

Predictive Analysis in Banking using Machine Learning

Gupta Ashwin Arunkumar¹, Chaurasiya Ravi Panchuram¹, Khambati Mohammed Aadil Afzal¹, Niraj Sureshchand Yadav¹, Uma Goradiya²

¹Department of Computer Engineering, Shree LR Tiwari College of Engineering, Mumbai, Maharashtra, India

²Assistant Professor Department of Computer Engineering, Shree LR Tiwari College of Engineering, Mumbai, Maharashtra, India

ARTICLE INFO

Article History:

Accepted: 01 April 2023

Published: 21 April 2023

Publication Issue

Volume 10, Issue 2

March-April-2023

Page Number

434-439

ABSTRACT

This paper discusses the utilization of machine learning for predicting loan approval and credit card fraud detection. Specifically, the paper proposes the use of the Random Forest Algorithm and Support Vector Machine Learning Algorithm for achieving better accuracy. The banking sector's main objective is to ensure their assets are in safe hands, and to achieve this, a verification process is carried out. However, the process takes a long time, and there is no guarantee of selecting deserving applicants. To address this problem, a system has been developed that predicts the suitability of an applicant for loan approval based on a model trained using machine learning algorithms. The system has achieved 92% accuracy using the Random Forest Algorithm. The system has a user interface web application where users can input necessary details for the model to predict. The system's drawback is that it considers multiple attributes, whereas, in real life, a loan application may be approved based on a single strong attribute, which the system cannot detect. With the increasing number of online transactions, credit card usage has become more prevalent. Losing physical credit cards or credit card information can result in significant financial loss. Therefore, there is a need to detect fraudulent transactions and secure them. To address this issue, the paper proposes the use of the Support Vector Machine Algorithm, which focuses on analyzing and preprocessing data sets and deploying fraud detection using Credit Card Transaction data.

Keywords: Machine Learning (ML), Random Forest, Loan Approval, Support Vector Machine Algorithm, Credit Card Fraud

I. INTRODUCTION

The credit line is the primary source of income for banks, as they earn interest on the loans they credit. In the market economy, banks play a crucial role, and their profits or losses are significantly influenced by the repayment or default of loans. To determine whether a borrower is a good (non-defaulter) or bad

(defaulter) before granting a loan is a crucial decision for banks. The credit risk refers to the likelihood of borrowers failing to meet their loan obligations. Determining whether a borrower will be a good or bad one is a difficult task for any organization. The manual process used by the banking system to check the creditworthiness of borrowers is accurate and effective. However, this process becomes ineffective

when there are a large number of loan applications to be processed simultaneously. In such cases, decision-making takes a long time, and a significant amount of manpower is required. Classifying borrowers as good or bad based on their ability to repay debts can be accomplished using machine learning algorithms. This approach can help applicants and bank employees by predicting loan outcomes.

In today's business landscape, companies across the globe are experiencing rapid growth and striving to deliver superior customer service. To achieve this, companies process vast amounts of data every day, including customers' personal and financial information. As a result, data security has become crucial for these companies. Online shopping has also gained popularity, making it easier for criminals to carry out malicious activities such as Trojan and pseudo-based station attacks. Fraudulent events have become a significant concern, especially when criminals steal cardholder's information. Machine learning and computational intelligence communities have proposed numerous automation solutions to tackle the issue of credit card fraud detection. Data is available all over the world, and organizations of all sizes are loading information with high volume, variety, speed, and value. Credit cards are a lucrative and straightforward target for fraudsters, enabling them to obtain significant amounts of money within a short period without any risk. This paper utilizes an online banking transaction repository dataset, training the dataset to analyse and classify transactions as fraudulent or normal.

II. LITERATURE

Several related works are done in the field of Loan Approval Prediction and Credit Card Fraud Detection. [1] The study utilized only one algorithm without comparing it with other algorithms. Specifically, the Logistic Regression algorithm was used, and the highest accuracy achieved was 81.11%. Based on the

results, it was concluded that loan applicants with good credit scores, high income, and low loan amount requirements have a higher likelihood of loan approval.

[2] Logistic Regression, Support Vector Machine, Random Forest, and Extreme Gradient Boosting algorithms were utilized in this paper. The accuracy percentages were similar for all the algorithms, but the support vector machine had the lowest variance. Having lower variance means there will be less fluctuation of scores, and the model will be more accurate and consistent.

[3] The objective of this paper is to investigate bank loan prediction and propose a model that utilizes machine learning algorithms such as Support Vector Machines (SVM) and Neural Networks. The study aims to determine how banks can make decisions in approving loans.

[4] The comparison of two machine learning algorithms was reported in this paper. The two algorithms evaluated were the two-class decision jungle and the two-class decision tree, which achieved accuracies of 77.00% and 81.00%, respectively. In addition to accuracy, precision, recall, F1 score, and AUC were also computed.

