International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 06 | June -2020 ISSN: 2582-3930

Credit Risk Assessment based Loan Defaulters for Commercial Bank using Machine Learning Techniques

NamrataKamble[1], Renuka Hole[2], AarohiGaikwad[3], Ankur Naik[4], Shrikant Kokate[5]

[1] *Department of Computer Engineering,PimpriChinchwad College of Engineering*
[2] *Department of Computer Engineering, PimpriChinchwad College of Engineering*
[3] *Department of Computer Engineering, PimpriChinchwad College of Engineering*
[4] *Department of Computer Engineering, PimpriChinchwad College of Engineering*
[5] *Department of Computer Engineering, PimpriChinchwad College of Engineering*

Abstract -In banking sector credit score plays a very important role. It is important to find which customer is valid and which is not valid for loan. Now to classify customer's credit score is used. Based on this credit score of customer the bank will decide whether to approve loan or not. In banks there are major failures due to credit risks. We can automate this by using various Machine learning algorithms to identify loan defaulters. To classify and predict the customers we have used various Machine Learning algorithms like Decision Tree, Random Forest, gradient boosting and Feature Selection technique. Using this algorithms we accurately classify valid and invalid customers for loan.

Key Words:Knowledge Discovery, Ensemble Learning, Feature Selection, Receiver Operating characteristics, Data mining, Behavioral Scoring.

1. INTRODUCTION

The tremendous growth in computing power and computing capacity has resulted in the growth of huge databases. Data Mining also popularly referred to as Knowledge Discovery from Data i.e. KDD. It is an automated extraction of patterns from large amounts of data. The KDD process consists of certain steps like data cleaning, data integration, data selection, transformation, data mining, and knowledge representation. Data mining is the process of KDD in databases in which we extract useful patterns, to extract these useful patterns we use intelligent techniques.

In loan lending system credit scoring of borrowers, creditworthiness is one of the most important problems to be addressed in the Banking Industry. To reduce illegal activities in banking industry credit scoring is used. Here, credit risk is defined as a risk that borrowers will fail to meet their loan obligations. In the banking industry success and failure are based on their credit risk. If the credit amount is not collected properly, then the bank will be at loss. So, a bank's profit is correlated to credit risk. Credit scoring is divided into two group applicants and behavioral scoring. Behavioral scoring is used to classify the existing customers based on their payment history and personal information, whereas Application scoring is used to classify the applicant into good and bad groups through decision tree classifier.

2. LITERATURE SURVEY

Syed ZamilHasanShoumo, Mir IshrakMaheerDhruba, SazzadHossain, NawabHaiderGhani, HossainArif, Samiul Islam in paper[1] comparative analysis is made using algorithms such as Random Forest, Extreme Gradient Boosting, Logistic Regression and Support Vector Machine for identifying defaulters. Principal Component Analysis & Recursive Feature Elimination with Cross-Validation have been used for dimensionality reduction.

WenyuQiu, Siwen Li, Yumeng Cao, Hua Li in paper [2] proposed the proportion of "good" and "bad" samples is usually extremely imbalanced. Therefore, it constructs an ensemble model for the imbalanced datasets of small enterprise with a pre-judging mechanism.

Prof. G. Arutjothi and co-author Dr. C. Senthamarai in paper [3] proposed a system where the credit scoring model has been prepared. In this paper, KNN technique and R programming language is used. For normalization of data min max normalization is used.

Prof. Ajay Byanjankar, MarkkuHeikkilä, and JozsefMezei in paper [4] proposed a credit scoring model using neural networks. This model groups the loan application into two types, default and non-default. This neural network-based credit scoring model performs effectively in screening default applications.

Prof. TrilokNathPandey, Alok Kumar Jagadev, Suman Kumar Mohapatra and Satchidananda in paper [5] for getting better risk analysis various types of machine learning algorithms are used for the evaluation.

Prof R.S.Ramya and S.Kumaresan in paper [6] proposed a system where feature selection techniques are used on high dimensional data, to reduce dimensionality by removing the irrelevant and redundant features. The prediction accuracy of data mining algorithms and chi-square correlation is used to reduce the feature dimensionality.

