IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

Credit Risk Modelling & Predictive Analytics for Loan Portfolio Optimization

Prof. Ujwala Khartad[1] and Mr. Shivam Sanjay Wadje[2]
[1] Professor, Department of Computer Science & Engineering (Data Science)
[2] Student, Department of Computer Science & Engineering (Data Science)
Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, Maharashtra, India
Ujwala.ds@tgpcet.com and wadjeshivam6@gmail.com

Abstract: *Credit risk assessment is a major challenge for financial institutions around the world. Traditional rule-based systems often struggle to understand complex and non-linear patterns in borrower data, which leads to incorrect predictions of default and higher risk in loan portfolios. This paper introduces a complete machine learning-based framework for credit risk modelling that helps predict the chance of loan default and improve the management of loan portfolios. Using a dataset from Kaggle (Home Credit Default Risk), the study includes full data preprocessing, exploratory data analysis (EDA), feature engineering, and three classification techniques: Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier. Class imbalance, which is common in credit datasets, is handled using the Synthetic Minority Over-sampling Technique (SMOTE) and adjusting class weights. The models are tested using metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Based on the predicted probabilities, borrowers are grouped into Low, Medium, and High-risk categories to guide risk-based decisions. The proposed approach shows better performance in identifying high-risk borrowers and provides clearer insights into decision-making. The results show that Gradient Boosting achieves the highest ROC-AUC score of about 0.94, which is better than other baseline models. The study offers a reproducible, easy-to-understand, and scalable solution for credit risk analysis that can be used in retail banking and FinTech environments.*

Keywords: Credit Risk, Loan Default Prediction, Machine Learning, Random Forest, Gradient Boosting, SMOTE, Logistic Regression, ROC-AUC, Feature Engineering, Risk Segmentation

I. INTRODUCTION

The global financial system depends a lot on being able to accurately and efficiently assess credit risk. A loan default happens when a borrower doesn't pay back the money they borrowed as agreed, which directly hurts the lender and can cause problems for the whole financial system. Reports from the industry show that non-performing loans (NPLs), where borrowers don't repay their debts, are a big issue for banks in both developed and growing economies. The Basel Accords (I, II, and III) have stressed the need for strong internal risk models that go beyond old ways of evaluating credit risk. Traditional methods for evaluating credit, like the FICO score, expert opinions, and rule-based scorecards, don't handle complex data well.

They struggle with high-dimensional data, relationships between different features, and changing borrower behaviour. These methods often result in too many false negatives, meaning they approve loans to people who later default, increasing the risk for the lender's portfolio. The introduction of machine learning (ML) and predictive analytics has opened up new possibilities for managing credit risk.

Algorithms such as Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting can find complex patterns in large datasets that include personal, financial, behaviour, and credit history data. Studies have shown these models perform better than traditional methods in many cases.

This paper introduces a structured credit risk modelling pipeline built using Python and the Scikit-learn framework.

DOI: 10.48175/IJARSCT-32670

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

The process includes getting data, cleaning it, doing exploratory data analysis, creating features, training models, evaluating their performance, and segmenting borrowers by risk. The final system is a ready-to-use, easy-to-understand tool that sorts borrowers into Low, Medium, and High risk categories. Visualizations created with Matplotlib, Seaborn, and Power BI dashboards help support decision-making.

The paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset; Section 4 explains the methodology; Section 5 details the implementation steps; Section 6 includes algorithm pseudocode; Section 7 shows a system flowchart; Section 8 discusses implementation specifics; Section 9 presents results and analysis; Section 10 outlines future work; Section 11 concludes the paper; followed by acknowledgments and references.

II. LITERATURE SURVEY

Early credit risk assessments relied on traditional statistical models. Logistic Regression became a common method because it was simple to understand, easy to explain, and could provide probabilities. One of the first and most important models was Altman's Z-score from 1968, which used financial ratios to predict if a company would fail. While these models laid the groundwork for credit scoring, they had limits. They assumed relationships between variables were linear, which made them less accurate in real-life situations where relationships are often more complex. The move from traditional methods to machine learning brought big improvements in credit risk analysis.

