A Research Paper on Loan Delinquency Prediction

Aditya Sarkar[1] adityasarkar937@gmail.com**,**
Karedla Krishna Sai[2] krishnasai0880@gmail.com,
Aditya Prakash[3] ad.prakash4@gmail.com,
Giddaluru Veera Venkata Sai[4] veeravenkatasai55@gmail.com,
***Manjit Kaur**[5] manjit.12438@lpu.co.in

Department of Computer Science , Lovely Professional University, Phagwara

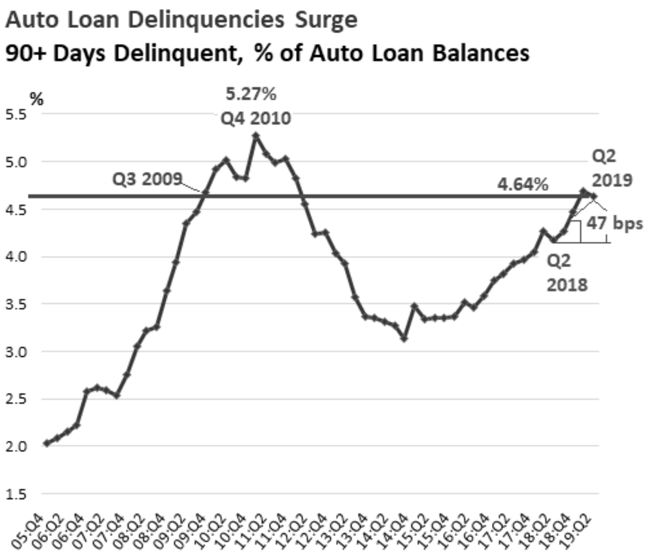
---

Abstract - *Evaluating and forecasting the willingness of borrowers to repay is critical for banks to minimize the possibility of defaults in the payment of loans. For this purpose, there is a mechanism set up by banks to manage a loan request depending on the status of the lender, like job status, history of credits and many more.But the new assessment scheme may not be sufficient for assessing the willingness of such borrowers to repay, such as students or non-credit historians. To better determine the willingness of all sections of people to repay, we tried different machine learning models on the Kaggle data model and assessed the value of all the features used. This paper uses Machine Learning to identify factors that influence mortgage defaults. Our main purpose is to figure out the delinquency status of loans for the any 'n' next month given the delinquency status for the previous 12 months. We apply the adjusted Gradient Boosting (XGBoost) approach to data. First, we prove that the precision of XGBoost's estimation is better than the logistic regression. Second, we use the Permutation Feature Importance approach to identify the most significant variables affecting delinquency. This paper uses loan statistics over an extension cycle and points out the value of the interest rate in the determination of the delinquency. Our findings vary from those found in the literature. In our Combined Loan-to-Value and Unemployment model, while significant, they are both dominated by the factors described above. This comparison in data draws attention to the importance of the market cycle in deciding the causes of delinquency*

Key Words*: *Delinquency, Gradient Boosting, Permutation Feature, Model based, Memory based.

1. INTRODUCTION

It is a long-standing concern of the reasons that influence mortgage default. As because of the poor credit history, a lot of individuals are waiting to get loan facilities from reliable lenders and banks. Those individuals are mostly are either students or adults who are unemployed and inexperienced to support the reputation of anonymous lenders. The lenders can get profit by taking advantage of very high interest rates ,having secret clauses included in deal. Despite assessing the creditor based on status of their previous data that they had with the bank, there are also other ways to calculate or estimate their ability to repay. Let us discuss an example, work became a major cause affecting the willingness of an individual to repay when a working adult possess fixed income and cash flow. Another considerations, such as real estate, marital status, and the place of residence, assist in the analysis of the potential to repay. Because of this, in our research, we are preparing to use deep learning algorithms to analyze the relationships between borrower and their willingness to repay. Loan default prediction is one of the most uncertain and vexing problems faced by many financial organizations as it creates many loopholes for exploitation as it marks a important effect on the profit of such institutes. . Therefore, to keep a clean slate, the banks put extra efforts into monitoring and evaluation measures in such cases as it is crucial for their profit and alarms them of any exploits. It is also a best practice to ensure timely repayment of loans by borrowers. Sometimes these carefully planned steps still fail to deliver and meet the result that was expected, and a major proportion of loans become delinquent.



