DOI: *10.53739/samvad/2021/v23/166261*

ISSN (Print) : 2249-1880

SAMVAD: SIBM Pune Research Journal, *Vol XXIII, 37-49, December 2021* *ISSN (Online) : 2348-5329*

Extrapolation of Loan Default using Predictive Analytics: A Case of Business Analysis

Riktesh Srivastava

City University College of Ajman, Ajman, United Arab Emirates;
r.srivastava@cuca.ae

Abstract

The research assesses the validity of a customer's appropriateness for a loan using a machine learning approach called predictive modeling. Banks and Non-Banking Financial Companies (NBFCs) are at danger of significant Non-Performing Assets (NPAs) due to customer non-payment of loans (Non-Performing Assets). The data for this study came from Kaggle, and eight different prediction models were employed to determine if the borrower would be able to repay the loan. Adaboost, k-Nearest Neighbors (k-NN), Logistic Regression, Support Vector Machines (SVM), Decision Tree, Naive Bayes, Neural Networks, and Random Forest (RF) are the eight models, respectively. The purpose is to back up decisions made on the basis of factual evidence rather than subjective reasons. Classification Accuracy, Precision, Recall, and F-1 scores are the four performance parameters used to determine the results. With 70% and 30% respectively, the dataset is separated into train and test datasets. The whole analysis is done in two phases, with the first being a full model that is trained on 70% of the train data and the second being observed on 30% of the test data. The purpose of this study is to see how objective characteristics influence borrowers to default on loans, to identify the most common reasons for default, and to predict which customers would default. There are two evaluations we did for the research, wherein, first we took overall train set and make predictions using predictive modeling. The Adaboost predictive model delivers the greatest results, with a recall rate of 0.384, classification accuracy of 59.2 percent, true-positive rate of 69.74 percent. Second, we performed feature selection and discovered that Credit History with 31 percent had the utmost impact on loan default detection. By partitioning the dataset into Credit_History 1 and 0, we discovered that Credit History 1 produces superior results, with a rate of 0.444, 60.5 percent classification accuracy, and a true-positive rate of 68.7%.

Keywords: Adaboost, Decision Tree, k-Nearest Neighbors (k-NN), Logistic Regression, Naïve Bayes, Neural Network, Non-Banking Financial Companies (NBFC), Support Vector Machine (SVM), Random Forest

1. Introduction

Banks and Non-Bank Financial Companies (NBFCs) give loans to borrowers and benefit from the interest they charge on repayment. Banks and NBFCs benefit if the borrower makes the loan payments on schedule, but they lose money if the borrower defaults (known as Non-Performing Assets (NPAs)). In India, banks and NBFCs have long taken a lenient attitude toward loan defaults, resulting in a rising loan delinquency rate (Chopra *et al.*, 2020) which doubled the declared loan delinquency rate. Relative economic stability during the exercise and the absence of a capital backstop together make it unique. We find that the expected reduction in information asymmetry does not automatically lead to the recapitalization of banks by markets. The consequent undercapitalization leads to underinvestment and risk-shifting through zombie lending. The impact flows to the real economy through borrowers, including shadow banks, and adversely impacts growth. These findings show that bank cleanup exercises not accompanied by policies aimed at recapitalization may be insufficient even during normal times. However, following the financial crisis of 2008, banks began to use technology to assess a borrower's ability to repay a loan.

In 2020, India had a gross NPA of 8% (Statista, 2021), with wilful defaults totaling over 28,784 crores for banks (Ghosh, 2021) and more than 50% for NBFCs (Shukla, 2021). According to Reserve Bank of India (RBI) reports, total NPAs for public and private sector banks in 2020 will be 230917.59 crores and 55745.87 crores, respectively (RBI, 2021). Despite the fact that the RBI permitted borrowers to put their payments on hold for three months, the number of people seeking to avoid paying their loans surged.

The accuracy of the results remained a challenge (Blöchlinger & Leippold, 2006; Dastile *et al.*, 2020; Einav *et al.*, 2013; Jia, 2018). Fintech and developments in machine learning prediction models, on the other hand, have made it simpler for banks to identify trustworthy borrowers and expedite money lending and debt collection (Moneycontrol, 2020). As a result, the majority of Indian banks and NBFCs have started using machine learning predictive analytics to forecast loan defaults (PTI, 2021).

This paper examines the variables that contribute to loan defaults in order to assist banks and NBFCs in taking early and appropriate remedies to avoid Non-Performing Assets (NPAs). For training data, there are 12 variables and 687 observations in the dataset. The study begins with a descriptive analysis of the training dataset, and then moves on to predictive models to choose the optimal model based on the four performance indicators. After then, the experiment is run on the test data.

The following are the five sections of the paper. We looked at numerous sorts of prediction models employed by other authors in Section 2. The usage of the Team Data Science Process (TDSP) as a Data Mining framework for analysis is explained in Section 3. The installation of several prediction algorithms for identifying loan defaults is described in Section 4. A comparison of all the models is presented in Section 5 based on classification accuracy, recall, precision, and F-1 score. Section 6 concludes with suggestions and a conclusion.

