Credit Risk Prediction Using Machine Learning

Fatima Kssayrawi
Department of Engineering and Computer Science
Oakland University, Rochester, USA
fatimakssayrawi@oakland.edu

I. Abstract

Credit risk prediction is critical in the banking and financial sector to assess whether a borrower is likely to default on a loan. In this project, we develop a supervised machine learning pipeline to classify borrowers as high or low risk based on demographic, financial, and behavioral attributes. The solution uses a benchmark dataset of over 250,000 observations and explores the effectiveness of three classification models: Decision Tree, Random Forest, and Gradient Boosting.

Given the significant class imbalance in the dataset, the SMOTE technique was applied to oversample the minority class and ensure fair training. Additional preprocessing steps included one-hot encoding of categorical features and feature scaling. The models were evaluated using accuracy, precision, recall, and F1-score, focusing on the model's ability to identify high-risk individuals correctly.

Among the tested algorithms, the Random Forest classifier achieved the best balance between precision and recall, making it the optimal choice for deployment. The project concludes with developing a reusable prediction function, enabling real-time credit risk classification based on user input or incoming borrower data.

II. Introduction

Credit risk is one of the most significant challenges faced by financial institutions, particularly in the banking sector. It refers to the likelihood that a borrower will default on a financial obligation such as a loan, bond, or credit line. Accurately predicting credit risk is essential for maintaining economic stability, minimizing losses, and ensuring regulatory compliance. Banks, lenders, and investment firms rely on robust credit risk assessment tools to make informed lending decisions, optimize their portfolios, and manage capital reserves efficiently.

Traditionally, credit risk has been assessed using statistical models that analyze historical financial data, including credit scores, income, and repayment history, to determine a borrower's creditworthiness. However, these models are often limited in capturing complex, non-linear relationships in borrower behavior. With the rise of machine learning and access to large, diverse datasets, credit risk prediction has evolved to include more sophisticated approaches that deliver higher predictive accuracy and better generalization.

In this project, we leverage a comprehensive benchmark dataset widely used in both academic and industry contexts to predict credit risk. Using a pipeline that includes data visualization, preprocessing, and machine learning, we train and compare three models: Decision Tree, Random Forest, and Gradient Boosting Machines (GBM). After applying techniques like SMOTE for class balancing and feature scaling for normalization, we evaluate the models using standard classification metrics to deliver a reliable and interpretable credit risk prediction system. Our results demonstrate that ensemble models, particularly Random Forest, achieve a strong balance between performance and interpretability, making them well-suited for deployment in real-world financial settings.

III. Related work

Credit risk modeling has a long-standing history rooted in statistical techniques. Traditional approaches such as logistic regression have been extensively used to model the Probability of Default (PD) by identifying linear relationships between borrower characteristics and default likelihood. These models often incorporate key financial indicators such as Loss Given Default (LGD) and Exposure at Default (EAD) to quantify and manage financial exposure [4], [5]. While effective in structured environments, these classical models are often constrained by their linearity and feature independence assumptions.

With data availability and computational power growth, modern research has shifted towards using machine learning for credit risk prediction. Recent studies have explored many models, including decision trees, random forests, gradient boosting, and even deep neural networks, demonstrating their ability to handle high-dimensional and non-linear datasets [6]–[9]. These models can integrate alternative data sources such as social behavior, mobile usage, and transactional patterns, providing a more nuanced view of borrower behavior and improving prediction performance.

Compared to previous work, our approach adopts a hybrid methodology that combines classical decision trees with ensemble learning techniques. This allows us to retain interpretability while improving performance through model aggregation. Unlike some studies focusing solely on accuracy, we place equal emphasis on recall and F1-score, especially for the minority (high-risk) class, which is more critical from a risk management perspective. We also address class imbalance using SMOTE, a commonly recommended technique in credit scoring literature, to ensure fair model training.