Source of data: New York Fed WOLFSTREET.com

Chart -1: Delinquencies Surge

What is Delinquency? Well Delinquency occurs when a borrower misses a payment against his/her loan. Given the information like mortgage details, borrowers related details and payment details, our main purpose is to find out the delinquency status of loans for any or every user of the bank for the next month when there given a data of delinquency status for the previous 12 months (in number of months). Moreover, we have implemented a data science and machine learning model which was based on the features that helps us to find the best fit for our problem. This model will enable us to find and predict repayment ability of the loaners.

1.1 Data Cleaning

The data set included in the project is from Kaggle official repository and the data set and named as train and test data sets. These datasets help us to firstly train or data then we have a data set that can be used and observe what the outcome of the test set is.

We have used cleaning techniques as there were multiple errors in the data that were to be cleaned such as repetitive values and the as usual ubiquitous NULL values. Python and R are the most preferred language for machine learning as it provides multiple and numerable inbuilt libraries that can be installed in the system of the user to make the problems easier to tackle. There are pandas, NumPy and scala which helps in potting graphs and getting predictions then there are neural networks that also work wonders when used. Thus it helps us by improving the accuracy of the model that is used in depicting the data.

Data cleaning is a normal term that means removing or altering those data that are incorrect values there can be incomplete data, misread data, or corrupted data. These data need to handled as these are unwanted data that need to be changed and our first task is to alter these values.

When combining many data from different data sources , there are many opportunities for data to be duplicated or mislabeled. If data is wrong, outcomes and algorithms are unreliable, albeit they'll look correct.

There are no strict regulations that are set which is to be followed to remove unwanted data. Data cleaning depends on the given data and only that much is done what is required of us. The steps are different for different datasets and thus cleaning needs to be written again for most of the data sets uniquely. But it's crucial to determine a template for your data cleaning process so you recognize you're doing it the proper way whenever.

Why do u require data cleaning? Well, we use it to remove unwanted data and see which data is being unwantedly repeated or has NA values in them.

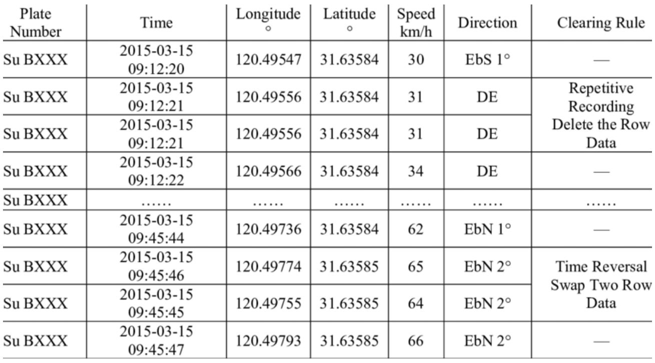


Plate Number	Time	Longitude °	Latitude °	Speed km/h	Direction	Clearing Rule
Su BXXX	2015-03-15 09:12:20	120.49547	31.63584	30	EbS 1°	—
Su BXXX	2015-03-15 09:12:21	120.49556	31.63584	31	DE	Repetitive Recording Delete the Row Data
Su BXXX	2015-03-15 09:12:21	120.49556	31.63584	31	DE	
Su BXXX	2015-03-15 09:12:22	120.49566	31.63584	34	DE	—
Su BXXX	……	……	……	……	……	……
Su BXXX	2015-03-15 09:45:44	120.49736	31.63584	62	EbN 1°	—
Su BXXX	2015-03-15 09:45:46	120.49774	31.63585	65	EbN 2°	Time Reversal Swap Two Row Data
Su BXXX	2015-03-15 09:45:45	120.49755	31.63585	64	EbN 2°	
Su BXXX	2015-03-15 09:45:47	120.49793	31.63585	66	EbN 2°	—

Fig. 1 Incorrect values/NA values

1.2 Related Works

We have studied several research papers based on loan delinquency projects on machine learning and analysed different aspects.

