E-ISSN 2581 – 7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 18, June 2020
DOI: 10.32377/cvrjst1818

Loan Delinquency Prediction using Machine Learning Techniques

B. Ashwin Kumar[1], Chadive Koushik Reddy[2], Chilkamarri Krishna Srinivas[3] and Koya Lokesh Reddy[4]
[1]Asst. Professor, CVR College of Engineering/CSE Department, Hyderabad, India
Email: ashwinvrk@gmail.com
[2] B. Tech Student, CVR College of Engineering/CSE Department, Hyderabad, India
Email: koushikchadive99@gmail.com
[3] B. Tech Student, CVR College of Engineering/CSE Department, Hyderabad, India
Email: krishnasrinivas3303@gmail.com
[4] B. Tech Student, CVR College of Engineering/CSE Department, Hyderabad, India
Email: lokeshreddy2501@gmail.com

Abstract: **Loan delinquency prediction is one of the most critical and crucial problems faced by financial institutions and organizations. It is a remarkable effect which could result in the demolition of the profitability rate that leads to the shattering of the organization. Delinquency is a condition that arises due to the failure of payment of loans by the borrowers, which shows tremendous effect on the evolution of financial institutions. It requires more authentications to track the periodic repayments of debts and adaptation of strategies that helps in proliferating the institutions. Based on the details like date of issuance, pay-back time, amount, account details, credit score further issuance of loan is assured this helps in disseminating the delinquency problem.**

Index Terms: **Delinquency; Prediction; Logistic Regression; Supervised Learning; Decision Tree; Random Forest; Ensemble.**

I. Introduction

According to the statistics, 40% of the economy of any organization is drowning out due to no repayments of debts by borrowers. This catastrophe can be eradicated by monitoring the stream of transactions and debts. This process of evaluation is not so easy for humans to carry out, it requires an automated system which can predict the loss percentage, delinquency rate and monitor the stream of transactions. A loan is considered "delinquent" when a borrower doesn't make a loan payment on time. Most lenders allow consumers a grace period to make up a missed payment and get their loan out of delinquency. However, once a loan is delinquent for a certain period, it gets into the risk of going into default. It's important to make timely payments to avoid defaulting, which can have negative impacts on credit score and the ability to receive credit in the future. Loans default will cause huge loss for the banks, so they pay much attention to this issue and apply various methods to detect and predict default behavior of their borrowers. Machine Learning is an efficient way to eradicate this disaster, it is an amalgamation of Classification and Clustering Algorithms with different capabilities, functionalities it can compute a simple-range - problems to a complicated decision-making.

The models use a supervised learning technique which produces a target variable, depending on the type of value produced by the target variable (discrete or continuous) type of algorithm is considered.Here the target variable produces discrete values so we use classification algorithms.

Based on the features of the Machine Learning environment it is clear that it is the best solution for Loan Delinquency prediction which helps an organization to regain the hike in profit percentage and also helps in the portfolio of financial organization.

II. Related Work

Literature Review

"Ref. [1]" Jian Chen, Ani Katchova, Chenxi Zhou in their paper "Agriculture loan delinquency prediction" they have used the "logistic regression model" to control delinquency their work on agriculture loan delinquency gave an insight on the efficiency of logistic regression algorithm to predict the accurate result.

*"Ref. [2]"*Uzair Aslam, Hafiz Ilyas, AsimSohail in their paper "Empirical study on loan default prediction" the credit score has become an important parameter nowadays for any financial organization to approve the loan, by considering credit score as a parameter different models like "SVM, Logistic regression, Neural networks" are used to predict the default rate, it helped us to explore new techniques, their efficiencies and complexities.

*"Ref. [3]"*Anastasios Petropoulos, Vasilis Siakoulis, EvaggelosStavroulakis and Aristotle Klamargias in their paper "A robust machine learning approach for credit risk analysis of loan" in their research they have shared the capabilities of a "Random Forest, Neural networks and Decision tree" and generated a solution to loan credit risk which gave an insight on efficiency of decision trees and Random forest.