This paper uses one of the most widely used datasets in credit risk prediction, a benchmark for evaluating and

developing predictive models. This dataset is extensively employed in both academic research and industry applications due to its comprehensive nature, containing detailed information about borrowers' financial history, demographics, credit behavior, and default outcomes. To build the models, we deploy decision tree algorithms and ensemble methods. We begin with some visualizations throughout the analysis to explore and understand the dataset, communicate findings effectively, and enhance model interpretability. Then, robust data preprocessing techniques clean and transform the raw dataset, addressing missing values, outliers, and class imbalances. This step ensures high-quality input data, significantly impacting model accuracy.

Following data preparation, we implement decision tree algorithms due to their simplicity and interpretability. We employ ensemble methods like Random Forest and Gradient Boosting Machines (GBM) to improve further predictive performance, which aggregate multiple decision trees to create a more robust and precise prediction model.

By integrating these elements, our approach aims to provide an accurate and interpretable framework for predicting credit risk. This will enable financial institutions to make informed decisions based on clear insights into borrower behavior and risk factors.

The remainder of this paper is organized as follows:

- **Section 2** introduces credit risk prediction and outlines the importance of this problem in the financial sector. **Section 3** reviews related work and discusses key concepts and prior research in traditional and machine learning-based credit risk modeling. **Section 4** describes the dataset used in this study and details the preprocessing steps to prepare the data for modeling, including encoding, scaling, and class balancing. **Section 5** presents the machine learning models implemented in this project, including Decision Tree, Random Forest, and Gradient Boosting. **Section 6** discusses the experimental results, compares model performance, and analyzes confusion matrices and classification metrics in depth. **Section 7** describes the deployment process of the final model, including the saving of essential components, the structure of the prediction pipeline, and the interactive user input system designed for real-time credit risk classification. **Section 8** concludes the paper by summarizing the main findings and contributions and highlighting potential directions for future work.

IV. Dataset Description and Preprocessing

A. *Dataset Description*

The "credit risk prediction" dataset contains 252000 observations with no missing values; the test data has the target variable "risk_flag" missing with 28000 observations collected by the univ.ai hackathon.[1] .

[1] https://www.kaggle.com/datasets/arunavgautam/credit-risk-prediction-by-univai-hackathon/data?select=training+data.csv

The dataset contains a variety of borrower-related features that can be grouped into the following categories:

1. **Demographic Features:** Variables such as **age, marital status, city, and state** provide insight into the borrower's personal and geographic background. These attributes can influence an individual's financial behavior and risk profile.
2. **Employment and Stability Indicators**: Experience**, current_job_years, and current_house_years** reflect a borrower's work history and residential stability. These metrics are essential proxies for reliability and long-term financial commitment.
3. **Financial Attributes:** The dataset includes **income**, a critical factor when evaluating a borrower's repayment capacity. Higher and more stable income levels often correlate with lower credit risk.
4. **Asset Ownership:** The features **house_ownership and car_ownership** indicate whether a borrower owns key assets. Ownership may signal greater financial responsibility or stability, influencing the risk evaluation.
5. **Professional Information:** The **profession** feature adds context about the borrower's occupational field, which may relate to income stability or default risk patterns across sectors.
6. **Target Variable:** The binary column **risk_flag** is the target variable. It indicates the borrower's risk category:
    a. **1 for high risk** (likely to default)
    b. **0 for low risk** (not likely to default)
    This label is used to train classification models for credit risk prediction.

B. *Dataset Preprocessing*

In this section, we provide the preprocessing techniques applied to the dataset to ensure it is in a suitable format for use in predictive models.

- **Data Visualization and Analysis**

Here, we present several data visualizations to make the dataset more understandable and to facilitate the analysis process. Visualization plays a crucial role in understanding the distribution and relationships between different features in the dataset. For example, pair plot visualizations examine pairwise relationships and correlations between features, providing insights into how the features interact.

