Financial Institutions Loan Default Prediction Using Machine Learning Models

Khoong Tai Wai[1], Nyo Zhi Ni[1], Chou Nyet Xing[1],
Ng Pei Chi[1], Lee Ruo Fan[1] and Au Yong Eu Jin[1]

[1]Faculty of Accountancy, Finance and Business, Tunku Abdul Rahman University of Management and Technology, Malaysia

Received: 06.02.2025, Revised: 25.06.2025, Accepted: 05.08.2025

Abstract

Loan default has posed significant challenges for lending institutions in the financial sector. However, the development of machine learning (ML) offers transformative approaches to improve credit risk assessment and decision-making accuracy. This study explores the factors influencing loan default and the application of ML models in the financial services industry. The effectiveness of various ML algorithms is analyzed using a Kaggle dataset of borrowers to identify the optimal model. The original dataset consists of 148,670 records with 34 features. After pre-processing, the dataset was refined to 121,203 records with 27 features. The data cleansing process improves quality and reliability, thus enhancing prediction accuracy. The study applies Random Forest, XGBoost, Decision Tree, K-Nearest Neighbors, and Logistic Regression to compare predictive accuracy. The results demonstrate that XGBoost and Random Forest achieve the highest accuracy, with 89.70% and 89.07% respectively. This research contributes to loan default prediction development by identifying more effective ML algorithms, enabling financial institutions to make informed lending decisions and mitigate financial risks. Furthermore, the study emphasizes challenges related to interpretability, class imbalance, and regulatory compliance. Future research should expand the dataset to cover multiple years, include diverse sources, and integrate demographic and macroeconomic variables to improve model accuracy and generalizability.

Keywords: Loan Default Prediction, Credit Risk Assessment, Machine Learning Algorithms, Ensemble Models (Random Forest & XGBoost), Financial Institutions

Introduction

In the financial sector, borrower's credit risk has posed significant challenges for lending institutions. Loan default can lead to reduced profitability, increased non-performing loans, stricter lending policies, and reputational damage. For example, a study highlighted that bad loans can reduce banks' profitability and limit their ability to issue new credit (Fredriksson & Frykström, 2019). An increase in loan defaults can also result in banks adopting more conservative lending practices, making it more challenging for individuals and businesses to access credit (Withers, 2025). Moreover, loan defaults can damage a bank's reputation, erode customer trust, and cause a decline in market share (Forvis Mazars, 2024).

The development of artificial intelligence (AI) and machine learning (ML) offers transformative tools for enhancing credit risk assessment. These technologies enable the analysis of vast datasets to identify patterns and predict the likelihood of loan defaults with greater precision than traditional methods. For example, AI-driven credit software can automatically approve or deny applicants based on model outputs, streamlining the loan approval process and reducing operational costs (Lee, 2024). Furthermore, AI and

Corresponding author: [1]khoongtw@tarc.edu.my

ML can facilitate real-time risk monitoring, which allows banks to detect early warning signs of default and take proactive measures (Nallakaruppan et al., 2024). Additionally, these advanced technologies can help identify unusual transaction patterns indicative of fraudulent activities, thereby protecting both the bank and its customers (KPMG, 2021).

However, integrating ML into credit risk assessment presents challenges that must be addressed. The interpretability of complex models remains a key concern, as financial regulators and stakeholders require transparency in credit decisions (Kremer et al., 2024). In addition, the effectiveness of ML models depends heavily on the quality of training data; biased or incomplete datasets can lead to unfair assessments (Cedar Rose, 2024). Ensuring compliance with regulatory frameworks and ethical standards is also critical to preventing systemic risks associated with automated credit decisions (S&P Global, 2025). While ML has the potential to revolutionize credit risk assessment, financial institutions must implement these technologies responsibly, balancing innovation with fairness and accountability. Importantly, dataset imbalance can bias predictions toward non-defaults. Although ensemble models such as Random Forest and XGBoost handle this better, future research should incorporate techniques like oversampling (SMOTE, ADASYN) or cost-sensitive learning to enhance fairness and accuracy.

According to Bank Negara Malaysia (BNM), a stress test showed that under adverse scenarios of income and employment shocks, between 3.8% and 4.0% of banking system loans could be at risk of default by 2024 due to borrowers having insufficient financial buffers (Mail, 2022). Thus, mortgage default prediction is crucial for financial institutions, as accurate insolvency forecasting lowers credit risk and enables proper provisioning planning (Krasovytsky et al., 2024). While loan defaults cannot be eliminated, predictive modelling can significantly reduce risk and losses. Defaults contribute to financial losses for banks, tighter lending policies, and wider economic instability. To mitigate these risks, banks increasingly rely on predictive analytics to identify potential defaulters early (Krasovytsky et al., 2024). These models analyze borrower income stability, credit scores, loan amounts, interest rates, and other key risk factors, allowing for more accurate assessments of creditworthiness. By integrating these insights, banks can optimize loan approvals, implement proactive risk management strategies, and customize repayment plans to reduce default rates. Ultimately, this enhances financial stability while promoting sustainable lending practices.

Thus, the objective of this study is to evaluate the most suitable ML model for loan default prediction. The significance of this research lies in its potential to support financial institutions in enhancing risk assessment accuracy, reducing non-performing loans, and improving credit allocation. By systematically comparing traditional and ensemble ML algorithms on a large borrower dataset, the study provides evidence-based guidance for selecting predictive models that balance accuracy, fairness, and interpretability. This is particularly important in Malaysia and other developing economies where regulatory scrutiny, financial inclusion, and systemic stability are pressing concerns. Motivated by the rising complexity of borrower behavior and the limitations of conventional credit scoring systems, the study addresses a critical need for more sophisticated, data-driven tools that can capture non-linear patterns in loan performance. In doing so, it contributes to both academic knowledge and practical policy discussions on the adoption of AI-driven technologies in financial risk management.

Literature Review

Loan Default

A study by An et al. (2020) highlights the importance of a borrower’s profile in predicting loan default risk. Factors such as favorable loan terms, high credit scores, stable yearly income, and responsible line recycling rates reduce an individual's risk of default. Besides, variables such as loan interest rates, debt-to-income (DTI) ratios, and the number of negative public records increase the likelihood of defaulting on a loan. The study concludes that as borrowing rates, DTI ratios and negative public records increase, the likelihood of loan default also increases. Besides, research showed interest rates indicate the cost of borrowing money and are often expressed as a percentage of the borrowed amount. Higher interest rates increase loan costs, making it harder for borrowers to repay their loans. This is because higher rates raise installment amounts, which can lead to poor loan performance. Loan default and interest rates are closely related, in which as higher interest rates make regular payments more difficult. Furthermore, a borrower's ability to repay a loan is strongly tied to income. A borrower with a high and stable monthly income from various sources is more likely to meet repayment obligations. Conversely, those with lower incomes are more likely to struggle with loan payments, leading to a higher risk of default (Uddin, 2019). Besides, a descriptive study by Ali (2021) examined factors influencing loan default risk among 176 respondents from the banking sector, using a “Yes” or “No” checklist questionnaire. The results showed that over 80% of respondents agreed on the importance of interest rate, income, loan size, loan payment tenure, and borrower history. Specifically, 89.77% cited loan size, 86.36% income, 84.66% interest rate, 83.52% loan tenure, and 80.68% borrower history as key determinants of default risk.

