International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Machine Learning Algorithms for Automation in Loan Prediction System

Prathamesh Khairnar
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India
pjkhairnar692000@gmail.com

Chetan Naphade
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India
chetannaphade859@gmail.com

Dr. Archana Kale
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India
archana.kale@mescoepune.org

Prasad Ghulanawar
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India
prasadghulanawar01@gmail.com

Aniket Patil
Computer Engineering
Modern Education Society's College of Engineering.
Pune, India
aniketpatilair600@gmail.com

Abstract: **In banking system, distribution of loans is the core business part of almost every bank. Banks have many sources of income and schemes to sell but main source of income of any banks is on the income received by giving loans to customers. So they can earn from interest of those loans which they credit. Technology has boosted the existence of the humankind the quality of the life they live. Every day we are planning to create something innovative and also plan for future so, for this we need money. With the advancement of technology, there are so many enhancements in the banking sector also. With the enhancement in the banking sector lots of people are applying for bank loans. Many people depend on bank loans for different purpose. Risk is constantly involved in approval of loans because banking officials are very acutely aware of the price of the mortgage quantity by its customers. Even after taking lot of precautions, approval choices are not correct every time. So, there is need of automation of this system so that loan approval will become less risky and incur less loss for banks because bank has its limited assets which it has to allocate to limited and deserving people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. Also, banks have to maintain their NPA(Non-Performing Asset) and CRAR(Capital to Risk Weighted Assets Ratio). So in this paper we are trying to reduce this risk factor behind selecting the right person so as to save lots of bank efforts and assets. This can be done via Machine Learning algorithms like Random Forest, Logistic Regression, Support Vector Machine, Naïve Bayes, KNN which provide higher efficiency in data classification. The main objective of this paper is to predict whether approving the loan to particular applicant will be the right choice or not using the Random Forest Algorithm.**

Keywords: **Credit line, Risk, Automation, Machine Learning Algorithms, NPA, CRAR**

I. INTRODUCTION

Most of the people depend on bank loans for different purposes like buying home, car, for education purpose. So they apply for loans in banks and provide their details. So, identifying the deserving applicant amongst all the applicants is very difficult task for banking officials and sometimes banking officials can be biased so there is need of automation in this loan approval system sector. With increasing population, the number of applicants are increasing day by day. So, for handling this huge information of candidates and identifying the correct applicant for approving loans can be done by using data science and machine learning techniques. As we are going to classify the large dataset as per the bank's criteria we need classifying algorithms. In those algorithms, all the supervised learning algorithms will be considered. It can classify data on the basis of bank's criteria and applicant's provided information.
Main steps of classification algorithm:

- Finalize the dataset that need to be classified.
- Pre-process data and train the training dataset using random forest algorithm.
- Apply that model on testing dataset.

After applying this trained model on test dataset, banks can take decision according to given output.
The prime objective of this paper is we have to use data classification technique to analyse train data and after that taking the decision on the basis of it. The main objective of paper is to predict whether the loan can be sanctioned or not. For selecting the best classification model, we compared different classification algorithms like Logistic regression, SVM, Decision tree, Random forest. Finally, we got highest accuracy in random forest. This automation will provide brief, quick, easy way to pick out the deserving applicants, maintain the cash flow in banking system, reduce NPA of banks.

II. LITERATURE SURVEY

Since few years, there is lot of research is going on classification algorithms that can be used to classify large

datasets. Some of the works that we have referred for this automation system.

An Approach for Prediction of Loan Approval using Machine Learning Algorithm:
Mohammad Ahmad Sheikh, Amit Kumar Goel, Tapas Kumar.
A bank's profit or a loss depends to a large extent on loans that means whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non-Performing Assets and can maintain CRAR. This makes the study of this phenomenon very important. Previous researches in this field has shown that there are so many methods to study the problem of controlling and identifying loan defaulters. But as the right predictions are very important for the maximization of profits, it is essential to study the different algorithms and compare their results. A very important approach in predictive analytics is used to study the problem of predicting loan defaulters: The Logistic regression model.

Loan Default Forecasting using Data Mining:
Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal
Estimation or assessment of default on a debt is a crucial process that should be carried out by every bank to help them to assess if a loan applicant can be a defaulter at a later phase so that they process the application and decide whether to approve the loan or not. The conclusion derived from such assessments helps banks and other financial institutions to reduce their losses and eventually increase the profit by giving loans to right person. Hence, it becomes very important to construct a model that will check for the the different aspects of an applicant's information and derive a result regarding the concerned applicant. Banks must check for the real purposes for demanding the loans. The increasing number of bad debts resulting from commercial banks' loans reflects in the failure of banks' economy and results in the imbalanced cash flow. We have used data mining algorithms to predict the likely defaulters from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future.

Prediction of Loan Status in Commercial Bank using Machine Learning Classifier:
G. Arutjothi, Dr. C. Senthamarai.
Banking systems always need a more accurate predictive modeling automation system which can take right decisions in short time. Predicting credit defaulters manually is a difficult task for the banking system. The loan status derived from analysis of applicant information is one of the quality indicators of the loan. It doesn't show everything immediately, but it is a first step of the loan lending process. The loan status is used for creating such automation systems. The credit scoring model is used for accurate analysis of credit data to find defaulters and valid customers. The objective of this paper is to create a credit scoring model for credit data. Various machine learning algorithms are used to develop the financial credit scoring model. In this paper, we propose a machine learning classifier-based analysis model for credit data. Paper uses algorithms like Min-Max normalization and K Nearest Neighbor(KNN) classifier, Decision tree classifier. The objective is implemented using the software package R tool.