2. Related Work

Machine learning and predictive modeling have been shown to be effective in detecting loan defaults by a number of academics. However, in the part on research, the most current research articles (from 2018 to 2021) are taken into account. The most recent study was chosen because it clearly demonstrates how academics are using machine learning models to predict the results of a particular dataset. The latest study also provides a detailed overview of machine learning methods that may be used to improve an application's intelligence and capabilities.

Chen *et al.* (2018) studies on how to cope with data imbalance in order to improve the performance of neural networks in loan default prediction. The accuracy rate obtained was more than 80%. For assessing credit risk, used Dependency Sensitive Convolutional Neural Networks (DSCNNs). With an F1-score of more than 0.86 and a prediction of more than 65 percent, experimental findings suggest that an appropriate bundle of approaches may achieve good prediction performance. For loan prediction, (Al-qerem *et al.*, 2019) employed the Nave Bayes machine learning system, which had an accuracy of about 87.2 percent. (Wu *et al.*, 2019) applied the Nave Bayes machine learning model with 63 percent accuracy for loan prediction. Zhu *et al.* (2019) conducted the loan default research using Random Forest predictive models with 95% accuracy. (Gurbani, 2019) used logistic regression and random forest to predict the loan defaults. There study states that though Random Forest predicts better than Logistic Regression, however in the banking business, one must be able to understand model findings and accurately explain why loans are being declined to clients, as per government rules and compliance standards. As a result, Logistic Regression should be used to create the true model before it is deployed in production. Simple Logistic Regression is a score that is made up of coefficients multiplied by features. It is possible to interpret it as probability. If users are turned down, the features with the lowest ratings may be discovered, and the account holder can be advised on how to improve their score. Aniceto *et al.* (2020) performed the credit risk evaluation for banks from Brazilian bank's loan database using various machine learning techniques. They adopted Support Vector Machine, Decision Trees, Bagging, AdaBoost and Random Forest models, and compare their predictive accuracy. There experiment show that Random Forest and AdaBoost perform better when compared to other models. Lai (2020) states that now we have more possibilities for categorizing and forecasting loan default than ever before, thanks to the onset of the big data age and the development of machine learning algorithms. The author show that the AdaBoost model can forecast loan default with 100% accuracy using a real-world dataset from a bank in China, beating alternative models such as XGBoost, random forest, k closest neighbors, and neural network. For airtime lending, applied a variety of credit scoring methodologies based on an acceptable machine learning model. Over three million loans belonging to over 41 thousand consumers with a three-month payback period were scrutinized. Several cross-validation procedures are used to assess the ability of logistic regression, decision trees, and random forest to categorize defaulters (Kriebel & Stitz, 2020) deep learning outperforms them in almost all cases. However, machine learning models combined with word frequencies or topic models also extract substantial credit-relevant information. A comparison of six deep neural network architectures, including state-of-the-art transformer models, finds that the architectures mostly provide similar performance. This means that simpler methods (such as average embedding neural networks analyzed the credit risk using the deep learning neural networks. The authors stated that machine learning is an effective tool to predict the loan defaults thereby avoiding the credit risk. Madaan *et al.* (2021) a large population applies for bank loans. But one of the major problem banking sectors face in this ever-changing economy is the increasing rate of loan defaults, and the banking authorities are finding it more difficult to correctly assess loan requests and tackle the risks of people defaulting on loans. The two most critical questions in the banking industry are (i used Random Forest and Decision Trees machine learning models to test the loan defaults on the same dataset, and the findings revealed that the Random Forest method beat the Decision Tree approach with significantly higher accuracy used four machine learning methods (Random Forest (RF)), extreme gradient boosting tree (XGBT), Gradient Boosting Model (GBM), and Neural Network (NN) to predict important factors affecting loan repayment in the Chinese P2P market. All four methods had an accuracy of over 90%, with RF outperforming the other classification models. Barbaglia *et al.* (2021) study the loan default behaviour in seven European nations using a dataset of 12 million residential mortgages. The authors used logistic regression to compare the outcomes of the nations and found that they could predict loan defaults with an accuracy of more than 90%. Chen & Zhang (2021) examined the classification performance of six machine learning algorithms: neural network, KNN, logistics, SVM, random forest, and decision tree. The experimental findings demonstrate that using the suggested approach, the model's prediction performance can be considerably improved, with the AUC value increasing from 0.765 to 0.929. The comprehensive prediction impact of the neural network is superior than the other five prediction models, according to the authors. Sarkar *et al.* (2021) used the adjusted Gradient Boosting (XGBoost) and Logistic regression machine learning models to predict the loan default. The authors stated that the precision of XGBoost's estimation is better than the logistic regression. Sunitha *et al.* (2021) employed logistic regression predictive modeling to accurately forecast loan default with an accuracy rate of 84.4 percent. Zhao (2021) also utilized merely a logistic regression model, achieving a 70% accuracy rate.

All of the related work on using a predictive model to identify loan default is based on current research. Almost all of the authors utilized several prediction models to identify loan defaults for the given dataset, according to the research. Almost all of the researchers were able to identify the best appropriate model and provide results based on certain performance indicators. The research gap discovered and addressed in the study is that the analysis is first performed on the entire train dataset before being evaluated on test data. Aside from that, a feature selection is carried out, which determines the most relevant parameter that influences loan default, and a separate test is carried out to determine correctness.

