IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

Performance Analysis of Machine Learning models for Credit Delinquency Prediction

Muktha Priya K S[1] and Dr. Sunitha G P[2]
4th Semester, Department of MCA[1]
Associate Professor, Department of MCA[2]
J N N College of Engineering, Shimoga, India, Karnataka, India

Abstract: *The activities of lending loan to those who are in the financial distress are implemented by the financial institutions like bank. The area which is becoming significant in the analysis of monetary risk is credit risk assessment. The distinct dataset's credit risk is considered using multiple machine learning methods. The review of all credit risk databases should be used to draw a conclusion on when to grant the loan to the particular customer or disapprove the individual's application, which is a complex job. The paper evaluates an in-depth inspection of the individual's credit dataset or check by the consumer. This study investigated various risk assessment methodologies that are used in the evaluation of credit datasets. Using trained ml algorithms, it is possible to find correlations between consumer preferences and characterize them for early action.*

Keywords: Credit risk, Loan, Machine learning, Algorithms, Delinquency, Risk Prediction

I. INTRODUCTION

Financial Institutions, especially banks are mainly building on individual and commercial loans for their evolving business. Banks are providing various other products like insurance, fixed deposit, recurring deposit etc. However loans remain to be central business for the major profit for the banks[1]. The banks get their majority of capital through lending loans. If any customer fails to repay the loan bank sustain majority of losses. It is imperative for financial firms to discern between "good" and "bad" consumers especially on the loan taken out by the applicant. In the lending process "good" consumer is considered to someone who has a low probability of default and bank will have surety to issue the loan. A "bad" consumer is someone who has high chances of default. A determined proportion of the loans made by the financial institutions can end up in default.

Banks are a major player in the market economy. They decide who gets all the money and under what constraints, and their decisions can make or shatter an investment. Individuals and businesses require cooperation in the market and society to function [4]. To choose whether to lend, banks utilize a key risk management algorithm that forecasts the default probability. Credit reports are strengthened by determining the odds of someone experiencing financial woes during next two years.

Based on certain criteria such as monthly earnings, age, the number of open credit lines and mortgages, and so on, we used multiple different classifiers to estimate whether someone may encounter budget woes in the next 2 years. A review of the reliability of the implemented models is also carried forth. This report also outlines the data purification processes used. When lending money, the model is used to assist make the best economic choices.

II. LITERATURE REVIEW

Vidhi Khanduja et.al analyzed the performance of various machine learning algorithm in [1]. It is observed that their proposed method is LDA outperforms PCA. Li et al., contribution's in [7] analyzed data from Credit Unions in the United States to execute a systematic similar study between five base - line models and ensemble algorithms. Also it is explored that Varsha Aithal et al., implemented machine learning techniques on German credit data in [2] and analyzed the same. Huichao Mi employs a set of new of machine learning algorithms in their work [8] to analyze the impact of machine learning algorithms in Small and medium - sized credit risk assessment.

In the paper [3] Trilok Pandey et al., discussed about various types of classifiers and ensemble classifiers using German data set. S. Sathya Bama et al., in their work [4] predicted that identifying the default accounts in advance based on

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

ensemble learning and artificial intelligence techniques. In paper [5] S.J. Shiv et al., have compared the performance of three machine learning algorithms in determining bank default risk.

III. PROPOSED METHODOLOGY

The Machine Learning models Logistic Regression, Naïve Bayes and Random Forest are applied to predict the credit delinquency.

The general description of Machine Learning algorithms and approach of algorithm flow to the given case study is described in the following section

3.1 Logistic Regression

A logistic regression output with parameter estimates can be patterned using the quantitative approach of regression models.

An equation between attributes and label is created with the help of logistic function also known as sigmoid function. Consider a model with a1, a2, a3,... ak features. Let Y represent the binary output, which can take the values 0 or 1.

We can denote p as the likelihood of Y = 1 as p = P(Y=1).