Prof Yinxiao Ma, Hong Liu in paper [7] proposed support vector machine (SVM) to establish a data classification model, which uses historical dataset. By using this algorithm we are going to analysed the ability of applicant paying a

debit which reduces the risk of a bank to provide a loan.

Prof AshleshaVaidya in paper [8] discusses logistic regression and its mathematical representation. Here logistic regression adheres to as a machine learning tool to actualize the predictive and probabilistic approaches to loan approval prediction.
Prof JozefZurada in paper [9] proposed the classification performance rate using various machine learning techniques like logistic regression (LR), support vector machine (SVM), decision trees (DTs), neural network (NN), case-based reasoning (CBR), radial basis function neural network (RBFNN).

Prof Sudhamathy G., JothiVenkateswaran C in paper [10] presented a framework to effectively identify the Probability of Defaulters of a Bank Loan applicant. This model is built using data mining functions available in R. To avoid time consuming pre-processing steps, classification and clustering techniques in R are used to make the data for further use. This pre-processed dataset is further used for building the decision tree classifier.

Prof ArchanaGahlaut, Tushar, Prince Kumar Singh in paper [11] we look at whether data mining techniques are useful to predict and classify the customer's credit score good or bad to overcome the future risks of giving loans to the clients who cannot repay. Here, we use historical dataset of a bank for predictive modelling. Thus, banks can use them for the better outcome of their overall credit system.

3. EXISTING SYSTEM AND RESULTS

In this model there is a prediction system which, first fetches data then once the data is fetched, then train the system against the features like Loan Amount, Credit History, Property Area, Income, etc. Here, The Logistic regression algorithm is used for predicting customer into class 0 or class. 1. After classifying the data the decision is made by using the Decision Tree algorithm. Then, the predicted status is displayed to the customer.