These new methods allowed for more flexible and powerful techniques that could find complex patterns in data. Bassens and others in 2003 were among the first to compare different machine learning models for credit scoring. They found that models like neural networks and support vector machines worked better than logistic regression across various datasets. Decision Trees became popular too because they were easy to use and interpret, but they were often criticized for overfitting, especially with messy or high-dimensional data. Today, ensemble methods are the go-to choice in credit risk prediction.

Techniques like Random Forest and Gradient Boosting work well because they combine multiple models to improve performance and reduce errors. A review of machine learning algorithms for loan defaults between 2020 and 2023 showed that Random Forest was widely preferred because it handled high-dimensional and imbalanced data well. Studies using datasets like the Home Credit Default Risk data found that Random Forest achieved AUC scores above 0.85. More advanced Gradient Boosting techniques like Boost and LightGBM performed even better, with AUC scores ranging from 0.93 to 0.97.

Hybrid models, such as the stacking ensemble from Almasy et al. in 2024, which combined XGBoost and Random Forest with SMOTE, reached an impressive accuracy of 0.97 on resampled credit data. Another study in The American Journals in 2025 found that Gradient Boosting outperformed Random Forest with an accuracy of 91.3% and an AUC-ROC of 0.94, thanks to its iterative approach of improving weak learners one after another. One big challenge in credit risk modeling is the imbalance between default and non-default cases.

In many cases, the number of defaults is much smaller, often in a ratio from 1:10 to 1:50. This can cause models to Favor predicting the majority class, making them less effective at identifying defaulters. To tackle this, Chawla et al. in 2002 introduced SMOTE, a technique that creates synthetic examples of the minority class by creating new instances between existing ones. Later research has shown that using SMOTE with ensemble models improves recall of the minority class without much loss in overall accuracy. For example, studies combining LightGBM with SMOTE have reported AUC-ROC scores of 0.832, which is better than baseline models on imbalanced data. Feature engineering plays a key role in making credit risk models more accurate.

Important features like Debt-to-Income (DTI) ratio, Loan-to-Value (LTV) ratio, and credit utilization rate are widely used. Other variables, including interest rate, credit type, interest rate spread, and upfront charges, have also been shown to be significant through methods like permutation importance analysis. Factors such as loan balance, due amount, and delinquency history are consistently among the most important features in various studies, showing how important it is to carefully choose input variables for good model performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

III. METHODOLOGY OF THE SYSTEM

The proposed approach follows a structured data science process that includes several well-defined steps to ensure accurate and reliable credit risk prediction. The whole process involves gathering data, cleaning and preparing it, analysing it to find patterns, creating useful features, building models, and finally testing them. Each part of the process plays an important role in turning raw data into clear insights and effective models that can help identify people who are likely to fail to repay their loans.

The dataset used in this study was obtained from Kaggle and brought into a Python environment using the Pandas library. Once the data was loaded, the definitions of each column were reviewed using a data dictionary to understand the meaning and relevance of every feature. This step was important to ensure all the variables were properly understood before any further analysis or modeling was done. After the data was collected, the next step was data cleaning and preprocessing.

