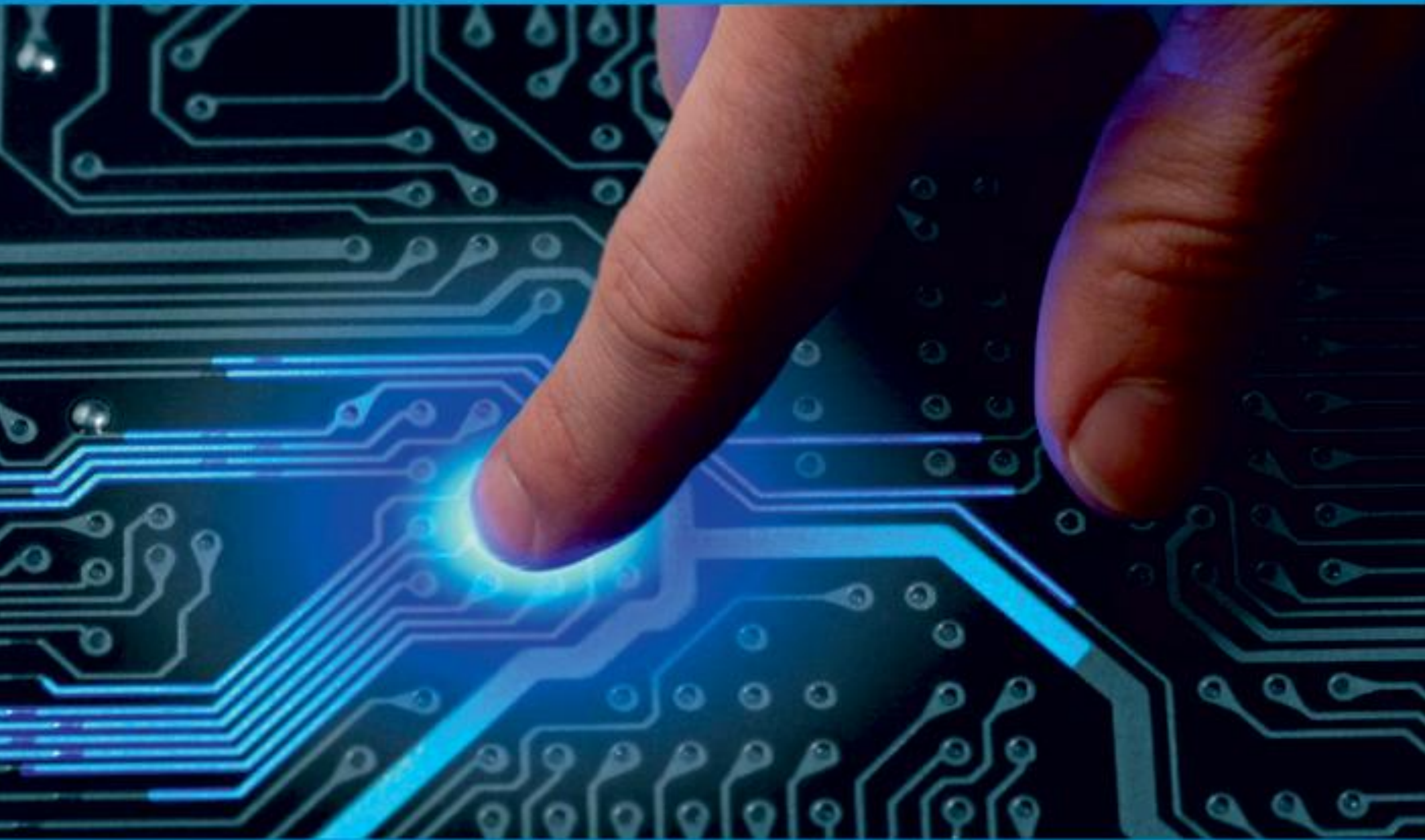
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205086 |

Predicting Bank Loan Eligibility Using Machine Learning Models

Narmadha T, Czan Dev, Shibu Kumar kanu, Motim Miya, Md Shakil ahsan mazumder

Associate professor, Dept. of CSE, Jain university, Bengaluru, India

UG Student, Dept. of CSE, Jain university, Bengaluru, India

UG Student, Dept. of CSE, Jain university, Bengaluru, India

UG Student, Dept. of CSE, Jain university, Bengaluru, India

UG Student, Dept. of CSE, Jain university, Bengaluru, India

ABSTRACT: The loan approval process is crucial for financial institutions as it involves the evaluation of borrowers' creditworthiness to determine whether to accept or reject loan applications. The study also makes an effort to examine the borrowing patterns and loan performance of subprime bank borrowers. In light of the subprime mortgage crisis of 2008, which highlighted the risks of lending to borrowers with bad credit, this study attempts to provide light on the factors that influence the loan performance of subprime borrowers. In this paper, we propose a loan approval prediction system that analyzes borrowers income status and other relevant information using machine learning techniques. We train and evaluate our machine learning model using a dataset of loan applications. Our system uses a variety of machine learning algorithms to predict loan approvals based on borrower data such as income levels, employment status and other factors. Our technology empowers financial institutions to make informed lending decisions based on reliable data.

