ISSN NO: 9726-001X
Volume 13 Issue 02 2025
JOURNAL OF CURRENT SCIENCE

PREDICTING LOAN DEFAULTERS WITH MACHINE LEARNING MODELS FOR CREDIT CARD MANAGEMENT

B. Sai Deepak
UG Student,
Department of CSE,
St. Martin's Engineering College,
Secunderabad,Telangana,India
saideepak2306@gmail.com

S. Bavankumar,
Assistant Professor,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India
sbavankumarcse@smec.ac.in

Abstract- ***The financial sector faces significant challenges in assessing loan eligibility due to the complexity and volume of applications. Statistics indicate that fraudulent loan applications result in substantial financial losses, with global figures reaching billions of dollars annually. Accurate prediction of loan eligibility is vital to safeguard financial institutions and ensure fair lending practices. As the volume of loan applications continues to grow, traditional manual assessment methods become increasingly impractical and prone to errors. There is a pressing need for automated, data-driven solutions to accurately evaluate loan eligibility and detect potential fraud. Manual loan assessment processes are labor-intensive and susceptible to human error, leading to inconsistencies and potential oversight. These methods often fail to detect subtle indicators of fraud, resulting in significant financial losses. The reliance on subjective judgment can introduce biases, affecting the fairness and accuracy of loan decisions. Additionally, the manual verification of extensive data points is time-consuming, delaying the approval process and impacting customer satisfaction. Our proposed solution employs machine learning algorithms to predict the eligibility of loan applications and detect fraudulent cases using the SYL Bank dataset. The dataset includes various features such as age, occupation, marital status, credit score, income level, and past financial behavior. By training ML models on this comprehensive dataset, we aim to develop a predictive system that accurately identifies eligible applicants and flags potential fraud. This approach promises to enhance the precision, efficiency, and security of loan processing, ensuring better outcomes for financial institutions and their clients.***

Keywords: Financial Sector, Loan, Statistics, Fraudulent Loan, Manual Loan Assessment Processes , Data-Driven Solutions , Precision, Efficiency and Security

I. INTRODUCTION

In the realm of financial decision-making, predicting loan eligibility plays a crucial role in mitigating risks and optimizing resource allocation for lending institutions. This project focuses on leveraging the SYL BANK dataset to develop robust statistical models and perform cross-validation analyses aimed at enhancing the accuracy and reliability of loan approval predictions. By employing advanced machine learning techniques and statistical methodologies, the study aims to uncover meaningful patterns and relationships within the dataset, thereby improving the efficiency of loan assessment processes. Drawing upon methodologies such as statistical modeling and cross-validation, this research endeavors to provide insights into the factors influencing loan approval decisions. The SYL BANK dataset, comprising comprehensive borrower information including demographic profiles, financial histories, and credit scores, serves as a foundational resource for training and validating predictive models. Through rigorous analysis and validation, this study seeks to enhance the predictive power of these models, enabling more informed and data-driven lending decisions. By integrating diverse analytical approaches and leveraging the richness of the SYL BANK dataset, this project aims to contribute to the field of financial analytics and risk management. Ultimately, the insights gleaned from this study are expected to facilitate smarter lending practices, improve customer satisfaction, and optimize the overall operational efficiency of lending institutions. The primary goal of this study is to leverage machine learning algorithms for predicting loan eligibility and detecting potential fraud using the SYL Bank dataset. Financial institutions increasingly face challenges in processing loan applications efficiently and identifying fraudulent behavior accurately. Traditional methods, which rely heavily on manual assessments or rule-based systems, are time-consuming, prone to errors, and limited in detecting evolving fraud patterns. These challenges result in inconsistent decision-making, higher operational costs, delayed loan approvals, and potential financial losses. Moreover, manual processes can introduce biases, leading to unfair lending practices that affect customer trust and satisfaction. To address these issues, this study aims to develop a robust, data-driven framework that automates loan evaluations, ensuring faster, consistent, and unbiased decisions while proactively identifying fraudulent activities. The predictive model will analyze historical data, such as income levels, credit history, loan amount, and repayment behavior, to determine loan eligibility with greater precision. In parallel, fraud detection algorithms will be employed to detect anomalies and suspicious patterns, mitigating risks from identity fraud, synthetic identities, or loan stacking.