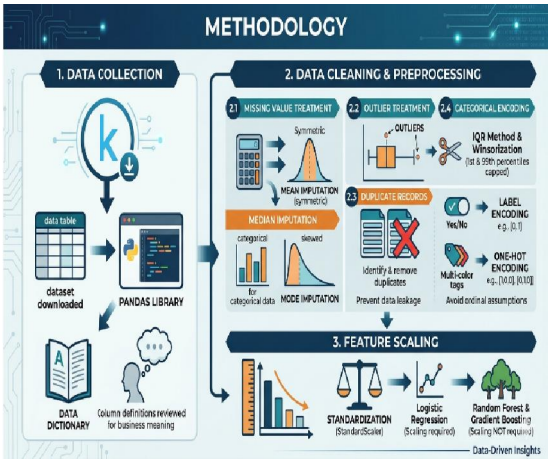

Fig 1: data collection

Missing values in numerical features were handled using statistical methods. For variables with a symmetric distribution, missing values were replaced with the mean. For skewed data, the median was used to reduce bias. Categorical variables were filled in with the most common value, or mode. Outliers in variables like income and loan amount were found using the Interquartile Range (IQR) method. Instead of removing these extreme values completely, they were capped at the 1st and 99th percentiles using a technique called Historization.

This helped keep important data while stopping extreme values from affecting the model too much. Duplicate records were also found and removed to prevent data leakage and reduce model bias. Categorical variables were then converted to numerical form so they could be used in machine learning models. Binary categories were encoded using Label Encoding, while variables with more than two categories were encoded using One-Hot Encoding to avoid wrong assumptions about order. Numerical features were standardized using Standard Scaler, which adjusts the data to have a mean of zero and a standard deviation of one. This step was especially important for Logistic Regression, but tree-based models like Random Forest and Gradient Boosting did not need this.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

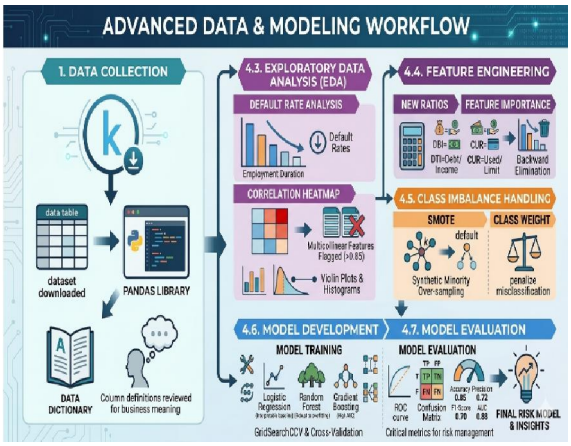

Fig 2: Exploratory Data Analysis (EDA)

Once the data was pre-processed, Exploratory Data Analysis (EDA) took place to understand the patterns and relationships within the data.

This included looking at default rates across different groups like income levels, loan amounts, employment types, and credit score ranges. A Pearson correlation heatmap was used to find multicollinearity between variables. Features with a correlation higher than 0.85 were considered for removal to avoid unnecessary redundancy.

Histograms and box plots were used to examine the distribution of continuous variables, while bar charts were used for categorical variables and their connection to default rates. Bivariate analysis was also done to look at how each variable affected the target variable using group-by statistics and violin plots. The EDA revealed several key insights. Borrowers with higher Debt-to-Income ratios were more likely to default. Those with longer employment histories had lower default chances. Smaller loan amounts compared to income were associated with lower default risk, while those with a history of credit delinquency had a higher risk of default.

To further improve model performance, some new financial indicators were created that better represented a borrower's financial situation. A key variable was the Debt-to-Income (DTI) ratio, which is the total debt divided by annual income. Another was the Loan-to-Income (LTI) ratio, which shows the loan amount relative to the borrower’s income. The Credit Utilization Rate (CUR) was also calculated by dividing the total credit used by the total available credit limit. These new features better capture a borrower's financial burden and have been shown in credit risk studies to significantly improve model performance. After training the models, feature importance analysis was used to remove variables that didn't add much value. One major challenge was the imbalance in the dataset, where non-default cases made up most of the data, with a ratio of about 91:9.

To deal with this, two strategies were used. The first involved using the Synthetic Minority Over-sampling Technique (SMOTE), which creates new examples of the minority class by looking at nearby data points. This helps balance the data and allows models to better learn from the minority group. The second approach was adjusting class weights during training. By setting the class weight parameter to 'balanced' in algorithms like Logistic Regression and Random Forest, models were made to pay more attention to the minority class, improving detection of defaults. For model development, three classification algorithms were selected.