The mathematical relationship between these variables is,

$$\ln(\frac{p}{1-p}) = b0 + b1a1 + b2a2 + b3a3 + \ldots.. + bkak$$

In this case, the term p/(1-p) is known as the odds and depicts the potential of the occurrence. Thus, the log odds are stated as ln(p/(1-p)) and are used to simply convert a probability between 0 and 1 to a range between (,+). The factors (or values) that will be assessed through training are indicated by the symbols b0, b1, b2...

To obtain the value of p, equation is simplified and it results in Sigmoid function. The function is given by

$$S(x) = \frac{1}{1 + e^{-x}}$$

The Logistic Regression model is applied on the proposed data set on its features. The data set description is provided in Table 2 and in the Table 1 the features used and the weightage value obtained is recorded.

Predictions are made using the above-mentioned equation. Before that, we'll train our model to find the values of our parameters b0, b1, b2... that result in the least amount of error.

In the Logistic Regression, 7 Features are used which are listed in the below table 1.

Table 1: Features or Attributes used for Logistic Regression

Features or Attributes	Weights or parameters in selected model
Revolving Utilization Of Unsecured Lines	12.0149
Age	-3.11605
Debt Ratio	2.20085
Monthly Income	-0.35006
Number Of Open Credit Lines And Loans	-0.0739745
Number Of Times 90 Day Late	3.27848
Number Real Estate Loans Or Lines	0.779585
Number Of Dependents	1.72528

3.2 Random Forest

Random Forest (RF) constructs and randomizes a forest. It is a classification algorithm that is monitored. The bagging procedure has been used to train a forest, which would be a collection of Classification algorithms Trees. Random forest contains a high percentage of decision trees and afterwards aggregates them to get a consistent and predictable assessment. The Randomized Forest procedure has the unusual merit of becoming useful including both regression and classification problems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

3.3 Naïve Bayes

A class of classification methods based on the Bayes Theorem is known as naive Bayes classifiers. In general, visually clustering methods such as Naive Bayes could be used to generate additional prediction models. They can be used in a wide range of disciplines for intrusion detection system, forecasts, and so on. The Bayesian argument says that the likelihood function is given by giving a conditional probability distribution over incidences G and H.

$$P(\frac{G}{H}) = \frac{P\left(\frac{H}{G}\right) * P(G)}{P(H)}$$

where G and H are events and $P(H) \neq 0$.

To compute the probability of event G provided that event H is true. Event H is also known as Evidence.

P(G) is the number of occurrences. The evidence is a result from a characteristic of an unidentified occurrence (here, it is event H). P(G|H) is the likelihood of an event happening once evidence is observed.

All the 12 features of the data set are used in all the proposed

Machine Learning models for credit delinquency prediction.

Methodology

The proposed Machine learning classification techniques is described through the following steps of Fig 1

Step 1: The data set is partitioned into training and validation using extraction of features.

Step 2: The training data are used to create a training model by applying different classification methods as Logistic Regression, Naive Bayes, Random Forest.

Step 3: Using the test data, a predictive model is created based on training data

Step 4: The output of the prediction model is correlated with the model created using trained data.

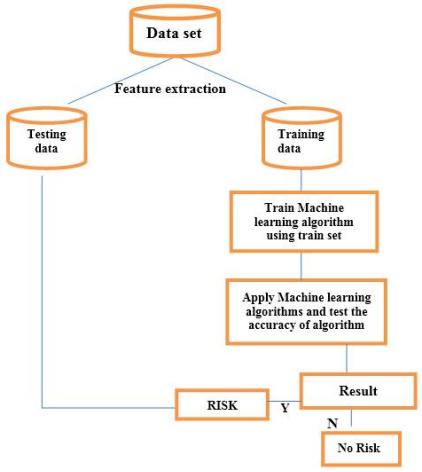

Fig .1 Flow chart of the proposed model

IV. CASE STUDY

4.1 Dataset

This section includes an observational research in which various classification methods such as Random Forest, Naive Bayes and Logistic Regression The data set is "Give Me Some Credit" from kaggle.com[14], and the intention is to determine and evaluate how well an individual may undergo financial turmoil (90-day delinquent or worse) over the next two years. SeriousDlqIn2Yrs is a binary response variable. Many of the explanatory variables will be reviewed, and most or all of them will be considered as predictor variables in the estimation method constructed.