"Ref. [4]" Research on "Loan default prediction using random forest classifier" by Lin Zuh, Dafeng Qui, DajiErgu, Cai Ying consists of a crucial technique of denoising data and effectively implement the random forest classifier algorithm to predict the loan default status with various input factors. According to the paper, the establishment of a random forest model implements two steps: building a decision tree and forming a random forest.

CVR Journal of Science and Technology, Volume 18, June 2020
DOI: 10.32377/cvrjst1818

"Ref. [5]" Research on "Loan Prediction Analysis using Decision Tree" by Nikhil Mandan, Siddharth Nanda discussed the functionalities of the decision tree and how the model will split the attributes into subsets to predict the results. This paper also enhanced the capabilities of decision trees by comparing with other classification models like SVM, Naive Bayes, Logistic Regression.

III. Problem Statement

Building an efficient model using machine learning techniques to predict the status of loan delinquency of a customer. It helps banks, financial institutions to monitor the status of the borrower that helps in deciding whether to approve the loan request or to reject the loan request and also it helps in increasing the profitability of an organization.

A finance providing organization, banks or private vendors gauge some parameters of a lender like age, income of the borrower, borrower credit score, purpose of loan, type of loan, the service time of borrower, previous transactions, previous loan repayment history, loan term, unpaid principal amount, debt to income ratio, Interest to approve loan request of a lender.

To build a machine learning model, key parameters or features of a borrower like Age, Loan term, principal amount, Interest rate, Debt to Income ratio, borrower credit score, co-borrower credit score, Loan purpose are considered. These attributes help a machine learning model to predict accurate outcomes and support lenders to decide whether to approve loan requests or to reject the loan request.

IV. Proposed Method

The main goal is to develop a solution that helps in decimating the delinquency problem. Machine learning provides decision-making classification algorithms to deal with such problems. Since the problem is a classification problem, classification decision-making algorithms are used, provided by machine learning.

Classification is a Supervised learning approach in which the program learns from the input data and then uses this learning to classify new observations.

The most widely used classification algorithms like Logistic Regression, Naive Bayes classifier, Nearest Neighbor, Support Vector Machines, Decision Trees, Random Forest and Neural Networks.Among these Logistic Regression, Random Forest and Decision Tree are considered in the model building because the response variable is categorical.

Logistic Regression: The idea of logistic regression is to find the relationship between features and the probability of particular outcomes.

The response variable has two values "0" or "1" that is "acceptance or rejection" which helps in decision making whether to approve the loan or disprove that controls the delinquency problem by rejecting the fault loan approvals.

Decision Tree: It uses tree representation to solve problems in which each leaf node represents a class label and internal nodes represent the attributes.

The decision-making process by response variable is quite different from logistic regression. In the Decision Tree, data is split multiple times based on the features provided with which subsets of data are generated; the final subsets are called "leaf nodes" or "terminal nodes".

To predict the outcome the average values of each acceptance case and rejection case are taken based on the values decision is taken. By this process efficiency increases and produces more accurate results.

Random Forest: Random forest is a supervised learning technique which uses ensemble learning methods for classification.

Ensemble learning is a technique in which predictions from multiple machine learning algorithms are made together to make accurate results than any individual models.

It operates by a multitude of decision trees at training time and outputs the classified response variable. Decision trees classifiers can be aggregated into a random forest ensemble which combines their inputs.

Results from different individual trees are aggregated through averaging then the decision is taken based on the averaged value. If the average value is nearer to the acceptance rate then the response or outcome of the model would be acceptance else the response would be rejection.

Based on the unique functionality and features of random forest algorithms it is more efficient and reliable which can produce results with a very high accuracy rate and capable of making precise predictions. Models developed using such an efficient algorithm for delinquency problems can generate very accurate outcomes with which the problem can be eradicated.