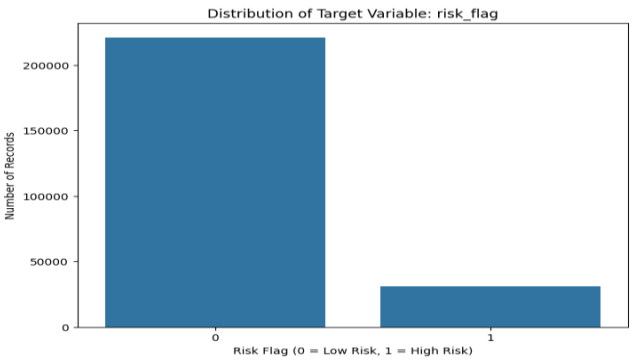

Figure 1 Distribution of Target Variable: Risk Flag

1. **Risk Flag Distribution (Target Variable)**: As described in Figure 1, the initial class distribution revealed a significant imbalance between the two classes, with most borrowers labeled as low risk (risk_flag = 0). This highlights the need for balancing techniques like SMOTE, as models trained on imbalanced data favor the majority class and perform poorly on the minority (high-risk) group.
2. **Income Distribution**: The income distribution plot offers a detailed overview of the income levels present in the training dataset, as described in Figure 2. By plotting the frequency of various income values, we can understand the overall spread and central tendency of income among the individuals. Identifying any skewness or outliers in the data is crucial, as it can significantly impact model performance if left unaddressed. Such a distribution plot also allows us to infer whether the dataset contains individuals with predominantly low, middle, or high income, which could be an essential factor in predicting credit risk. The income distribution appears uniformly spread across the dataset, with no extreme peaks or dips. This indicates that the dataset includes various income levels and avoids heavy skewness.

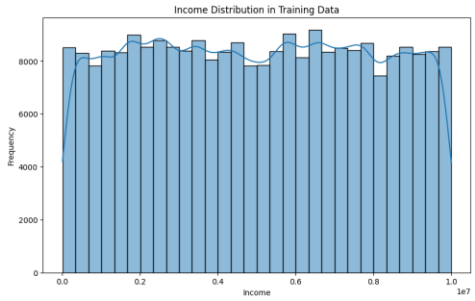

Figure 2 Income distribution

3. **Age Distribution by Risk Flag**: This visualization provides insights into the relationship between age and the credit risk flag as described in Figure 3. By distinguishing age distributions for those with and without the risk flag, we can determine if specific age groups are more prone to high risk. For instance, younger individuals might exhibit a different risk profile than older individuals, which could be linked to financial stability, job experience, or spending habits. This understanding could be used to create age-based features that enhance model predictability.

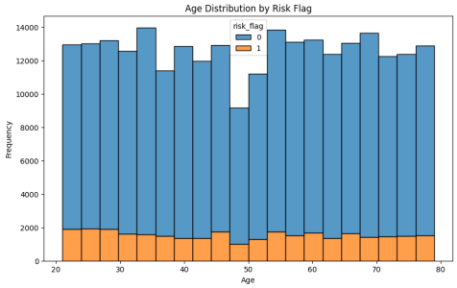

Figure 3 Age Distribution Vs Risk flag

4. **Correlation Heatmap**: The correlation heatmap for numerical features provides an overview of the interrelationships between various numerical attributes in the dataset, as described in Figure 4. By observing the correlation coefficients, we can determine which features highly correlate with one another and the target variable. This information is vital in feature selection, as highly correlated features might introduce multicollinearity, adversely impacting model performance. The heatmap also identifies features most strongly correlated with the risk flag, guiding which attributes could be most informative for predictive modeling.