There are 150,000 rows in the dataset and 12 categories such as product utilization rates, and demographic and behavioral characteristics, these are often used to predict business strategies in many organizations or borrowing companies. The following table 2 displays a description of the data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

Table 2: Attributes used for the proposed methodology

Attribute Number	Description	Type
1	Serious Delinquency in 2yrs	Y/N
2	Revolving Utilization Of Unsecured Lines	Percentage
3	age	Integer
4	Number Of Time 3059 Days Past Due Not Worse	Integer
5	Debt Ratio	Percentage
6	Monthly Income	Real
7	Number Of Open Credit Lines And Loans	Integer
8	Number Of Times 90Days Late	Integer
10	Number Real Estate Loans Or Lines	Integer
11	Number Of Time 60-89 Days Past Due Not Worse	Integer
12	Number Of Dependents	integer

4.2 Exploratory Data Analysis

The data and variables are thoroughly analyzed in this section. Distributions of target ratio and variable values are also being looked into. Null value analysis is performed to eliminate them from drastically damaging modeling, and analytical techniques.

Following the descriptive inputs, missing values are also checked, and some missing values are found for 2 inputs ('Monthly Income' and 'Number of Dependents'). In addition to these reasons, some modeling methods cannot function with missing values, thus they must be controlled if reliable models are to be produced.

4.3 Data pre-processing

At first appearance, the dataset looks to have adequate information, but the available information is minimal to train the ml algorithm and required additional basic data pre-processing.

4.4 Feature Selection

This section completes the correlation matrix and looks for any input correlations. in a logical, statistical manner, using associated variables. Target ratio to investigate which can be used to build a better predictive model and it is found that the target ratio is 6.7%. Followed by an analysis of the target attribute, all other inputs/features are reviewed to determine whether there are irregularities or deviations in the input value ranges. For example, some numbers in the 'DebtRatio' parameter are more than one, so they are included when generating the simulations.

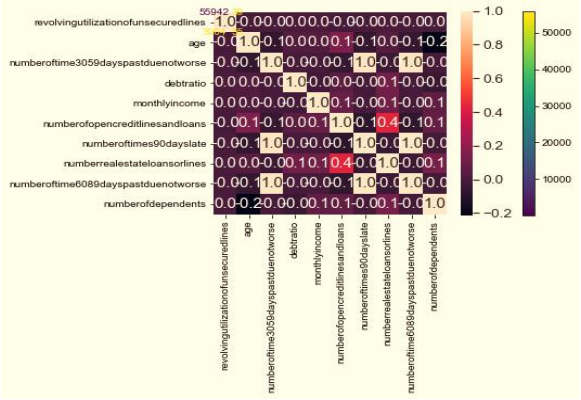

Fig 2. Correlation matrix

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

The correlation matrix is used to evaluate the correlation between the features, and it is determined that the quantity of real estate loans or lines and the number of open credit lines and loans are strongly associated.
The independent features are very less correlated with dependent variable as shown in Fig 2.

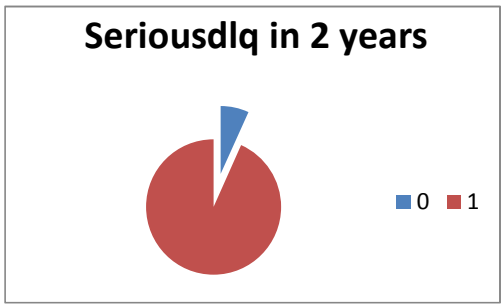

Fig 3 Target variable in 2 years

From the graph it can be concluded that roughly 6.7% of the borrowers have become defaulters within 2 years.

V. ANALYSIS AND OUTCOMES

5.1 In-depth Review

Varied contexts and different parameters were used to examine the credit risk assessment systems. Common metrics including F1 score, specificity, accuracy, sensitivity, error rate, and precision were used to compare these approaches. Below is a quick explanation of each measurement used to assess a techniques.

True Positive (TP) – categorize of good credit risk as good
True Negative (TN) – categorize of `bad credit risk as bad
False Positive (FP) – categorize of bad credit risk as good
False Negative (FN) – categorize of good credit risk as bad

Accuracy is generally determined by dividing the number of correct forecasts by the total dataset.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Recall-The ratio of the number of precise positive predictions to all positives can be used to assess sensitivity/recall.

