IRJCS: International Research Journal of Computer Science
Volume 12, Issue 04, April 2025
ISSN: 2393-9842
https://www.irjcs.com/archives

Predicting Loan Repayment: A Machine Learning Approach

Prashant Tiwari, Nitesh Kumar
Student, Department of Information Technology,
Greater Noida Institute of Technology,
Greater Noida (Engg. Institute), Uttar Pradesh, India
prasaaant21122002@gmail.com, nitesh98016.nk@gmail.com
Dr.Vikas Singhal
Professor & Head, Department of Information Technology,
Greater Noida Institute of Technology (Engg. Institute),
Greater Noida, Uttar Pradesh, India
hodit@gmail.com

Publication History
Manuscript Reference: IRJCS/RS/Vol.12/Issue04/APCS10082 | Research Article | Open Access | Double-Blind Peer Reviewed Article ID: IRJCS/RS/Vol.12/Issue04/APCS10082
Received: 06, April 2025, Revised: 12, April 2025 Accepted: 21 April 2025 Published Online: 30 April 2025
http://www.irjcs.com/volumes/Vol12/iss-04/02.APCS10082.pdf
Article Citation: Prashant,Nitesh,Dr.Vikas(2025).Predicting Loan Repayment: A Machine Learning Approach. IRJCS: International Research Journal of Computer Science, Volume 12, Issue 04 of 2025 pages 125-130
doi:> https://doi.org/10.26562/irjcs.2025.v1204.02
BibTeX Dr.Vikas@2025Predicting

Abstract: In order to reduce the risk of loan payment default, banks must assess and forecast the loaners' capacity to repay. As a result, a system has been established by the banks to handle the loan application according to the loaner's circumstances, including work status, credit history, etc. However, some loaners, including students or those without credit histories, may not be able to be evaluated for repayment capacity using the present rating approach. We trained a variety of machine learning models on the Home Credit Default Risk Kaggle dataset and assessed the significance of each feature in order to accurately determine the repayment capacity of every group of individuals. Next, using the features relevance score as a guide, we examine and choose the most recognizable characteristics to forecast the loaner's capacity for payback.

Keywords: credit, loan, home credit, loan payment defaulters, loan capacity.

I. INTRODUCTION

One of the most significant goods offered by financial institutions is the loan. Every institution is working to develop effective business plans to encourage more clients to apply for their loans. Nevertheless, some clients are unable to repay the loan once their application has been accepted. As a result, before granting a loan, many financial institutions consider a number of factors. It is challenging to ascertain if a particular borrower will repay the loan in full or result in charges. There will be less interest to collect if the lender is overly stringent since fewer loans will be authorized. However, they wind up authorizing loans that default if they are overlyenient. Several machine learning models are used in this work to assess loan habits. One of the most significant goods offered by financial institutions is the loan. Every institution is working to develop effective business plans to encourage more clients to apply for their loans. Nevertheless, some clients are unable to repay the loan once their application has been accepted. As a result, before granting a loan, many financial institutions consider a number of factors. It is challenging to ascertain if a particular borrower will repay the loan in full or result in charges.

There will be less interest to collect if the lender is overly stringent since fewer loans will be authorized. However, they wind up authorizing loans that default if they are overlylenient. Several machine learning models are used in this work to assess loan habits. Many are having trouble obtaining loans from reliable sources, such banks, because of inadequate credit histories. These individuals, who are typically students or jobless adults, may lack the expertise necessary to support the legitimacy of the unnamed lenders. These borrowers may be exploited by dishonest lenders that charge ex orbit ant interest rates or include unstated clauses in the contract. There are several additional methods to gauge or forecast the borrower's capacity for repayment besides relying just on their credit score. For instance, since a working adult has more steady wages and cash flow, employment may have a significant impact on a person's ability to repay. For this study, we selected the Home Credit Default Risk dataset from Kaggle.com. This is open 308K anonymous customers with 122 distinct attributes are included in the dataset. By examining the relationship between these characteristics considering customer repayment capacity, our algorithm can assist lenders in assessing borrowers from additional angles while also assisting borrowers particularly those with insufficient credit histories in locating reliable lenders, creating a win-win scenario.

II. LITERATURE REVIEW

[1] One of the most significant goods offered by financial institutions is the loan. Every institution is working to develop effective business plans to encourage more clients to apply for their loans. Nevertheless, some clients are unable to repay the debt following their applications have been accepted. As a result, while granting a loan, many financial institutions consider a number of factors .It might be challenging to ascertain if a particular borrower will repay the loan in full or will result in charges for partial repayment. There will be less interest to collect if the lender is overly stringent since fewer loans will be authorized.

