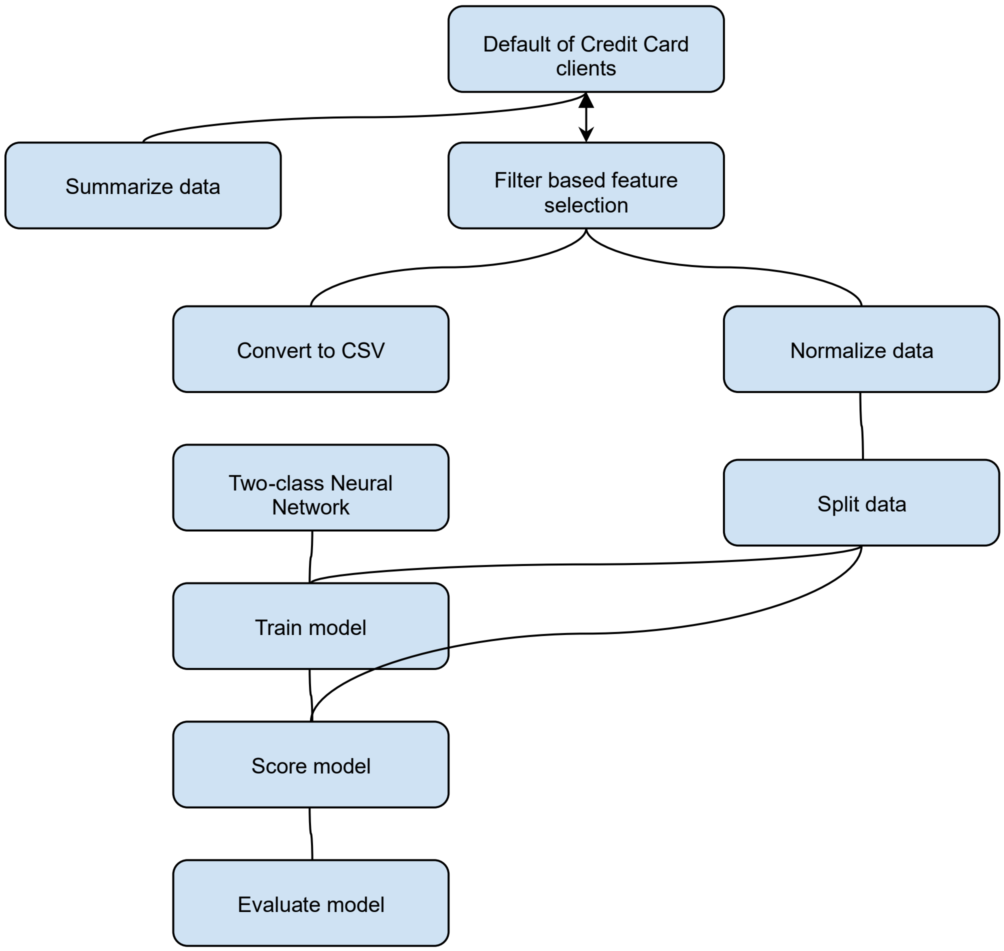

Figure 2: A prediction model

After data collection, to gain useful information different data analytics techniques such as Excel, R programming, and Tableau to process data. Then, in filtering and recognizing meaningful patterns, artificial intelligence, data mining, machine learning, and modeling help in determining prediction in this project. In this project, I will use sample data of loan application forms, and the identity of applicants approved to the lending institution in the past few years ago. I will use data from the financial institution data repository database to predict if a loan applicant will fail to repay or not based on the objective data and whether a lending institution should lend to a loan application or not.

Project variables

An AI system that readily works on historical borrowing data to precise information for decision making.

- R scripts having fitted models from the data.
- The results of the study and the research publication.
- Collected data in a CSV file format.
- Dashboard for the updated creditworthiness of a loan applicant
- Recommendation based on the classification of a loan applicant on the predetermined variables.

- An efficient and fast predictive model that improves loan processing speed.