These studies collectively indicate that loan default risk is shaped by both borrower characteristics and loan attributes. However, they also reveal the limitations of traditional assessment approaches, which may overlook non-linear relationships between factors. For instance, the interaction between income and loan size may amplify risk in ways not captured by simple statistical models. This gap highlights the motivation for using advanced ML techniques capable of handling complex patterns, interactions, and non-linearities in borrower data, providing a more comprehensive and accurate prediction of default risk.

Machine Learning Models

ML has significantly improved loan credit analysis by leveraging big data and advanced computational techniques. Financial institutions can assess credit risk more accurately by analyzing borrower’s personal information, credit history, and other relevant factors, leading to more informed loan decisions (Raheem, 2024). A key advantage of ML in credit risk assessment is its ability to learn patterns from historical data and generate predictions without explicit programming. Among various ML techniques, supervised learning is the most widely used, as it trains a model on a labeled dataset to recognize relationships and predict outcomes for new, unseen data. Unlike traditional econometric models that explain relationships between variables, ML models prioritize predictive accuracy, enabling more dynamic and adaptable decision-making. In addition, digital transactions generate massive amounts of valuable data. If effectively processed, this big data enhances credit risk assessment. The Covid-19 pandemic further underscored the need for faster and more reliable credit evaluation methods, as financial institutions faced increased loan applications and rising default risks. In contrast to traditional models, ML standardizes decision-making, improving efficiency and reducing inconsistencies (Hoang & Wiegratz, 2023). While ML adoption in banking is still in its early stages, its rapid growth signals a transformative shift in financial risk management.

Logistic Regression (LR) is a fundamental algorithm for binary classification. It assigns observations to distinct categories by estimating the probability of class membership. The model applies the logistic sigmoid function to a linear combination of input features, mapping predictions into the range [0,1] (Patel et al., 2020). The sigmoid, with its characteristic S-shaped curve, transforms raw values into probabilities, enabling decision-making based on a specified threshold. Due to its simplicity, interpretability, and efficiency, LR remains a widely used baseline model in predictive analytics, although it is limited in capturing complex non-linear relationships.

Random Forest (RF) is an ensemble learning method designed to improve predictive accuracy and reduce overfitting by aggregating multiple decision trees (Patel et al., 2020). Each tree is trained on a random subset of both the dataset and its features, ensuring diversity among learners. For classification, RF predicts outcomes through majority voting, where the final class label (ŷ) is determined by the most frequent output across all trees:

$$\hat{y} = \text{mode}\ (T_1(x), T_2(x), \ldots, T_n(x)) \quad (1)$$

For regression tasks, it calculates the average prediction from all decision trees:c

$$\hat{y}= \frac{1}{n}\sum_{i=1}^{n} T_i(x) \quad (2)$$

where $T_i\ (x)$ represents the prediction from the i_{th} decision tree. This ensemble approach enhances stability and improves predictive performance.

Decision Trees (DTs) are supervised machine learning algorithms used for both classification and regression tasks. They structure data in a hierarchical, tree-like form, where internal nodes represent decisions based on specific features, branches indicate possible outcomes, and leaf nodes denote final predictions or class labels (Aslam et al., 2019). The tree recursively splits the dataset into smaller, more homogeneous subsets according to attribute-based conditions, resulting in an intuitive flowchart-like structure.

A key advantage of DTs is their interpretability: the visual tree structure makes it easy to trace how individual predictions are derived, which is valuable for decision-making processes in domains such as credit risk analysis, medical diagnosis, and recommendation systems.

To identify the most informative splits, DTs employ criteria such as **Entropy, Information Gain,** and the **Gini Index:**

Entropy Equation

$$H(S) = -\,\Sigma\, p_i \log_2 p_i \quad (3)$$

where P_i represents the probability of each class in the dataset. Entropy measures the impurity in a dataset, and a lower entropy value indicates a purer node.

Information Gain (IG)

$$IG = H(S) - \ \Sigma \frac{|S_v|}{|S|} H(S_v) \quad (4)$$

where $H(S)$ is the entropy before the split, S_v is a subset of S and $H(S)$ is the entropy of that subset. The feature with the highest IG is selected for splitting.

Gini Index

$$Gini(S) = 1 - \ \Sigma\, p_i^2 \quad (5)$$

where p_i is the probability of a class in the dataset. A lower Gini Index value indicates better split purity.

K-Nearest Neighbors (KNN) is a supervised ML algorithm used for both classification and regression tasks. It makes predictions by identifying the k nearest data points (neighbors) based on a selected distance metric. For classification, KNN assigns the majority class among the neighbors, while in regression, it calculates the average value. The choice of the k parameter significantly impacts the results, as different values of k can lead to varying

outcomes (Lai, 2020). The distance between a given test point x and each training point x_i is typically calculated using
Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{ij})^2} \tag{6}$$

where x_j represents the feature values of the test sample, and x_{ij} represents the feature values of the training sample. The algorithm selects the k closest neighbors based on this distance and predicts the output as:

Classification: The majority class among the k neighbors (mode of the labels).

Regression: The average of the k nearest values.

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning algorithm that enhances the traditional gradient boosting algorithm while minimizing computational resource usage. It improves traditional gradient boosting by incorporating regularization to mitigate overfitting, optimizing sorting through parallel processing to speed up execution, and pruning trees based on maximum depth to reduce runtime. XGBoost constructs decision trees sequentially, where each new tree corrects the errors of the previous ones, making it highly effective for structured data problems such as classification and regression (Kanaparthi, 2023). The XGBoost algorithm follows the gradient boosting framework and optimizes the following objective function:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

where $l(yi,\ \hat{y}i)$ represents the loss function that measures prediction error, and $\Omega(f_k)$ is the regularization term that controls model complexity. This formulation ensures a balance between model accuracy and generalization, making XGBoost one of the most effective algorithms in predictive modeling.

Various scholars have explored the effectiveness of different ML models for predicting loan defaults, offering significant insights for financial institutions. Among these models, LR showed moderate performance overall. Notably, Orji et al. (2022) and Lin (2024) reported the highest accuracy, exceeding 80%, whereas Soni and Shankar (2022) recorded a lower accuracy of 65.34%. In contrast, the RF model consistently demonstrated high performance. Studies by Orji et al. (2022), Satheeshkumar et al. (2024), and Lin (2024) all reported accuracies above 90%, highlighting the model's robustness. For the DT model, performance varied. Orji et al. (2022) achieved the highest accuracy of 91.11%, whereas Zhu et al. (2023) reported the lowest at 63.17%, indicating moderate reliability overall. Regards the KNN model also exhibited moderate accuracy across studies. The highest accuracy of 93.33% was achieved by Orji et al. (2022), while Soni and Shankar (2022) reported the lowest at 78.17%. XGBoost emerged as another strong performer, with Satheeshkumar et al. (2024) and Lin (2024) achieving accuracies above 95%. Zhu et al. (2023) also reported a respectable accuracy of 80.98%. This comparison highlights the growing effectiveness and reliability of ensemble learning techniques, particularly RF and XGBoost, in predicting loan defaults. These models consistently outperformed traditional algorithms, underscoring their potential as robust tools for risk assessment in financial institutions.