A. System Architecture

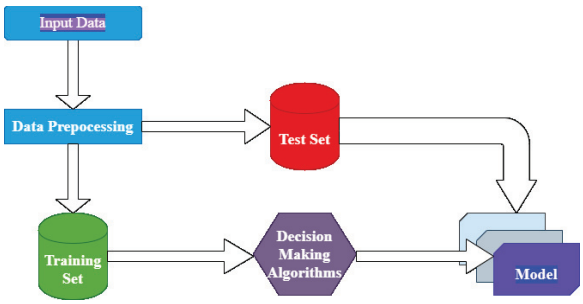

Figure 1. System Architecture.

The dataset is imported from the resource; it undergoes several phases before it is given as input to the algorithm.

Initially, data is checked for any null values or missing values then the entire data is scaled into a single format to eradicate the type errors.

In the data prep-processing stage the entire data is processed to remove anomalies or any redundancy which acts as a barrier for a model to predict accurate results. After the successful data pre-processing the entire data is divided into train and test sets, train set is used to make the model acquire knowledge regarding the features, changes, inter-

CVR Journal of Science and Technology, Volume 18, June 2020
DOI: 10.32377/cvrjst1818

dependencies in the dataset it is also called the learning phase or knowledge phase for a machine learning algorithm.

The appropriate algorithm is used and the training data is fit into the algorithm which helps in building an efficient model.

Testing phase of a model is also known as the application phase, where some sample data is provided to verify the rate of accurate predictions by the model, if the model predictions are accurate then the model can be used for future predictions as shown in figure1.

B. Dataset Used

The loan Delinquency dataset is considered from Analytics ML Hackathon 2019. This dataset consists of over 116059 records with 28 attributes. We have divided the dataset into train data and test data.

C. Procedure

1. IMPORTING LIBRARIES:

Library is a collection of functions and methods that allows many actions to take place without writing the code.

Libraries that are used to develop a machine learning model are:

1. Numpy
2. Pandas
3. Scikit learn
4. Seaborn
5. Matplotlib

Numpy is used to support large multidimensional data and also helps in high-level mathematical functions to operate on these arrays.

Pandas library is used to read different types of datasets and store them as a data frame and also it helps in structuring the unstructured data. It is built on Numpy.

Scikit learn is also known as sklearn it is a premier machine learning package which contains all machine learning algorithms

Seaborn is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

Matplotlib is a very powerful plotting library useful for working with Python and NumPy. The most used module of Matplotlib is Pyplot, which provides an interface like MATLAB but instead, it uses Python and it is open source.

2. READING THE DATASET:

It is an important step in model building, the data may be available in any format like (.txt,.xlsx,.csv). Based on the availability of format of the dataset, an appropriate method is selected to read the data.

3. SPLITTING THE DATA:

In this phase the acquired data is divided into **train** and **test sets**, the train set has 87044 records and the test set has 29015 records.

The train data helps the model to learn the patterns and test data is used to check how precisely the model has been trained.

4. DATA PRE-PROCESSING:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.

Data Preprocessing is a technique that is used to convert the raw data into a standard data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

5. MODEL BUILDING:

Model Building is a key step in which by using the appropriate algorithms respective models are built which help in solving simple statistical problems to complex decision-making problems.

(I) LOGISTIC REGRESSION ALGORITHM:

Logistic regression algorithm is a module which is imported from the**linear_module** package in **sklearn library.**

LogisticRegression() is a function which is used to apply the logistic regression algorithm to the data.

fit () function is used to fit the train data into the logistic regression algorithm on which the model is built and trained so that future predictions can be made precisely.

(II) DECISION TREE ALGORITHM:

Decision tree algorithm is a decision-making algorithm which helps in dealing with complex problems.

This algorithm is inherited from the DecisionTreeClassifier module which is imported from the **tree** package present in the **sklearn**library.

DecisionTreeClassifier() method is used to apply decision tree algorithms to a machine learning model with decision tree functionalities.

fit () is used to fit the train data into the Decision tree algorithm and the algorithm will split accordingly to generate the subsets of leaf nodes which helps in predicting the precise results.

After model building, the test data is fitted into the algorithm to verify the accuracy of the model and to check for precise results.