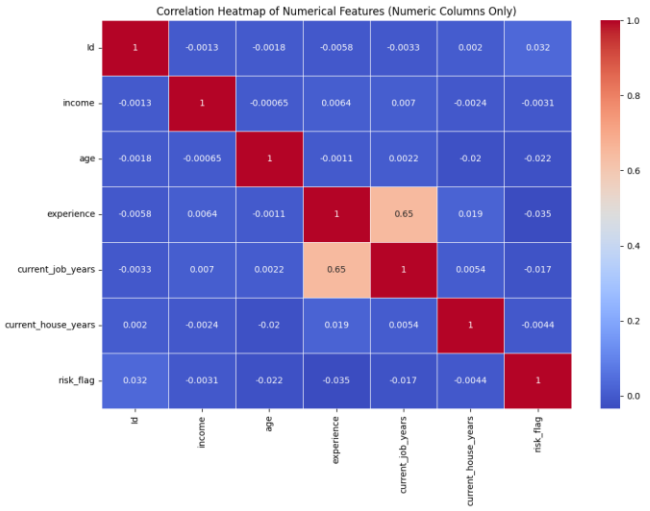

Figure 4 Correlation Map

These visualizations provide a foundational understanding of the dataset and help guide decisions regarding feature engineering, data preprocessing, and model selection. Including these visualizations and analyses in the paper will enhance the dataset's interpretability and the rationale behind subsequent modeling decisions.

Furthermore, Principal Component Analysis (PCA) is also applied as described in Figure 5, which reduces the dataset's dimensionality and enables data visualization in two components. This visualization helps assess the classes' separability and provides insight into the dataset's underlying structure.

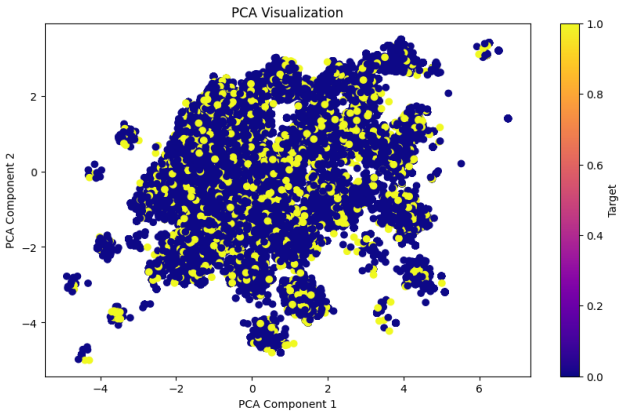

Figure 5 PCA Visualization

- **Handling Duplicate Data**

As a first step, the dataset was examined for duplicate entries. Duplicate records can introduce data leakage, skew model performance, and lead to overfitting by reinforcing patterns the model has already seen. After running the duplicate check, it was confirmed that the dataset contains **no duplicated rows**, indicating clean and unique samples across all observations.

- **Encoding Categorical Variables**

The dataset includes several categorical features: **profession, married, house_ownership, car_ownership, city, and state**. To convert these into a machine-readable format, **one-hot encoding** was applied using **pd.get_dummies ().** This transformation allows models to interpret categorical variables without introducing ordinal bias.

- **Separating Features and Target Variable**

Following the encoding process, the dataset was split into **independent features (X)** and the **target variable (y),** the risk_flag, which indicates whether a borrower is high or low. A count plot was used to visualize the class distribution, revealing a **severe class imbalance** in favor of low-risk borrowers.

- **Addressing Class Imbalance with SMOTE**

To correct the severe class imbalance observed in the dataset**, the Synthetic Minority Over-sampling Technique (SMOTE)** was applied. SMOTE generates synthetic examples for the minority class by interpolating between existing samples, effectively increasing its representation and ensuring balanced class distribution during training. SMOTE was applied only to the training set, not the test set. This is intentional and important:

- **Applying SMOTE to the training set only** ensures that the model learns from a balanced dataset and can generalize well.
- Applying SMOTE to the test set would introduce synthetic (artificial) data into the evaluation, compromising the integrity of the model's performance metrics.
- **The test set must remain untouched** and reflect real-world class distributions to estimate model generalization accurately.

After applying SMOTE, both classes in the training set were balanced at **176,803 samples** each. A bar chart comparison before and after SMOTE confirmed the successful redistribution of class labels, improving the model's ability to identify high-risk borrowers without bias toward the majority class.