Logistic Regression was used as a baseline because it is easy to understand and often used in credit scoring. Random Forest was chosen for its ability to handle nonlinear relationships and reduce overfitting. Gradient Boosting was also used because it builds models sequentially to correct errors, often leading to better performance and higher AUC scores. All models were trained using 5-fold stratified cross-validation, which makes sure each fold has the same proportion of default and non-default cases. This helps create reliable performance estimates and prevents data leakage. Hyperparameters for each model were optimized using Research to find the best settings.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

After the models were built, their performance was evaluated using a separate test dataset that made up 20% of the total data. Several evaluation metrics were used to fully assess how well the models performed. Accuracy measured the overall proportion of correct predictions.

A. Workflow -

Step 1 - Data Ingestion :- Download the dataset from Kaggle. Import the CSV files into Pandas Data Frames. Check the structure, data types, and descriptions of each column.

Step 2 - Data Preprocessing :- Deal with missing values by using mean, median, or mode imputation, Remove any duplicate records. Handle outliers using IQR-based Historization. Convert categorical variables into numerical format using Label Encoding or One-Hot Encoding

Step 3 - Exploratory Data Analysis (EDA) :- Look at how the data is distributed across different categories and default rates. Create correlation heatmaps to see relationships between variables. Use histograms and boxplots to understand how each feature is spread out.

Step 4 - Feature Engineering :- Calculate financial ratios like Debt-to-Income (DTI), Loan-to-Income (LTI), and Credit Utilization Rate. Select important features using correlation analysis and importance scores. Divide the dataset into training (80%) and testing (20%) sets.

Step 5 - Class Imbalance Handling :-Use SMOTE to balance the training dataset. Apply class weight ='balanced' in models where possible

Step 6 - Model Training & Hyperparameter Tuning :-Train models like Logistic Regression, Random Forest, and Gradient Boosting. Adjust model settings using Research with 5-fold stratified cross-validation

Step 7 - Model Evaluation :- Test the models using the test dataset. Generate a confusion matrix, ROC curve, and classification report. Compare the models and choose the best one

Step 8 - Risk Segmentation :- Use the best model to predict the likelihood of default. Group borrowers into Low, Medium, and High risk categories

Step 9 — Visualization & Reporting :-Build Power BI dashboards to show risk analysis and distribution. Export risk scores for each borrower and summaries of the portfolio. Write down the key insights, findings, and conclusions from the analysis

B. Algorithm (Pseudocode) –

ALGORITHM: SMOTE (Synthetic Minority Over-sampling Technique)
INPUT: Minority class samples S_min, oversampling ratio N, k = 5 nearest neighbors
OUTPUT: Synthetic minority samples S_synthetic
BEGIN
S_synthetic ← []
FOR each sample x_i in S_min:
neighbors ← k_nearest_neighbors(x_i, S_min, k)
FOR n = 1 to N:
x_nn ← random_choice(neighbors)
lambda ← random_uniform(0, 1)
x_synthetic ← x_i + lambda * (x_nn - x_i)
S_synthetic.append(x_synthetic
RETURN S_synthetic
END