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision-The proportion of accurate positive forecasts divided by the total number of positive predictions is the precision.

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

F1-Score-The F1-Score maintains a balance between specificity and recall.

$$\mathrm{F1} - \mathrm{S} = \frac{2 * \mathrm{PREC} * \mathrm{REC}}{\mathrm{PREC} + \mathrm{REC}}$$

5.2 Result Achieved

With all of the assessment done with the help of machine learning classifiers it was found out that Random Forest has achieved maximum accuracy of 94%. To illustrate the difference in the results, the table below summarizes the accuracy, precision, recall, and F1-score of the machine reinforcement learning used.

Table 3: Comparison of the results of several machine algorithms when applied to the credit dataset "give me some credit" from kaggle.com

TYPE	ACC	PREC	RECALL	F1-SCORE
Logistic regression	0.93	0.49	0.02	0.03
Random forest	0.94	0.45	0.19	0.28
Naïve Bayes	0.93	0.59	0.01	0.03

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

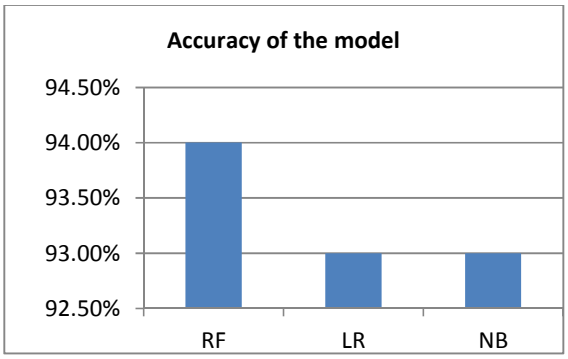

Fig. 4 Graph of Accuracy of the model

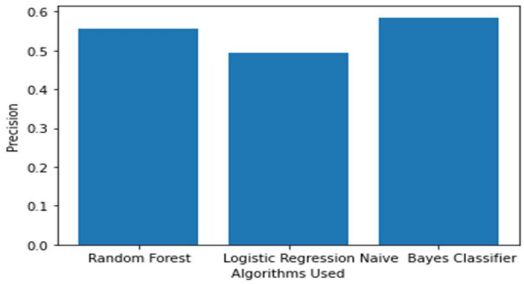

Fig. 4 Graph of Precision of the model

Table 4: Comparison of confusion matrix of different algorithms.

Type	TN	FP	FN	TP
Logistic regression	55360	621	3273	746
Random forest	55912	69	3952	67
Naïve Bayes	55942	39	3964	55

5.3 Comparison

Some papers were compared that use German credit data set from Kaggle. It was found out that that were comparatively less accurate than this paper. One of the paper got Random forest has the best algorithm and have got a accuracy of 78% which is 9% less than this paper. Another paper has concluded Logistic Regression has the most accurate which is 12% less than this paper.

The observations between the findings of the paper and other mentioned paper is tabulated in the following table 5.

Table 5: Accuracy Comparison with relevant reference works

Algorithm	Findings	Ref1[2]	Ref2[10]
Random forest	94%	78%	89%
Logistic Regression	93%	75%	87%
Naïve Bayes	93%	77%	-

VI. DISCUSSION

The model inputs must be classified in settings like banks or other lending organizations, and they should not have been shared with the business units. Because a business unit can give a clue to a branch or outside customers to purposefully or unintentionally take over the system if they fully understand the input list. It may irreparably harm the model's performance and lead the bank down the incorrect road. When a model's performance doesn't change mid-term as expected, it should be reevaluated. Model performance should be tracked periodically. Because this approach also seeks to alter consumer behavior and instructs users on how to act after receiving a bank credit.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, June 2022
Impact Factor: 6.252

VII. CONCLUSION AND FUTURE WORK

In this study, by using Machine learning classifiers such as Random Forest, Naïve Bayes and Logistic Regression the borrower's delinquency prediction is solved to achieve better precision. It is used to find out whether the customer will cause default in the next two years or not. Using this method bank can decide to whom loan has to be granted. In contrast to expert-based approaches or weekly and monthly performance reports for monitoring, the project's results are trustworthy, accurate, and most crucially, based on statistics. The used method is simple to comprehend and use, making it ideal for banking environments. The model's performance is decent, and given how straightforward the inputs are, related managers and the rest of senior management should have no trouble understanding it. To construct a more accurate prediction model, the number of data may be increased in the following stage, both in terms of volume and variety.