Methodology

Data Collection

This study employs a Kaggle dataset comprising borrower-related features that serve as the foundation for loan default prediction (Yesser, 2022). The original dataset contained 148,670 records across 34 features. After a structured preprocessing procedure, the dataset was refined to 121,203 records with 27 features, ensuring improved quality and consistency.

The dataset includes a broad range of borrower-specific attributes, such as personal details, loan characteristics, and credit history information. The dependent variable is the **status** column, where a value of *1* indicates a defaulted loan and *0* represents a fully paid loan. The accuracy of predictive modeling is highly dependent on the quality of input data; hence, careful preprocessing was necessary to enhance the reliability and performance of the machine learning models applied in this study.

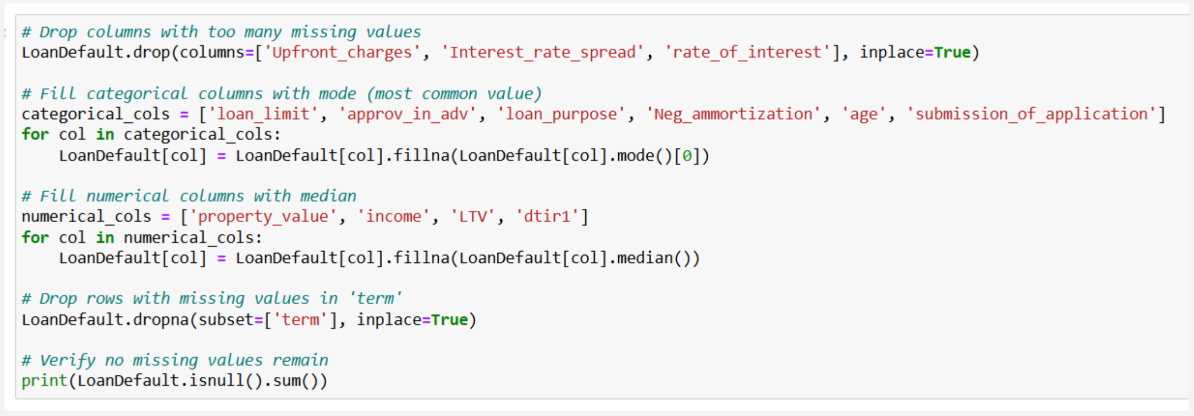
Data Preprocessing

To ensure data reliability, several steps were undertaken, as summarized in Figure 1. First, missing values were assessed, revealing that 14 of the 30 retained variables contained incomplete data. Variables with excessive missing values, such as *upfront charges*, *interest rate spread*, and *rate of interest*, were excluded to reduce bias and prevent unreliable conclusions.

For categorical variables (e.g., *loan limit*, *approv in adv*, *loan purpose*, *neg amortization*, *age*, and *submission of application*), missing entries were imputed using the **mode**. This approach preserved the categorical structure without artificially inflating variability. For numerical variables (e.g., *property value*, *income*, *loan-to-value ratio [LTV]*, and *debt-to-income ratio [DTI]*), missing data were replaced with the **median**, a measure less sensitive to extreme outliers than the mean.

Figure 1

Handle Missing Value



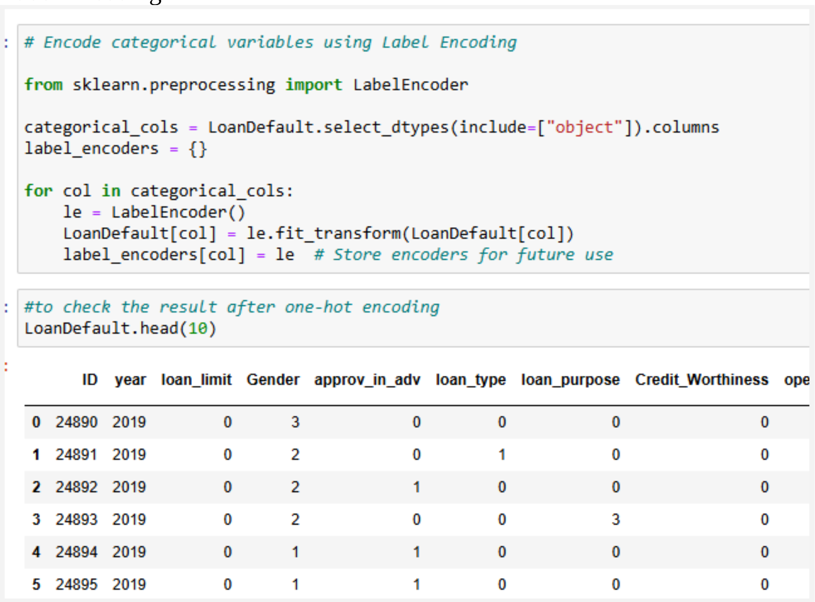
```
# Drop columns with too many missing values
LoanDefault.drop(columns=['Upfront_charges', 'Interest_rate_spread', 'rate_of_interest'], inplace=True)

# Fill categorical columns with mode (most common value)
categorical_cols = ['loan_limit', 'approv_in_adv', 'loan_purpose', 'Neg_ammortization', 'age', 'submission_of_application']
for col in categorical_cols:
    LoanDefault[col] = LoanDefault[col].fillna(LoanDefault[col].mode()[0])

# Fill numerical columns with median
numerical_cols = ['property_value', 'income', 'LTV', 'dtir1']
for col in numerical_cols:
    LoanDefault[col] = LoanDefault[col].fillna(LoanDefault[col].median())

# Drop rows with missing values in 'term'
LoanDefault.dropna(subset=['term'], inplace=True)

# Verify no missing values remain
print(LoanDefault.isnull().sum())
```

Data Transformation

Figure 2 show label encoding was applied to convert categorical variables into numerical format, making them suitable for ML models. Label encoding assigns a unique integer to each category, ensuring that algorithms can effectively process and interpret the categorical data.

Figure 2

Label Encoding



```
# Encode categorical variables using Label Encoding

from sklearn.preprocessing import LabelEncoder

categorical_cols = LoanDefault.select_dtypes(include=["object"]).columns
label_encoders = {}

for col in categorical_cols:
    le = LabelEncoder()
    LoanDefault[col] = le.fit_transform(LoanDefault[col])
    label_encoders[col] = le  # Store encoders for future use
```

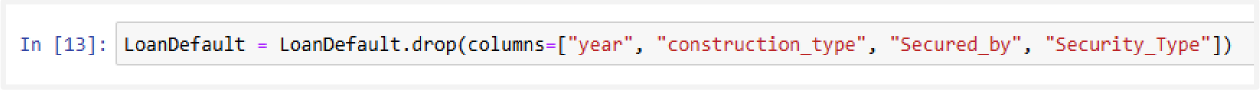
```
#to check the result after one-hot encoding
LoanDefault.head(10)
```

	ID	year	loan_limit	Gender	approv_in_adv	loan_type	loan_purpose	Credit_Worthiness	ope
0	24890	2019	0	3	0	0	0	0	
1	24891	2019	0	2	0	1	0	0	
2	24892	2019	0	2	1	0	0	0	
3	24893	2019	0	2	0	0	3	0	
4	24894	2019	0	1	1	0	0	0	
5	24895	2019	0	1	1	0	0	0	

Figure 3 showed remove the unnecessary or irrelevant variables to simplify the dataset and enhance model performance. The year variable was dropped because it is unlikely to directly affect loan default and may introduce noise. Additionally, columns such as construction type, secured by and security type were removed because they contained the same value for all records, providing no useful information to the model.