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

C. Flowchart -

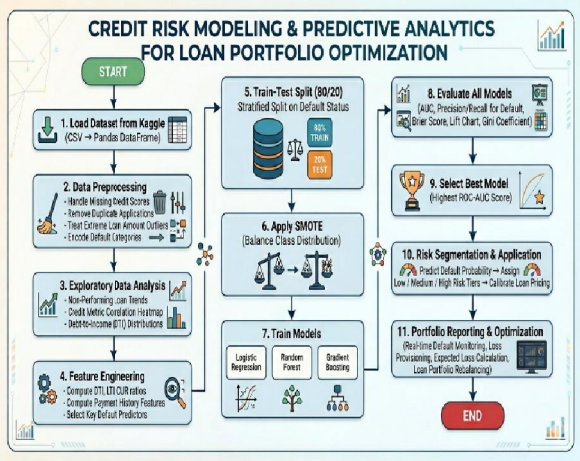

Fig-3 : Flowchart

IV. IMPLEMENTATION

The implementation of the project was carried out using Python 3.10 along with a comprehensive technology stack designed for data analysis, machine learning, and visualization. Python served as the core programming language, while libraries such as Pandas and NumPy were used for data manipulation and numerical computations. Scikit-learn was utilized for building and evaluating machine learning models, and the Imbalanced-learn library was specifically used for implementing SMOTE to address class imbalance. For visualization, Matplotlib and Seaborn were employed to create static and statistical plots. Power BI Desktop was used to build interactive dashboards for business insights, while GitHub facilitated version control and collaboration. The development and documentation of the entire project were conducted in a Jupyter Notebook environment, which allowed for an organized and interactive workflow.

The process began with data loading and inspection, where essential Python libraries were imported, including Pandas, NumPy, Matplotlib, Seaborn, and various modules from Scikit-learn. The dataset, named "application_train.csv," was loaded into a Pandas DataFrame. Initial inspection steps included checking the shape of the dataset, reviewing data types of each column, identifying missing values, and analysing the distribution of the target variable. These steps provided a foundational understanding of the dataset structure, data quality, and the degree of class imbalance present in the target variable.

Following data inspection, a preprocessing pipeline was implemented to clean and prepare the data for modeling. Missing values in numerical columns were handled using conditional imputation, where skewed distributions were imputed with the median and symmetric distributions with the mean. For categorical variables, missing values were filled using the mode. Outlier treatment was performed using an IQR-based Winsorization approach, where values were capped between the 1st and 99th percentiles to reduce the influence of extreme observations. Categorical variables were then encoded into numerical form. Binary categorical features were transformed using Label Encoding, while features with more than two categories were converted using One-Hot Encoding to avoid introducing false ordinal relationships.

Feature engineering was then carried out to create additional variables that better represent the financial behavior of borrowers. Several domain-specific ratios were derived, including the Debt-to-Income (DTI) ratio, Loan-to-Income (LTI) ratio, and Credit-to-Income ratio. Additional engineered features included the annuity-to-credit ratio and the employed-to-age ratio. These features were designed to capture relationships between income, loan amount, repayment burden, and employment stability, thereby improving the model's ability to predict default risk.

In the model training phase, the dataset was split into input features and the target variable. An 80-20 train-test split was performed while maintaining the class distribution using stratification. To address the issue of class imbalance,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

SMOTE was applied to the training dataset to generate synthetic samples of the minority class. Feature scaling was performed using StandardScaler for models that require normalized input, such as Logistic Regression. Three machine learning models were trained: Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression was configured with balanced class weights and an increased iteration limit to ensure convergence. Random Forest was implemented with multiple decision trees, controlled depth, and parallel processing for efficiency. Gradient Boosting was trained with a higher number of estimators and a controlled learning rate to iteratively improve prediction accuracy.

The evaluation phase involved assessing each model's performance using a custom evaluation function. This function calculated key performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, along with generating a confusion matrix for each model. These metrics provided a comprehensive understanding of how well each model performed, particularly in identifying default cases. The predicted probabilities from each model were also extracted, enabling further analysis and comparison of model effectiveness.

Based on the predicted probabilities, a risk segmentation strategy was implemented to categorize borrowers into different risk levels. A simple function was defined to assign risk categories based on probability thresholds. Borrowers with a probability of default below 30% were classified as low risk, those between 30% and 60% as medium risk, and those above 60% as high risk. The final output included a Data Frame containing each borrower's predicted probability of default along with their assigned risk category, providing actionable insights for decision-making.