[5] The paper solely employs the K-Nearest Neighbor Classifier. To standardize the attribute values, the researchers utilized the Min-Max Normalization technique, which involves decomposing the attribute values. They achieved a maximum accuracy of 75.08% when the dataset was divided equally between training and testing sets and k was set to 30.

[6] The paper evaluated different machine learning algorithms to detect credit card fraud in real-time. The results show that Support Vector Machines (SVM) had the highest accuracy of 91%, while K-Nearest Neighbors (KNN) had an accuracy of 72%, Logistics Regression achieved 74% accuracy, and Naive Bayes had 83% accuracy. SVM and Naive Bayes are promising algorithms for credit card fraud detection, but the effectiveness of these algorithms may depend on various factors.

[7] This paper evaluated that used a dataset from Kaggle and employed various machine learning algorithms to predict credit card defaults. The study found that the Naive Bayes model had the highest accuracy of 82%, followed by the SVM and Regression models, both with an accuracy of 81%. The Logistic model had an accuracy of 80%, and the KNN and Random Forest models had an accuracy of 79%. These results suggest that machine learning techniques can be useful for predicting credit card defaults.

[8] This paper compares the performance of various supervised machine learning algorithms for credit card fraud detection. The authors use a publicly available dataset and evaluate the performance of several algorithms, including decision tree, k-nearest neighbor (KNN), logistic regression, random forest, support vector machine (SVM), and artificial neural network (ANN). They compare the accuracy, precision, recall, F1-score, and area under the curve (AUC) for each algorithm and find that SVM and ANN perform the best. The authors conclude that machine learning algorithms can be effective in detecting credit card fraud and that SVM and ANN are promising methods for this task.

III. DATASET

Table -1: Dataset Variables and their Description for Loan Approval Prediction

Variable Name	Description
Loan_ID	Unique ID
Gender	Male/Female
Marital_Status	Applicant Married (Yes/No)
Dependents	Number of Dependents
Education_Qualification	Graduate/Undergraduate
Self_Employed	Self-Employed (Yes/No)
Applicant_Income	Applicant Income
Co_Applicant_Income	Co-Applicant Income
Loan_Amount	Loan Amount in lacs
Loan_Amount_Term	Term of loan in months
Credit_History	Credit History meets guidelines
Property_Area	Urban/Sem-Urban/Rural
Loan_Status	Loan Approved (Yes/No)

The dataset consists of 615 rows and 13 columns, with 12 columns serving as independent variables and 1 column serving as the dependent variable. Some of the entries in the dataset are null, which have been filled using the mean and mode method. Additionally, the label encoder has been utilized to convert string values into binary values of 1 and 0. The attribute "Loan_Status" is the target variable.

Variable Name	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit
GENDER	Male/Female
EDUCATION	Graduate School/University/High School/Others
MARITAL_STATUS	Married/Single
AGE	Age in Years
PAY_0	Repayment Status in September, 2005
PAY_2	Repayment Status in August, 2005
PAY_3	Repayment Status in July, 2005
PAY_4	Repayment Status in June, 2005
PAY_5	Repayment Status in May 2005
PAY_6	Repayment Status in April, 2005
BILL_AMT1	Amount of bill statement in September, 2005
BILL_AMT2	Amount of bill statement in August, 2005
BILL_AMT3	Amount of bill statement in July, 2005
BILL_AMT4	Amount of bill statement in June, 2005
BILL_AMT5	Amount of bill statement in May, 2005
BILL_AMT6	Amount of bill statement in April, 2005
PAY_AMT1	Amount of previous payment in September, 2005
PAY_AMT2	Amount of previous payment in August, 2005
PAY_AMT3	Amount of previous payment in July, 2005
PAY_AMT4	Amount of previous payment in June, 2005
PAY_AMT5	Amount of previous payment in May, 2005
PAY_AMT6	Amount of previous payment in April, 2005
DEFAULT_PAYMENT	YES/NO

Table -2: Dataset Variables and their Description for Credit Card Fraud Prediction

Details of dataset values-

Scale for PAY_0 to PAY_6: (-2 = No consumption, -1 = paid in full, 0 = use of revolving credit (paid minimum only), 1 = payment delay for one month, 2 = payment delay for two months, ... 8 = payment delay for eight months, 9 = payment delay for nine months and above) We have records of 30000 customers. In our dataset we got customer credit card transaction history for past 6 month, on basis of which we have to predict if customer will default or not.

