International Journal of
Engineering Research and Science & Technology

ISSN:2319-5991

www.ijerst.org

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

PREDICTING LOAN DEFAULTERS WITH MACHINE LEARNING MODELS FOR CREDIT CARD MANAGEMENT

Dr. S Venkata Achuta Rao[1*], N. Nikhila[2], D. Yashwanth[2], G. Deepak[2], Y.V.V.Vardhan kumar[2]

[1] Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering (AIML)

[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana

ABSTRACT

The financial sector faces growing challenges in assessing loan eligibility due to the increasing complexity and volume of applications. Fraudulent loan submissions continue to result in significant financial losses, with global impacts amounting to billions of dollars each year. Accurate and efficient prediction of loan eligibility is therefore essential to protect financial institutions and promote fair lending practices. Traditional manual assessment methods are becoming obsolete, as they are time-consuming, error-prone, and often fail to identify subtle signs of fraud. These manual processes rely heavily on subjective judgment, which can introduce inconsistencies and biases, ultimately affecting the reliability of loan decisions. Moreover, verifying a large number of data points manually delays approvals and reduces customer satisfaction.To address these issues, we propose a machine learning-based solution that automates loan eligibility prediction and fraud detection. Using the SYL Bank dataset—which includes features such as age, occupation, marital status, credit score, income level, and historical financial behavior—we aim to train predictive models capable of accurately identifying eligible applicants and detecting fraudulent activity. This data-driven approach enhances the accuracy, consistency, and speed of loan assessments, providing a more secure and streamlined process for financial institutions and their clients.

Keywords: Loan Eligibility Prediction, Machine Learning, Statistical Modeling, Risk Assessment, Cross-Validation

1.INTRODUCTION

Accurate prediction of loan eligibility is a critical component of financial decision-making, helping lending institutions minimize risk and optimize resource allocation. This project centers on utilizing the SYL BANK dataset to build reliable statistical and machine learning models, supported by cross-validation techniques, to improve the precision of loan approval predictions. By applying advanced analytics, the study aims to uncover significant patterns and correlations within the data, ultimately streamlining the loan assessment process.

The research employs a combination of statistical modeling and cross-validation to identify key factors that influence loan approval outcomes. The SYL BANK dataset, which includes detailed borrower data—such as demographics, financial history, and credit scores—forms the basis for training and evaluating predictive models. Through careful analysis and validation, the study enhances the models' ability to make accurate, data-driven predictions.

By integrating multiple analytical techniques and leveraging the depth of the SYL BANK dataset, this project contributes to advancements in financial analytics and risk assessment. The findings are expected to support more efficient, equitable, and informed lending practices, while also improving customer satisfaction and operational performance for financial institutions.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

2.LITERATURE SURVEY

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.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

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 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.

3.PROPOSED SYSTEM

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Step 1. Importing Libraries:

- Pandas and Numpy: For data manipulation and numerical operations.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

- 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.

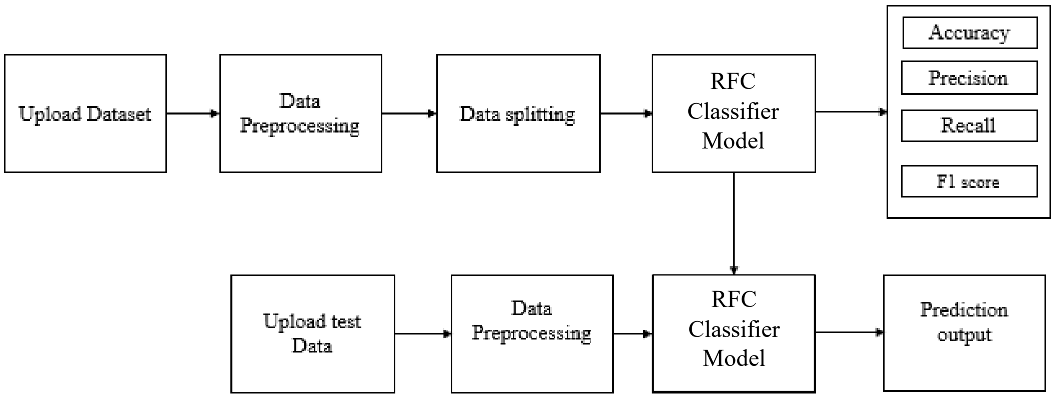

Figure 1: Block Diagram

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:

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

- 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.
- Tune Hyperparameters: Optimize model performance by tuning hyperparameters using cross-validation results.

Advantages of Proposed System

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

4. RESULTS AND DISCUSSION

This paper 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. Figure 2 shows the count plot of the output column and Figure 3 shows the count plot after applied smote

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

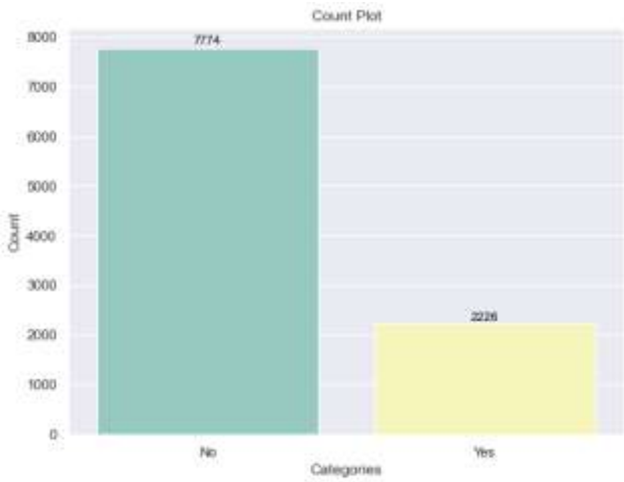

Figure 2: Count plot of isFraud Column

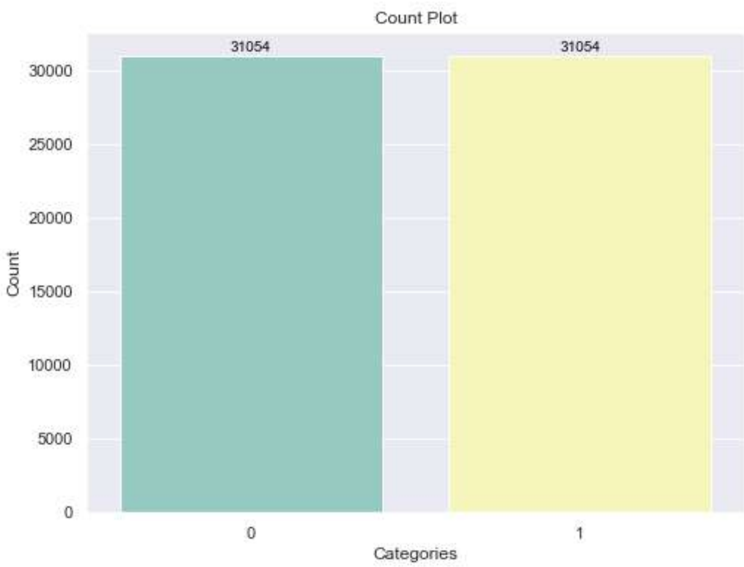

Figure 3: Count plot after applied SMOTE

```
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 RFC

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

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. F1-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.

5.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.

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

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

[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