Table 1 gives the detail listing of the research gap, question and objective that is carried out in the study.

3. Loan Default Prediction: A Case study

Machine learning can give significant support to predict loan defaults for banks and NBFCs. This paper aims to demonstrate a possible application for machine learning in banks and NBFC's to indicate the loan defaults by borrowers. Furthermore, a proposal for objective sorting on potential reasons for loan default is predicted using classification algorithms. The work reveals how machine

Table 1. Research gap, questions and objectives for the study

Research Gap	Research Questions	Research Objectives
RG: There appears to be a paucity of literature on the impact of several predictive models in loan default prediction.	RQ 1: Can we use multiple prediction algorithms to forecast loan default?	RO1: To find the best appropriate predictive model for detecting loan default.
	RQ 2: Is the feature selection better at predicting loan default?	RO 2: To determine the relevance of feature selection and then to pick the best appropriate predictive model for loan default prediction.

learning can be a breathtaking innovation to predict loan defaults.

3.1 TDSP Framework

The research follows the steps from the TDSP framework, Team Data Science Process (Microsoft, 2020), and includes the following phases (Figure 1).

- Collect the dataset (divide it into 70% train and 30% test data set), perform the basic descriptive analysis for the dataset. The step is called as Data Explanation (Section 3.2).
- Use machine learning models to predict the loan default (Section 4).
- Identify the machine learning models with the best outcomes for the overall train and test dataset using performance measures. We also performed a feature selection set among the set of characteristics, as well as a performance measure, to assess the accuracy of the credit history (0, 1) (Section 5).

3.2 Data Explanation

The downloaded dataset has 13 features with 687 observations for train data and 294 observations for the test data. All components are related to the bank borrowers' who took a loan from the bank (Table 2).

Table 2. Dataset features

Loan_ID
Gender
Married
Dependents
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
Property_Area

The dependent variable, Loan_Status, identifies 0 when a borrower defaults on the loan and one otherwise.

Data Preparation is another crucial phase in the Data Explanation process. The most important stage in conducting machine learning experiments is data preparation. Experiments for train datasets may be undertaken after the data preparation to test the model (Redman, 2018). As a result, data preparation accounts for 80% of the time spent in predictive modeling (Press, 2016). The data preparation task in the study is to turn qualitative characteristics into quantitative ones. Property_Area and Education are the variables, with numeric values ranging from 1 to n.

3.3 Basic Descriptive Analysis

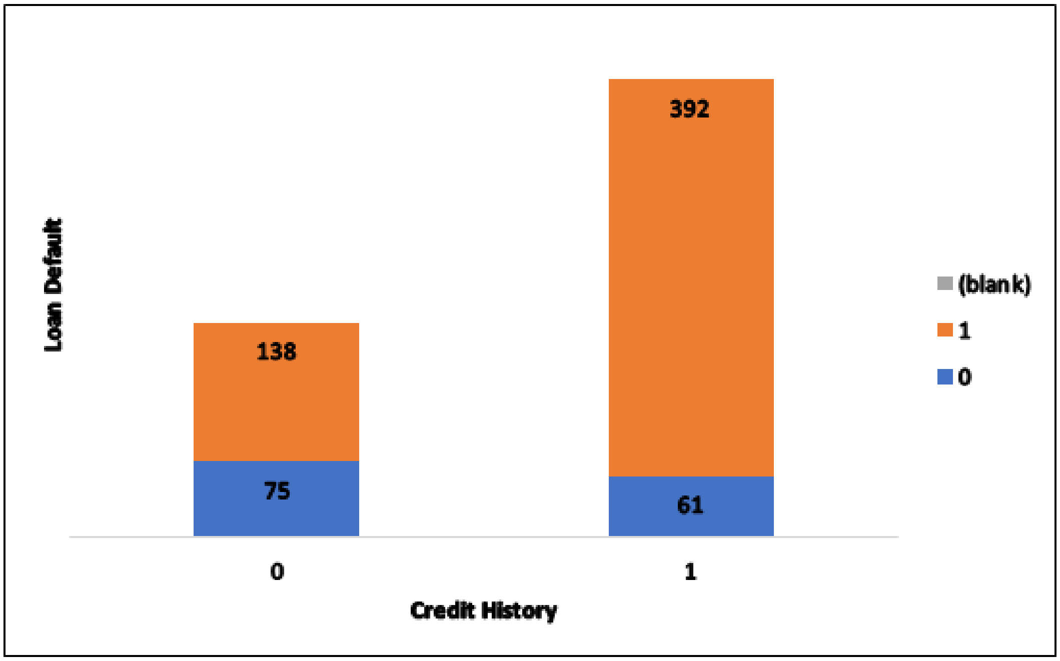
The basic descriptive analysis is used to figure out how the target variable is distributed across the dataset. 32 percent (218 customers) of the 687 consumers in the training sample defaulted on their loans, while the remaining 68 percent (469 customers) did not.

Each attribute was linked to the target variable Loan Status in the descriptive analysis of dataset features. Only the five most essential characteristics (with feature significance of less than 3%) were examined in this section. Credit history appears to be the most important element in loan default, as "Credit History" ranked first with 31 percent of the vote. Borrowers with a credit history of 0 have a default rate of 35%, compared to 13% for borrowers with a credit history of 1 (Figure 3).