IV. METHODOLOGY

The proposed system comprises a web application that utilizes a machine learning model to predict loan approval. The model is trained using various machine learning algorithms, and it is deployed in the web

application. The user is required to fill out an 11-field form, which corresponds to the 11 attributes used to train the model. Before training the model, the dataset undergoes pre-processing to replace null values using the mean and mode method and to convert string values to binary using label encoding. The pre-processed dataset is then split into two parts, with 80% being used for training and 20% for testing the model's accuracy using various algorithms. The Random Forest algorithm was applied to the dataset after splitting, yielding an accuracy of 92%. Once the model is trained, a pickle file is created. To predict loan approval, the user must first complete the form on the web application and click the "MAKE PREDICTION" button. Based on the trained model or pickle file, the system predicts whether the loan will be approved or not. This system streamlines the loan approval process for banks and organizations.

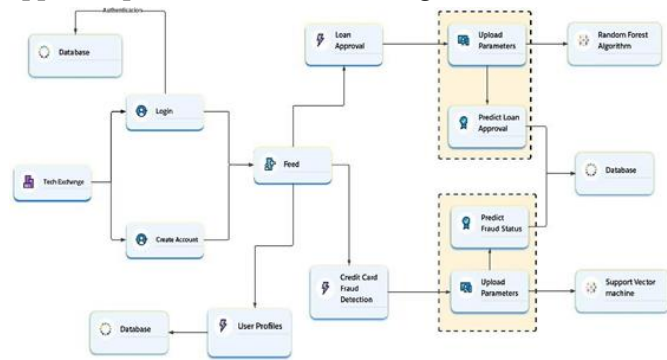


Figure 1. Architecture Diagram of TechExchange

Also, the proposed system is a web application that utilizes a machine learning model to detect credit card fraud. The dataset for this project was obtained from Kaggle, a data analysis website that provides datasets. This dataset contains 30,000 customer credit card transaction histories for the past six months, which was used to train the model. Users are required to fill out a form on the web application to predict credit card fraud transactions. The dataset was then pre-processed and divided into two parts, with 90% for training and 10% for testing the model's accuracy using various algorithms. The Support Vector Machine algorithm was applied to the dataset after splitting, resulting in an accuracy of 90%. After the

model is trained, a pickle file is created, which is used to predict credit card fraud transactions. When a user wants to predict a fraudulent transaction, they must complete the web application form and click the "MAKE PREDICTION" button. The trained model is then used to determine if the transaction is fraudulent or not.

V. IMPLEMENTATION

A. RANDOM FOREST ALGORITHM

The random forest classifier is a popular and powerful algorithm in machine learning. It is composed of multiple individual decision trees that work together to make decisions. This algorithm is known for its simplicity and diversity, which enables it to perform both classification and regression tasks.

How Random Forest Algorithm works:

The random forest algorithm begins by creating multiple decision trees. The final decision of the tree is based on the majority of the trees, which are selected randomly. A decision tree is a tree diagram that focuses on determining the course of action, where each branch represents a possible decision, occurrence, or reaction. When there are many subtrees in the forest, this algorithm is used to avoid overfitting of the model and reduce the time required for training. Additionally, it provides highly accurate results and can operate productively on large databases, predicting missing data. Steps of Random Forest Algorithm:

1. Select N random records from the dataset.
2. Use these N records to create a decision tree.
3. Specify the number of trees you want to include in the algorithm and repeat steps 1 and 2
4. For a new record in a regression problem, each tree in the forest predicts a value for the output variable Y.

B. SUPPORT VECTOR MACHINE ALGORITHM

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification and regression tasks. It is a powerful

algorithm that is often used in research papers to solve complex problems. The primary objective of SVM is to find a hyperplane that best separates the data into different classes. The hyperplane is chosen in such a way that the margin, which is the distance between it and the closest points from each class, is maximized. The points that lie closest to the hyperplane are called support vectors, which gives the name to the algorithm.

The SVM algorithm works by transforming the data into a higher-dimensional space using a kernel function. This allows the algorithm to find a hyperplane that can separate the data in a nonlinear way. The most commonly used kernel functions are linear, polynomial, and radial basis function (RBF).

The training of the SVM model involves minimizing a loss function that penalizes misclassifications and maximizes the margin. This is done by solving a quadratic optimization problem, which can be computationally expensive for large datasets. However, there are several optimization techniques that can be used to speed up the process, such as stochastic gradient descent and sequential minimal optimization.

Once the SVM model is trained, it can be used to predict the class of new data points by calculating their distance to the hyperplane. If the distance is greater than a certain threshold, the data point is classified as belonging to one class, otherwise, it is classified as belonging to the other class.

SVM has been widely used in research papers for classification tasks, such as image and text classification, as well as regression tasks, such as predicting stock prices and house prices. SVM has also been used in combination with other techniques, such as ensemble methods and feature selection, to improve the accuracy and efficiency of the model.

Overall, SVM is a powerful algorithm that can be used to solve a variety of machine learning problems. Its ability to handle both linear and nonlinear data, as well as its high accuracy and flexibility, make it a popular choice for researchers and practitioners alike.