[2] In order to reduce the risk of loan payment default, banks must assess and forecast the loaners' capacity to repay. Because of this, banks have developed a mechanism to handle loan requests according to the loaners' status, including job status, credit history, etc. However, some loaners, including students or those without credit histories, may not be able to be evaluated for repayment capacity using the presentrating approach. We trained a variety of machine learning models on the Home Credit Default Risk Kaggle dataset and assessed the significance of each feature in order to accurately determine the repayment capacity of every group of individuals.

[3] In today's world, borrowing from financial institutions has become normal. Every day, a large number of people apply for loans for a variety of reasons. However, not all of these applicants are accepted, and not all of them are trustworthy. Every year, a sizable portion of bank loans are not repaid, causing the bank to suffer huge losses. The decision to authorize a loan carries a lot of risk. Thus, this project's objective is to collect credit data from several sources and then extract important information using a range of machine learning approaches. Businesses can use this model to determine whether to grant or deny consumer loan requests.

[4] A profitable niche market or micro market may arise if a model is able to find creditworthy clients who conventional credit scores did not uncover while lowering their risk of loan default , increasing the financial institution's or investor's profit margin. Despite the seeming benefits of having more clients, it's crucial to exercise caution when lending to those who are likely to default on their payments. As a result, meticulous evaluation criteria and a cautious attitude were maintained throughout the project. Whether or not an investor should lend is a binary classification problem that affects loan default prediction. For this issue, an appropriate model is logistic regression.

[5] The distinctive features of bank loans develop naturally to improve effectiveness. In a renegotiation model, moral hazard might occur on both sides of the relationship between a borrower and a lender. Renegotiated interest rates on the loan do not have to be monotonous in firm risk since firm risk is endogenous. The debt's original conditions are designed to effectively balance negotiating strength in a subsequent renegotiation, not to price default risk. Initial transfers from the borrower to the bank or from the bank to the borrower may be a part of nonlinear loan pricing. In its magazine, The Review of Financial Studies, Oxford University Press published an article on behalf of the Society for Financial Studies. Financial institutions have to consider several variables while approving a loan. To determine whether a given borrower can fully pay off the loan or cause it to be charged off is difficult. If the lender is too strict, fewer loans get approved, which means there is less interest to collect. But if they are too lax, they end up approving loans that default. [6] Machine learning can help us predict which loans will be charged off. Through various machine learning methods this project gives the predictive patterns in the financial data that can be used to ensure the clients that are capable of repayment are not rejected.

III. METHODOLOGY

The project's objective is to forecast the borrowers' capacity to repay using variables other than their credit history. It may be expressed as a two-class classification issue. The techniques we employed to pre-process the data and the machine learning algorithms we will employ to address the issue are described in the section that follows.

Gender	mean	median	min	max
Female	105912.26	111194.5	5985	150000
Male	105620.37	111209.0	6221	150000
Other	105265.06	109957.5	6891	150000

Table1: Based on Gender

A. Data Pre-Processing

We apply several data pre-processing techniques to our dataset before to its use for training and testing because of the intricacy of our raw data.

Employment Profile	Credit Score
Freelancer	565.9
Salaried	594.62
Self-Employed	580.05
Student	553.7
Unemployed	553.65

Table 2: Credit Score

- **Concatenation of features:** The features in the original data set originate from several sources. Displays a concise synopsis of the data files.

ISSN: 2393-9842
https://www.irjcs.com/archives

- Concatenating all of the characteristics together is the initial stage in our data pre-processing procedure. Using each borrower's unique ID number is the method to integrate all the features. For instance, SKIDCURR may be used to link the entries in the bureau.csv file with the equivalent rows in the application train.csv. To create the training and testing sets with the best possible use of the available data, we concatenated all of the characteristics together. Each data point has a total of 217 features after feature concatenation.