Figure 4 histogram shows that rural borrowers are more likely to fail on their loans. Rural borrowers have a 17 percent default rate on their loans (18 out of 109 customers). Borrowers from semiurban and urban property regions, on the other hand, had loan default rates of 11 percent and 13 percent, respectively.

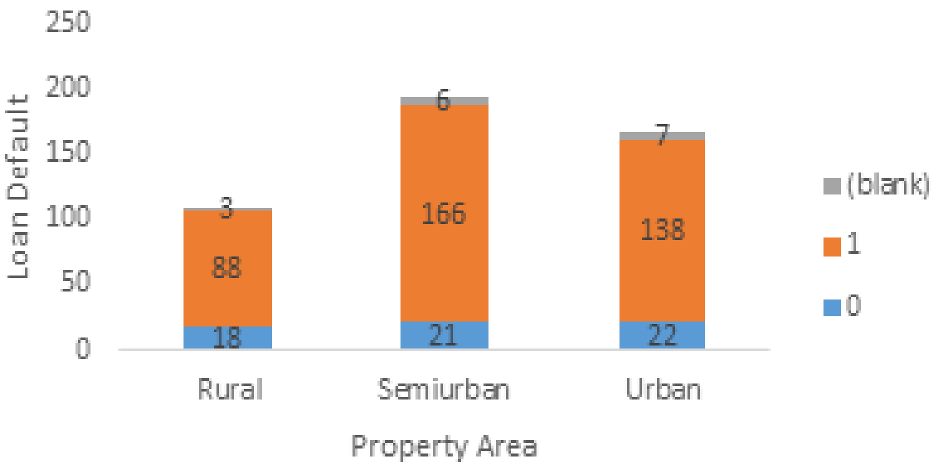
The loan default based on loan amount and period may be seen in the table in Figure 5. It has been discovered that 84 percent of consumers requesting for loans have a term of 310-409 days, implying a one-year payback period. Furthermore, one-fourth of clients who had loan periods of more than one year (18 months or more) failed on their payments (25 percent). Figure 5 also shows that if the loan period is more than 210 days, more over half of the clients (56 percent) default.

We plotted the loan default as a function of the loan applicant's income in Figure 6. Low-income borrowers are more likely to apply for a loan than those with higher incomes (93.6 percent of borrowers in the income range of 0-9999 applied for a loan). In addition, the lower-income category has a 13.2 percent loan default rate (58 out of 439



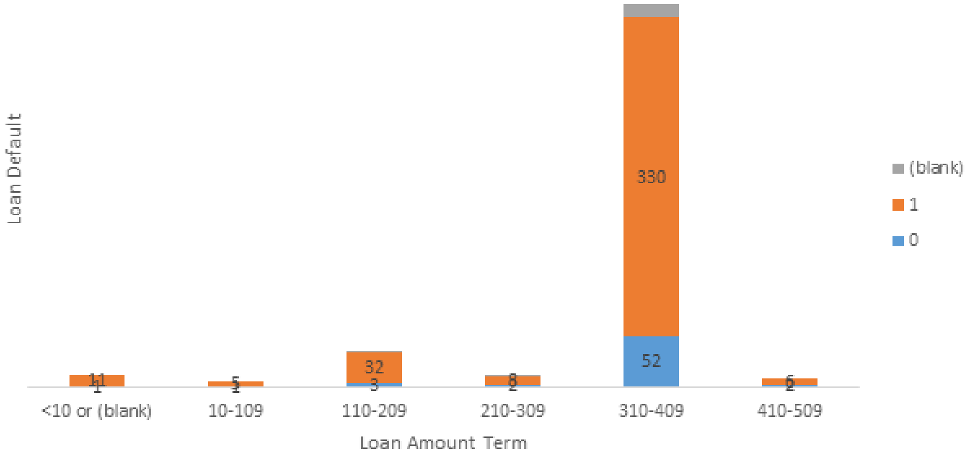
Credit History	Loan Defaults		% Loan Defaults
	0	1	
0	75	138	35%
1	61	392	13%

Figure 1. Distribution of Loan Default by Credit History.



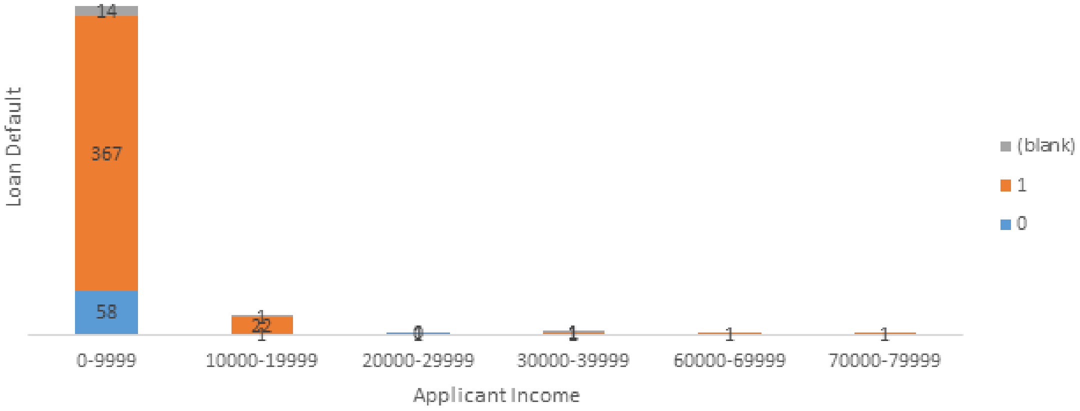
Property Area	Loan Default			% Loan Defaults
	0	1	Data not available	
Rural	18	88	3	17%
Semiurban	21	166	6	11%
Urban	22	138	7	13%

Figure 2. Distribution of Loan default based on Property Area.