A research paper of Loan delinquency based on Machine Learning using OptiML we have pointed some research gap that the proposed model could not meet the expectations and the respective proposed model mainly concentrates on predicting the credit score effects because of loan defaults [1].

Another research paper we have studied was by University Of California; Riverside .This paper mainly tells about the determinants of mortgage loan delinquency and predicts the mortgage defaults [2].

Some research papers show their impact on aspects such as low credit histories, people who are stubborn to get loan facilities from unknown lenders etc [3].

Another main issue that we tend to found in some analysis papers is employment rate that affects some borrowers whose loans find yourself in due. With the assistance of our project we tend to contribute our uttermost to fulfil the analysis gap planned in higher than models.

During this paper, the random forest algorithmic program is adopted to create a model for predicting loan default within the disposal club and therefore the results square measure compared with different 3 algorithms of logistical regression, call tree and support vector machine [4].

Demonstrated the assistance of machine learning models to forecast loan repayment ability on a difficult dataset. We demonstrated that data pre-processing stands at a careful collection of datasets balancing methods and classification models are all critical to achieving the best results [5].

Financial institutions are increasingly turning to AI and machine learning not only for credit risk management, but also to manage other risks like financial fraud, money laundering, the risk of not being checked in with the regulations and thus that can effect the income as it can result in financial loss from the client side.

If the danger related to a given loan record is foreseen well previous time, monetary establishments will do acceptable communication or different appropriate action to forestall loan default.

The part of these risks is managed exploitation varied AI and machine learning techniques. All these risks can be managed using various AI and machine learning techniques [6, 7].
This study is going the delinquent behaviour of borrowers employing a variable Korean account-level hierarchy of dataset. [8].

We chose to outline 3 delinquency states (i.e., no delinquency, delinquency, and high delinquency) and then to draw the attention on the probability of transitions between these states. From the assistance of panel knowledge, ascertained quarterly,we discover that changes in recipient characteristics throughout the loan amount have an serious effect on the transition chances, as do loan characteristics and behaviour , the borrower's characteristics at the time of account application process, and economics variables [9, 10].

2. Proposed methodology

Our main goal is to design and develop a system that assists in detecting and formulating the loan delinquency in a detailed way that can be easily understood by borrowers. Since it involves prediction work.
Machine learning provides a vast range of classification problems and decision-making algorithms which has become our best way for this problem. Supervised learning approach which we are going to use provides a perfect solution since the program learns from the input data and uses the output results to analyze new observations. There are many algorithms such as Naïve Bayes Classifier algorithm, Logistic Regression model, K-Means Clustering model, Decision trees model and Random Forest algorithm, Support Vector machines and many more.

Here we are providing a brief explanation of algorithms we have used.

2.1 Random Forest Algorithm:

To start, Random Forest is a machine learning technique that is also used in data science. It is a reliable technique that's capable of performing both regression and classification tasks with the use of multiple decision trees and how called Bootstrap and Aggregation, commonly mentioned as bagging. the essential idea behind this is often often to combine multiple decision trees in determining the last word output rather than relying on individual decision trees.
Random Forest has many decision trees that are treated because the base learning models. We randomly perform row sampling and have sampling from the dataset forming sample datasets for every model.
The Decision trees which are made is sensitive to the precise data on whichever data that they're trained. If there's any change within the training data then the resulting decision tree happens to be quite different and therefore the next upcoming predictions are often quite different.

Also Decision trees are computationally expensive to teach , carry a huge risk of overfitting, and have a bent to hunt out local optima because they can't return after they have made a split.

To address these weaknesses, we take help of Random Forest which enables us to use the blending of many decision trees into one model.