Figure 3

Drop Unnecessary Column

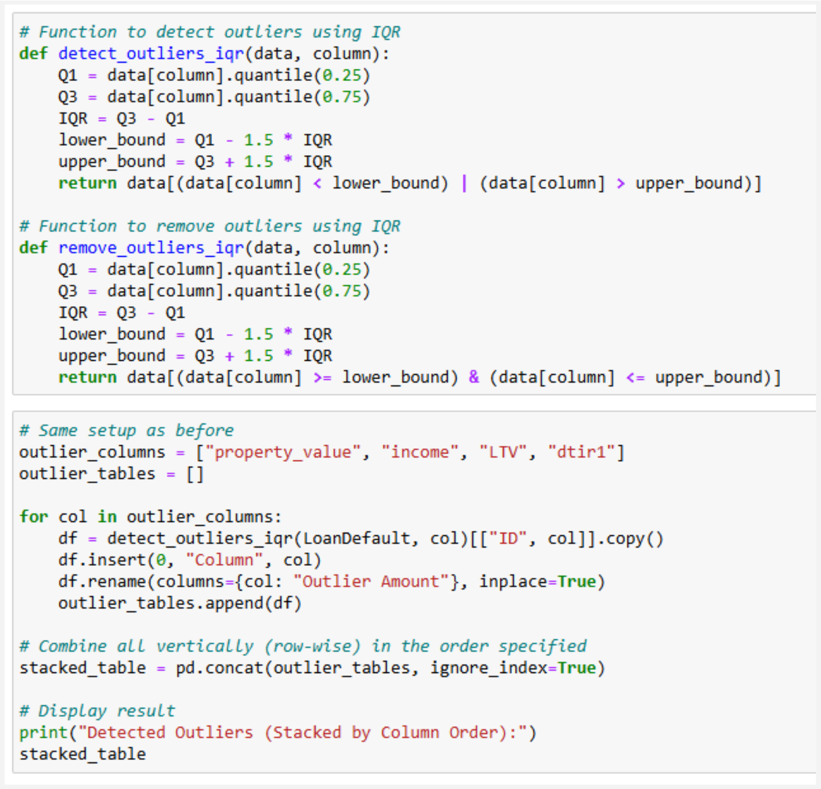


```
In [13]: LoanDefault = LoanDefault.drop(columns=["year", "construction_type", "Secured_by", "Security_Type"])
```

Figure 4 illustrates the process of detecting outliers using the Interquartile Range (IQR) method. This technique was applied to four key numerical columns, including property_value, income, LTV, and dtir1. For each column, a table was generated to list the identified outlier values for further review.

Figure 4

Detect Outlier



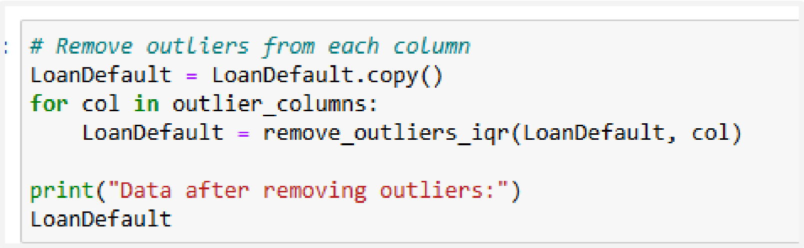
```
# Function to detect outliers using IQR
def detect_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return data[(data[column] < lower_bound) | (data[column] > upper_bound)]

# Function to remove outliers using IQR
def remove_outliers_iqr(data, column):
    Q1 = data[column].quantile(0.25)
    Q3 = data[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return data[(data[column] >= lower_bound) & (data[column] <= upper_bound)]
```

```
# Same setup as before
outlier_columns = ["property_value", "income", "LTV", "dtir1"]
outlier_tables = []

for col in outlier_columns:
    df = detect_outliers_iqr(LoanDefault, col)[["ID", col]].copy()
    df.insert(0, "Column", col)
    df.rename(columns={col: "Outlier Amount"}, inplace=True)
    outlier_tables.append(df)

# Combine all vertically (row-wise) in the order specified
stacked_table = pd.concat(outlier_tables, ignore_index=True)

# Display result
print("Detected Outliers (Stacked by Column Order):")
stacked_table
```

Figure 5 shows the removal of these outliers from the dataset. Outliers are unusual values that can negatively impact model performance by introducing noise or skewed results. Removing them helps improve the model's accuracy and reliability.

Figure 5

Remove Outlier



```
# Remove outliers from each column
LoanDefault = LoanDefault.copy()
for col in outlier_columns:
    LoanDefault = remove_outliers_iqr(LoanDefault, col)

print("Data after removing outliers:")
LoanDefault
```

Model

In this study, the LR, DT, RD, XGBoost, and KNN algorithms used to analyze borrower data from Kaggle and predict the likelihood of default (Table 1). These algorithms are widely used for classification problems in financial risk analysis.

Table 1

Libraries Used for Models

Model	Library	Class
Random Forest	scikit-learn	RandomForestClassifier
XGBoost	XGBoost	XGBClassifier
Decision Trees	scikit-learn	DecisionTreeClassifier
KNN	scikit-learn	KNeighborsClassifier
Logistic Regression	scikit-learn	LogisticRegression

First, the RF model was selected to predict loan default due to its ability to handle large datasets and complex data structures effectively. It can minimize the impact of noise and outliers, which is beneficial given the diverse nature of the dataset. The model was implemented using the RandomForestClassifier from the scikit-learn library with 100 decision trees (n_estimators=100) to enhance prediction accuracy. Using 100 decision trees because it provides a balance between reliable predictions and manageable processing time, especially for large datasets with diverse patterns. Second, the XGBoost was chosen for its efficiency in handling large datasets and its strength in managing imbalanced data through gradient boosting. The model was built using the XGBClassifier from the XGBoost library with 100 estimators (n_estimators=100), which is similar to the Random Forest.

Third, the DT model was applied for its simplicity and ease of interpretation. It effectively handles both numerical and categorical data, making it flexible for diverse datasets. The model was implemented using the DecisionTreeClassifier from the scikit-learn library, with a set random state (random_state=42) for consistent results. The value 42 was chosen because it is widely used as a standard seed to ensure reproducibility when splitting the dataset results. In addition, the hyperparameter tuning was conducted using GridSearchCV from the scikit-learn library to optimize the model. The param_grid explored different values for max_depth, min_samples_split, and min_samples_leaf. Setting max_depth limits the tree's depth, preventing overfitting; min_samples_split controls the minimum samples required to split a node, helping the model generalize better; and min_samples_leaf ensures a minimum number of samples at a leaf node, reducing overfitting. GridSearchCV used 5-fold cross-validation (cv=5), scoring='accuracy' to focus on prediction precision, and n_jobs=-1 to speed up processing. This approach aimed to find the best combination of hyperparameters, enhancing model performance and reducing overfitting.