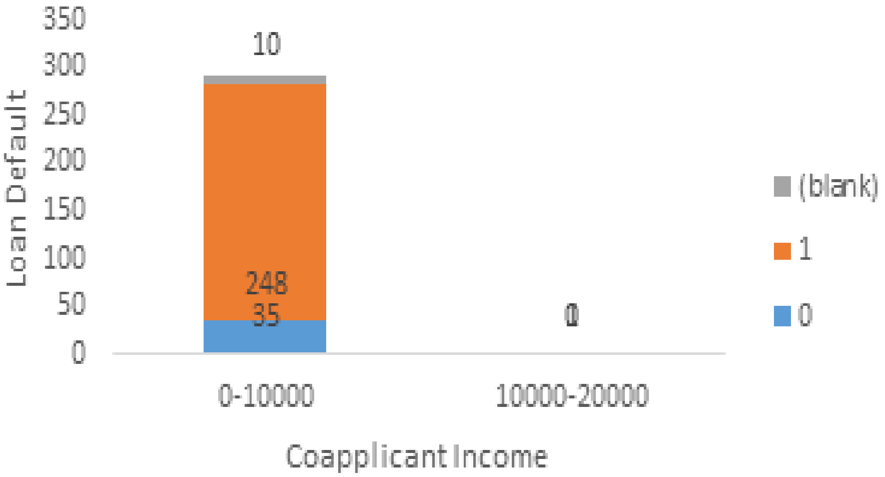
Loan Amount Term	Loan Defaults			% Loan Defaults
	0	1	blank	
<10 or (blank)	1	11	0	8%
10-109	1	5	0	17%
110-209	3	32	1	8%
210-309	2	8	1	18%
310-409	52	330	14	13%
410-509	2	6	0	25%

Figure 3. Loan Default on Loan Amount Term (Group by 100).



Applicant Income	Loan Defaults			% Loan Defaults
	0	1	blank	
0-9999	58	367	14	13.2%
10000-19999	1	22	1	4.2%
20000-29999	1	0	0	100.0%
30000-39999	1	1	1	33.3%
60000-69999	0	1	0	0.0%
70000-79999	0	1	0	0.0%

Figure 4. Distribution of loan default by applicant income.



Co-applicant Income	Loan Defaults			% Loan Defaults
	0	1	blank	
0-10000	58	367	14	12.42%
10000-20000	1	22	1	100%
20000-30000	1	0	0	100%
40000-50000	1	1	1	0%

Figure 5. Distribution of loan default by co-applicant income.

borrowers). Borrowers in the high-income segment, on the other hand, do not fail on their loans.

We showed the distribution of loan default based on co-applicant income in Figure 7. The income of a business partner or spouse is considered co-applicant income. Figure 7 demonstrates that with a co-applicant in the low-income band (0-10000), the total loan default rate is greater, with a 12.42 percent default rate. Loan default is 14.62 percent when the co-applicant is a business partner (172 out of 469) and 11.95 percent when the co-applicant is a spouse, according to the statistics (294 out of 469).

4. Data Modelling

The modeling procedure includes selecting models based on the research's various predicted models. Adaboost, k-NN, Logistic Regression, Support Vector Machines (SVM), Decision Tree, Naive Bayes, Neural Networks, and Random Forest are the study's eight distinct prediction models (RF). The quantity of data provided during training increases the prediction model's accuracy (Rácz *et al.*, 2021; Yadav *et al.*, 2021). The dataset is split into two portions, one for training and the other for testing, with 70:30 ratios (Figure 8).

- The train set contained 70% of the dataset with observations; and
- The test set contained the remaining 30% with 294 observations.