A bar chart comparing class distributions before and after SMOTE confirmed the successful balancing of the dataset.

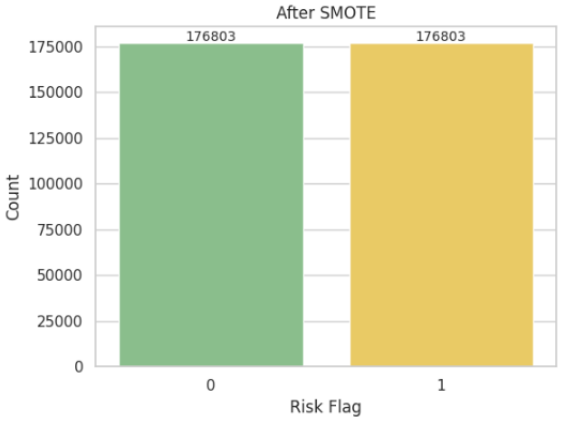

Figure 6 Class distribution after applying SMOTE

- **Standardizing the Data**

After balancing the data, **feature scaling** was performed using StandardScaler to ensure that all numeric features have the same scale and distribution. This step is particularly important for distance-based algorithms and improves convergence and consistency across machine learning models.

V. PREDICTIVE MODELS

This section introduces the classification algorithms used to predict credit risk, highlighting their underlying mechanisms, reasons for selection, and relevance to the specific challenges of the task. Selecting appropriate classifiers is essential, as each algorithm brings distinct advantages in processing structured data, capturing complex patterns, and addressing class imbalances—common characteristics in credit risk prediction problems. The goal is to assess each model's ability to accurately classify borrowers based on their default risk while uncovering key factors contributing to high or low creditworthiness.

A. Decision Tree (DT)

Decision Trees are a non-linear, rule-based method that splits the data into branches based on feature values. Each node in a decision tree represents a feature or attribute, and the branches from the node indicate the possible outcomes based on that feature's values. This branching approach allows decision trees to identify complex relationships and interactions between features. They are particularly valued for their simplicity and interpretability since the decision-making process can be visualized in an intuitive tree structure, making it easy for stakeholders to understand the results. Decision Trees can effectively handle categorical and numerical data, making them widely applicable in credit risk prediction and other classification problems [10].

B. Gradient Boosting (GB)

Gradient Boosting builds sequential decision trees, where each subsequent tree focuses on correcting the errors of the previous ones. By iteratively refining the predictions, Gradient Boosting Models (GBM) become highly effective in capturing subtle patterns and

relationships within the data. The algorithm minimizes a specified loss function, ensuring continuous improvement throughout the iterations. This feature makes GBM particularly powerful in achieving high predictive accuracy, often outperforming simpler models in real-world applications [11], [12].

C. *Random Forest (RF)*

Random Forest is an ensemble learning method aggregating multiple decision trees, where each tree is trained on a different subset of the data. By combining the outputs of these trees through majority voting or averaging, the model provides more stable and accurate predictions. Random Forest reduces the likelihood of overfitting, as combining diverse trees mitigates the risk of focusing too heavily on specific features or noise in the data. It is highly effective in handling large datasets and dealing with missing values, making it a popular choice in credit risk prediction tasks [13], [14].

Finally, we assess the effectiveness of these classifiers using a range of evaluation metrics, including accuracy, precision, recall, and F1-score. This allows us to compare the classifiers' performance comprehensively and determine the most suitable model for predicting credit risk in this context.

VI. Results and Discussions

The performance of the three classification models: Random Forest, Decision Tree, and Gradient Boosting, was evaluated using four key metrics: **accuracy**, **precision**, **recall**, and **F1-score**. The results are summarized in **Table 1**.