Fourth, the KNN was chosen for its simplicity and effectiveness in handling classification tasks such as predicting loan defaults. The model was implemented using the KNeighborsClassifier from the scikit-learn library. Before training the model, the StandardScaler was used to normalize the dataset (X_balanced_scaled and X_test_scaled). This scaling process was necessary because KNN relies on distance calculations, and having features on a similar scale ensures fair comparisons. The number of neighbors (n_neighbors=5) was set to balance bias and variance. Using 5 neighbors helps the model generalize better, avoiding overly complex decision boundaries while maintaining accuracy. Fifth, the LR was chosen because of its simplicity and efficiency in binary classification tasks such as predicting loan defaults. It is effective in situations where the relationship between variables is mostly linear. The model was implemented using the LogisticRegression from the scikit-learn library. Before fitting the model, the StandardScaler was applied to normalize the dataset (X_train_scaled and X_test_scaled), stabilizing the model's performance and speeding up convergence. Hyperparameter tuning was conducted using GridSearchCV with a parameter grid (param_grid = {'C': [0.01, 0.1, 1, 10, 100]}) to optimize the regularization strength.

Results and Discussion

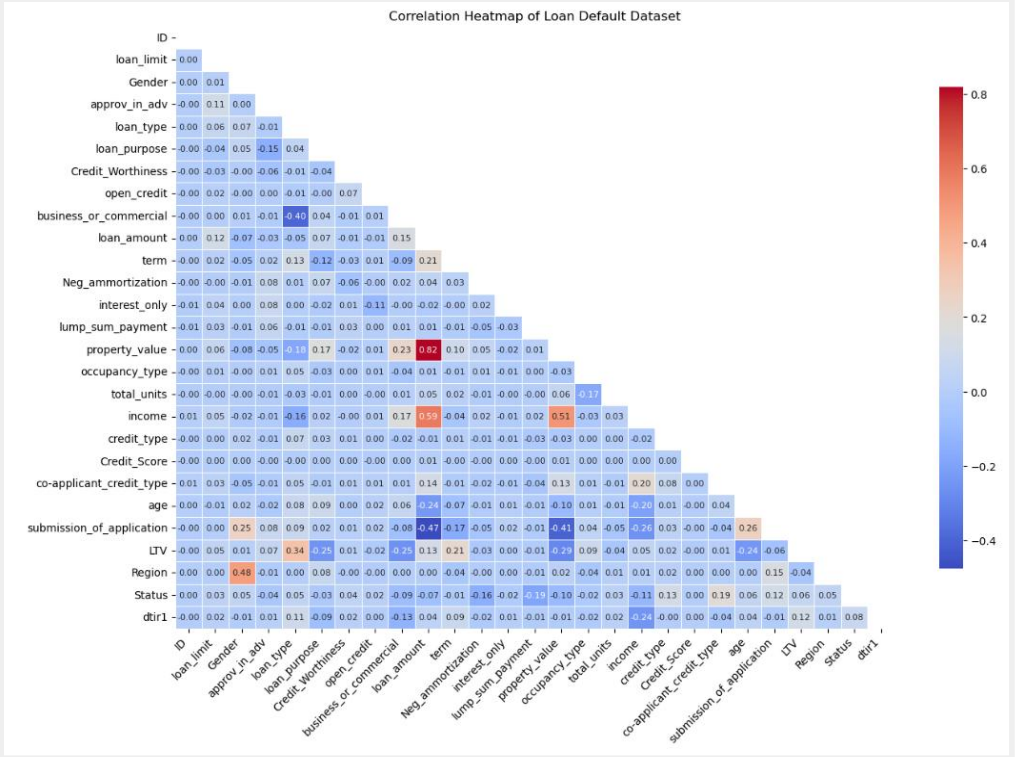
Descriptive Analysis

The heatmap visualizes the correlation coefficients between variables, providing an intuitive way to examine relationships and dependencies within the dataset. Colors represent both the strength and direction of correlations, making it easier to detect patterns (Jain, 2024). In Figure 6, the correlation values range from –0.4 to 0.8, indicating weak to strong linear relationships. A strong positive correlation is observed between *property value* and *loan amount* (0.82), indicating that applicants requesting larger loans are generally purchasing higher-value properties. Likewise, *income* is positively correlated with both *loan amount* (0.59) and *property value* (0.51), suggesting that higher-income individuals are more likely to purchase expensive properties and apply for larger loans. A moderate positive relationship is also identified between *gender* and *region* (0.48), which may reflect demographic trends across different geographical areas.

On the negative side, the *submission of application* variable shows a moderate negative correlation with both *loan amount* (–0.47) and *property value* (–0.41). This indicates that applications not submitted through institutional channels are generally associated with smaller loan sizes and lower-value properties. Similarly, *business or commercial* shows a negative correlation with *loan type* (–0.40). These correlation patterns provide meaningful insights into applicant behavior and loan characteristics. However, it is important to note that no variable exhibits a strong correlation with the dependent variable, Status (loan default outcome). This finding implies that default behavior is unlikely to be explained by a single predictor but rather by a combination of multiple factors or complex non-linear interactions. Consequently, advanced predictive modeling techniques are essential to uncover the underlying drivers of loan default more accurately.

Figure 6

Correlation of Key Variables in the Dataset



Source: Anaconda Jupyter Notebook

Data Splitting

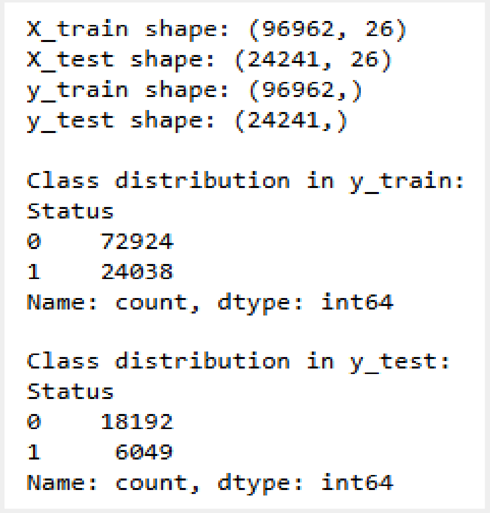
The dataset was partitioned into training and testing subsets using an 80:20 split ratio implemented via the *train_test_split* function in *scikit-learn* (Figure 7). In this approach, 80% of the data (96,962 records) was used to train the models, enabling them to learn underlying patterns, while the remaining 20% (24,241 records) served as the test set to evaluate performance on unseen data. To ensure reproducibility, the split was performed with a fixed random state (=42), guaranteeing that repeated runs would yield consistent partitions.

An examination of the target variable (*Status*) revealed a class imbalance: the majority of cases were non-defaults (0), while defaults (1) constituted a smaller proportion. Specifically, the training set contained 72,924 non-defaults and 24,038 defaults, whereas the test set comprised 18,192 non-defaults and 6,049 defaults. This imbalance reflects the real-world lending environment, where defaults occur less frequently than successful repayments.