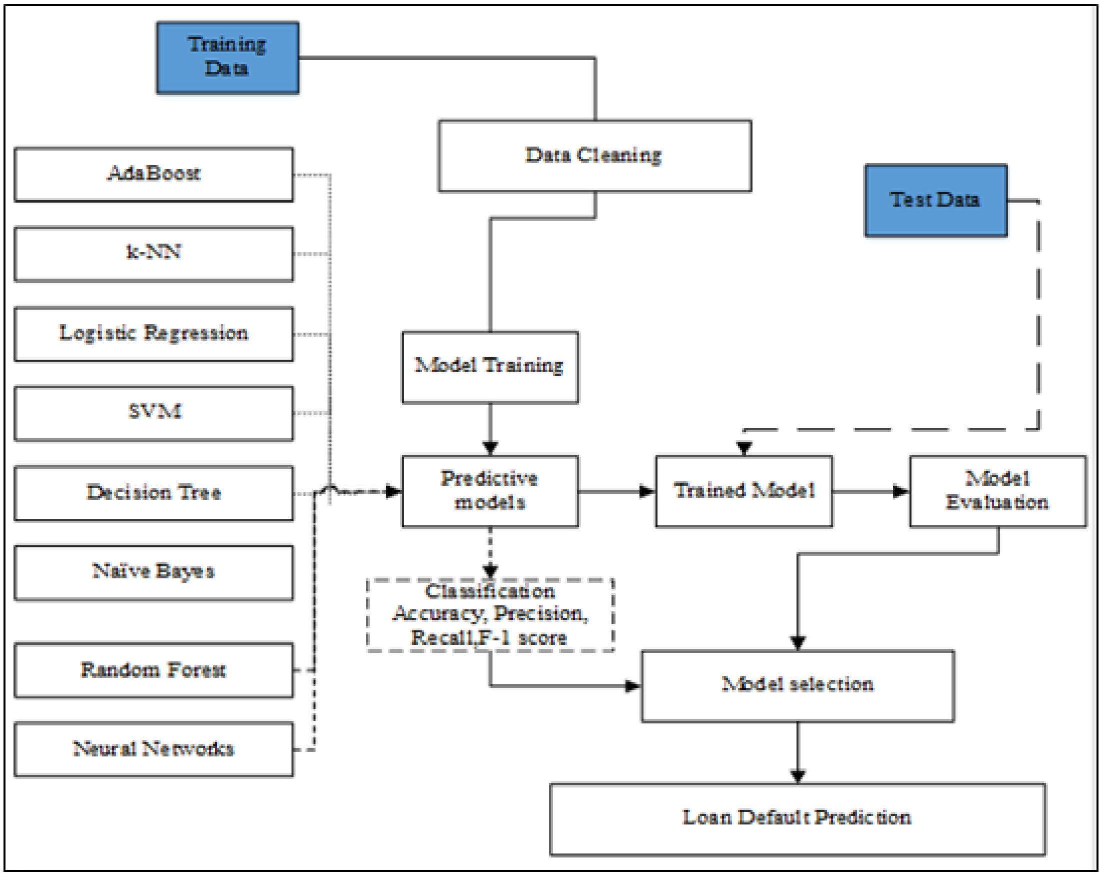


Figure 6. Data Modeling process.

5. Results

This phase assesses the predictive models' abilities using four accuracy criteria (classification accuracy, precision, recall, and F1-score) from the confusion matrix, summarized in Table 2.

As mentioned in Table 2, the AdaBoost was recognized as the best predictive model to accomplish the objective of the analysis.

Table 3 testifies the confusion matrix of the AdaBoost predictive model, which suitably classified 124 out of 195 instances.

AdaBoost predictive model obtained:

- The lowest False Positive Rate (FPR) of around 61.6% means that the model merely failed to detect 59 defaulters with the recall score of 0.384; and
- The True Positive Rate (TPR) of almost 69.74% means the model acceptably forecasts 136 out of 195 defaulters.

5.1 Feature Selection

The loan default dataset has a lot of characteristics, and it was discovered during feature selection that not all of these factors are meaningful all of the time. Adding extra attributes to a model during training diminishes overall accuracy, increases complexity, limits generalization capabilities, and biases the model. We performed a feature selection for the dataset and discovered that Credit History has a significant role in detecting loan default. Figure 9 displays the list of factors in sequence of its prominence to loan default.

Table 3. Evaluation metrics

	Classification Accuracy	Precision	Recall	F-1
AdaBoost	1	1	1	1
k-NN	0.728	0.628	0.349	0.448
Logistic Regression	0.699	0.343	0.557	0.248
SVM	0.658	0.408	0.453	0.372
Decision Tree	0.921	0.877	0.869	0.885
Naïve Bayes	0.713	0.405	0.593	0.307
Random Forest	0.932	0.938	0.839	0.932
Neural Network	0.792	0.590	0.786	0.472

Table 4. AdaBoost predictive model outcomes for test data

AdaBoost	Classification accuracy	Precision	Recall	F-1 score
	0.592	0.392	0.384	0.388

After that, we split the original dataset into two halves, one with Credit History 1 and the other without, and ran the prediction model again to see whether it could detect loan default with more accuracy. The purpose is to see which is better by training the model under two different scenarios: one in which the model is trained when the client hasn't defaulted on the loan (Credit History 1), and another in which the model is trained when the client has defaulted on the loan (Credit History 0).

5.1.1 Outcomes for Credit_History

The feature selection strategy is used to reduce over fitting in these classification algorithms' curse of dimensionality. The four performance measures are used to discuss the performance of these eight prediction models once again. The data indicate that Credit History 0 produces superior outcomes. Figure 10 shows the results for both Credit History settings.

It is observed that when we divide the overall dataset based on feature selection of Credit_History, AdaBoost algorithm still gives the best results with 100% correct observations. However, when we measure the outcomes on the four mentioned performance metrics, the outcomes is better, when we select the dataset with Credit_History 1 (Table 4).