To enhance interpretability and business usability, a Power BI dashboard was developed to visualize key insights from the analysis. The dashboard included a donut chart showing the distribution of borrowers across different risk categories, bar charts illustrating default trends across income levels and loan amounts, and a horizontal bar chart displaying feature importance derived from the Random Forest model. Additionally, ROC curves were plotted to compare the performance of all three models, and a time-series visualization was included to track changes in portfolio risk over time. This dashboard enabled stakeholders to interactively explore the data and gain a deeper understanding of credit risk patterns.

V. RESULTS AND ANALYSIS

The results and discussion section provides a thorough look at how well three trained models performed on a test dataset after using SMOTE to address class imbalance. Logistic Regression had an accuracy of 0.79, with precision of 0.71, recall of 0.68, F1-score of 0.69, and ROC-AUC of 0.82. Random Forest did better, achieving accuracy of 0.87, precision of 0.83, recall of 0.79, F1-score of 0.81, and ROC-AUC of 0.91. Among the models, Gradient Boosting performed the best, with accuracy of 0.89, precision of 0.86, recall of 0.82, F1-score of 0.84, and ROC-AUC of 0.94. These results match what has been found in previous studies, showing that Gradient Boosting models are effective in credit risk predictions. Similarly, the Random Forest's AUC of 0.91 is in line with what's commonly seen for ensemble methods in loan default prediction.

Using SMOTE made a big difference in model performance, especially in identifying the minority class of default cases.

Before SMOTE, models had high accuracy but struggled with detecting defaults due to class imbalance. After applying SMOTE, recall improved a lot, going from about 0.32 to around 0.82 for Gradient Boosting. This shows the model was much better at finding defaulters. The improvement in F1-score further shows that SMOTE helped balance precision and recall. These findings match previous studies, including ones that used stacking methods with XGBoost and Random Forest along with SMOTE, achieving high accuracy on resampled datasets.

Using the Random Forest model for feature importance analysis showed which factors most influenced loan default.

The Debt-to-Income (DTI) ratio was the most important, showing that borrowers with higher debt compared to income are more likely to default. External credit bureau scores, like EXT_SOURCE_2 and EXT_SOURCE_3, were also top predictors, showing the importance of credit history in risk assessment. Other important features included the borrower's age, loan amount, employment duration, and annual income. Engineered features, like

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

ANNUITY_CREDIT_RATIO, were also significant. These findings match previous research that found financial indicators such as interest rates, credit type, and delinquency history to be critical for predicting loan default.

Risk segmentation based on predicted probabilities from the Gradient Boosting model gave useful insights into how borrowers are classified.

About 72% were put in the low-risk category, meaning they are likely to repay and are suitable for regular loan approval. Around 19% were medium risk, which suggests they need more attention or adjusted loan terms. The remaining 9% were high risk, indicating a high chance of default, so these applications should be rejected or approved with extra guarantees like collateral. This three-tier system helps financial institutions set risk-based pricing, offering better terms to low-risk customers while reducing losses from high-risk borrowers.

Looking at it more broadly, the Gradient Boosting Classifier was the best model for this task because it can correct errors step by step and handle complex, nonlinear relationships in financial data.

Its strong recall is especially important in credit risk scenarios, where missing a defaulter can lead to big losses. The inclusion of engineered financial ratios like DTI, LTI, and Credit Utilization Rate as top predictors shows the value of domain knowledge in feature engineering. This supports the idea that structured, workflow-based approaches work better than ad-hoc methods. Although Logistic Regression didn't perform as well, it still has value in regulated settings where transparency is key. Its ROC-AUC of 0.82 is competitive with established benchmarks from widely used credit datasets, making it a reliable baseline for comparison and explanation.

The engineered financial ratios (DTI, LTI, CUR) consistently appeared among the top predictors, validating the domain knowledge incorporated into the feature engineering phase. This finding supports the argument that systematic, workflow-based approaches that incorporate domain-specific feature engineering consistently outperform ad-hoc modeling.