Rather than artificially altering the dataset distribution through resampling, the study focused on selecting models that are naturally more robust to imbalance. Ensemble methods such as Random Forest and XGBoost are particularly well-suited in this context, as they can better handle skewed datasets by capturing complex patterns and weighting misclassifications more effectively. This approach ensures that predictive modeling remains aligned with practical applications, where imbalanced credit datasets are the norm.

Figure 7

Training and Testing Datasets After Splitting



```
X_train shape: (96962, 26)
X_test shape: (24241, 26)
y_train shape: (96962,)
y_test shape: (24241,)

Class distribution in y_train:
Status
0    72924
1    24038
Name: count, dtype: int64

Class distribution in y_test:
Status
0    18192
1     6049
Name: count, dtype: int64
```

Test Results

Table 2 presents the evaluation outcomes for the five machine learning models using accuracy, precision, recall, and F1-score. Among them, XGBoost demonstrates the strongest overall performance, achieving the highest accuracy of 89.70% with a strong balance across precision, recall, and F1-score. This indicates its effectiveness in handling the imbalanced dataset and capturing complex feature interactions. Random Forest closely follows with an accuracy of 89.07% and a Class 1 F1-score of 0.73; however, its recall of 0.60 is slightly lower than that of the Decision Tree model. Decision Tree (DT) achieves a competitive accuracy of 88.47%, though its recall of 0.64 highlights some difficulty in consistently identifying all default cases, reflecting its vulnerability to overfitting despite being highly interpretable. By contrast, K-Nearest Neighbors (KNN) performs less effectively, with an accuracy of 84.78% and a Class 1 F1-score of 0.64, suggesting difficulty in distinguishing between defaults and non-defaults due to its sensitivity to class imbalance and distance metrics. Logistic Regression (LR) shows the weakest performance, achieving only 77.32% accuracy, with a recall of 0.22 and a Class 1 F1-score of 0.32, indicating significant misclassification of default cases. Overall, the results highlight the superiority of ensemble methods such as XGBoost and Random Forest, which not only deliver the highest accuracies but also provide the best trade-off between precision and recall. For financial institutions, this translates into models that more effectively capture true defaults, thereby reducing potential credit losses, while minimizing false positives and avoiding unnecessary loan rejections. In contrast, the weaker performance of simpler models such as KNN and LR underscores the challenges of applying traditional approaches to complex and imbalanced credit datasets.

Table 2

Performance Evaluation of Machine Learning Models for Loan Default Prediction

Model	Class	Precision	Recall	F1-score	Accuracy
Random Forest	0	0.88	0.99	0.93	89.07%
	1	0.94	0.60	0.73	
XGBoost	0	0.89	0.98	0.93	89.70%
	1	0.92	0.64	0.76	
Decision Trees	0	0.88	0.98	0.93	88.47%
	1	0.92	0.59	0.72	
KNN	0	0.86	0.95	0.90	84.78%
	1	0.77	0.55	0.64	
Logistic Regression	0	0.79	0.96	0.86	77.32%
	1	0.63	0.22	0.32	

Receiver Operating Characteristic (ROC) Curve Analysis

The Receiver Operating Characteristic (ROC) curve was used to evaluate the trade-off between sensitivity (true positive rate) and specificity (true negative rate), with the Area Under the Curve (AUC) serving as a summary measure of each model's discriminatory power. The Random Forest (RF) model achieved an AUC of 0.89, indicating strong ability to distinguish between defaulters and non-defaulters while maintaining a good balance between identifying true positives and minimizing false positives. Similarly, the XGBoost model also recorded an AUC of 0.89, reflecting excellent discriminatory capability, largely due to its advanced boosting technique that optimizes learning from misclassified cases. The Decision Tree (DT) model performed slightly lower with an AUC of 0.86, suggesting decent performance but with potential risks of overfitting that may reduce generalization. The K-Nearest Neighbors (KNN) model achieved a moderate AUC of 0.81, showing reasonable but less effective differentiation, likely due to its sensitivity to class imbalance and distance metrics. Logistic Regression (LR) recorded the lowest AUC of 0.74, indicating limited ability to separate the two classes, though it remains valuable as a simple, fast, and interpretable baseline. Taken together, the ROC analysis reinforces that ensemble models, particularly RF and XGBoost, provide the most reliable performance for loan default prediction, whereas simpler models demonstrate weaker classification ability under imbalanced conditions.

Importance Feature in Loan Default Prediction

Figure 8 shows that the most crucial feature for Random Forest (RF) in predicting loan defaults is the loan-to-value (LTV) ratio. Gonzalez et al. (2016) found that higher initial LTV ratios are associated with greater default risk. Other significant features include property value, credit type, and the debt-to-income (DTI) ratio, which emphasize the borrower's financial standing and ability to manage debt. In addition, features such as income and credit score also play important roles, reinforcing the borrower's earning capacity and creditworthiness. Zhou et al. (2018) reported that the total assets of the borrower have no significant impact on loss given default (LGD). However, a better credit score and lower DTI ratio help reduce loan default risk, while longer loan tenures and larger loan amounts increase it. Therefore, when evaluating a borrower's repayment ability, greater attention should be given to the structure of assets rather than the overall size of total assets (Zhou et al., 2018). Overall, the RF model appears to prioritize financial stability and traditional risk factors when assessing loan default risk.