Overdue Prediction of Bank Loans Based on LSTM-SVM, Random Forest:
Xin Li, Xianzhong Long, Guozi Sun, Geng Yang and Huakang Li.
In the aspect of bank loans, the accuracy of traditional user loan risk prediction models, such as KNN, Naïve Bayes, DNN, are not beneficial with the data growth. This article is based on the work of Overdue prediction of bank loans based on Deep Neural Network technique. And we propose to analyze the dynamic behavior of users by LSTM algorithm, and use the SVM algorithm to analyze the user's static information to solve the current prediction problems. This article uses user's basic information, bank records, user browsing behavior, credit card billing records, and loan time information to evaluate whether users are delinquent. These static datasets are the primary input for SVM technique. For LSTM model, we extract user's recent transaction type as input to LSTM, to predict the probability of users' overdue behavior. In last stage, we calculate the average of the two algorithms for getting the required decision. From the comparison of these predictions, LSTM-SVM model gives better efficiency than other traditional algorithms.

Prediction Defaults for Networked-guarantee Loans:
Dawei Cheng, Zhibin Niut, Yi Tu and Liqing Zhang.
Networked-guarantee loans may cause the systemic risk related concern of the government and banks in China.. Since the guaranteed loan is a debt obligation promise, if one enterprise in the guarantee network becomes defaulter or becomes financially weak then debt risk may spread like a virus across the guarantee network, even lead to a big financial crisis for bank and crisis for country's financial security. In this paper, we propose an imbalanced network risk diffusion model to pridict the enterprise default risk in a short future. Positive weighted k-nearest neighbors (pw-KNN) algorithm is developed for the stand-alone case – when there is no default contagious; then a data-driven default diffusion model is integrated to further improve the prediction accuracy. We perform the empirical study on a real-world three years' loan record from a major commercial bank. The results show that our proposed method outperforms conventional credit risk methods in terms of AUC.

III. Need For Classification

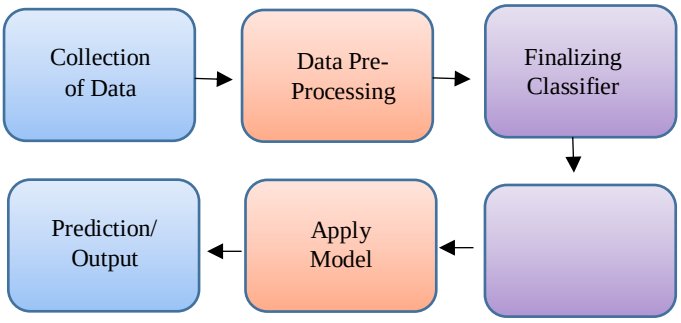
Classification algorithm is a supervised learning technique that is used to identify the category of new observations on basis of training data. It is very difficult to classify the large data by taking all attributes in consideration. So, classification algorithms provide quick classification and provides accurate results in very less time. So, in data science and machine learning, various classifiers like decision tree, logistic regression, random forest, SVM, etc. are provided to

complete classification of large dataset having many attributes.

IV. Architecture Diagram and Process

1. *Input Data:*
   Download train and test dataset consisting the loan applicant's information such as application id, name, loan amount, income, co-applicant income, gender, service years. This data is stored in csv files. These csv files can be downloaded from kaggle website.

2. *Pre-Processing:*
   The csv file contains some null values and irrelevant information which needs to be cleansed. The iloc() and shape() functions can be used to remove the null values. Noise cancellation and data cleansing is done in this step.



Architecture Diagram

3. *Finalize Classifier:*
   Compare all the classification algorithms and after comparing, finalize the appropriate algorithm on basis of efficiency. So, here we finalize the random forest algorithm.

4. *Train Model:*
   Train the dataset using finalized random forest classifier.

5. *Apply Model:*
   Apply this built model on test dataset.

6. *Output:*
   Classify the applicants based on the applicant's information and bank's criteria using random forest classifier.

V. Machine Learning Algorithms

1) *Random Forest:*
Random forest is a supervised learning algorithm for classification purpose. Random forest is a classifier that contains a number of decision trees on various subsets of given dataset and takes average to improve the classification accuracy. Random forest takes the output from each tree in it and based on the maximum votes on predictions, it classifies data. In this classifier, more the number of trees, more is the accuracy because every attribute is taken into consideration while classifying.
It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing.

2) *Decision Tree:*
Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

3) *Support Vector Machine:*
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.

4) *Logistic Regression:*
Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

VI. CONCLUSION

Data science, machine learning and the Artificial Intelligence are the rising technologies in this modern world. The more and more research is going on to make it efficient. In this system we referred 4 to 5 algorithms and finalized random forest algorithm which provides highest efficiency among all algorithms that we have referred. More research is going on the other algorithms also which can be used or added to the system to make it more efficient.

VII. REFERENCES

[1]. Yadav, O. P., Soni, C., Kandakatla, S. K., Sswanth, S. (2019). Loan prediction using decision tree. International Journal of Information and computer Science, 6(5).

[2]. Arutjothi, G., Dr. Senthamarai, C. (2017). Comparison of feature selection methods for credit risk assessment. International Journal of Computer Science, 5(I), No 5.

[3]. Aida Krichene," Using a naive Bayesian classifier methodology for loan risk assessment," Journal of Economics, Finance and Administrative Science, 2017.

[4]. Pidikiti Supriya, Myneedi Pavani, Nagarapu Saisushma,Namburi Vimala Kumari, k Vikash, "Loan Prediction by using Machine Learning Models", International Journal of Engineering and Techniques.Volume 5 Issue 2, Mar-Apr 2019.