- **Random Forest** delivered the highest overall performance, achieving an **accuracy of 0.90**, with a **precision of 0.93**, a **recall of 0.96**, and an **F1-score of 0.94**. These metrics indicate that the model is both precise in its predictions and highly effective in identifying true cases, making it the most balanced and reliable classifier among the three.
- **Decision Tree** also performed strongly, with an **accuracy of 0.89**, **a precision of 0.92, a recall of 0.93**, and an **F1-score of 0.93**. While slightly below Random Forest regarding recall and F1-score, it still provides a solid performance with interpretable decision rules.
- **Gradient Boosting** showed comparatively lower results, with an **accuracy of 0.83**, **a precision of 0.89, a recall of 0.72, and an F1-score of 0.80**. The notably lower recall suggests that the model struggles to effectively identify high-risk borrowers, which may be critical in credit risk applications where false negatives carry significant consequences.

Based on these results, **Random Forest** is the most effective model for credit risk prediction, combining high accuracy with strong performance across all evaluation metrics.

Model	Accuracy Score	Precision	Recall	F1 Score
Random Forest	0.90	0.93	0.96	0.94
Decision Tree	0.89	0.92	0.93	0.93
Gradient Boosting	0.83	0.89	0.72	0.80

Table 1 performance results for the three models studied

In summary, the Random Forest model outperformed the other models in all aspects, as described in Figure 7, demonstrating the best balance between precision and recall. The Decision Tree model showed moderate performance, making it a viable alternative, though slightly less robust. With its lower recall and F1 score, the Gradient Boosting model appears less suitable for this classification task, especially when accurately identifying positive cases is critical.

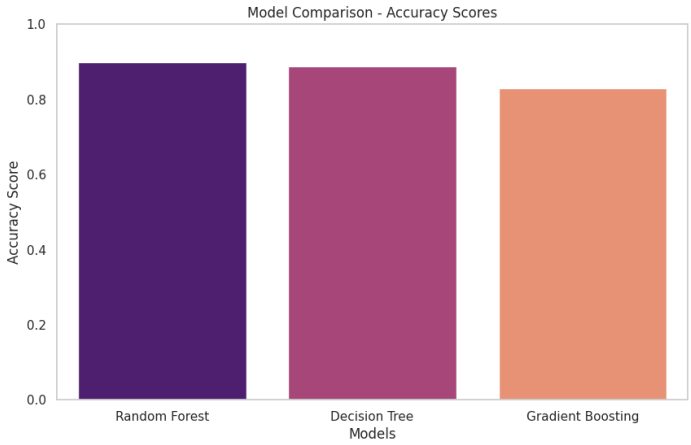

Figure 7 Model comparison in terms of accuracy

The confusion matrices for the Random Forest, Decision Tree, and Gradient Boosting models are illustrated in **Figure 8**, **Figure 9**, and **Figure 10**, respectively. These visualizations highlight each model's strengths and weaknesses in identifying both high-risk and low-risk borrowers.

1. **Random Forest** exhibited the most balanced performance, achieving **42,313 true negatives** and **2,937 true positives**, while maintaining relatively **low false positives (1,888)** and **false negatives (3,262)**. This suggests the model effectively minimizes missed high-risk cases and false alarms for low-risk borrowers.
2. **Decision Tree** performed reasonably well, with **41,144 true negatives** and **3,597 true positives**. However, it had a **higher false positive count (3,057)** than Random Forest, indicating a greater tendency to misclassify low-risk borrowers as high-risk.
3. **Gradient Boosting**, while showing decent results on low-risk predictions (**40,773 true negatives**), struggled significantly with identifying high-risk borrowers. It produced **only 1,038 true positives** and a high **false negative count of 5,161**, meaning

it failed to classify many risky borrowers correctly. This underperformance may lead to a serious underestimation of credit risk in real-world scenarios.

Overall, the **Random Forest model** achieved the best trade-off between **precision, recall, and F1-score**, making it the most **robust and dependable choice** for accurate credit risk prediction across both classes.