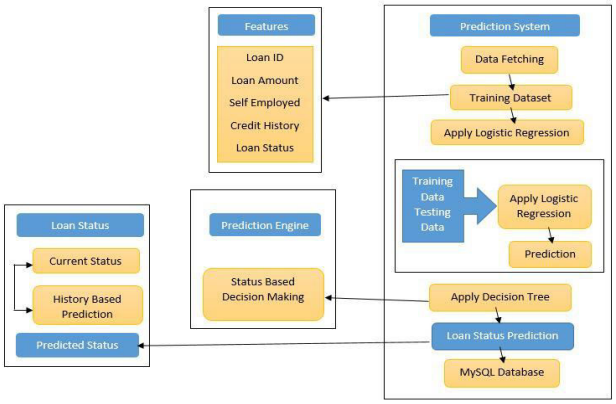

Fig -1: Existing System

4. ANALYSIS OF DATASET

1. Shape of dataset

```
(614, 13)
```

2. Data Pre-processing

i. Read csv file into a pandasdataframe.
ii. Count of missing values in each column through isnull().sum().

```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig -2: Count of missing value before Handling

iii. Handling missing values for numerical variables, this values are replaced with their mean associate with each attribute in dataset.
For example:
mean_loan=df['LoanAmount'].mean()
df["LoanAmount"].fillna(mean_loan, inplace=True)
iv. LabelEncoder is used for converting catergorical data into numerical data.
For example:

Property_Area	(Label Encoding)
Urban	2
Semiurban	1
Rural	0

Table -1: Label Encoding

v. Repeat step iii and iv to handle missing values as well as converting categorical vales to numerical values.

5. PROPOSED SYSTEM

The proposed model focuses on predicting loan defaulters. The model consists of four main parts i.e. Data, Controller, View and Model.

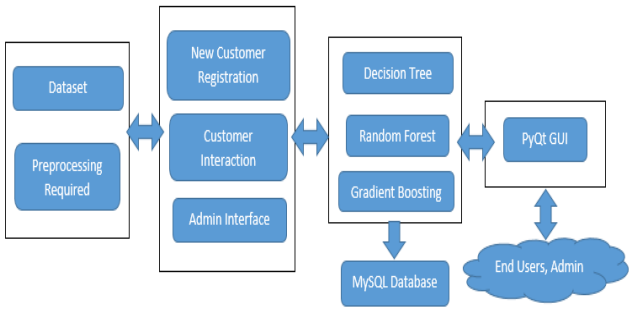

Fig -2: Proposed System

1. **Data:** This consists of a dataset, pre-processing will be applied to the dataset and new pre-processed data set will be used for further model building.

2. **Model:** This consists of 4 modules:
   - **New Customer Registration:** There are new customer i.e. (person who wants to create his account in the bank or want to apply for a loan amount). This customer will fill the registration form details and submit it. Once submitted these details will be verified by the Admin.
   - **Customer Interaction:** This module takes name and email id as input from customer and then mail is sent to customer whether he/she is eligible for the loan or not eligible for loan.
   - **Admin Interaction:** In this module, the admin verifies all the details sent or submitted by the customer through the registration form. Once the admin gets information of customers he comes to know how many customers are applicable for loan.
   - **Business Logic (Controller):** This consists of model building and prediction using two main algorithms these are Decision tree. Random Forest and Gradient Boosting. The pre-processed dataset will be used in training the model and testing it. First modelling will be done i.e. recursive feature elimination will be done. After Feature modelling, the feature selection will be done i.e. Chi-Square and then, modified decision tree will be applied. Random Forest is used combines multiple algorithms of same type that is multiple decision trees which further results in a forest of trees. Gradient Boosting is used for produces model of weak predicted values for ensemble learning usually decision trees. It also helps to improve our model accuracy.

3. **View:** View module mainly deals with the interaction between applicant who apply for loan and Administrator of our system.

6.METHODOLOGY

1. Decision tree

A decision tree is a graph to represent choices and their results in the form of a tree. The event in the graph is nothing but a node and the edges are decision rules. The name 'Decision Tree' tells us, that it builds a tree structure and learns the tree structure through the built model. Decision tree is just like a flowchart and it consists of logical decisions. Further the logical decisions split into branches that indicate choices. The combination of decisions is denoted by leaf nodes which are the termination point of decision tree.

2. Random Forest

Random forest is a supervised machine learning algorithm. It is based on ensemble learning. Ensemble Learning is a supervised machine learning algorithm. In ensemble learning you join different types of algorithms or same algorithm multiple times to form a powerful prediction model. The name "Random Forest" is given because, the algorithm combines multiple algorithm of same type that is multiple decision trees which further results in a forest of trees. This algorithm can be used for both regression and classification tasks. This algorithm uses bagging and feature randomness when building each individual tree trying to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Steps:

1. Pick N random records from the dataset.
2. Build a decision tree based on these N records.
3. Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
4. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output).The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

3. Gradient Boosting

In any machine learning technique if we try to predict the target variable there are main causes of difference in actual and predicted values. The main causes are noise, variance, and bias. Ensemble learning help to reduce the factors like noise, variance, and bias except noise. An ensemble learning is a collection of predictors that is it is the mean of all predictors. The mean of all predictors give a final prediction. We use Ensemble learning in which many predictors try to predict same target variable. These many predictors perform a better job than a single predictor to predict the target variable. Ensemble techniques are classified as Bagging and Boosting. In Boosting the consequent predictors learn from the mistakes of the preceding predictors.

Thus, the observations have an unequal probability of appearing in consequent models and ones with the superlative error appear the max. The predictors are chosen from different models like classifiers, decision trees, repressor, etc. The predictors are chosen from different models because new predictors learn from past mistakes committed by preceding predictors, this takes less iterations to reach to close actual predictors. But, choosing stopping criteria is very important here as it could lead to overfitting of data.

Gradient Boosting is an example Boosting algorithm. This algorithm is a technique for regression and classification problems. This technique produces model of weak predicted values for ensemble learning usually decision trees. The main objective of this algorithm is to define a loss function and minimize it.

7. RESULT ANALYSIS

1. Best Feature Selection

Feature Selection is the technique where we select best features from our dataset. Irrelevant can make model learn based on irrelevant features and also decrease the accuracy of the models.

There are various methods for feature selection we use Feature Importance method which gives score of each feature where the feature having highest score is more important towards our predicted variable. Here ApplicantIncome, LoanAmount, CoapplicantIncome, CreditHistory and Loan_Amount_Term are more important towards Loan_Status that is output variable.

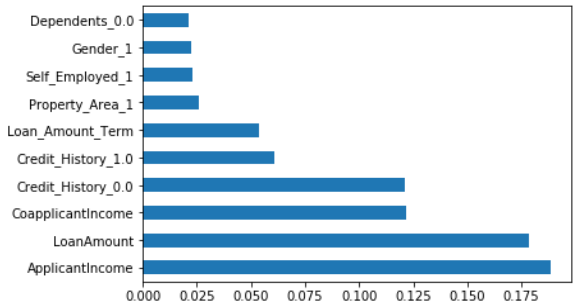

Fig -3: Best Feature Selection

2. Confusion Matrix

Confusion Matrix is an error matrix which shows the performance of classification model.Correct and incorrect predictions are summarized by using confusion matrix with their count.

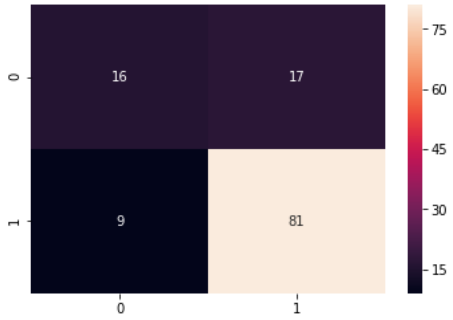

Fig -4: Confusion Matrix

Above figure shows confusion matrix for y_test and y_pred variables.

Table -2: Prediction Terms with Values

Prediction Terms	Value
True Positive (TP)	16
False Negative (FN)	17
True Negative (TN)	81
False Positive (FP)	09

3. Classification Report

The precision is the ratio of correctly classified true positive examples to the ratio of sum of true positive and false positive. For No (Bad) class it gives precision as 0.64 and for Yes (good) class it gives precision as 0.83.

The recall is the ratio of correctly classified true positive examples to the ratio of sum of true positive and false negative. For No (Bad) class it gives recall as 0.48 and for Yes (good) class it gives recall as 0.90.

F-score is the ratio of product of twice of recall and precision to the sum of recall and precision. For No (Bad) class it gives F-score as 0.55 and for yes (good) class it gives f-score as 0.86.

The support is the number of transaction in true response to the total number of responses. For No (Bad) class it gives support as 33 and for Yes (good) class it gives support as 90.

	precision	recall	f1-score	support
N	0.64	0.48	0.55	33
Y	0.83	0.90	0.86	90
micro avg	0.79	0.79	0.79	123
macro avg	0.73	0.69	0.71	123
weighted avg	0.78	0.79	0.78	123

Fig -5: Classification Report

4. ROC Curve