KEYWORDS: KNN; logistic regression; decision tree; employee; machine learning

I. INTRODUCTION

Bank employees are currently reviewing application documents and deciding to grant loans to eligible applicants. This manual process is long due to the large number of applications. To solve this problem, a neural network model has been proposed to predict the credit risk of banks. The chosen model is a neural network support specifically designed to predict loan defaults. In this study, joint method, logistic regression and support vector classifiers were used to increase the accuracy of prediction. The purpose of using these classifiers is to increase data performance and obtain better results. The current loan agreement faces several shortcomings that hamper its effectiveness and efficiency. First, the process is lengthy due to manual verification of each loan applicant's information. This situation not only delays the disbursement of credit but also harms customers. Second, relying on human judgment can lead to errors in evaluating claims. These mistakes can lead to bad credit decisions, such as giving loans to unsuitable people. Thirdly, the manual process is inefficient and resource-intensive, driving up operational costs for the bank as significant human effort is required to process a large volume of loan applications. Furthermore, without automated systems, there's a heightened risk of assigning loans to applicants who don't meet eligibility criteria, potentially increasing the rate of loan defaults and financial losses for the bank. Additionally, manual processing is not easily scalable, posing challenges for bank employees as the number of loan applications grows. Lastly, inconsistent decision-making may arise due to varying interpretations of eligibility criteria among different bank employees, further undermining the reliability of the loan approval process.

II. RELATED WORK

The study in [3] conducted a systematic literature review to compare the suitability of ML models for credit risk assessment, specifically in the context of rural borrowers with limited loan history. The authors in [4] employed various ML algorithms (RF, XGBoost, GBM, and Neural Network) to predict loan defaults in the Chinese peer-to-peer (P2P) market, with RF exhibiting the highest accuracy. In [5], ensemble ML techniques (AdaBoost, LogitBoost, Bagging, and Random Forest) were used to predict loan approval in bank direct marketing data, with AdaBoost achieving the highest accuracy. The study in [6] utilized ML algorithms (neural network, naive Bayes, KNN, decision tree, and ensemble

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205086 |

learning) to predict customer creditworthiness and establish an automated risk assessment system, achieving accuracy ranging from 80% to 76%.

III. PROPOSED WORK

To overcome the complexities of the current loan approval system, we have deployed an innovative automated loan prediction system powered by machine learning algorithms. This system functions in two distinct phases: initial training using historical loan application data, which enables the machine learning model to extract patterns and complexities embedded within the approval process. Following training, the model autonomously evaluates loan applications, leveraging its acquired knowledge to assess applicant eligibility. The adoption of this system offers several benefits. First, it substantially shortens loan approval timelines, enabling applicants to receive prompt responses and enhancing their overall experience. Secondly, automation significantly reduces the likelihood of human error by relying on data-driven algorithms, resulting in more accurate eligibility assessments. Thirdly, the system ensures consistent and efficient decision-making by adhering to predefined criteria and algorithms, guaranteeing that qualified applicants receive loan approvals promptly and without inconsistencies. Through this automated loan prediction system, we aim to streamline the loan approval process, enhance efficiency, minimize errors, and ultimately accelerate loan approvals for eligible applicants.

IV. METHODOLOGY

This research was conducted using Python on Kaggle's Jupyter Notebook cloud environment. The proposed model predicts the customer's credit eligibility based on the provided information. The inputs to this model include the attributes from the dataset as shown in Table 1. The next section dives into the dataset, explaining the process used to clean and preprocess the data for modelling.

A. Dataset

The data used in this study is the historical data of "Profitable credit data", which can be accessed through Kaggle [7]. The banking sector, in particular, has adopted this technology in the field of data science and analytics. In this framework, data of 615 lines and 14 attributes, mostly related to the classification problem, are given. The purpose is to determine whether the loan application should be accepted or rejected based on the various information provided by the user during the online application. These details include gender, marital status, education, number of residents, income, loans, credit history, and more.

B. Data preprocessing and Analysis

1. Synthetic Minority Oversampling Technique (SMOTE):

This technique is useful in solving the problem of uneven distribution, which is the main source of error in machine learning models. When the number of classes in the data set is small, inconsistencies will arise and it will be difficult for the model to learn the decision boundary effectively [8]. In this study, we use the SMOTE technique to solve this challenge by sampling more in fewer classes. We achieve this by creating copies of small classes in the training data before fitting the model.

2. A one-time coding process helps convert the categorical variables in the dataset into binary form so that the ML model can understand the data.