The comparative study is mentioned as:

- The lowest false positive rate (FPR) of around 55.% and 57.5% respectively for Credit_History 1 and 0 with the recall rate of 0.431 and 0.343; and
- The true positive rate (TPR) of almost 68.7% and 46.6% for Credit_History 1 and 0.

The findings clearly suggest that the dataset for Credit History 1 produces superior results, with 20 percent more true outcomes. The details are mentioned in table 5.

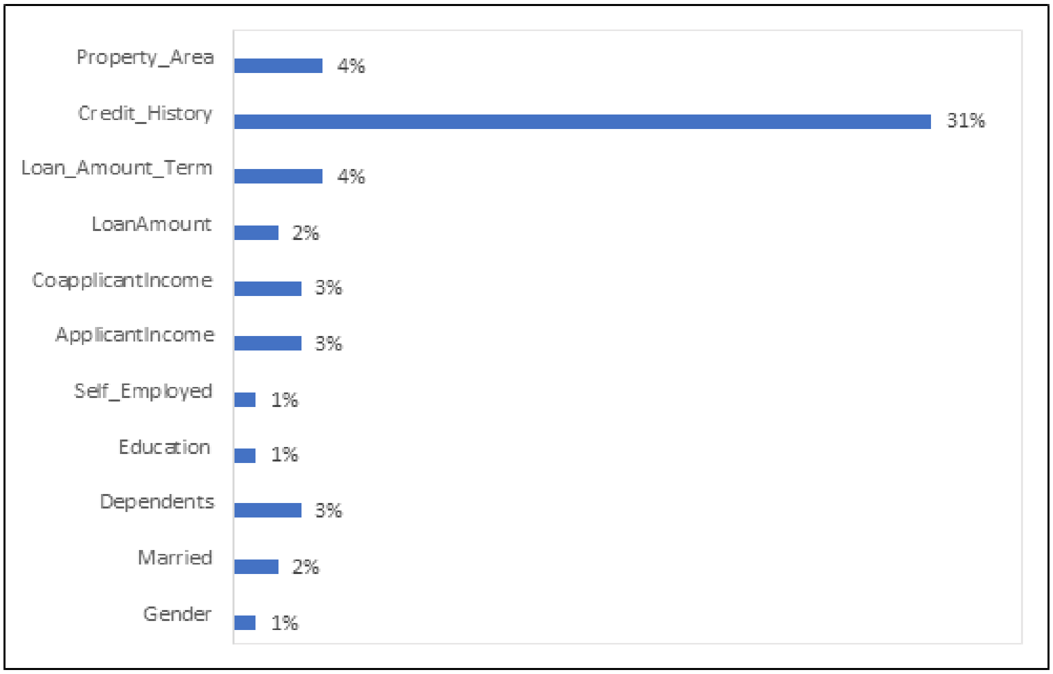


Figure 7. Feature Selection.

	Classification Accuracy		Precision		Recall		F-1	
	Credit_ History 1	Credit_ History 0	Credit_ History 1	Credit_ History 0	Credit_ History 1	Credit_ History 0	Credit_ History 1	Credit_ History 0
AdaBoost	1	1	1	1	1	1	1	1
k-NN	0.777	0.699	0.692	0.702	0.261	0.787	0.379	0.742
Logistic Regression	0.738	0.566	0.444	0.587	0.029	0.720	0.054	0.647
SVM	0.740	0.743	0.500	0.733	0.138	0.840	0.216	0.783
Decision Tree	0.926	0.926	0.846	0.911	0.877	0.960	0.861	0.783
Naïve Bayes	0.736	0.699	0.462	0.750	0.087	0.680	0.146	0.713
Random Forest	0.940	0.949	0.949	0.947	0.812	0.960	0.875	0.954
Neural Network	0.823	0.912	0.855	0.932	0.384	0.907	0.530	0.919

Figure 8. Feature Selection outcomes for Credit_History 0 and 1.

Table 5. AdaBoost predictive model outcomes for test data

	Classification accuracy	Precision	Recall	F-1 score
AdaBoost (Credit_ History 1)	0.605	0.419	0.444	0.431
AdaBoost (Credit_ History 0)	0.452	0.288	0.424	0.343

6. Conclusion

The conclusions obtained are related to those in the literature, and their managerial implications are analyzed. In this paper, the predictive model approach is used to study the certainty of loan default by the defaulter to forecast creditworthiness.

Eight predictive models were examined to determine the best match for prediction for Research Question 1 (RQ1), Can we employ several prediction algorithms to forecast loan default? The experiment answered the first research objective (RO 1), and we discovered that the AdaBoost predictive model produces the best results, scoring 100% on all four performance indicators. The model produced (CA=0.592, precision=0.877, recall=0.384, and F-1 score=0.388) and the critical factors that stimulate consumers' creditworthiness for the test dataset. For Research Question 2 (RQ2), Is the feature selection better at predicting loan default, we obtained that Credit_History is the most important variable that predicts the loan default with 31%. For the Research Question 2 (RQ2), we partitioned the dataset into two portions, with Credit History 1 and 0. We tested all eight prediction models again and found that AdaBoost provided the best results with 100% accuracy for both datasets. Credit History 1 findings, on the other hand, are better (CA=0.605, precision=0.419, recall=0.444, and F-1 score=0.431). The prospective future work for this study will be a further development of the model by deepening analysis on variables used in the models. Data available have restrictions in terms of specifics of defaulters and timeline, which stipulates that the behaviour of defaulters outside the timeline may not track the same outline.