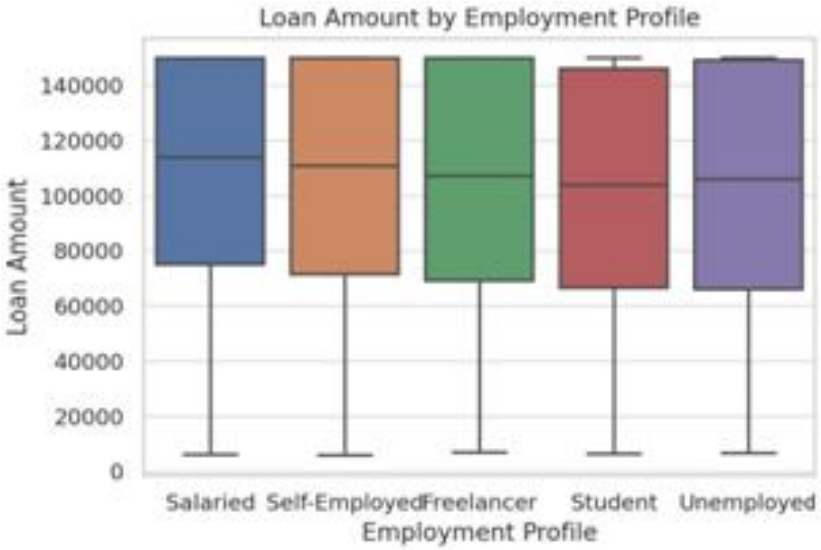

Fig 1: Loan Amount by Employment Profile

- **Feature Encoding and Normalization:** Our features are available in a number of formats, including Boolean values, unbounded integers, floating numbers between 0 and 1, and phrase strings. We face a problem since the features aren't immediately used for training. We factorize these characteristics using label encoding, which maps the string values to category values, each of which is represented by an integer, in order to avoid classification biases towards certain features. Nevertheless, there are too many categories for some aspects, making label encoding challenging. The single feature is then extended into several features using one- hot encoding for these kinds of features, with each expanded feature having values that are restricted to just 0 and 1.

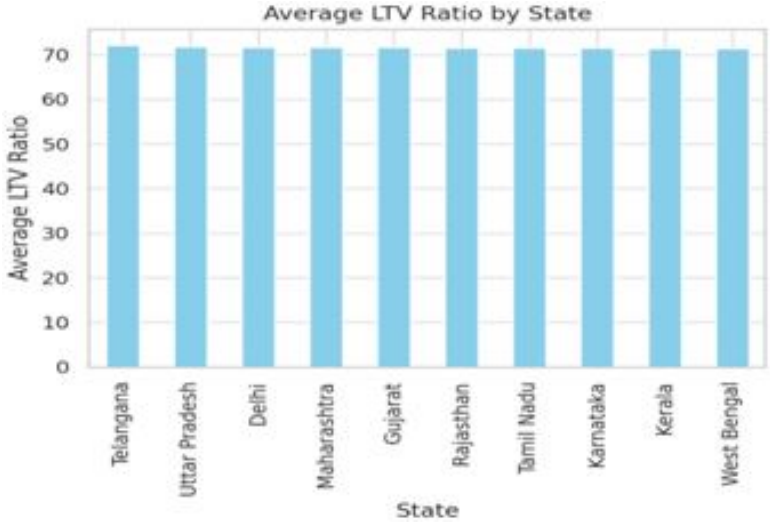

Fig 2: Average LTV Ratio By Ratio

- **Replacement of Invalid/Empty Entries:** [12] In addition to feature processing, invalid and empty entries are another issue that hinders adequate machine learning algorithm training. There is a discernible quantity of data in the dataset that has either empty or incorrect items (like Nan, for example, an extremely big integer). [10] Taking the mean of the feature values and using into fill in the blanks is one approach we used. Based on the proportion of erroneous values in each column, we employ an approach to eliminate rows or columns from entries that include a high number of invalid values. We establish a threshold number, which in our instance is 30% at first.