3. Normalization: The purpose of machine learning model profile normalization is to transform features and ensure that they are all at the same level. Normalization helps improve the training stability and performance of the model.

4. Data analysis (EDA) involves examining data sets to identify patterns, trends, and inconsistencies while cleaning data by imputing missing or incomplete data. The dataset analysis revealed:

- A higher proportion of male applicants compared to female applicants.
- A majority of applicants are married.
- The dataset implies there more number with good credit(1) and less number with bad credit(0).

As illustrated in Figure 1, Applicant_Income exhibits the strongest positive correlation with Loan_Amount among the key dataset variables.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205086 |

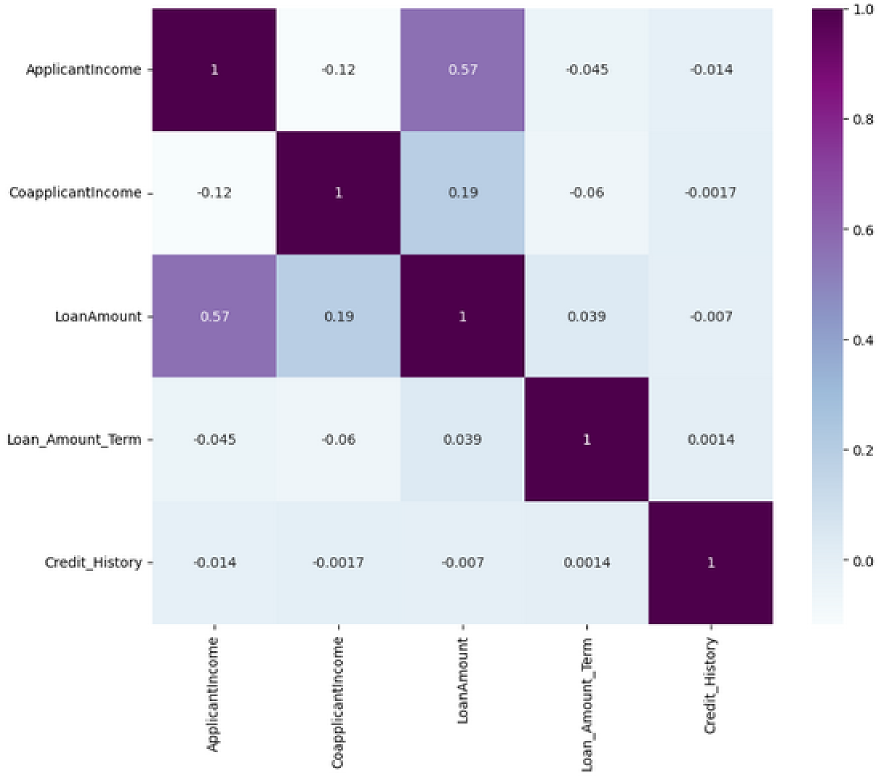


Fig 1: correlation of key variables in the dataset

V. RESULTS AND ANALYSIS

Evaluation metrics quantify the effectiveness of a machine learning (ML) model. The Confusion Matrix provides a detailed breakdown of the number of correct and incorrect predictions made by an ML model, categorized by class:

- True Positives: Actual positive cases correctly predicted.
- False Positives: Actual positive cases incorrectly predicted.
- True Negatives: Actual negative cases correctly predicted.
- False Negatives: Actual negative cases incorrectly predicted.

A. Logistic Regression (LR) Algorithm

LR is a simple distribution algorithm used to model binary (0,1) variables. LR estimates the probability of the response/variable based on one or more variables called the predictor/independent variable [10].

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Where;
L= maximum value
K=growth rate
X=value of midpoint

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205086 |

The given below figure 2 shows the evaluation result of logistic regression model

Classification Report For LogisticRegression():

	precision	recall	f1-score	support
0	0.91	0.39	0.55	54
1	0.75	0.98	0.85	100
accuracy			0.77	154
macro avg	0.83	0.68	0.70	154
weighted avg	0.81	0.77	0.74	154

Fig 2: LR model evaluation

B. K-Nearest Neighbor (KNN) Algorithm

KNN is a supervised ML algorithm that uses Euclidean distance to calculate the distance between features and then uses the “feature similarity” in the dataset to match the features. The formula is as follows:
Dist ((x,y),(a,b))=$\sqrt{(x-a)2+(y-b)2}$

Where: (x, y) and (a, b) are the coordinates of two points in the plane.

