IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Bank Customer Credit Approval using Predictive Analytics Techniques

Ramanjinamma G
Assitant Professor
Department of Computer Science and Engineering
Sai Vidya Institute of Technology, Bengaluru
Visvesvaraya Technological University, Belagavi

Sowmya H N
Assitant Professor
Department of Computer Science and Engineering
Sai Vidya Institute of Technology, Bengaluru
Visvesvaraya Technological University, Belagavi

Deepika G
Assistant Professor
Department of Computer Science and Engineering
Sai Vidya Institute of Technology, Bengaluru
Visvesvaraya Technological University, Belagavi

Yashaswini D M
Assistant Professor
Department of Computer Science and Engineering
Sai Vidya Institute of Technology, Bengaluru
Visvesvaraya Technological University, Belagavi

Abstract - One of the major services provided by banks and other financial institutions is loan lending. Many bank customers seek loan from banks and approving loan requires various manual procedure checks. Loan approval requires the bank to check whether the customer is reliable and if the customer can pay it back within a stipulated time. A small manual error in this procedure can result in a huge amount of loss to the bank. Therefore making use of some automated techniques is more reliable. In this paper we make use of machine learning algorithm to predict whether the loan can be given to a particular customer or not, we initially look for the required features from the dataset and for the selected data we apply machine learning algorithm. This paper concentrates on Learning Vector Quantization algorithm for prediction. LVQ can be used for both classification and Regression Problems, in case of classification it is applicable to both binary (two class) and multiclass classification problems.

Key words - Loan, Machine Learning, Learning Vector Quantization.

I. INTRODUCTION

Bank Customer Credit Approval using Predictive Analytics Techniques involves the application of advanced algorithms to analyze historical data and predict the probability of a borrower defaulting on a loan. This process begins with the collection and preprocessing of relevant features such as credit score, income, and debt-to-income ratio. The supervised label based machine learning models, such as logistic regression or decision trees, are then trained on datasets, where past loan outcomes serve as the basis for prediction. These models learn previous patterns by the dataset and to make accurate predictions about whether a loan applicant is likely to repay the loan or pose a higher risk of default. This predictive capability enables banks to make informed and data-driven decisions, optimizing their lending processes by efficiently assessing creditworthiness and minimizing the potential for financial losses.

In compliance with the regulatory requirements enforced by the government or its agencies, the bank adheres to its own internal guidelines. Loan eligibility requirements, loan types to be offered, loan terms, loan security, and procedures are all

included in the lending guidelines. Not everyone meets the lending requirements set forth by the particular lending organization in order to be eligible for a loan. Lending organizations examine the client's financial statements during a loan analysis to ascertain the client's financial stability and capacity to repay the loan without difficulty.

Table 1 - List of attributes with its data type

ID	Integer value
Age	Integer value
Experience	Integer value
Income	Integer value
Zip code	Integer value
Family	Integer value
CCAvg	Floating point value
Education	Integer value
Mortgage	Integer value
Personal Loan	Integer value
Securities account	Integer value
CD Account	Integer value
Online	Integer value
CreditCard	Integer value

This concentrates on the statistics of previous clients of various banks to whom on a hard and fast of parameters mortgage have been approved. In order to obtain reliable results, the machine learning model is trained on that record. The main objective of this work is to predict whether the loan can be given to a particular customer or not. To predict loan approval, learning vector quantization algorithm is used. First step is to clean the data so that we can avoid the missing values in the data set. To train our model data set of 5000 records and 14 numerical attributes has been taken. To approve a loan to a customer various parameters such as age, experience, income etc., has been considered. List of attributes are recorded in the table 1.

II. LITERATURE SURVEY

Below is a description of relevant studies about the use of machine learning on loan approval and financial data:

Bank loan prediction is a critical task in financial institutions, and the use of classifiers in this context has garnered significant attention in the literature. Various machine learning and statistical techniques have been employed to enhance the accuracy of loan prediction models. Several studies emphasize the importance of feature selection and engineering in improving the predictive performance of loan classifiers. Commonly considered features include credit score, income, debt-to-income ratio, employment history, and other financial indicators. Feature scaling and normalization techniques are often employed to ensure that each feature contributes appropriately to the classifier's decision-making process.