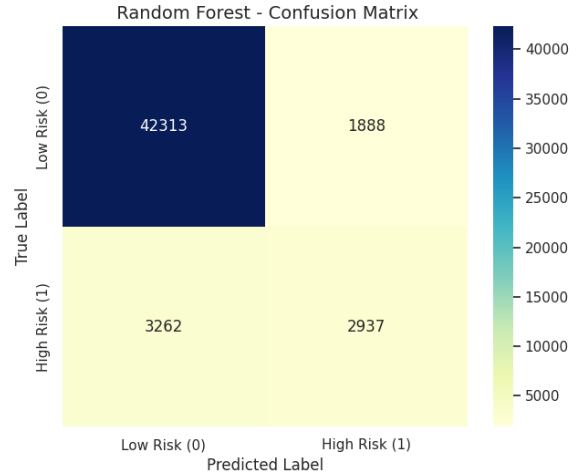

Figure 8 Confusion Matrix -Random Forest Model

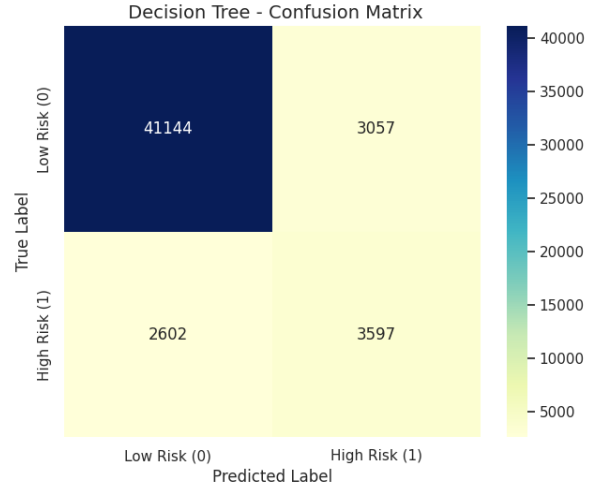

Figure 9 Confusion Matrix for Decision Tree Model

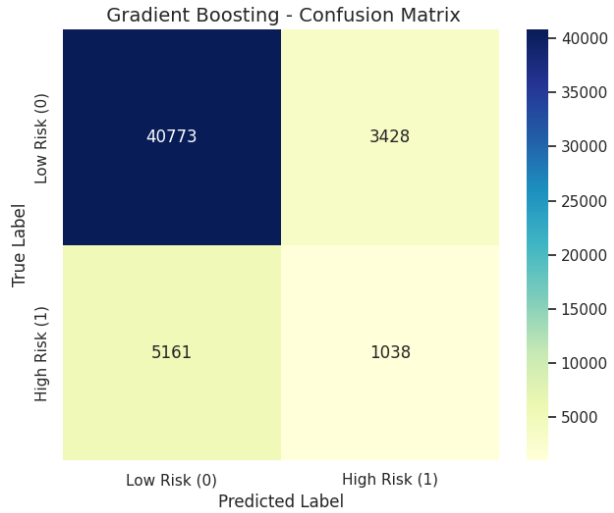

Figure 10 Confusion Matrix of the Gradient Boosting Model

VII. Model Deployment

To enable real-time credit risk prediction, the best-performing model—**Random Forest**—was prepared for deployment alongside the necessary preprocessing tools. The deployment pipeline was designed to ensure consistency between training and inference by saving the key components used during the modeling phase.

1. Model and Preprocessing Artifacts

The following components were saved using **joblib** for later reuse:

- **random_forest_model.pkl**: The trained Random Forest classifier.
- **scaler.pkl**: The StandardScaler instance used to standardize input features.
- **model_columns.pkl:** The list of feature columns used after one-hot encoding to ensure alignment during prediction.

These files allow the system to transform and evaluate new borrower data exactly as done during model training, maintaining integrity and prediction reliability.