(III) RANDOM FOREST ALGORITHM:

Random forest algorithm is an ensemble learning algorithm in which inputs from different algorithms are processed to generate accurate results. This algorithm generates a

E-ISSN 2581 – 7957
P-ISSN 2277 – 3916
CVR Journal of Science and Technology, Volume 18, June 2020
DOI: 10.32377/cvrjst1818

decision tree and the output of each decision tree is considered to be an input to the random forest algorithm.

The Random Forest algorithm is a module which is imported from the **ensemble** package of the**sklearn** library **RandomForestRegressor ()** is a function which is used to apply random forest algorithms to the data.

fit () function is used to fit the train data to generate the model and train the data according to the changes, correlation, covariance of variables in the dataset and it is tested by using test data to verify that desired outcomes are generated or not.

6.DATA VISUALIZATION:

VISUALISATION OF LR MODEL:

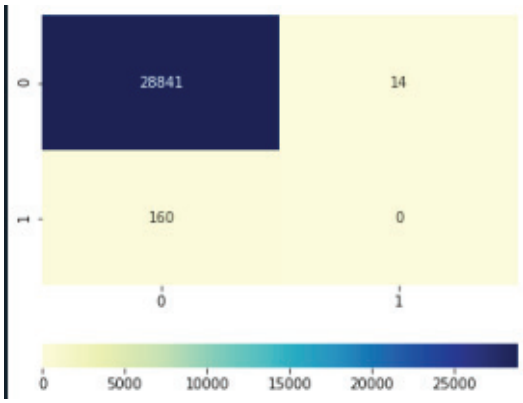

Figure 2. Heatmap diagram representing the nearness of the data predictions by the model using Logistic regression Algorithm.

The above heatmap represents the number of records falling in the acceptance range and number of records falling in the rejection range. Heatmap is generated with the help of a confusion matrix which is developed by a logistic regression algorithm.

The blue color region represents the number of accepted records (28841) other shades of the heatmap represents the number of rejected records (174) as shown in figure2.

VISUALISATION OF DECISION TREE:

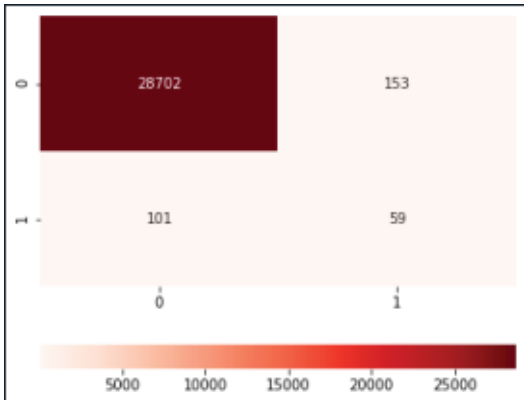

Figure 3. Heatmap diagram representing the nearness of the data predictions by the model using the Decision Tree Algorithm.

The above heatmap is generated from the decision tree algorithm. It represents the number of acceptance records based on predictions made by the algorithm. Based on the confusion matrix generated by algorithm (28702) records are accepted and remaining (313) records rejected as shown in figure3.

VISUALISATION FOR RANDOM FOREST:

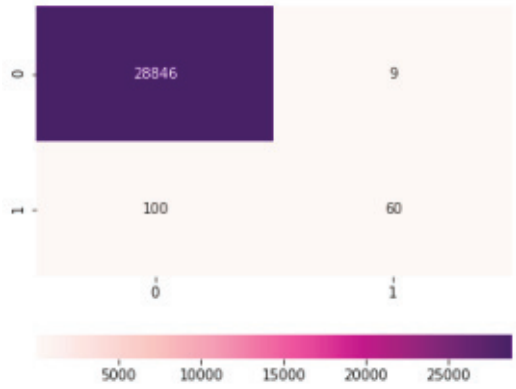

Figure. 4. Heatmap diagram representing the nearness of the data predictions by the model using the Random Forest Algorithm.