VI. RESULT

This research paper presents a new web application that utilizes machine learning algorithms to predict loan approvals and detect fraudulent transactions made through credit cards. The system evaluates input data provided by the user to predict loan approval status and detect potential fraud. The system was evaluated on a training dataset consisting of 615 rows, resulting in a 92% accuracy rate for loan approval predictions. The system was also tested on a training dataset of 30,000 customers, with a resulting accuracy rate of 94% for credit card fraud detection.

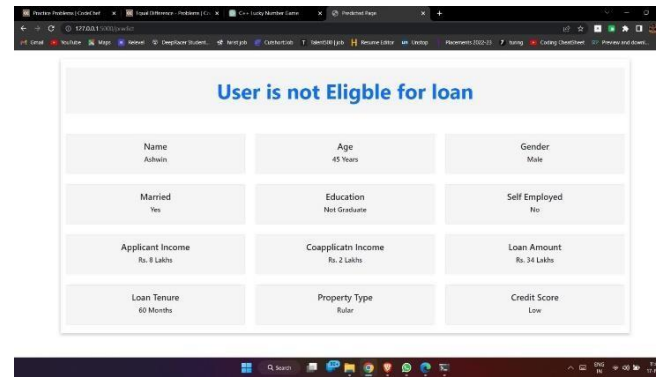
Normal Transaction

Credit Card Fraud Detection

User is Eligible for loan

Name ASHWIN GUPTA	Age 22 Years	Gender Male
Married Yes	Education Graduate	Self Employed Yes
Applicant Income Rs. 15 Lakhs	Coapplicant Income Rs. 12 Lakhs	Loan Amount Rs. 3 Lakhs
Loan Tenure 60 Months	Property Type Rural	Credit Score Medium

Loan Approval Prediction



VII. CONCLUSION

In this project, to predict the approval status of a loan application, we opted for a machine learning approach using the bank dataset. We evaluated Random Forest Machine Learning algorithms provided the accuracy of 92%. We also identified the most important features that affect loan approval

status. This model can help banks understand the crucial factors for loan approval.

The issue of credit card fraud is a significant concern for businesses, as it can result in substantial losses. With the increasing use of credit cards for online transactions, the number of credit card fraud cases has also risen. Therefore, there is a pressing need to detect fraudulent transactions and ensure the security of credit card users. The primary goal of this paper is to prevent and detect fraudulent transactions during credit card transactions. While previous systems have faced issues with accuracy, this paper examines the performance of the Support Vector Machine Algorithm, which has been shown to produce good results for both small and large data sets.

VIII. REFERENCES

- [1]. Sheikh, Mohammad & Goel, Amit & Kumar, Tapas. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. 490-494. 10.1109/ICESC48915.2020.9155614.
- [2]. Shoumo, Syed Zamil Hasan, Mir Ishrak Maheer Dhruba, Sazzad Hossain, Nawab Haider Ghani, Hossain Arif and Samiul Islam. "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking." TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON) (2019): 2023-2028.
- [3]. Arun, K., Garg Ishan and Kaur Sanmeet. "Loan Approval Prediction based on Machine Learning Approach." (2016).
- [4]. K. Alshouli, A. AlGhamdi and D. P. Agrawal, "Azure ML Based Analysis and Prediction Loan Borrowers Creditworthy," 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020.
- [5]. 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.
- [6]. A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 488-493, doi: 10.1109/CONFLUENCE.2019.8776942.
- [7]. Saurabh Arora, Sushant Bindra, Survesh Singh, Vinay Kumar Nassa, "Prediction of credit card defaults through data analysis and machine learning techniques", scientific committee of the 1st International Conference on Computations in Materials and Applied Engineering – 2021
- [8]. Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal, Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison, 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 978-1-7281-2791-0/20/\$31.00 © 2020 IEEE, 2020.
- [9]. Ch. Naveen Kumar, D. Keerthana, M Kavitha, M Kalyani, "Customer Loan Eligibility Prediction using Machine Learning Algorithms in Banking Sector", 2022 7th International Conference on Communication and Electronics Systems (ICCES), pp.1007-1012, 2022.
- [10]. Jayan Kokru, Abhijeet Shrikant Ghodke, Prathmesh Chavan, Siddharth Chand, Prof Sagar Mane, "Bank Loan Approval Prediction System Using Machine Learning Algorithms", International Journal of Advanced Research in Science, Communication and Technology, pp.132, 2022.

Cite this article as :

Gupta Ashwin Arunkumar, Chaurasiya Ravi Panchuram, Khambati Mohammed Aadil Afzal, Niraj Sureshchand Yadav, Uma Goradiya, "Predictive Analysis in Banking using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.434-439, March-April-2023. Available at doi : <https://doi.org/10.32628/CSEIT2390247> Journal URL : <https://ijsrcseit.com/CSEIT2390247>