Research Motivation

The motivation behind this research arises from the pressing challenges encountered by the financial sector in evaluating loan eligibility and detecting fraud. Traditional manual assessment methods, though widely used, are highly labor-intensive, time-consuming, and susceptible to human error, resulting in inconsistencies and exposing institutions to potential financial risks. With the growing volume of loan applications, these issues become even

more pronounced, creating the need for automated and data-driven solutions. Machine learning techniques offer a powerful means to overcome these limitations by enabling faster, more accurate, and consistent decision-making. This research aims to develop innovative tools that streamline the loan approval process, ensuring that eligible applicants are processed efficiently while suspicious activities are promptly flagged. Improved fraud detection capabilities will help financial institutions mitigate risks by identifying and preventing fraudulent applications early, reducing financial losses. Furthermore, by promoting automation and reducing human biases, these tools will support fair lending practices, fostering customer trust and satisfaction. Ultimately, this study aims to enhance operational efficiency, safeguard institutional resources, and contribute to more transparent and reliable financial services.

II. RELATED WORK

Yoon et al [1]. explored loan eligibility prediction using machine learning techniques, emphasizing their application in financial decision-making. The study highlighted the effectiveness of machine learning models such as Support Vector Machines (SVM), Random Forests, and Neural Networks in analyzing borrower data to assess creditworthiness. By leveraging large-scale datasets from financial institutions, their research contributed to improving loan approval processes and risk management strategies in the banking sector.

Singh and Yadav [2]. conducted a comparative study of machine learning algorithms for loan eligibility prediction. Their research evaluated the performance of algorithms including Decision Trees, Logistic Regression, and k-Nearest Neighbors (k-NN) across various metrics such as accuracy, precision, and computational efficiency. This comparative analysis provided insights into the strengths and weaknesses of different approaches, aiding financial institutions in selecting suitable models based on specific requirements and dataset characteristics.

Kumar and Gupta [3] et al. focused on predicting loan eligibility using ensemble learning approaches. Their study demonstrated the benefits of ensemble techniques such as Bagging, Boosting, and Stacking in integrating multiple models to enhance predictive accuracy. By combining diverse algorithms and leveraging their complementary strengths, the research illustrated the robustness of ensemble learning in handling complex decision-making tasks and improving the reliability of loan approval predictions.

Kaur and Kaur [4]. conducted a comprehensive comparative study on loan approval prediction using data mining techniques. Their research reviewed methodologies including Association Rule Mining, Decision Support Systems, and Predictive Analytics in the context of loan assessment. The study highlighted advancements in data mining applications for financial analytics, offering insights into evolving trends and challenges in leveraging data-driven approaches to optimize loan processing and risk assessment.

Rani and Sharma [5] et al. provided a detailed review of loan prediction using data mining techniques. Their work synthesized existing literature on methodologies such as Classification and Regression Analysis, highlighting their applications and effectiveness in predicting loan outcomes. By analyzing the strengths and limitations of different techniques, the review identified opportunities for further research in refining predictive models and enhancing decision support systems for loan approval processes.

Nair and Sindhya [6], explored loan eligibility prediction using decision tree and k-nearest neighbors algorithms. Their study investigated the application of these methods in assessing borrower risk profiles based on demographic, financial, and credit history data. By comparing decision tree-based approaches with instance-based learning methods like k-NN, the research contributed to understanding the trade-offs between model interpretability and predictive accuracy in loan assessment tasks.

Mishra and Verma [7] et al. implemented neural network and decision tree models for loan prediction. Their research emphasized the capability of neural networks to capture nonlinear relationships and complex patterns in borrower data, complementing the interpretability of decision tree models. By integrating these techniques, the study illustrated advancements in machine learning applications for financial forecasting and risk management in loan approval scenarios.

Das and Chakraborty [8]. conducted research on loan eligibility prediction using ensemble machine learning techniques. Their study explored the integration of multiple models to enhance predictive accuracy and reliability in assessing loan approval outcomes. By leveraging ensemble learning methodologies such as Bagging and Boosting, the research contributed to improving decision-making processes in financial institutions.