Random forest could also be a bagging technique and not a boosting technique. The trees which are made in random forests will always run in parallel to every other. There does not exist any interaction between these trees while one constructs these decision trees.

These works/operate mainly by making multiple trees or decision trees. During the sessions where we put it at training, the time and outputting the category that is the mode of the classes or prediction, it uses those multiple trees the prediction are mean prediction that is also called regression of the individual trees.

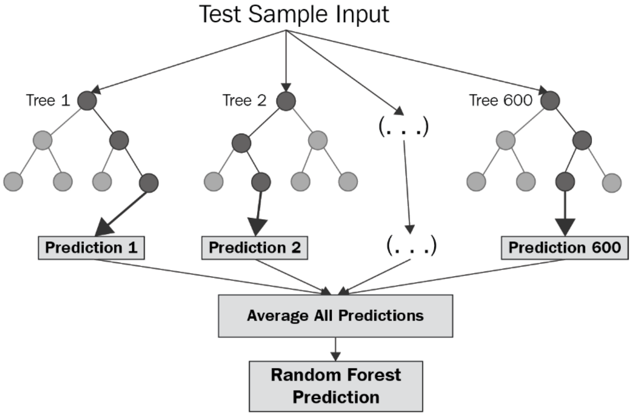


Fig -2: Basic Design of Trees with Random Forest

Disadvantages of Decision Trees:-

1. They totally relies on the data set that has been provided to it and any change in the data set means that the whole prediction has turned out to be in vein because the prediction is completely wrong for that data set.
2. Decision trees are quite expensive as well. They are harder to train because of the expense in larger

projects and always have that problem of overfitting.

3. Overfitting is a problem that one runs into when making decision trees.

This is why we use Random forest.

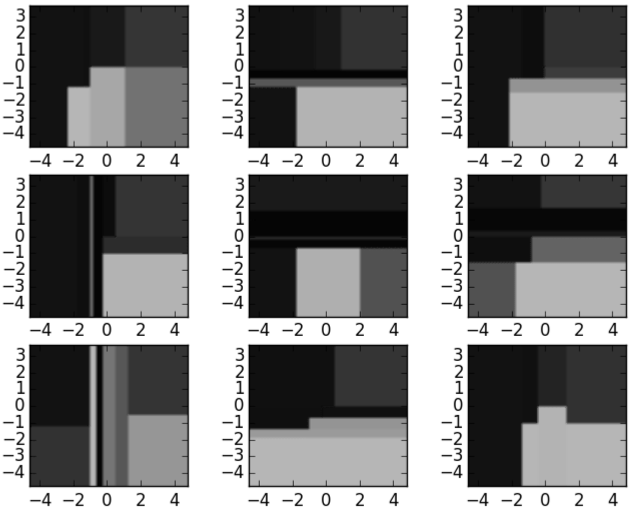


Fig -3: Data Set of Trees

This is a set of 9 data set of decision tree and in order to make a Random Forest we combine these data sets.

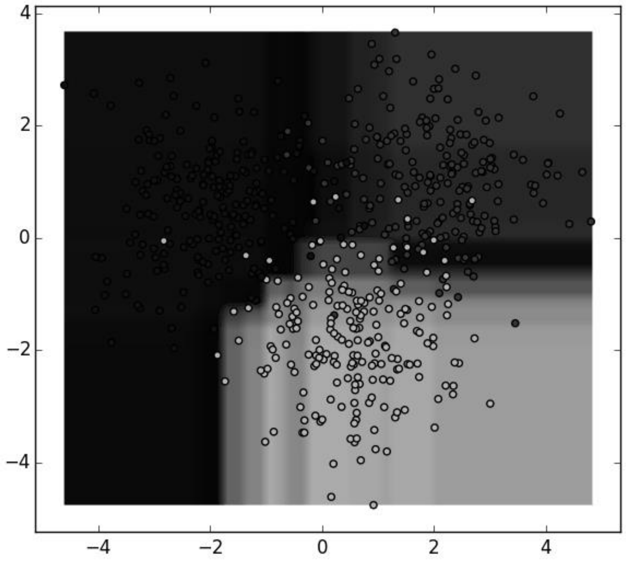


Fig -4: Random Forest Aggregated Result

This is the random forest aggregated result.