 https://doi.org/10.26562/irjcs

ISSN: 2393-9842
https://www.irjcs.com/archives

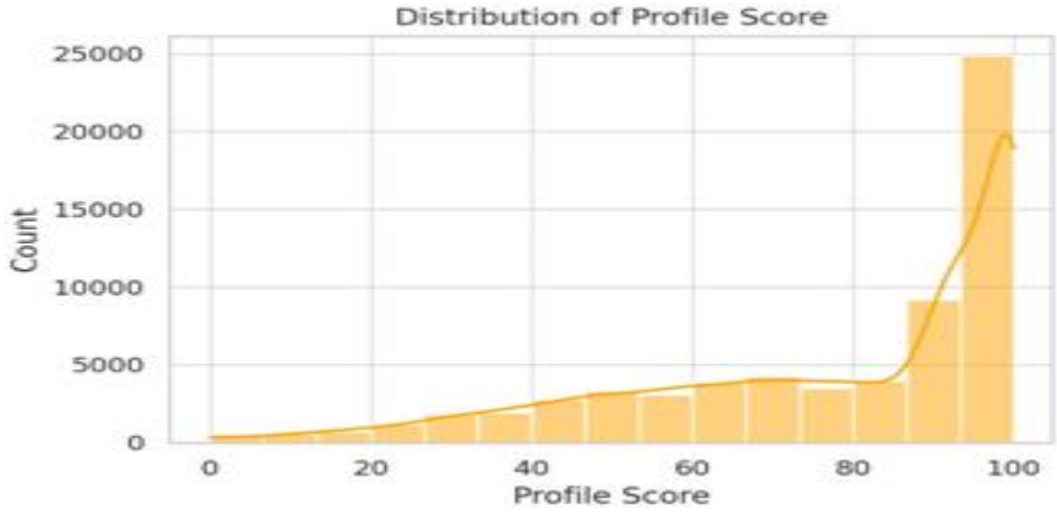

Fig 3: Distribution of Profile Score

- **Polynomial feature transformation:** We also applied polynomial transformation to our feature values to incorporate a polynomial combination of the features in order to maximize the performance of our linear classifiers.

Existing Customer	Loan Tenure
No	86.24
Yes	210.08

Table 3: Existing Customer

B. MACHINE LEARNING TECHNIQUES

In this part, we experiment with several machine learning models to estimate the ability of borrowers to repay loans. Among the machine learning algorithms are neural networks, logistic regression, random forests and Naive Bayes. A few of the algorithms are ones we studied in class, while others we looked up online that would work well with the dataset. We outline the rationale behind our decision to employ these algorithms in this section.

Logistic Regression: The fact that logistic regression does not impose strict assumptions on the distribution of the underlying data makes it a useful baseline model to test on machine learning challenges. Logistic regression is a fantastic model to use as our initial approach for our classification challenge.

Random Forests: Random Forests typically work effectively with datasets that are unbalanced. Using a random forest classifier on our dataset also allows us an understandable method to see our features by displaying the relative relevance of each feature, which helps us understand the variables influencing a borrower's capacity to repay a loan. Additionally, we aim to determine whether adding a certain amount of unpredictability to the classification issue might increase the precision of our findings.

Naive Bayes: Naive Bayes applies the"naive" assumption on the features, which indicates that the features are conditionally independent of each other given the class variable. In contrast to the other two algorithms, Naive Bayes is a generative approach, as we discovered in class. The performance of this generative approach and the fundamental discriminative approaches may then be compared.

Neural Networks: Multi-layer perception, often known as neural networks, are among the most widely used techniques for classification issues. It is a function approximate that, thanks to the non-linearity introduced by the activation functions, can classify data with non- linear decision boundaries in addition to modeling the distribution of linear data. In our project, we carefully choose hyper parameters such the number of layers and neurons in order to fit a multi-layer perception.

IV. APPLICATION

Credit Risk Assessment in Banking: By examining past loan data, credit scores, income levels, and other pertinent financial indicators, machine learning models allow banks to precisely evaluate the credit risk of loan applicants. Conventional credit rating techniques frequently ignore subtle patterns in data and are inflexible. Banks may use predictive modeling to automate and customize risk assessment, identifying high-risk candidates for more scrutiny while granting loans to low-risk applicants. Better financial decisions are made as a result, and defaults are reduced and profitability is increased. Additionally, models that can be continually trained on fresh data, like as logistic regression, random forests, and gradient boosting, guarantee dynamic adaptation to shifting borrower behaviors and economic trends.

Financial Inclusion and Microfinance: In underprivileged regions where official credit records may be missing, machine learning has the potential to revolutionize financial inclusion. Predictive models can evaluate the ability of people who are not often included in standard credit systems to repay loans by using alternative data, such as social media activity, utility payments, and mobile phone usage.

 https://doi.org/10.26562/irjcs

This empowers small firms and entrepreneurs by allowing microfinance institutions to safely offer modest loans. As these models improve, they not only lower default rates but also increase confidence in the official financial system. In the end, this helps the unbanked obtain credit, which promotes economic growth.

Loan Portfolio Management: Accurately predicting repayment patterns is essential for financial institutions managing a sizable loan portfolio in order to preserve liquidity and lower non-performing assets (NPAs). By classifying debtors into risk groups and forecasting the probability of future defaults, machine learning models are useful. Institutions are able to create personalized repayment programs, modify credit limits, and optimize interest rates thanks to this predictive information. Additionally, proactive client interaction or debt recovery methods are made possible by real-time monitoring via ML dashboards, which guarantees early warning systems for delinquency. Portfolio managers may ensure regulatory compliance and sustainable development by keeping a healthy loan book via the use of clustering algorithms and predictive scoring.