2. Real-Time Prediction Function

A reusable Python function, **predict_credit_risk(input_data)**, was developed to automate making predictions for new users. It takes a dictionary of borrower features (e.g., income, age, house_ownership) as input, performs one-hot encoding and scaling, and returns a binary prediction:

- **0:** Low Risk (borrower likely to repay)
- **1:** High Risk (borrower likely to default)

This function loads the saved model, scaler, and feature columns to ensure compatibility between the training and inference pipelines. It also automatically handles unseen or missing features by filling in with default values when necessary.

3. User Interaction Interface

To simulate real-time usage, a **command-line interface (CLI)** was implemented using a helper function, **get_user_input(),** which prompts the user to input values for all relevant borrower features. These inputs are passed into the prediction function, and the model returns an interpretable result indicating whether the borrower is at high or low risk.

VIII. Conclusion

This study focused on developing an effective credit risk prediction model by evaluating and comparing the performance of Random Forest, Decision Tree, and Gradient Boosting classifiers. The data preprocessing involved several essential steps to ensure robust model performance. Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance, enhancing the model's ability to accurately predict minority class instances, which is crucial in credit risk scenarios. Additionally, categorical features were processed using appropriate encoding techniques to ensure compatibility with the machine learning algorithms. The results showed that the Random Forest model, with an accuracy score of 94.93%, provided the most accurate and balanced predictions, outperforming Decision Tree and Gradient Boosting in precision, recall, and F1 score.

In future work, we aim to further model optimization through hyperparameter tuning and feature engineering to enhance predictive accuracy. Additionally, incorporating more diverse data sources, such as alternative credit data or macroeconomic indicators, could provide a more comprehensive risk assessment. Exploring advanced ensemble techniques or integrating deep learning models may also improve performance. Finally, deploying and testing these models in a real-world credit risk management system would provide valuable insights into their practical applicability and reliability in dynamic environments. These avenues of exploration could further strengthen the model's robustness and effectiveness for credit risk prediction in real-world applications.

IX. Appendix

The Jupyter Notebook with the corresponding Python code and the dataset used to complete this report can be found on **GitHub** at the following address:

https://github.com/fatimakssayrawi9/Credit-Risk-Prediction-Using-Machine-Learning

References

[1] D. P. Louzis, A. T. Vouldis, and V. L. Metaxas, "Macroeconomic and bank-specific determinants of non-performing loans in Greece: A comparative study of mortgage, business and consumer loan portfolios," *Journal of Banking & Finance*, vol. 36, no. 4, pp. 1012–1027, 2012.

[2] C. Zhang, X. Zhang, and L. Yang, "Credit risk prediction model combining SMOTE with deep learning-based feature selection method," *Journal of Financial Stability*, vol. 61, p. 101088, 2022.

[3] S. Dey, S. Bhattacharya, and P. Dey, "A review of machine learning techniques for credit risk modeling and predicting default risk," *Financial Innovation*, vol. 6, no. 1, pp. 1–28, 2020.

[4] V. García, A. I. Marqués, and J. S. Sánchez, "An insight into the experimental design for credit risk and corporate bankruptcy prediction systems," *Journal of Intelligent Information Systems*, vol. 54, no. 2, pp. 341–370, 2020.

[5] A. Resti and A. Sironi, *Risk Management and Shareholders' Value in Banking: From Risk Measurement Models to Capital Allocation Policies*, Routledge, 2021.

[6] G. Wang, J. Ma, L. Huang, and K. Xu, "Two credit scoring models based on dual feature selection and ensemble learning," *Knowledge-Based Systems*, vol. 89, pp. 274–285, 2021.

[7] H. Huang, J. Zhou, L. Sun, and G. Zou, "Credit risk evaluation with kernelized soft-margin softmax regression," *Applied Soft Computing*, vol. 89, p. 106080, 2020.

[8] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[9] H. Zhu, L. Du, and Z. Li, "A hybrid deep learning model for predicting credit risk," *IEEE Access*, vol. 10, pp. 101125–101136, 2022.

[10] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Boca Raton, FL, USA: CRC Press, 2017.

[11] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.

[13] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[14] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.