1. Random forest algorithm is still considered a highly accurate algorithm. The predictions it provides are accurate and always reliable.
2. It brightly shines where the amount of data is large.
3. It provides us the values of the variables that are important and the variables.
4. It always maintenance accuracy and has compensates for the missing data in a model.

There are still some problems that is faced by Random forest algorithm and overfitting is one of them.

2.2 Linear Regression:

Line Regression algorithms are simple algorithms. They are mainly used to find out the best fit of a data set, or to find the mid ground of the data that one uses. The equation of a line is y=mx+c where the variables y and x and dependent and independent respectively.

Now, line regression is of two types which have either single dependency and multiple dependencies.

Making Predictions -

Let us try to understand prediction with an example where we will talk about height and weight. As we can generally assume that as the height increases then the body-weight also increases therefore we will make a normal dependency regression where we will predict the data.

Our linear regression would somewhat represent like this for this problem where y is the weight and x is the height then it would be:

$y = a1 + a2 * x$

or

weight =a1 +a2 * height

Where a2 is the bias coefficient and a2 is the coefficient for the height column. We try and utilize a good learning techniques that help us to find the optimum set of values. These values will represent the coefficient values. Once we get that or obtain it then we can plug in different height values to predict the weight like any normal prediction.

For this example, lets use a1 = 0.1 anda2 = 0.5. Let's use them and calculate the weight (in kilograms) for a person with a height of 182 centimeters.

weight = 0.1 + 0.5 * 182

weight = 91.1

From the above equation, we can use different heights and weights and get more of the points in the data. The a1 is our starting point and we do not take into consideration what the a2 is. We can use multiple heights from 100 to 250 centimeters and put them to the equation and get weight values, creating our line.

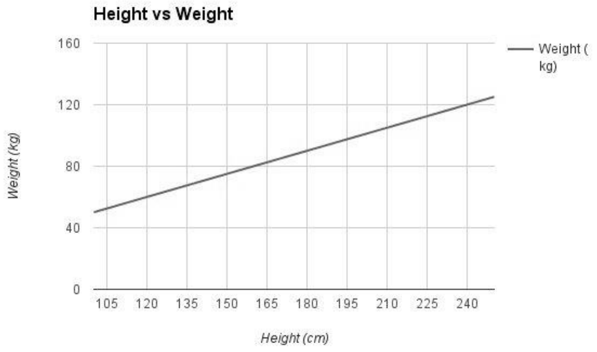


Fig -4: Height and Weight Graph

2.3 Confusion Matrix:

A confusion matrix is a table that is used to find out what is the description of the performance is where the trues values are known. These use classification models to provide the findings and are operated on the test data set mostly. The test data set has known true values thus on those values it creates its outputs. The confusion matrix is simple to understand and the true values can be visually be seen to be quite understandable. The part that becomes confusing are the related terms that come with it

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig -5: Matrix Representation of outcomes

Now, what do we learn from the above-given matrix? Let us start with simple "Yes" and "No". There are only two possible predictions where one can be "yes" and the other can be "no". Let us assume that the classifier detects which patient is suffering from the disease. So yes would represent he/she has a disease and "no" would represent they do not have any disease and are well.
The model shows that the number of predictions that were made was 165 that is a total of 165 patients were checked if they had any disease and the outcome showed.
The classifier shows that out of 165 cases, it suggested that 110 people were inflicted and 55 were not. But when checked in reality it was found that 105 patients in the total sample were actually positive and were given a false negative while and 60 patients did not have any disease but were given a false negative.