The given below figure 3 shows the evaluation result of K nearest neighbor model

Classification Report For KNeighborsClassifier(n_neighbors=3):

	precision	recall	f1-score	support
0	0.63	0.44	0.52	54
1	0.74	0.86	0.80	100
accuracy			0.71	154
macro avg	0.69	0.65	0.66	154
weighted avg	0.70	0.71	0.70	154

Fig 3: KNN model evaluation

C. Decision Tree (DT) Algorithm

Decision tree algorithms use features and attributes in data sets to make informed decisions. The main purpose of the decision tree algorithm is to increase the information gain rate. This goal is achieved by classifying the characteristics (nodes) starting from the highest data. The calculation formula for the increment data is as follows:

$IG\ (T, a) = HT - H(T \mid a)$ [16]

Where: H(T | a) is the conditional entropy of , T and is the value of the attribute.

The given below figure 4 shows the evaluation result of Decision tree algorithm model

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205086 |

```
Classification Report For DecisionTreeClassifier():
              precision    recall  f1-score   support

           0       0.61      0.57      0.59        54
           1       0.78      0.80      0.79       100

    accuracy                           0.72       154
   macro avg       0.69      0.69      0.69       154
weighted avg       0.72      0.72      0.72       154
```

Fig 4 : DT model evaluation

VI. CONCLUSION AND FUTURE WORK

After a detailed analysis of the advantages and limitations of the product according to its users, one can make sure that the product is a good member. It works well and meets the financial needs of the user. Additionally, members can work together in different systems. Despite occasional bugs, content violations, and concerns about certain features in predictive technology, steps may be taken in the future to improve security, reliability, and software updates. The above software has the potential for further development and can work well with automated workflows. Currently the system is trained using historical data, but as the software progresses it will be beneficial to incorporate new test data into the training process at a specific time. The random forest classifier provides the best accuracy with 82% accuracy on the test data. A combination of learning methods such as Bagging and Boosting can be used to achieve better results.

REFRENCES

[1] "Most commonly used A.I. application in investment banking worldwide 2020, by types." Statista, 15-Sept-2021 [Online]. Available: https://www.statista.com/statistics/1246874/ai-used-in-investment-banking worldwide-2020/ [Accessed: 29-Jan-2022]

[2] G. Dorfleitner, E.M. Oswald, & R. Zhang. "From Credit Risk to Social Impact: On the Funding Determinants in Interest-Free Peer-to-Peer Lending." J Bus Ethics. 2021 Vol.170, pp. 375–400. https://doi.org/10.1007/s10551-019-04311-8

[3] A. Kumar, S. Sharma, & M. Mahdavi, "Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review." Risks 9.11 (2021): 192

[4] J. Xu, Z. Lu, and Y. Xie, "Loan default prediction of Chinese P2P market: a machine learning methodology." Scientific Reports, 2021, Vol. 11(1), pp. 1 19.

[5] H. Meshref, "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms."International Journal of circuits, systems, and signal processing. 2020, Vol. 14, pp. 914-922 DOI: 10.46300/9106.2020.14.117

[6] A.S. Aphale, and S.R. Shinde, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval." International Journal of Engineering Research & Technology (IJERT). 2020, Vol. 9 pp. 991-995

[7]"Loan Eligibility Dataset." Kaggle, 15-Aug-2020. Available online

[8] A.S. Hussein, T. Li, C.W. Yohannese, & K. Bashir. "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE." International Journal of Computational Intelligence Systems. 2019, Vol. 12(2), PP.1412.of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA. 2003, pp. 129-136.

[9] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness arXiv:2010.16061 (2020). and correlation." arXiv preprint

[10] J.M. Hilbe. "Logistic Regression." International encyclopedia of statistical science. 2011, Vol 1: pp. 15-32.

[11] A. Saini. "Logistic Regression | What is Logistic Regression and Why do we need it 26-Aug-2021[Online] Available: https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding of-logistic-regression-for-data-science-beginners/#h2_5 [Accessed: 28-Jan 2022]

[12] M. Shouman, T. Turner, and R. Stocker. "Applying k-nearest neighbour in diagnosing heart disease patients." International Journal of Information and Education Technology. 2012 Vol. 2(3), pp. 220-223.

[13] L.K. Ramasamy, S. Kadry, Y. Nam, & M.N. Meqdad. "Performance analysis of sentiments in Twitter dataset using SVM models. International Journal of Electrical and Computer Engineering (IJECE). 2021 Vol. 11, No. 3, pp.2275 2284 https://doi.org/10.11591/ijece.v11i3.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2023.1205086 |

[14] R. Kunchhal. “Mathematics Behind SVM | Math Behind Support Vector Machine.” 28-Dec-2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind svm/ [Accessed: 27-Jan-2022]

[15] K. Yadav, and R. Thareja. "Comparing the performance of naive bayes and decision tree classification using R." International Journal of Intelligent Systems and Applications. 2019, Vol.11(12), p.11.

[16] K. Ramya, Y. Teekaraman, & K.R. Kumar. "Fuzzy-based energy management system with decision tree algorithm for power security system." International Journal of Computational Intelligence Systems. 2019, Vol.12(2), pp.1173.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com