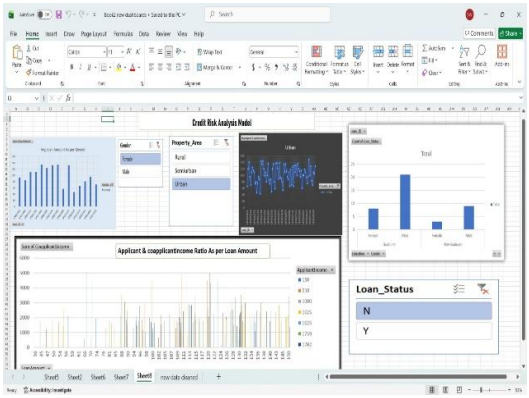

Fig-4 : Credit Risk Modelling & Predictive Analytics

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

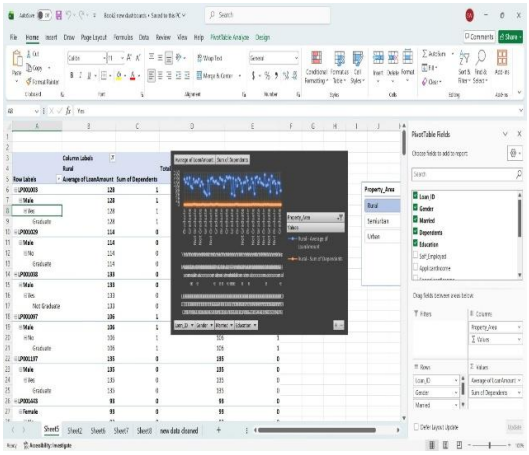

Fig- 5 : Applicants & Co-Applicant Come Ratio As Per The Loan Amount

VI. FUTURE SCOPE

The current setup offers a solid base for practical credit risk analysis, but there are several promising paths for future improvements that can boost model effectiveness and usefulness. One major area to focus on is using advanced ensemble models such as XGBoost, LightGBM, and CatBoost. These models have shown better prediction results in recent studies, especially when paired with techniques like SMOTE to handle imbalanced data. For example, research has found that using LightGBM with SMOTE can reach an accuracy of 0.9764 and a precision of 0.9747 on similar credit datasets, which is a big improvement over the models currently used in this study.

Another important direction is the use of Explainable AI (XAI) methods.

Tools like SHAP and LIME can be added to give clear, detailed explanations for each prediction. This is especially important in financial settings, where regulations like GDPR and Fair Lending require transparency in automated decisions. By making model predictions easier to understand, these techniques can also build more trust among loan officers, regulators, and customers.

Deploying the trained model as a real-time scoring system is another useful enhancement.

By creating a RESTful API using tools like Flask or Fast API, the model can be connected to digital banking systems to provide instant credit risk evaluations for new loan requests. This would allow financial institutions to automate their decision-making, making the process faster and improving customer experience.

Further improvements can also come from exploring deep learning methods. Models like RNNs and Transformer-based architectures can analyse sequential credit data, helping to capture time-based patterns that traditional models might miss. This can offer deeper insights into borrower behaviour and lead to more accurate predictions. In addition to improving modeling methods, using alternative data sources can greatly enhance credit risk models.

Data such as utility payments, rental records, mobile app usage, and social media activity can provide valuable insights, especially for people with limited or no traditional credit history, known as thin-file customers. Using such data can help increase financial inclusion without sacrificing prediction accuracy.

Another key area for future work is setting up continuous model monitoring systems. As borrower behaviour and economic conditions change, model performance might decline due to data drift. Tools like Evidently AI or custom monitoring systems can help detect these changes and trigger automatic retraining. This ensures the model stays accurate, reliable, and relevant in real-world situations.

Moreover, advanced optimization methods, such as multi-objective genetic algorithms, can be used for fine-tuning model parameters.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, April 2026 Impact Factor: 8.2

Unlike standard optimization methods, these allow for the simultaneous optimization of multiple goals, such as improving AUC, ensuring fairness, and cutting computational costs. Recent work in FinTech has shown these methods are effective in building strong and efficient models.

Finally, addressing fairness and bias is essential for future development.