Let us understand the basic terminology of the matrix.
true positives (TP): The cases where we predicted that the person had a disease and when checked it was found that they in fact had a certain disease.
true negatives (TN): The cases where we predicted that the person did not have any disease and when checked it was found that we were correct in our prediction.
false positives (FP): The cases where we predicted that the person had a disease and when checked it was found that we were WRONG in our prediction and they did not have any disease. Also called as Type I error.
false negatives (FN): The cases where we predicted that the person did not have any disease and when checked it was found that we were WRONG in our prediction and they did have a certain disease. Also called as Type II error.

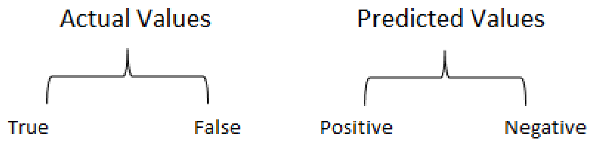


Fig -6: Different Outcomes

The confusion matrix that we used was part of the end prediction.

y	y pred	output for threshold 0.6	Recall	Precision	Accuracy
0	0.5	0	1/2	2/3	4/7
1	0.9	1			
0	0.7	1			
1	0.7	1			
1	0.3	0			
0	0.4	0			
1	0.5	0			

Fig -7: Confusion matrix predictions

For a quick glance let us look at all the terms.

- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)
- F-measure =

(2* Recall + Precision) / (Recall + Precision)

2.4 Neural Networks:

A neural network could be a series of algorithms whose main purpose is to try and to acknowledge all the hidden agendas and relationships. It is done in a subtle manner where the information is handled in a way that somewhat represents how the human brain also works around a problem. Its purpose is to try and mimic the same way that the human brain does and thus there are multiple layers that work before the output.

The duration during input the neural networks communicate that sends data and make predictions with the other system of the neurons that are present the underlying layers which are either organic or are purely artificial in its nature. One of its best features is that it can quickly adapt to the dataset that has been provided to it, a change in the dataset or the input that is provided by the user also gets quickly reflected in the output of the prediction model. It is said to generate the most effective results without having any criteria of the output nor there are any overfitting like the Random Forest algorithm.

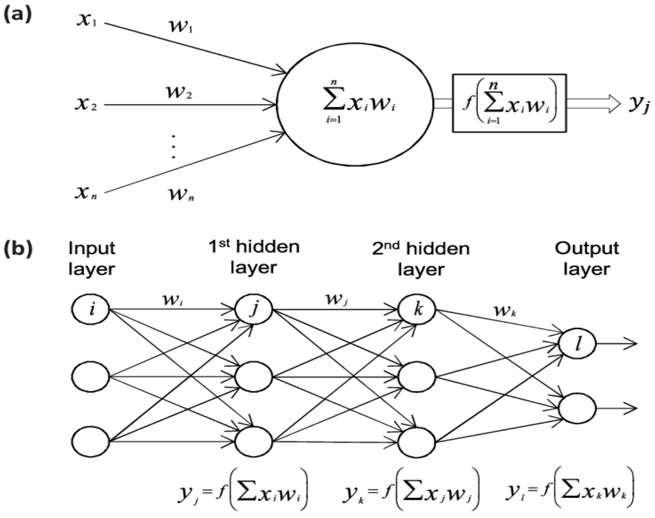


Fig -8: Neural Network

It is a representation of the human brain the dendrites the neurons and the axios, and it has a series of algorithm that helps it recreate that mimicry. It shows the relationship that exists between the knowledge that is present in the human brain.

It is extremely helpful into utilizing the applications and their spread in any financial service. It helps in the assessment of any risk involved and forecasting it thus we have used it in our paper.

A neural network works just how the human brain works. The neural network present in it is what it tries to mimic. A “neuron” in a very neural network could be a mathematical relation that collects and classifies information in step with a selected architecture. The network contains a powerful projection to those methods that are like curve fitting or have multiple variable analysis.

A neural network contains layers of interconnected nodes. Each node could be a perceptron and is analogous to multiple rectilinear regression. The perceptron feeds the signal produced by a multiple regression toward the mean into an activation function which will be nonlinear.