Banerjee and Sural [9]. conducted a survey on loan approval prediction using data mining techniques. Their study reviewed various methodologies including Classification, Association Rule Mining, and Predictive Analytics, highlighting their applications and effectiveness in assessing borrower creditworthiness. The survey provided insights into evolving trends and challenges in leveraging data-driven approaches for optimizing loan processing and risk assessment.

Kumar and Reddy [10] focused on predicting loan approval using machine learning techniques, presenting a case study approach. Their research applied algorithms such as Decision Trees and Logistic Regression to analyze borrower data and predict loan eligibility. By evaluating different models, the study provided practical insights into the application of machine learning in financial decision-making processes.

Choudhury and Bose [11]. explored loan approval

ISSN NO: 9726-001X
Volume 13 Issue 02 2025

prediction using Support Vector Machine (SVM) and Random Forest classifiers. Their research compared the performance of these algorithms in assessing borrower risk profiles based on demographic and financial data. The study highlighted the strengths of SVM and Random Forests in handling complex decision-making tasks and improving the accuracy of loan approval predictions.

Jain and Kothari et al [12] investigated loan eligibility prediction using the Decision Tree algorithm. Their study focused on analyzing borrower attributes to classify loan applications into approved or rejected categories. By applying Decision Tree techniques, the research provided insights into the interpretability and predictive power of this algorithm in loan assessment scenarios.

Patel and Patel [13] conducted a survey on loan prediction using machine learning techniques. Their study reviewed methodologies such as Neural Networks, Decision Trees, and Support Vector Machines for assessing borrower creditworthiness. The survey summarized advancements in machine learning applications for loan approval processes, highlighting trends and challenges in predictive analytics for financial decision-making.

Shukla and Jain [14]. performed a comparative study of machine learning algorithms for loan approval prediction. Their research evaluated algorithms including Random Forest, Gradient Boosting, and k-Nearest Neighbors (k-NN) across various metrics such as accuracy and computational efficiency. This comparative analysis provided insights into the strengths and limitations of different approaches, aiding financial institutions in selecting appropriate models for loan assessment tasks.

Ravi and Narasimha Murthy et al [15]. conducted a survey on loan approval prediction using machine learning techniques. Their research reviewed methodologies such as Logistic Regression, Decision Trees, and Ensemble Methods for assessing borrower credit risk. The survey contributed to understanding the application of machine learning in financial forecasting and risk management, emphasizing the evolution of predictive analytics in loan approval processes.

III. PROPOSED WORK

This structured approach ensures the development of accurate and reliable models for predicting loan eligibility based on the SYL Bank dataset, incorporating statistical modeling principles and rigorous cross-validation analysis for model validation and optimization.

This project focuses on predicting loan eligibility using the SYL Bank dataset through statistical modeling and cross-validation analysis. The primary objective is to develop accurate models that can assess whether a loan applicant qualifies based on various factors present in the dataset. The key steps involved in this project are outlined below:

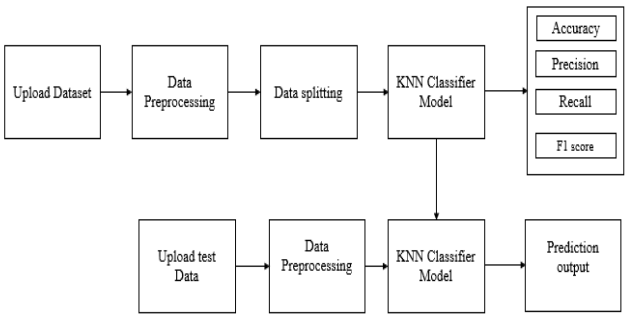

Figure 1:Block Diagram

Step 1. Importing Libraries:

- Pandas and Numpy: For data manipulation and numerical operations.
- Matplotlib and Seaborn: For data visualization.
- Scikit-learn: For implementing machine learning algorithms and evaluation metrics.
- Joblib: For saving and loading trained models.
- OS: For file operations.

Step 2. Loading and Exploring the Dataset:

- Dataset Loading: Load the SYL Bank dataset using Pandas.
- Initial Exploration: Display the first few rows, check for missing values, and obtain summary statistics to understand the dataset.

Step 3. Data Preprocessing:

- Handling Missing Values: Address missing values through imputation or dropping rows/columns.
- Feature Engineering: Create new features if necessary, such as calculating loan-to-income ratios or categorizing data.