Loan approval was analyzed using logistic regression models, and distinct performance metrics were calculated. Based on overall performance metrics like sensitivity and specificity, these models are compared. The version produces unique effects, as demonstrated by the most recent affects. The model is only slightly better because it includes variables (private attributes of the borrower such as age, purpose, credit score history, credit score amount, credit score duration, etc.) in addition to bank account information (which implies the borrower's wealth) that must be taken into account in order to accurately calculate the likelihood of a mortgage default. The version comes to the conclusion that a bank must now do more than just target wealthy customers when issuing mortgages; it must also ascertain

The credit score dataset from financial institutions is examined using machine learning techniques to assess consumers' creditworthiness and loan-paying potential. The financial institution credit score dataset is read using various machine learning methods. The results of the test indicate that other than the Nearest Centroid and Gaussian Naive Bayes, they outperform in terms of accuracy and several metrics related to overall performance evaluation. The accuracy charges for each of the methods ranged from 76% to more than 80%. Furthermore, the most important factors that influence a customer's credit score worthiness are taken into account. Then, these maximal critical functions are applied to a few selected algorithms, and their total accuracy of performance is compared with the example of applying all

The model makes a prediction about how safe it is to assign a loan to a specific consumer. Because there are missing values in the dataset, the decision tree technique and Python data cleaning are used to

implement this loan prediction problem. For the missing values, the map function is employed. A bank customer's eligibility for a loan is determined by a number of factors, including credit history and installment payments, among others. This methodology offers a quick, easy, and instant method for selecting loan applicants who deserve the money. The bank benefits in a number of ways from the model. Every feature involved in loan processing has its weight automatically determined by the Loan Prediction System, and the same attributes are handled based on their associated weight when processing new test data.

III. METHODOLOGY

The methodology typically begins with the collection of historical data pertaining to loan applicants, encompassing diverse features such as income, credit score, employment status, debt-to-income ratio, loan amount, and loan term, along with corresponding repayment outcomes. Subsequently, data pre-processing steps are employed to clean the data, addressing issues like missing values, outliers, and inconsistencies, while converting categorical variables into numerical representations through techniques such as one-hot encoding or label encoding. Normalization or scaling of numerical features is often performed to ensure uniformity in scale across variables. Following this, feature engineering techniques may be applied to derive new features or transform existing ones, aiming to enhance the predictive capability of the model. This could involve creating ratios, binning numerical features, or extracting pertinent information from text fields.

Through the analysis of numerous characteristics, the suggested model seeks to forecast bank clients' creditworthiness for loan payback. The list of features from the dataset is the input that the model receives. The dataset contains 14 attributes in total and contains all numerical values. The output from the classification algorithm is, predicting whether or not to approve the customer request to lend loan. In this proposed method the model is need to be trained with customer data to predict the approval of loan. The training dataset is then provided to machine learning model and on upon this dataset the model is trained. Each new customer's details filled at the time of application form acts as test data set. After the process of testing, the model selected predicts whether the loan can be approved for the particular customer or not.

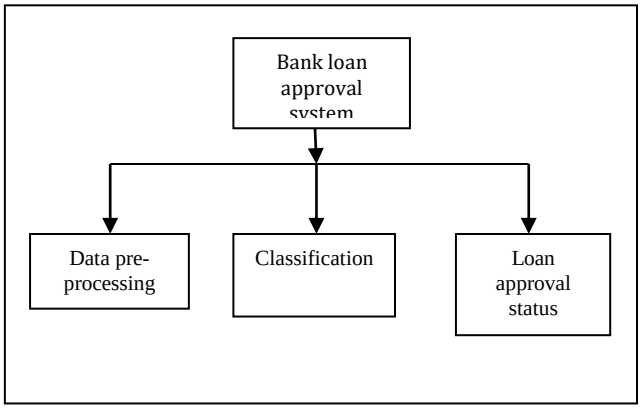

Figure 1 - Structural chart of proposed model

A. Dataset Description

In the context of machine learning and data analysis, a dataset is often used to train, test, or evaluate algorithms. For example, in a bank loan prediction task, a dataset might include information about various applicants, such as their credit scores, income, and employment history, along with a label indicating whether each applicant was approved or denied a loan. The algorithm uses this dataset to learn patterns and relationships within the data, enabling it to make predictions on new, unseen data.

Table 2 - Description of attributes in the dataset

ATTRIBUTES	DESCRIPTION
ID	Customer ID
Age	Customer's age in years
Experience	professional experience in Number of years
Income	Customer Annual income
ZIP Code	PIN code of city
Family	Number of persons in family
CCAvg	Average spending on credit cards per month
Education	Educational qualification of customer
Mortgage	Value of house mortgage if any
Personal Loan	Has the customer received loan or not?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the client have an account with the bank for a certificate of deposit (CD)?
Online	internet banking usage
CreditCard	Bank Credit card usage

B. Data Pre-processing

In the data pre-processing stage we look for missing values in the dataset. In the dataset taken there were no missing values and the plot below shows the check of missing values.