3. SMOTE Algorithm:

Imbalanced Data Distribution is one of the main problems of data sets which are used to analyse using Machine Learning models. To solve this problem, we are using SMOTE algorithm which is a widely used handling imbalanced class distribution algorithm.

- SMOTE (Synthetic Minority Oversampling Technique) algorithm balances class distribution by maximizing minority examples by suing replication methodology. SMOTE assists in producing new minority instances ahead of existing minority instances. The algorithm generates virtual records of training with the assistance of linear interpolation method.
- Formula: for each p belongs to X, new example: p’ = p + rand (0, 1) * | p – pz|

Here the function (0, 1) represents a number which is random number between 0 and 1.

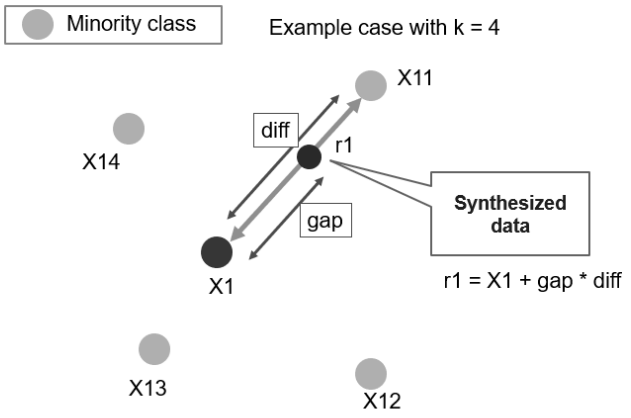
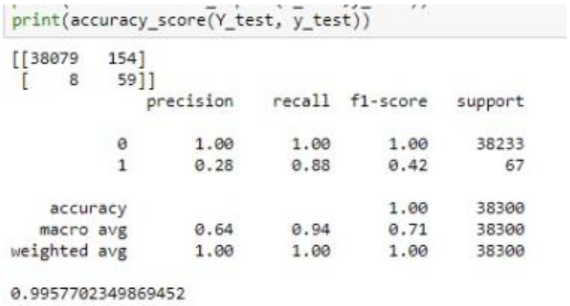


Fig -9: Representation of Smote

Our goal is to determine the delinquency status of loans for the next month based on the delinquency status for the previous 12 months using information such as mortgage details, borrowers related details, and payment details (in number of months). Our new approach tries to fill the gap and overcome the challenges that a borrower facing at present. He can accurately predict the delinquency and acts accordingly to protect his credit score by using the result.

4. Result:

Several studies have analyzed the bank-specific and economic science determinants of non-performing loans, each at the cross-country level as additionally among countries.



```
print(accuracy_score(Y_test, y_test))

[[38079   154]
 [    8    59]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     38233
           1       0.28      0.88      0.42        67

    accuracy                           1.00     38300
   macro avg       0.64      0.94      0.71     38300
weighted avg       1.00      1.00      1.00     38300

0.9957702349869452
```

Fig -10: Our Prediction of Loan Delinquency

Our prediction through neural network and confusion matrix comes as close as 99%. Through this code we have been able to identify and predict loan delinquency which an accuracy of 99% .

We have made developed a system that assists in detecting and formulating the loan delinquency in a detailed way that can be easily understood by borrowers. The outcomes are the direct result of the user's history and it takes data of the user and provides a status on it.

This paper uses Machine Learning algorithm of SMOTE to identify factors that influence mortgage defaults. The evidence also points to a differential impact across bank's model.

To be additional specific, we have found a better way of finding errors in data of such users who tries to create confusion and cheat the system of bank. This in turns provides a clear status for every user that wants a loan from the bank and is easier to deal with for any user.

Our results are therefore consistent with evidence that highlights the beneficial impact of the finding out impact of delinquency system on problem loans.