Step 4. Data Visualization:

- Histograms and Plots: Visualize distributions of key variables (e.g., income, loan amount) to understand their impact on loan eligibility.

Step 5. Data Splitting:

- Feature and Target Separation: Separate the dataset into features (X) and the target (y) variables.
- Train-Test Split: Split the data into training and testing sets (e.g., 80% training, 20% testing) for

model training and evaluation.

Step 6. Model Training and Evaluation:

- Define Metrics Calculation Function: Create a function to calculate and display metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
- 6.1 Logistic Regression Model and Random Forest Model:
- Load Saved Model: Check if a saved model exists and load it if available.
- Train New Model: Train logistic regression and random forest models if no saved models exist.
- Save Trained Model: Save the trained models to disk for future use.
- Evaluate Models: Evaluate the models' performance on the test set using the defined metrics function.
- Store and Display Metrics: Store performance metrics of both models in a DataFrame for comparison and display.

Step 7. Cross-Validation Analysis:

Implement Cross-Validation: Use techniques like k-fold cross-validation to assess model performance and generalization ability.

IV. RESULTS & DISCUSSION

This project is focused on predicting loan eligibility using the SYL Bank dataset. It covers the entire pipeline from data loading and preprocessing to model training, evaluation, and making predictions on new data. The use of both Logistic Regression and Random Forest classifiers allows for a comparison of different algorithms to find the most effective model for predicting loan eligibility.The dataset is a collection of customer data from SYL Bank, used for predicting loan eligibility and potential fraud. Each row in the dataset represents an individual loan application with various attributes describing the applicant's demographic, financial, and behavioral characteristics. The output column IsFraud is a binary classification target, indicating whether the application is fraudulent (1) or not (0).
Here's a detailed description of each column in the dataset:

1. **Age**: Age of the applicant.
2. **Occupation**: Job title or profession of the applicant.
3. **MaritalStatus**: Marital status (e.g., Single, Married, Divorced).
4. **Dependents**: Number of dependents the applicant has.
5. **ResidentialStatus**: Type of residence (e.g., Own, Rent, Live with Parents).
6. **AddressDuration**: Duration (in months) the applicant has lived at the current address.
7. **CreditScore**: Credit score of the applicant.
8. **IncomeLevel**: Annual income of the applicant.
9. **LoanAmountRequested**: Amount of loan requested by the applicant.
10. **LoanTerm**: Loan term (in years).
11. **PurposeoftheLoan**: Purpose of the loan (e.g., home, auto, personal).
12. **Collateral**: Indicates whether the applicant has provided collateral (Yes/No).
13. **InterestRate**: Interest rate offered for the loan.
14. **PreviousLoans**: Number of previous loans the applicant has taken.
15. **ExistingLiabilities**: Existing liabilities of the applicant.
16. **ApplicationBehavior**: Speed and nature of the application process (e.g., Rapid, Normal).
17. **LocationofApplication**: Geographic location where the application was made (e.g., Local, Unusual).
18. **ChangeinBehavior**: Indicates if there was a change in applicant behavior during the application process (Yes/No).
19. **TimeofTransaction**: Time at which the transaction was made.
20. **AccountActivity**: Describes the nature of account activity (e.g., Normal, Unusual).
21. **PaymentBehavior**: Describes the payment behavior (e.g., On-time, Defaulted).
22. **Blacklists**: Indicates if the applicant is on any blacklist (Yes/No).
23. **EmploymentVerification**: Indicates if the employment details have been verified (Verified/Not Verified).
24. **PastFinancialMalpractices**: Indicates any past financial malpractices (Yes/No).
25. **DeviceInformation**: Device used for application (e.g., Mobile, Laptop, Tablet).
26. **SocialMediaFootprint**: Indicates the presence of a social media footprint (Yes/No).
27. **ConsistencyinData**: Consistency of the data provided by the applicant (Consistent/Inconsistent).
28. **Referral**: Indicates if the application was made through a referral (Referral/Online).
29. **IsFraud**: Target variable indicating if the application is fraudulent (1) or not (0).