4 - Project Analysis

4.1 Dataset

Data used in this study comes from LendingClub.com, which is a peer-to-peer lending organisation based in San Francisco, California. It consists of details of 2,925,493 loan records for the period between 2007 and the third quarter of 2020. For each loan record, the data consist of 141 attributes, measuring individual and group borrowing and repayment behaviour. The data is available for download on the Kaggle machine learning repository.

4.1.1 Data cleaning and pre-processing

Most machine learning methods use listwise deletion to deal with missing cases. This means cases (loan records) with at least one missing attribute would be excluded from the analysis. Some attributes on the data had so many missing cases and would lead to exclusion of many cases. To avoid such a scenario attributes with more than 40% missing cases were excluded. Conversely, the data includes both individual and joint loans. Some attributes like details of the co-borrower are only possible for group loans. Such details are not available for individual loans and would lead to exclusion of all individual loans if listwise deletion happens; we dropped such variables too to avoid losing much data.

4.1.2 Data partitioning

The data was partitioned randomly into 80% training and 20% testing sets. The training set will be used to fit the models while the remaining 20% will be used to evaluate and compare the performance of the models.

4.2 Exploratory data analysis

Summary statistics and graphing methods were used to understand better borrowers on this sample and their borrowing behaviours.

4.2.1 Frequency distribution of loan outcomes

Figure 1 below shows that 90% of the loan records used to train our models were properly serviced. The remaining 10% of the loans were either on default or had been charged off. The difference between the two loan statuses is; the organisation treats loans with more than 120 without payments as defaulted and charges off defaulted loans if there are no hopes of receiving further payments.

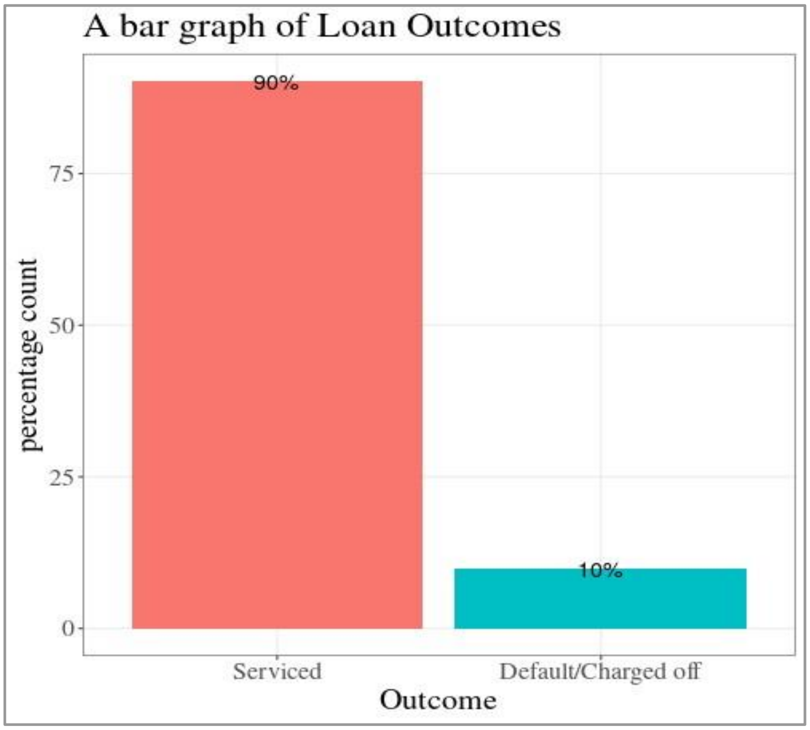

Figure 1: Distribution of loan outcomes

4.2.2 Distribution of loan outcome across independent variables

By comparing loan outcomes across application types we see that defaults/charge offs were slightly higher on individual loans (9.96%) as opposed to group loans (7.80%). On the other hand, borrowers who had a charge off and were working with debt settlement companies had a significantly higher chance of defaulting (99.04%) compared to those who were not working with a settlement company (8.38%). Borrowers on hardship plans had a lower chance of repaying their loan (0.04%) compared to those not on hardship plans (10.40%). For homeownership, people with rented homes had the highest risk of not repaying loans (11.54%) followed by people who own houses (10.06%) and then people on mortgaged (8.30%). Long-term loans had a higher chance of being defaulted (12.18%) compared to loan term loans (8.31%). Verified clients on the other hand had higher default rates (15.02%) compared to source verified (10.44%) and unverified ones (6.48%). See *figure 2* below.