5. Conclusion:

This paper analyzes the factors causing the delinquency of the loan. We also have shown the use of algorithms that are related to machine learning on very complex datasets to analyse the prediction potential of the loan to delineate. To achieve the best results, we have shown the pre-processing data, careful extraction of data set balancing techniques and classification algorithms are all very important. The Boost approach is an easy-to-implement and interprets machine learning approach. In that way, it is very appropriate to be used in scientific economic studies. Applying this method to a dataset providing further details on local variables can further explain the impact of these variables on delinquency. The project's possible future work will include further refinement of the model, including further in-depth study of the model's variables as well as the creation of new variables to improve predictions.

ACKNOWLEDGEMENT :

We sincerely thank Lovely Professional University, Punjab for providing us with the opportunity to write a research paper.

We would also like to thank Ms.Manjit Kaur for guiding us in every stage of the research paper that we have presented. With this research paper we have learnt about the problems in loan undertaking and the risks involved by the bank when providing load to any person. I hope this research paper will help others and come use to those in need.

REFERENCES

[1] Ghosh, S. (2019). Loan delinquency in banking systems: How effective are credit reporting systems. *Research in International Business and Finance*, *47*, 220-236.

[2] Kumar, B. A., Reddy, C. K., Srinivas, C. K., & Reddy, K. L. (2020). Loan Delinquency Prediction using Machine Learning Techniques. *CVR Journal of Science and Technology*, *18*(1), 105-109.

[3] Granström, D., & Abrahamsson, J. (2019). Loan default prediction using supervised machine learning algorithms.

[4] Liang, Y., Jin, X., & Wang, Z. Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods.

[5] Ereiz, Z. (2019, November). Predicting Default Loans Using Machine Learning (OptiML). In *2019 27th Telecommunications Forum (TELFOR)* (pp. 1-4). IEEE.

[6] Jayadev, M., Shah, N., & Vadlamani, R. (2019). Predicting Educational Loan Defaults: Application of Artificial Intelligence Models. *IIM Bangalore Research Paper*, (601).

[7] Kim, H., Cho, H., & Ryu, D. (2018). An empirical study on credit card loan delinquency. *Economic systems*, *42*(3), 437-449.

[8] Qiu, W. (2019, July). Credit Risk Prediction in an Imbalanced Social Lending Environment Based on XGBoost. In *2019 5th International Conference on Big Data and Information Analytics (BigDIA)* (pp. 150-156). IEEE.

[9] Shrivastava, S., Jeyanthi, P. M., & Singh, S. (2020). Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting. *Cogent Economics & Finance*, *8*(1), 1729569.

[10] 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.

[11] Loan delinquency in community lending organizations: Case studies of neighborworks

organizations Esmail Baku &Marc Smith Pages 151-175 Published online: 31 Mar 2010.

[12] An empirical study on credit card loan delinquency Author Hyeongjun Kima, Hoon Chob Doo, jin Ryu.

[13] Loan officers and loan 'delinquency' in Microfinance: A Zambian case Author Rob Dixon, John Ritchie, Juliana Siwale

[14] Li, P., & Han, G. LendingClub Loan Default and Profitability Prediction.

[15] Loan Delinquency in Banking Systems: How Effective Are Credit Reporting Systems? Author Jugnu Ansari and Saibal Ghosh

[16] Analyzing the Causes and Evolution of Loan Delinquency/Arrears within Microfinance Institutions. A Critical Path of Action. 17 Pages Authors -Leonard Ajonakoh Fotabong: University of Buea, Ndenka Aaron: University of Buea Finance, Tasoh Toh: Independent.

[17] Predicting mortgage loan delinquency status with neural networks Authors- Ksenia Ponomareva, Paul Epstein and David Knight.

[18] Loan Default Prediction using Machine Learning Techniques Authors Vikash.V and Mohammed Aamir Ahmed

[19] Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud Authors Anand Motwani Vellore Institute of Technology, VIT Bhopal University, M.P. India and Prabhat KUMAR Chaurasiya.

[20] LendingClub Loan Default and Profitability Prediction Aurthors -Peiqian Li and Gao H