Performance measurement is an essential task in Machine learning. In classification problems we can count on an AUC-ROC curve. AUC-ROC curve is used to check/visualize the performance of multi-classification problem. To check classification model's performance AUC-ROC curve is an important metrics. AUROC is nothing but Area under the Receiver Operating Characteristics. If value of AUC is higher the model is better at predicting 0's as 0's and 1's as 1's.ROC curve represents TPR against FPR. A model which is near to 1 is known to be excellent model, which means it has a good measure of separability. AUC near to zero means a poor model, which means it has worst measure of separability. When AUC is 0.5 means there is no class separation capacity whatsoever. The True Positive Rate (TPR) is the ratio of correctly classified true positive examples to the ratio of sum of true positive and false negative. The False Positive Rate (FPR) is the ratio of incorrectly classified false positive examples to the ratio of sum of true negative and false positive.

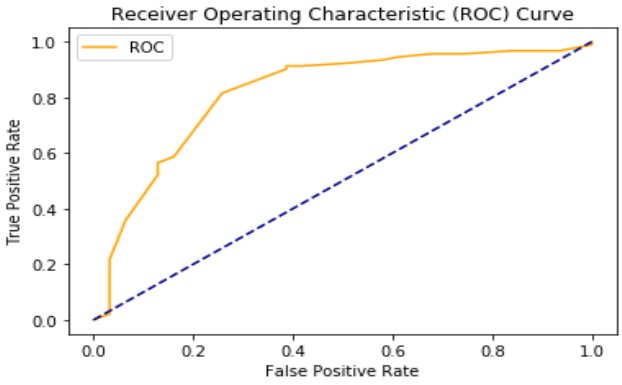

Fig -6: Receiver Operating Characteristic (ROC) Curve

5. Decision Tree

In below graph "Credit_History" is represented as a root node because it has highest entropy that is 0.908 which is nothing but measure of disorder for 491 samples from dataset. The "Credit_History" node has two more branches (Dependents_0.8 & Property_Area_1). Leaf node represents decision which is applicant is applicable or not for loan.

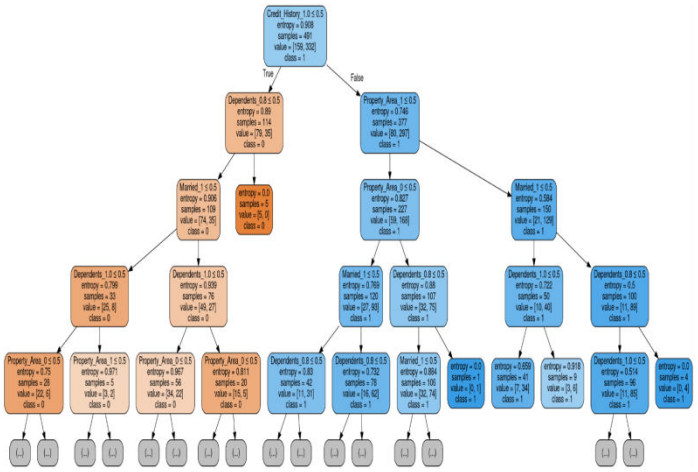

Fig -7: Decision Tree

8. CONCLUSION AND FUTURE WORK

In this project, we have proposed a credit score model to predict the loan applicant as a good applicant or bad applicant based on a credit score. In this model we have used different classifiers Random Forest, Decision Tree, Gradient Boosting is used to make comparison study. The decision tree algorithm in this model gives an accuracy of 80%. Further Gradient Boosting algorithm is used which increases the accuracy of our model. This model can be used in commercial applications and banks to classify their customers into good and bad applicants and approve those customers with loan.

REFERENCES

[1] S. Z. H. Shoumo, M. I. M. Dhruba, S. Hossain, N. H. Ghani, H. Arif and S. Islam, "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 2023-2028, doi: 10.1109/TENCON.2019.8929527.

[2] W. Qiu, S. Li, Y. Cao and H. Li, "Credit Evaluation Ensemble Model with Self-Contained Shunt," 2019 5th International Conference on Big Data and Information Analytics (BigDIA), Kunming, China, 2019, pp. 59-65, doi: 10.1109/BigDIA.2019.8802679.

[3] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

[4] A. Byanjankar, M. Heikkilä and J. Mezei, "Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach," 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 2015, pp. 719-725, doi: 10.1109/SSCI.2015.109.

[5] T. N. Pandey, A. K. Jagadev, S. K. Mohapatra and S. Dehuri, "Credit risk analysis using machine learning classifiers," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 1850-1854, doi: 10.1109/ICECDS.2017.8389769.

[6] R. S. Ramya and S. Kumaresan, "Analysis of feature selection techniques in credit risk assessment," 2015 International Conference on Advanced Computing and Communication Systems, Coimbatore, 2015, pp. 1-6, doi: 10.1109/ICACCS.2015.7324139.

[7] Y. Ma and H. Liu, "Research of SVM Applying in the Risk of Bank's Loan to Enterprises," 2010 2nd International Conference on Information Engineering and Computer Science, Wuhan, 2010, pp. 1-5, doi: 10.1109/ICIECS.2010.5678225.

[8] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[9] J. Zurada, "Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions?," 2010 43rd Hawaii International Conference on System Sciences, Honolulu, HI, 2010, pp. 1-9, doi: 10.1109/HICSS.2010.124.

[10] G. Sudhamathy and C. J. Venkateswaran, "Analytics using R for predicting credit defaulters," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, 2016, pp. 66-71, doi: 10.1109/ICACA.2016.7887925.

[11] A. Gahlaut, Tushar and P. K. Singh, "Prediction analysis of risky credit using Data mining classification models," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-7, doi: 10.1109/ICCCNT.2017.8203982.