Figure 8

Importance Feature in Loan Default Prediction for RF

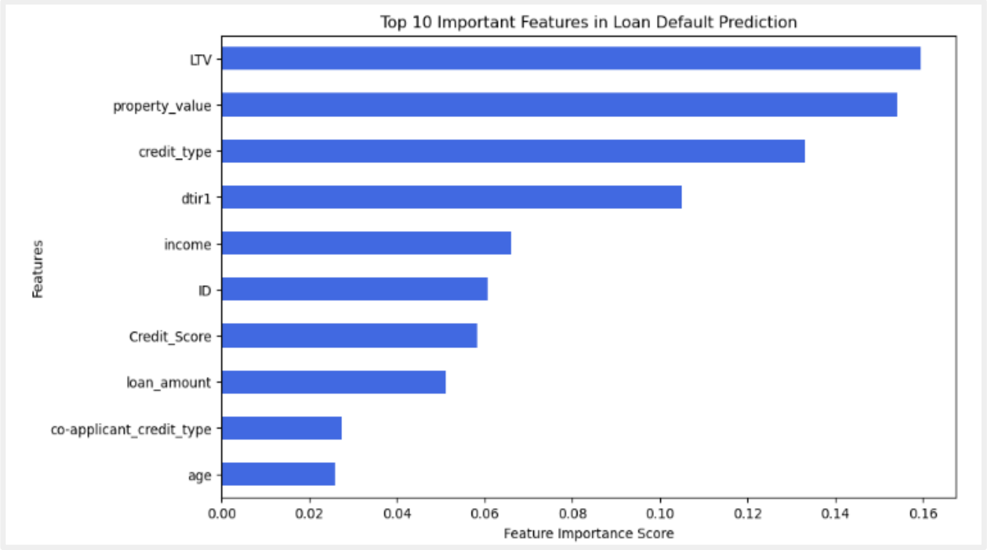
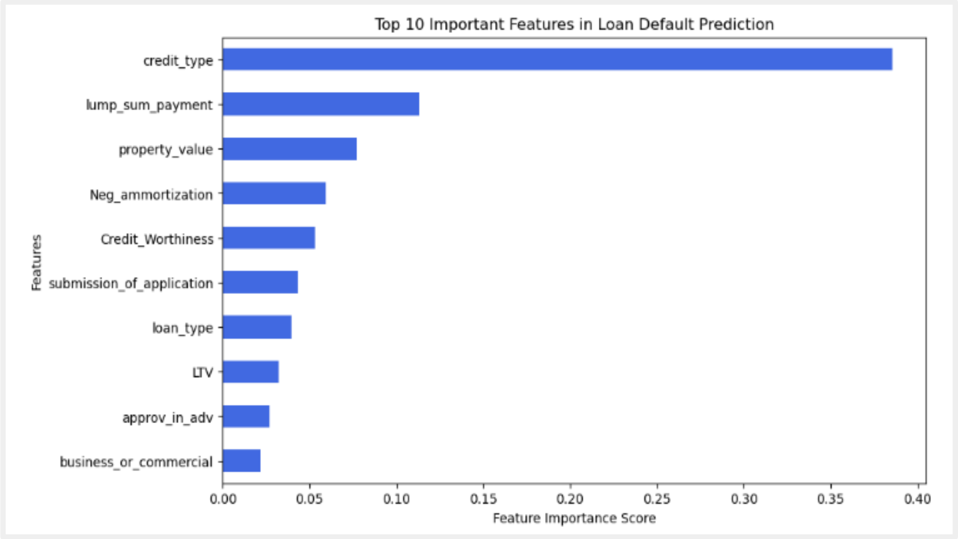


Figure 9 shows that the most crucial feature for XGBoost in predicting loan defaults is the credit type, indicating that the kind of credit a borrower holds plays a central role in assessing default risk. In addition, lump sum payment emerges as an important factor, suggesting that a borrower's ability to make substantial payments may reduce the likelihood of default. The inclusion of property value further highlights the relevance of the asset's worth in prediction. The study also shows that loan type and negative amortization are important features in the XGBoost model, demonstrating its ability to capture detailed loan and borrower behaviors. Negative amortization occurs when a borrower's monthly loan payment is insufficient to cover the interest on the loan. In contrast, the Random Forest model ranks LTV as one of the top features, whereas LTV is less important in XGBoost. This suggests that XGBoost emphasizes loan-specific details and borrower behavior rather than relying primarily on traditional financial ratios.

Figure 9

Importance Feature in Loan Default Prediction for XGBoost



Conclusion

Loan default significantly impacts banks' profitability, credit circulation, and market confidence. High default rates not only lead to an increase in non-performing loans but also prompt banks to adopt more stringent lending practices, thereby limiting credit access for both households and businesses. Therefore, enhancing credit risk evaluation is essential for supporting healthy lending activities and maintaining the stability of the financial system.

This study aimed to develop an effective predictive model for loan default using a publicly available dataset from Kaggle, which originally contained 34 features. A series of comprehensive data preprocessing steps were conducted to ensure data quality and model readiness. First, missing values were addressed to ensure the dataset was clean and suitable for analysis. Second, categorical variables were transformed using label encoding, converting them into numerical values that ML algorithms can process effectively. This method assigns a unique integer to each category, ensuring proper interpretation by the models. In addition, the "year" column was dropped, as it was unlikely to contribute meaningfully to loan default prediction and could introduce unnecessary noise. Outlier detection was then performed using the Interquartile Range (IQR) method on four key numerical columns—property value, income, LTV, and DTI ratio. For each of these columns, outliers were identified and removed to enhance model performance. Eliminating outliers helped reduce noise and potential distortions in model training, thereby improving accuracy and reliability.

The original dataset consisted of 148,670 records with 34 features. After preprocessing, the dataset was refined to 121,203 records with 27 features. To predict loan default, five ML algorithms were applied: LR, DT, RF, XGBoost, and KNN. The dataset was split into training and testing sets using an 80:20 ratio before model training began. Among the five ML models, XGBoost and RF delivered the best performance, both achieving high accuracy (89%), strong F1-scores, and AUC scores of 0.89. The DT model also showed good performance (88.47% accuracy, 0.86 AUC score), while KNN (84.78% accuracy) struggled with false positives, lowering its precision. LR performed the weakest, particularly in recall and AUC score (0.74). Overall, XGBoost and RF emerged as the most effective models for credit risk assessment. In summary, this study demonstrates the value of ML models—particularly ensemble methods such as XGBoost and RF—in improving the accuracy and reliability of loan default prediction. These models enable more effective identification of high-risk borrowers, leading to better lending decisions, reduced bad debt, and stronger credit risk management. As the financial industry continues to move toward data-driven decision-making, the application of these models in loan assessments can enhance the stability and resilience of lending systems.

From a theoretical perspective, previous research on loan default prediction has primarily relied on traditional methods such as rule-based credit scoring, expert judgment, and basic statistical models. These approaches often fail to capture the complex patterns in borrower behavior. This study addresses these limitations by applying advanced ML algorithms to improve prediction accuracy. By comparing model performance using metrics such as accuracy, recall, and F1-score, the study demonstrates the value of data-driven approaches in enhancing credit risk assessment and contributes to the advancement of financial risk management techniques.

From a practical perspective, this study recommends implementing the RF model into financial institutions' automated credit scoring systems. This integration would allow institutions to identify high-risk applicants early in the process, thereby increasing efficiency and reducing subjectivity in loan decisions. The model offers a scalable and dependable solution that supports loan officers in making faster, more informed decisions while minimizing manual workload and potential bias.

At the policy level, this study calls for government support through targeted research and development (R&D) subsidies and innovation funding. By establishing dedicated funds and offering grants or low-interest loans, policymakers can incentivize collaboration between financial institutions and technology firms to develop ML tools tailored for credit risk evaluation. Such initiatives foster innovation, strengthen the industry's risk control capabilities, and encourage the responsible and inclusive adoption of data-driven lending practices.

References

Alejandrino, J. C., Bolacoy Jr, J., & Murcia, J. V. B. (2023). Supervised and unsupervised data mining approaches in loan default prediction. *International Journal of Electrical & Computer Engineering* (2088-8708), *13*(2).

Ali, S. (2021). Factors Determining the Loan Default Risk of Individual Borrowers of Banks: An Empirical Study. *Journal of Cardiovascular Disease Research, 12*(6), 2156-2163. https://jcdronline.org/admin/Uploads/Files/6490982693adc9.46718331.pdf

An, X., Cordell, L., & Tang, S. (2020, May). *Extended Loan Terms and Auto Loan Default Risk* (Working Paper No. 20-18). Federal Reserve Bank of Philadelphia. https://www.philadelphiafed.org/-/media/FRBP/Assets/working-papers/2020/wp20-18.pdf

Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3483-3488.