The above heatmap is generated from the random forest algorithm. It represents the number of acceptance records based on predictions made by the algorithm. Based on the confusion matrix generated by algorithm (28846) records are accepted and remaining (169) records are rejected as shown in figure4.

V. RESULTS AND DISCUSSION

TABLE I.
EVALUATION METRICS

S.no	Model	Accuracy	Precision	Recall
1	LOGISTIC REGRESSION	0.9741	0.9744	0.9997
2	DECISION TREE	0.9725	0.9878	0.9839
3	RANDOM-FOREST	0.9846	0.9868	0.9975

Accuracy, precision, recall of a model using a confusion matrix. Confusion matrix is a summary of the number of correct and incorrect predictions made by a classifier and broken down by each class.

Accuracy is the ratio of the number of correct predictions to the total number of input samples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision is calculated by dividing the positive examples with the total number of positive examples.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall is defined as the ratio of the number of correctly classified positive values to the total number of positive values.

$$\text{Recall} = \frac{TP}{TP+FN}$$

True Positive (TP): Observation is positive, and is predicted to be positive.

False Negative (FN): Observation is positive, but is predicted negative.

False Positive (FP): Observation is negative, but is predicted positive.

True Negative (TN): Observation is negative, and is predicted to be negative.

VI. Conclusion And Future Work

The proposed solution is implemented and tested successfully. This paper speaks about how to eradicate the delinquency situation by developing different machine learning models using existing algorithms and comparing them to pick the best model. The results produced by every model are accurate with respect to their algorithmic properties. In the future, the generated model can be enhanced to increase the efficacy to deal with similar problems.

References

[1] Aslam, Uzair & Aziz, Hafiz Ilyas Tariq &Sohail, Asim& Batcha, Nowshath. (2019). "An Empirical Study on Loan Default Prediction Models". Journal of Computational and Theoretical Nanoscience. 16. 3483-3488. 10.1166/jctn.2019.8312.

[2] Premkumar B., Lakshmi R., Behera B. "Performance analysis and evaluation of machine learning algorithms in rainfall prediction, International Journal of Advanced Science and Technology", Volume 29, 2020

[3] Nali´c, J. and Švraka, A., 2018. "Using Data Mining Approaches to Build Credit Scoring Model: Case Study—Implementation of Credit Scoring Model" in Microfinance Institution. 2018 17th International Symposium Infotech-Jahorina (INFOTECH), IEEE. pp.1–5.

[4] Baesens, B. Roesch, D. and Schedule, H., "2016.Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. United States, John Wiley & Sons".

[5] Sarma, K.S., "2013.Predictive Modeling with SAS Enterprise Miner."

[6] "Practical Solutions for Business Applications. SAS Institute."

[7] Abdou, H.A. and Pointon, J., "2011. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2–3), pp.59–88".

[8] Jun Hao, Qianqian Feng, Weilan Suo, Guowei Gao, Xiaolei Sun, "Ensemble forecasting for electricity consumption based on nonlinear optimization, 2019, pp."

[9] Mingyue Jiang, Guowei Gao, Yirui Deng, Chenglong Wang. "Market Risks Prevention and Control of "Going Global" for Chinese Electrical Enterprises, 2019, pp."

[10] Ben Hassen H., Elaoud A., Masmoudi K., "Modeling of agricultural soil compaction using discrete Bayesian networks."

[11] International Journal of Environmental Science and Technology, Volume 17, 2020, Zhu L., Qiu D., Ergu D., Ying C., Liu K., "A study on predicting loan default based on the random forest algorithm, Procedia Computer Science, Volume 162, 2019."

[12]Song.Y., Wang Y., Ye X., Wang D., Yin Y., Wang Y. "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending, Information Sciences, Volume 525, 2020"

[13] Huang Y.-P., Yen M.-F, "A new perspective of performance comparison among machine learning algorithms for financial distress prediction Applied Soft Computing Journal, Volume 83, 2019"

[14] Chen R., Zhou H., Jin C., Zheng W. "Modeling of recovery rate for a given default by non-parametric method, Pacific Basin Finance Journal, Volume 57, 2019."