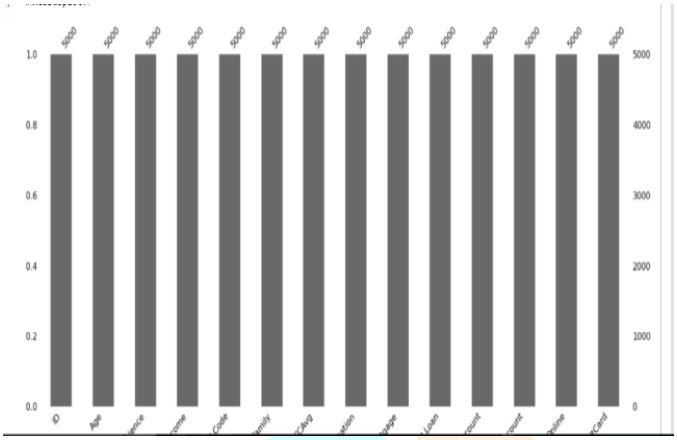

Figure 2 - Plot showing no missing values in the given dataset

C. Classification

In the process of classification, we have used three machine learning algorithms. The first one is K-NN (K- nearest neighbors) algorithm, the second one is Random forest algorithm and the last one is learning vector quantization (LVQ) algorithm. The detail of each algorithm is given below:

1. K-NN algorithm

K-Nearest Neighbour is one of the best Machine Learning algorithms primarily based on Supervised Learning technique. This algorithm considers the similarity among the new instance/records and to the available instances and places the new instance into the class that is most similar to the available categories. K-NN algorithm may be used for Regression in addition to this it can also be used for Classification however primarily it's used for the Classification problems. This algorithm is a non-parametric one, which implies that it does not work on assumption of the underlying data. It is also referred to as a lazy learner algorithm as it does not learn from the training dataset immediately, instead it stores the dataset and it performs an action on the dataset at the time of classification.

2. *Random Forest Classifier*

Random Forest is a well-known machine learning classifier that comes under the supervised learning technique.This classifier may be used for either Classification or Regression problems in Machine Learning. It works on the concept of collective learning, which is a process of gathering multiple classifiers to find a solution for a complex problem and to enhance the performance of the model. It has a number of decision trees on different subsets of the dataset and calculates the average to enhance the accuracy of prediction of that dataset. It not only depends on one decision tree, but it also takes the prediction from each decision tree and based on the majority of outcomes of predictions, it then predicts the final output.

3. *Learning vector Quantization (LVQ)*

The Learning Vector Quantization (LVQ) belongs to an artificial neural network algorithm that helps us to pick out what number of training instances to hold onto and learns precisely what those instances must appear like. LVQ can be used for both classification and Regression Problems, in case of classification it is applicable to both binary (two class) and multiclass classification problems.

Linear Vector Quantization (LVQ) is a data compression and clustering technique that operates on a given dataset by representing it using a set of representative vectors known as codewords. In the context of LVQ, the term "linear" pertains to the adjustment process during training, where the codewords are updated linearly based on the input vectors. In essence, LVQ refines the codewords to capture the inherent structure and patterns within the dataset. This involves iteratively presenting input vectors to the algorithm, selecting the most similar codeword, and adjusting it linearly to better match the input. This training process results in codewords that effectively represent clusters in the dataset, facilitating tasks such as data compression or classification. LVQ's application to a dataset allows for efficient representation and understanding of complex data distributions through the adaptation of codewords in a linear fashion.

The results of the above algorithms are compared and the best classifier is selected to

predict the loan approval analysis. The below figure 3 shows the comparison plot of the three algorithms.

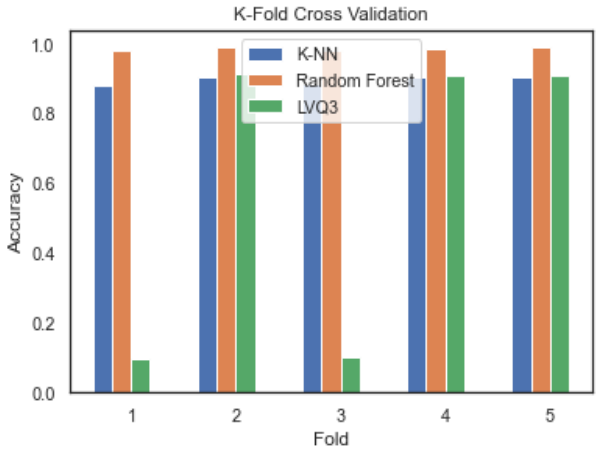

Figure 3 - Plot showing the accuracy comparison of LVQ, K-NN and Random Forest

IV. EVALUATION MERTRICS USED

The Random Forest classifier, a popular machine learning algorithm, Because of its ensemble nature, the Random Forest classifier, a well-liked machine learning method, produces promising results in a range of applications. Random Forest prevents overfitting and boosts model robustness by building several decision trees during training and combining their predictions. This makes it especially appropriate for classification jobs as it results in better generalization performance on unknown data. The algorithm excels in handling large datasets with numerous features, and its ability to rank features based on their importance provides valuable insights into the underlying data patterns.