Fraud Detection and Anomaly Monitoring: In addition to evaluating borrower reliability, predictive algorithms may identify fraudulent activity during the loan application and repayment processes. Algorithms that use machine learning can spot minute irregularities or discrepancies in application data that could indicate fraud. In order to identify suspect patterns, such as counterfeit papers, identity discrepancies, or unusual transaction activity, techniques including ensemble methods, neural networks, and outlier detection may examine hundreds of data in real time. By putting these processes in place, financial institutions may avoid suffering large financial losses as well as harm to their image. Furthermore, these technologies improve the overall resilience of lending platforms by continually learning from novel fraud behaviors.

V. FUTURE DIRECTION

There are a number of improvements we might apply to this study in the future. For instance, the exploratory data analysis does not take the outlier problem into account. The prediction model's findings will not be as reliable if the dataset contains outliers. Furthermore, for forecasting the loan's repayment status, the deep learning algorithm approach ought to be used. Additionally, we would have more training samples if our dataset were larger. It might solve the large variance issue and strengthen the validity of our findings. We showed how to forecast loan repayment capabilities using machine learning techniques on a very difficult dataset. We demonstrated that data pre-processing, a thoughtful selection of dataset balancing strategies and classification algorithms are all crucial for achieving optimal performance. On our dataset, neural networks and logistic regression perform admirably, and k-means is also useful. To further enhance model performance on this crucial prediction job, we intend to investigate increasingly complex learning algorithms and dimension reduction strategies in the future.

VI. CONCLUSION

These days, a lot of individuals request for loans for a variety of reasons as the lending industry grows in popularity. Nonetheless, there are instances in which borrowers fail to pay back the majority of their bank loans, leading to significant financial losses. Therefore, it would significantly reduce the financial loss if there was a method to effectively categorize the loaners beforehand. Prior to doing exploratory data analysis and feature engineering, the dataset in this study was cleansed. It was discussed how to handle both missing values and unbalanced datasets. Next, in order to determine if the borrower might repay the loan, we suggest four machine learning models: Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbours. When adjusting parameters, the Grid Search Cross Validation and Randomized Search Cross Validation techniques are used in different scenarios. Experiments revealed that the random forest model is the model that fits the dataset the best in terms of accuracy, while the logistic regression model with L2 penalty has the greatest AUC score. We anticipated that borrowers with better FICO scores and yearly incomes would be more likely to return the loan in full; moreover, borrowers with lower interest rates and smaller payments would also be more inclined to do so.

REFERENCES

1. "Home Credit Default Risk."Kaggle, https://www.kaggle.com/c/homecredit-default-risk/data "Smart parking system happiest minds "Aditya Basu,2014
2. Happiest Minds https://www.happiestminds.com/Insights/smart-parking/
3. Abhishek Bhagat et al. Predicting Loan Defaults using Machine Learning Techniques. PhD thesis, California State University, Northridge, 2018.
4. Hongri Jia Bank Loan Default Prediction with Machine Learning.pp.137163, Apr 10, 2018.
5. Gorton, Gary, and James Kahn. "The design of bank loan contracts."The Review of Financial Studies 13, no. 2 (2000): 331-364.
6. Kolo, Brian, Thomas Rickett McGraw, and Dathan Gaskill. "Systems and methods for using data metrics for credit score analysis." U.S. Patent Application 13/456,532, filed November 1, 2012.
7. Laurens vander Maaten, Geoffrey Hinton. "Visualizing Data using t-SNE." Journal of Machine Learning Research, 2008.
8. Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, 2007.
9. Chawla, Nitesh V., et al."SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

10. Ke, Guolin, et al."Lightgbm: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems.2017.
11. Liaw, Andy, and Matthew Wiener. "Classification and regression by random Forest." R news 2.3 (2002): 18-22.
12. Mierzewski, Michael B., Christopher L. Allen, Jeremy W. Hochberg, and Kevin Hall. "CFPB Finalizes Ability-to-Repay and Qualified Mort gage Rule." Banking LJ 130 (2013): 611.
13. Ivashina, Victoria, and David Scharfstein. "Bank lending during the financial crisis of 2008." Journal of Financial economics 97, no. 3(2010):319-338.
14. Murdock, C.W., 2011. The Dodd-Frank Wall Street Reform and Consumer Protection Act: What Caused the Financial Crisis and Will Dodd-Frank Prevent Future Crises. SMUL Rev., 64, p.1243.
15. Wongnaa, C. A., and Dadson Awunyo-Vitor. "Factors affecting loan repayment performance among yam farmers in the Sene District,Ghana." Agris on-line Papers in Economics and Informatics 5, no.665 2016-44943 (2013): 111-122.