Cedar Rose. (2024, October 30). *How AI and machine learning revolutionise credit risk assessments*. Cedar Rose. https://www.cedar-rose.com/blog/how-ai-and-machine-learnings-revolutionise-credit-risk-assessment

Chitambira, B. (2022, June 4). *Credit scoring using machine learning approaches* (Master's thesis, Mälardalen University). DiVA—Academic Archive Online. https://www.diva-portal.org/smash/get/diva2:1664698/FULLTEXT01.pdf

FasterCapital. (2025, April). *Budget constraint: Living within means – The role of budget constraints in consumer choices*. FasterCapital. https://www.fastercapital.com/content/Budget-Constraint--Living-Within-Means--The-Role-of-Budget-Constraints-in-Consumer-Choices.html#Understanding-the-Basics-of-Consumer-Choice-Theory

Forvis Mazars. (2024, January). *Navigating the loan default process for businesses: How to handle it*. Forvis Mazars. https://www.forvismazars.us/forsights/2024/01/navigating-the-loan-default-process-for-businesses-how-to-handle-it/

Fredriksson, O., & Frykström, N. (2019, March 14). *"Bad loans" and their effects on banks and financial stability* (Economic Commentary No. 2). Sveriges Riksbank. https://www.riksbank.se/globalassets/media/rapporter/ekonomiska-kommentarer/engelska/2019/bad-loans-and-their-effects-on-banks-and-financial-stability.pdf

Gonzalez, L. O., Santomil, P. D., Bua, M. V., & Sestayo, R. L. (2016). The impact of loan-to-value on the default rate of residential mortgage-backed securities. *Journal of Credit Risk, 12*(3), 1-13.

Hoang, D., & Wiegratz, K. (2023). Machine learning methods in finance: Recent applications and prospects. *European Financial Management*, *29*(5), 1657-1701.

Jain, A. (2024, January 21). *Unveiling the power of heatmaps in data visualization*. Medium. https://www.medium.com/@abhishekjainindore24/unveiling-the-power-of-heatmaps-in-data-visualization-8b370efca935

Kanaparthi, V. (2023, April 26). *Credit risk prediction using ensemble machine learning algorithms*. In 2022 International Conference on Inventive Computation Technologies, ICICT). IEEE. https://doi.org/10.1109/ICICT57646.2023.10134486

KPMG Lower Gulf. (2021, September). *Artificial intelligence in risk management*. KPMG. https://kpmg.com/ae/en/home/insights/2021/09/artificial-intelligence-in-risk-management.html

Krasovytskyi, D., & Stavytskyy, A. (2024). Predicting Mortgage Loan Defaults Using Machine Learning Techniques. *Ekonomika*, *103*(2), 140-160.

Kremer, A., Luget, A., Mikkelsen, D., Soller, H., Strandell-Jansson, M., & Zingg, S. (2024, July 1). Embracing generative AI in credit risk. McKinsey & Company. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/embracing-generative-ai-in-credit-risk

Lai, L. (2020, August). *Loan default prediction with machine learning techniques*. In 2020 International Conference on Computer Communication and Network Security (CCNS) (pp. 5-9). IEEE. https://ieeexplore.ieee.org/abstract/document/9240729

Lee, J. (2024, March 6). *AI-driven credit risk decisioning: What you need to know*. Experian Insights. https://www.experian.com/blogs/insights/ai-driven-credit-risk-decisioning/

Lin, J. (2024). *Research on loan default prediction based on logistic regression, RandomForest, XGBoost and Adaboost.* In SHS Web of Conferences (Vol. 181, p. 02008). EDP Sciences. https://www.shs-conferences.org/articles/shsconf/pdf/2024/01/shsconf_icdeba2023_02008.pdf

Mail, M. (2022, March 30). Bank Negara sees bank loan default hitting 4pc by end-2024 in event of employment shocks. Malay Mail. https://www.malaymail.com/news/malaysia/2022/03/30/bank-negara-sees-loan-default-hitting-4pc-by-end-2024-in-event-of-employmen/2050441

Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V. P., & Hameed, I. A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. *Risks*, *12*(10), 164. https://doi.org/10.3390/risks12100164

Nureni, A. A., & Adekola, O. E. (2022). Loan approval prediction based on machine learning approach. *Fudma Journal of Sciences*, *6*(3), 41-50.

Omogbhemhe, M. I., & Momodu, I. B. A. (2021, October). *Model for predicting bank loan default using XGBoost*. *International Journal of Computer Applications*, *183*(32).

Orji, U. E., Ugwuishiwu, C. H., Nguemaleu, J. C. N., & Ugwuanyi, P. N. (2022, April 17). *Machine learning models for predicting bank loan eligibility*. In *Proceedings of the 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)* (pp. 1–5). IEEE. https://doi.org/10.1109/NIGERCON54645.2022.9803172

Patel, B., Patil, H., Hembram, J., & Jaswal, S. (2020, June). *Loan default forecasting using data mining*. In *2020 international conference for emerging technology (INCET)* (pp. 1-4). IEEE. https://ieeexplore.ieee.org/abstract/document/9154100

Raheem, M. (2024). Loan Default Prediction using Machine Learning: A Review on the techniques. *Journal of Applied Technology and Innovation*, *8*(2), 1.

S&P Global Market Intelligence. (2023, December 4). *Artificial intelligence and alternative data in credit scoring and credit risk surveillance*. S&P Global. https://www.spglobal.com/en/research-insights/special-reports/artificial-intelligence-and-alternative-data-in-credit-scoring-and-credit-risk-surveillance

Satheeshkumar, S., Dakshana, M., Gunalan, K., Anandan, P., Saveetha, R., & Nithya, M. (2024, October). *Leveraging machine learning and forecasting techniques to enhance credit risk analysis and prediction*. In *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems* (ICSSAS) (pp. 781-786). IEEE. https://ieeexplore-ieee-org.tarc.idm.oclc.org/document/10760746

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020, July). *An approach for prediction of loan approval using machine learning algorithm.* In 2020 *International Conference on Electronics and Sustainable Communication Systems* (ICESC) (pp. 490-494). IEEE. https://ieeexplore-ieee-org.tarc.idm.oclc.org/document/9155614

Soni, A., & Shankar, K. P. (2022, May*). Bank Loan Default Prediction Using Ensemble Machine Learning Algorithm.* In 2022 *Second International Conference on Interdisciplinary Cyber Physical Systems* (ICPS) (pp. 170-175).

Uddin, M. (2019). Determinants of Loan Default of Low-Income Borrowers in Urban Informal Credit Markets: Evidence from Dhaka City. *European Journal of Business and Management*, *11*(26).

Withers, I. (2025, January 28). *Looser mortgage rules in Europe raise risks for lenders, warns Moody's*. Reuters. https://www.reuters.com/markets/europe/looser-mortgage-rules-europe-raise-risks-lenders-warns-moodys-2025-01-28/

Yasser. (n.d.). *Loan Default Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/yasserh/loan-default-dataset/data

Zhou, G., Zhang, Y., & Luo, S. (2018). P2P network lending, loss given default and credit risks. *Sustainability*, *10*(4), 1010.

Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503–513.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023). Explainable prediction of loan default based on machine learning models. *Data Science and Management*, *6*(3), 123-133.