REFERENCES

[1]. VidhiKhanduja; Simran Juneja," Defaulter Prediction for Assessment of Credit Risks using Machine Learning Algorithms",2020 4th International Conference on Electronics, Communicate and Aerospace Technology (ICECA),DOI:10.1109/ICECA49313.2020.9297590

[2]. Varsha Aithal, Roshan David Jathanna, " Credit Risk Assessment using Machine Learning Techniques",ISSN:2278-3075,Volume-9 Issue-1, 2019,DOI: 10.35940/ijitee.A4936.119119.

[3]. Trilok Nath Pandey,Suman Kumar Mohapatra, Alok Kumar Jagadev, Satchidananda Dehuri," Credit risk analysis using machine learning classifiers", International Conference on Energy, Communication, Data Analytics and Soft Computing,Aug.2017 (ICECDS-2017), DOI: 10.1109/ICECDS.2017.8389769

[4]. S. Sathya Bama, A. Maheshwaran, S. KishoreKumar, K. RaghulKumar, M. Yogeshwaran "Identification of Default Payments of Credit Card Clients using Boosting Techniques", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

[5]. S.J. Shiv; Srinivasa Murthy; Krishnaprasad Challuru," Credit Risk Analysis Using Machine Learning Techniques", 2018 Fourteenth International Conference on Information Processing (ICINPRO), DOI: 10.1109/ICINPRO43533.2018.9096854

[6]. T. Pandey, A. Jagadev, S. Mohapatra, and S. Dehuri, "Credit risk analysis using machine learning classifiers," Aug. 2017, pp. 1850–1854. DOI: 10.1109/ICECDS.2017.8389769. [3]

[7]. Yiheng Li * and Weidong Chen,"A Comparative Performance Assessment of Ensemble Learning for Credit Scoring", Oct. 2020

[8]. Huichao Mi. (2021). "Research on the Application of Machine Learning Algorithms in Credit Risk Assessment of Minor Enterprises",Converter Magazine,Volume 2021,no. 4 ISSN: 0010- 8189,2021

[9]. A. Fenerich, M. Steiner, P. Steiner Neto, E. Tochetto, D. Tsutsumi, F. Assef and B. Samways dos Santos, Use of machine learning techniques in bank credit risk analysis, Rev. int. métodos numér. cálc. diseño ing. (2020). Vol. 36, (3), 40

[10]. Tahmid, Nazre & Haque, Nasimul & Faruque, Umar & Keya, Mumenunnessa & Khushbu, Sharun & Marouf, Ahmed(2021) "A Concern of Predicting Credit Recovery on supervised learning algorithm",1-5. 10.1109/ICCCNT51525.2021.9579706.

[11]. Song Wen; Bi Zeng; Wenxiong Liao; Pengfei Wei; Zhihao Pan," Research and Design of Credit Risk Assessment System Based on Big Data and Machine Learning", IEEE 6th International Conference on Big Data Analytics (ICBDA),2021.

[12]. Su, C., Tu, F., Zhang, X., Shia, B., & Lee, T. (2021). A Ensemble Machine Learning Based System for Merchant Credit Risk Detection in Merchant Mcc Misuse. Journal of Data Science, 17(1),81-106.doi:10.6339/JDS.201901_17(1).0004

[13]. Guerra, Pedro & Castelli, Mauro. (2021). Machine Learning Applied to Banking Supervision a Literature Review. Risks. 9. 136. 10.3390/risks9070136.

[14]. https://www.kaggle.com/c/GiveMeSomeCredit

[15]. https://bigml.com/user/jbosca/gallery/dataset/5a7def3d2a83476e09000456