7. References

Alojail, M., & Bhatia, S. (2020). A Novel Technique for Behavioral Analytics Using Ensemble Learning Algorithms in E-Commerce. *IEEE Access*, *8*, 150072–150080. https://doi.org/10.1109/ACCESS.2020.3016419

Al-qerem, A., Al-Naymat, G., & Alhasan, M. (2019). Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection. *2019 International Arab Conference on Information Technology (ACIT)*, 235–240. https://doi.org/10.1109/ACIT47987.2019.8991084

Arutjothi, G., & Senthamarai, C. (2017). Prediction of loan status in commercial bank using machine learning classifier. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 416–419. https://doi.org/10.1109/ISS1.2017.8389442

Blöchlinger, A., & Leippold, M. (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, *30*(3), 851–873. https://doi.org/10.1016/j.jbankfin.2005.07.014

Chopra, Y., Subramanian, K., & Tantri, P. L. (2020). Bank Cleanups, Capitalization, and Lending: Evidence from India. *The Review of Financial Studies*, *hhaa119*. https://doi.org/10.1093/rfs/hhaa119

Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, *91*, 106263. https://doi.org/10.1016/j.asoc.2020.106263

Einav, L., Jenkins, M., & Levin, J. (2013). The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, *44*(2), 249–274. https://doi.org/10.1111/1756-2171.12019

Ghosh, S. (2021). *Wilful defaults took a turn for the worse in Apr-Dec amid pandemic*. Mint. https://www.livemint.com/industry/banking/wilful-defaults-took-a-turn-for-the-worse-in-apr-dec-amid-pandemic-11619030170683.html

Hassan, A. K. I., & Abraham, A. (2013). Modeling consumer loan default prediction using neural netware. *2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)*, 239–243. https://doi.org/10.1109/ICCEEE.2013.6633940

Jia, H. (2018, April 10). *Credit Scoring with Machine Learning*. Medium. https://medium.com/henry-jia/how-to-score-your-credit-1c08dd73e2ed

Krichene, A. (2017). Using a naive Bayesian classifier methodology for loan risk assessment:Evidence from a Tunisian commercial bank. *Journal of Economics, Finance and Administrative Science*, *22*(42), 3–24. https://doi.org/10.1108/JEFAS-02-2017-0039

Microsoft. (2020). *What is the Team Data Science Process?* https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

Moneycontrol. (2020). *How Machine Learning Is Reducing Loan Defaults And Easing Debt Recovery*. Moneycontrol. https://www.moneycontrol.com/news/technology/fintech-how-machine-learning-is-reducing-loan-defaults-and-easing-debt-recovery-4798461.html

Press, G. (2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. Forbes. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

PTI. (2021). *HDFC, ICICI Bank, SBI, among top-10 lenders in 2020; Google Pay, PhonePe top wallets: Report - Times of India*. The Times of India. https://timesofindia.indiatimes.com/business/india-business/hdfc-icici-bank-sbi-among-top-10-lenders-in-2020-google-pay-phonepe-top-wallets-report/articleshow/79844080.cms

RBI. (2021). *Need list of top the 10 Banks with lowest NPA*. https://tradingqna.com/t/need-list-of-top-the-10-banks-with-lowest-npa/100231

Reddy, M. V. J., & Kavitha, B. (2010). Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis. *2010 International Conference on Signal Acquisition and Processing*, 274–277. https://doi.org/10.1109/ICSAP.2010.10

Redman, T. C. (2018, April 2). If Your Data Is Bad, Your Machine Learning Tools Are Useless. *Harvard Business Review*. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless

Shoumo, S. Z. H., Dhruba, M. I. M., Hossain, S., Ghani, N. H., Arif, H., & Islam, S. (2019). Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking. *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2023–2028. https://doi.org/10.1109/TENCON.2019.8929527

Shukla, S. (2021). *Payment defaults rise 50% in May for NBFCs—The Economic Times*. https://economictimes.indiatimes.com/industry/banking/finance/payment-defaults-rise-50-in-may-for-nbfcs/articleshow/82725399.cms?from=mdr

Statista. (2021). *India: Gross non-performing loan ratio 2021*. Statista. https://www.statista.com/statistics/1013267/non-performing-loan-ratio-scheduled-commercial-banks-india/

Sunitha, T., M, C., M, R., G, S. sri, T. V.s., J., & A, T. (2021). *Predicting the Loan Status using Logistic Regression and Binary Tree* (SSRN Scholarly Paper ID 3769854). Social Science Research Network. https://doi.org/10.2139/ssrn.3769854

Wu, M., Huang, Y., & Duan, J. (2019). Investigations on Classification Methods for Loan Application Based on Machine Learning. *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, 1–6. https://doi.org/10.1109/ICMLC48188.2019.8949252

Zhao, S. (2021, March 8). *Predicting Loan Defaults Using Logistic Regression*. Medium. https://selenaezhao.medium.com/predicting-loan-defaults-using-logistic-regression-71b7482a8cf7

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503–513. https://doi.org/10.1016/j.procs.2019.12.017