1) Accuracy Score:

The term Accuracy Score is generally used for Defining Classification Accuracy.It refers to the ratio between the number of right predictions to that of the number of total instances available.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

(1)

2) Precision Score:

It is the ratio of number of right positive outcomes to that of the number of positive outcomes predicted by the classifier.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

(2)

3) Recall Score:

It is the ratio of number of right positive outcomes to that of number of all the relevant instances available.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

(3)

4) F1 Score

F1 Score is the parameter that is required to measure the test's accuracy.

The Harmonic Mean between precision and recall is termed as F1 score. Its value ranges between 0 to 1. It specifies how accurate and robust the classifier is

$$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

(4)

The values of the Evaluation metrics of the classifiers used are as follows:

The precision score, accuracy score, F1 score and the recall score of the K-NN classifier is given below:

- Precision score = 0.8016217777777778
- Recall score = 0.8953333333333333
- Accuracy score = 0.8953333333333333
- F1 score = 0.8458900222769375

The precision score, accuracy score, F1 score and the recall score of the LVQ classifier is given below:

- Precision score = 0.8479701694915254
- Recall score = 0.892
- Accuracy score = 0.892
- F1 score = 0.8553646183482947

The precision score, accuracy score, F1 score and the recall score of the random forest is given below:

- Precision score = 0.9866566467834609
- Recall score = 0.9866666666666667
- Accuracy score = 0.9866666666666667
- F1 score = 0.9863482017628835

Therefore from the above results it is quite clear that among the three classifiers mentioned, the random forest classifier has higher accuracy and precision score. Hence random forest classifier is used to train the model and make the loan analysis prediction.

V. CONCLUSION

In this paper, machine learning technique is used to study bank loan dataset so as to predict customer's loan approval and their ability to repay the loan if taken. If the predictive model demonstrates high accuracy, precision, and recall rates, it indicates that the model is effective in predicting whether a loan applicant is likely to default or not. This could be valuable for banks in making informed decisions about loan approvals and managing credit risk. Additionally, insights gained from the predictive model can inform banks about the key factors influencing loan default and help in developing strategies to mitigate risks associated with lending. We have used different machine learning algorithms on this dataset so as to determine which algorithm is the best one for the purpose of predicting bank loan approval. With a 98% accuracy rate, the random forest algorithm proved to be the most suitable model for predicting the approval status of bank loans based on evaluation metrics such as accuracy, precision, recall, and F1 scores that were computed for each algorithm in use.

REFERENCES

[1] C. Prasanth, R. P. Kumar, A. Rangesh, N. Sasmitha and D. B, "Intelligent Loan Eligibility and Approval System based on Random Forest Algorithm using Machine Learning," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 84-88, doi: 10.1109/ICIDCA56705.2023.10100225.

[2] Anshika Gupta, Vinay Pant, Sudhanshu Kumar, and Pravesh Kumar Bansal, "Bank loan prediction system using machine learning," 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART).

[3] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[4] V. E, P. Ravikumar, C. S and S. K. M, "An Efficient Technique for Feature Selection to Predict Customer Churn in telecom industry," 2019 1st International Conference on Advances in Information Technology (ICAIT), Chikmagalur, India, 2019, pp. 174-179, doi: 10.1109/ICAIT47043.2019.8987317.

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

[6] 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, India, 2017, pp. 416-419, doi: 10.1109/ISS1.2017.8389442.

[7] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier," *International Conference on Intelligent Sustainable Systems (ICISS)*, 2017.

[8] Ashlesha Vaidya, "Predictive and Probabilistic approach using logistic regression: Application to prediction of loan approval," 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2017.

[9] Gowda, Madhu Belur Gopala, Naveen Kumar Boraiah, Varun Eshappa, and Gopala Krishna

Chandra Shekara. "Classification of Epileptic EEG Signals Using Improved Atomic Search Optimization Algorithm." International Journal of Intelligent Engineering & Systems 16, no. 6 (2023).