5 - Conclusion

5.1 Conclusion

There is ever-increasing lending in finance as one of the ways of receiving financial support to cater to personal needs in the absence of bank unions or old banks. Currently, various financial institutions have established online platforms that offer money lending services to new loan applicants as a way of minimising the potential risks of money loss to loan defaulters. Besides, the microfinance institutions have also introduced various mobile-based systems that utilise spatial data, such as travel and expenditure behaviour to help in predicting an individual's creditworthiness as well as determine and classify any customer to a credit level. Even though most of the financial institutions are currently leveraging the benefits of credit score as an influential metric in loan processing and approval, the absence of fair and successful loan approval systems with the least minimal ratio value of loan defaulters is still a major stumbling block relating to loan processing and approval in the financial institutions.

The escalating instances of loan defaults cause massive losses in money lending companies thus creating an urgency to introduce effective strategies for addressing the identified issue. Developing a model for predicting loan default is critical in minimising the risks related to loan defaults after giving loans to individuals who end up not paying back the money. Emerging technologies such as Machine learning techniques are at the heart of addressing the issue. The techniques are helping in developing a practical predictive model which utilises an individual's historical data to predict their behaviours and classify them either as a loan defaulter or non-defaulters before giving them loans. Such approaches are significant in making useful decisions in financial institutions as far as minimising losses from loan defaults is concerned.

The current research study is associated with various limitations. Firstly, the study utilised a secondary data source that may contain inaccurate information thus producing unreliable results. Secondly, the dataset contained loan records with various attributes with missing cases thus affecting the final results of the study. Thirdly, the study focused on the data source from one money lending company which affected the outcome of the study. Lastly, the data used included both personal loans and joint loans which affected the study outcomes due to the unavailability of attributes such as the details of co-borrowers on individual loans.

5.2 Recommendations

- Based on the limitations of the study, I would recommend the use of primary data sources instead of secondary data sources since it gives first-hand information which could produce more accurate results.
- Also, I would recommend the use of data sources with more records as this would produce more reliable outcomes.
- Finally, I recommend the use of various data sources for different money lending companies to enable a comparison of the results.

5.3 Future Work

The current study was associated with various limitations which creates a need for future studies to address the identified shortcomings. I suggest the following future studies;

- Firstly, a study should be conducted using primary data sources to obtain first-hand information as this is likely to give more accurate results compared to secondary data sources, usually associated with various limitations.
- Secondly, it is critical to carry out another study that utilises a dataset with more records as this would give more reliable study results.
- Finally, there is a need to conduct another study that utilises datasets from various money lending companies to attain reliable results after comparing the outcomes.

Bibliography

Aditya Sai Srinivas, T., Ramasubbareddy, S., & Govinda, K. (2022). Loan Default Prediction Using Machine Learning Techniques. *Innovations in Computer Science and Engineering*, 529–535. https://doi.org/10.1007/978-981-16-8987-1_56

AIS Electronic Library (AISeL) - ICIS 2016 Proceedings: Credit-worthiness Prediction in Microfinance using Mobile Data: A Spatio-network Approach. (2016). AIS Electronic Library. https://aisel.aisnet.org/icis2016/EBusiness/Presentations/28/

Aniceto, M. C., Barboza, F., & Kimura, H. (2020). Machine learning predictivity applied to consumer creditworthiness. *Future Business Journal*, *6*(1). https://doi.org/10.1186/s43093-020-00041-w

Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An Empirical Study on Loan Default Prediction Models. *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3483–3488. https://doi.org/10.1166/jctn.2019.8312

Bagherpour, A. (2017),Predicting Mortgage Loan Default with MachineLearning Methods;