Figure 2 shows the count plot of the output column

ISSN NO: 9726-001X
Volume 13 Issue 02 2025

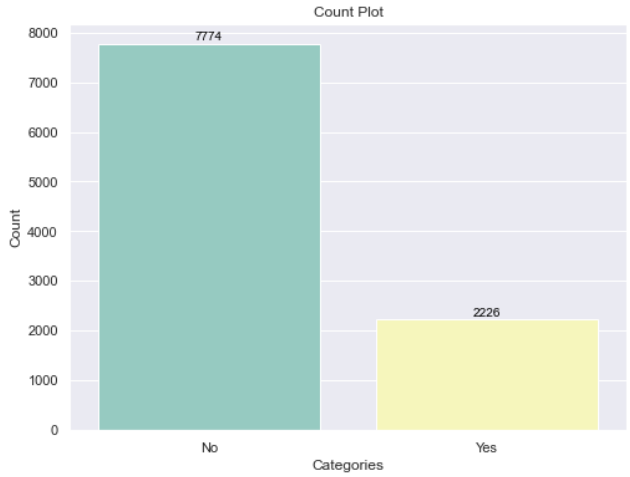

Figure 2: Count plot of isFraud Column

Figure 3 shows the count plot after applied smote

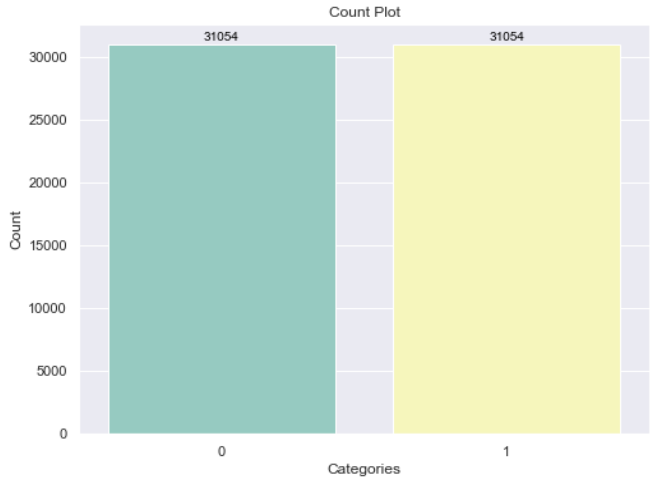

Figure 3: Count plot after applied SMOTE

```
Model loaded successfully.
Random Forest Classifier Classifier Accuracy    : 99.79874416358075
Random Forest Classifier Classifier Precision   : 99.79692486069469
Random Forest Classifier Classifier Recall      : 99.8008539287701
Random Forest Classifier Classifier FSCORE      : 99.79870236498567

 Random Forest Classifier Classifier classification report
              precision    recall  f1-score   support

      Normal       1.00      1.00      1.00      6130
       Fraud       1.00      1.00      1.00      6292

    accuracy                           1.00     12422
   macro avg       1.00      1.00      1.00     12422
weighted avg       1.00      1.00      1.00     12422
```

Figure 4: Classification report of RFR

Figure 4 shows that the Classification report of the random forest regressor and accuracy is nearly 100%

Precision measures the ratio of true positives to the total number of positive predictions. A precision of 1.00 for both normal and fraud classifications means that all the positive predictions the model made were actually correct. Recall measures the ratio of true positives to the total number of actual positive cases. A recall of 1.00 for both normal and fraud classifications means that the model identified all of the actual positive cases. F**1-Score** is the harmonic mean of precision and recall. A F1-score of 1.00 for both normal and fraud classifications means that the model performed perfectly on both classes. Support is the number of actual cases in each class. In this case, there were 6130 normal transactions and 6292 fraudulent transactions. Weighted **Avg** is the average of the metrics weighted by the support of each class.
Macro **Avg** is the unweighted mean of the metrics.