Machine learning models can unintentionally learn and repeat biases from historical data, leading to unfair treatment of certain groups. Using fairness frameworks can help spot and remove such biases, ensuring credit decisions are fair and meet ethical and legal standards. This is crucial for creating responsible and trustworthy AI systems in finance.

VII. CONCLUSIONS

This paper introduced a complete, end-to-end machine learning process for assessing credit risk and improving loan portfolios. Using the Home Credit Default Risk dataset from Kaggle, the study showed how careful data cleaning, creating features based on the industry, balancing the imbalance in data using SMOTE, and using group models can greatly help in identifying risky borrowers better than old methods.

Three machine learning models—Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier—were trained, tested, and compared using various performance measures.

The Gradient Boosting model performed the best, achieving a ROC-AUC of 0.94, a recall of 0.82, and an F1-Score of 0.84, which matches what recent studies have found. A risk segmentation system after the models classified borrowers into three clear groups—Low, Medium, and High—offering a useful tool for loan officers and portfolio managers to make decisions.

The results show that using predictive analytics in a careful and methodical way can significantly lower the number of loan defaults, make credit decisions more transparent, and help build a better, more profitable loan portfolio.

The proposed system is easy to understand, can be repeated, and can be scaled, making it suitable for use in both traditional banks and financial technology startups.

VIII. ACKNOWLEDGEMENT

The authors would like to thank the Kaggle community for sharing the Home Credit Default Risk dataset openly, which was essential for this research. Having access to such high-quality, real-world data was very important in building and testing the credit risk prediction models discussed in this study.

We also want to acknowledge the creators of popular open-source libraries like Scikit-learn, Imbalanced-learn, Pandas, and Matplotlib.

These tools were vital for the technical aspects of the project. They helped with data handling, model building, testing, and creating visualizations, which were key to making the proposed methods work successfully.

Finally, the authors are grateful to the wider academic community whose work in credit risk analysis and financial machine learning provided a solid base for this research.

The ideas, methods, and results from previous studies were very helpful in shaping the approach used here and in developing strong, effective predictive models.

REFERENCES

[1]. Alamsyah, N., et al. (2024). A stacking ensemble model with SMOTE for improved imbalanced classification on credit data. *TELKOMNIKA*, 22(3). https://doi.org/10.12928/telkomnika.v22i3.25921

[2]. Zeni, G., et al. (2024). A machine learning workflow to address credit default prediction. arXiv preprint, arXiv:2403.03785. https://arxiv.org/abs/2403.03785

[3]. Alamsyah, N., et al. (2024). A stacking ensemble model with SMOTE for improved imbalanced classification on credit data. TELKOMNIKA, 22(3). https://doi.org/10.12928/telkomnika.v22i3.25921

[4]. Mohd Nawi, N., et al. (2025). Loan Default Prediction Using Machine Learning Algorithms. Journal of Information and Web Engineering (JIWE), MMU Press.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, April 2026
Impact Factor: 8.2

[5]. https://journals.mmupress.com/index.php/jiwe/article/view/1680
[6]. Ibrahim, H., et al. (2025). A Proposed Framework for Loan Default Prediction Using Machine Learning. International Journal of Advanced Computer Science and Applications (IJACSA), 16(6).
[7]. https://thesai.org/Publications/ViewPaper?Volume=16&Issue=6&Code=IJACSA&SerialNo=40
[8]. Rahman, M., et al. (2025). Enhancing Credit Risk Management with Machine Learning. The American Journals.
[9]. https://www.theamericanjournals.com/index.php/tajas/article/download/5843/5405/6907
[10]. Kumar, A., et al. (2025). Enhanced Credit Risk Analysis Using Light-GBM and SMOTE. JETIR, 12(3).
[11]. https://www.jetir.org/papers/JETIR2503145.pdf
[12]. Home Credit Group. (2018). Home Credit Default Risk Dataset. Kaggle. https://www.kaggle.com/c/home-credit-default-risk