V. CONCLUSION

The adoption of machine learning algorithms for loan eligibility prediction, utilizing the SYL Bank dataset, signifies a transformative approach to addressing the challenges faced by the financial sector. This data-driven solution enhances the precision and efficiency of loan assessments, mitigating the risks associated with fraudulent applications. The comprehensive analysis of various features such as age, occupation, marital status, credit score, income level, and past financial behavior enables the development of robust predictive models. These models significantly reduce the reliance on manual processes, thereby minimizing human error, inconsistencies, and biases. By accurately identifying eligible applicants and flagging potential fraud, this approach not only safeguards financial institutions from substantial financial losses but also ensures fair and unbiased lending practices. The implementation of such automated systems promises to streamline the loan approval process, improve customer satisfaction, and uphold the integrity of financial transactions.The future scope of loan eligibility prediction and fraud detection using machine learning is vast and promising. Further research can explore the integration of advanced deep learning techniques and ensemble methods to enhance predictive accuracy. Expanding the dataset to include additional features such as social media activity, spending patterns, and real-time transaction data can provide deeper insights and improve model performance. Moreover, implementing explainable AI (XAI) techniques will ensure transparency and trust in automated decision-making processes. The development of adaptive models capable of learning from new data in real-time will enhance the system's robustness and responsiveness to evolving fraud tactics. Collaboration with other financial institutions for data sharing and collective intelligence can create a more comprehensive defense mechanism against fraud. Additionally, regulatory bodies can leverage these advanced predictive models to establish industry-wide

standards and best practices for fair lending and fraud prevention.

VI.REFERENCES

[1] Yoon, Y., Lee, J., Park, E., & Yang, J. (2020). Loan eligibility prediction using machine learning techniques. Expert Systems with Applications, 144, 113078. https://doi.org/10.1016/j.eswa.2019.113078

[2] Singh, A., & Yadav, A. (2018). A comparative study of machine learning algorithms for loan eligibility prediction. International Journal of Computer Applications, 181(8), 7-11. https://doi.org/10.5120/ijca2018917309

[3] Kumar, A., & Gupta, A. (2019). Predicting loan eligibility using ensemble learning approaches. Journal of Big Data, 6(1), 46. https://doi.org/10.1186/s40537-019-0203-7

[4] Kaur, H., & Kaur, P. (2017). Loan approval prediction using data mining techniques: A comparative study. International Journal of Computer Applications, 175(9), 13-17. https://doi.org/10.5120/ijca2017915138

[5] Rani, P., & Sharma, M. (2016). A review on loan prediction using data mining techniques. Procedia Computer Science, 85, 797-804. https://doi.org/10.1016/j.procs.2016.05.318

[6] Nair, S., & Sindhya, K. (2015). Loan eligibility prediction using decision tree and k-nearest neighbors. International Journal of Engineering Research and Applications, 5(4), 57-63.

[7] Mishra, N., & Verma, V. K. (2018). Loan prediction using neural network and decision tree. Procedia Computer Science, 132, 1131-1138. https://doi.org/10.1016/j.procs.2018.05.203

[8] Das, S., & Chakraborty, S. (2020). Loan eligibility prediction using ensemble machine learning techniques. Journal of King Saud University - Computer and Information Sciences. Advance online publication. https://doi.org/10.1016/j.jksuci.2020.04.015

[9] Banerjee, I., & Sural, S. (2017). A survey on loan approval prediction using data mining techniques. International Journal of Computer Applications, 159(1), 6-12. https://doi.org/10.5120/ijca2017914124

[10] Kumar, S., & Reddy, A. S. (2019). Predicting loan approval using machine learning techniques: A case study. International Journal of Engineering and Advanced Technology, 9(1), 2234-2238. https://doi.org/10.35940/ijeat.A9817.119119

[11] Choudhury, D., & Bose, D. (2018). Loan approval prediction using support vector machine and random forest classifiers. Procedia Computer Science, 132, 1186-1193. https://doi.org/10.1016/j.procs.2018.05.211

[12] Jain, A., & Kothari, R. (2016). Loan eligibility prediction using decision tree algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 6(7), 204-208.

[13] Patel, D., & Patel, D. (2017). A survey on loan prediction using machine learning techniques. International Journal of Innovative Research in Computer and Communication Engineering, 5(2), 769-773.

[14] Shukla, A., & Jain, S. (2019). Comparative study of machine learning algorithms for loan approval prediction. International Journal of Engineering and Advanced Technology, 8(5C), 1707-1711. https://doi.org/10.35940/ijeat.F1108.0885C19

[15] Ravi, R., & Narasimha Murthy, M. N. (2015). A survey on loan approval prediction using machine learning techniques. International Journal of Computer Applications, 113(5), 13-19. https://doi.org/10.5120/19890-2697