Auto Loan Defaults: Predictive Modeling and Key Drivers

Yuning He
Faculty of Arts
The University of Hong Kong
Hong Kong, China
hyn1105206457@gmail.com

Abstract

Loan default prediction plays a vital role in credit risk management, especially in emerging markets where financial inclusion is expanding rapidly. This study compares the effectiveness of four modeling approaches—Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—using a real-world dataset from India. It addresses challenges such as data imbalance through SMOTE and enhances interpretability with SHAP analysis. Results show that while tree-based models yield high overall accuracy, logistic regression achieves the best recall in identifying defaulters. Key predictors include external credit score, education level, income, and gender. Samples with higher score, education, income and male samples are predicted as groups that default less. The findings provide practical implications for lenders aiming to optimize risk screening.

CCS Concepts

• **Computing methodologies** → Machine learning; Machine learning approaches; Classification and regression trees.

Keywords

Loan default prediction, Machine learning, Logistic regression, XGBoost, Credit scoring, SMOTE, SHAP, Financial Risk Modeling

ACM Reference Format:
Yuning He. 2025. Auto Loan Defaults: Predictive Modeling and Key Drivers. In *2025 International Symposium on Machine Learning and Social Computing (MLSC 2025), October 16–18, 2025, Hongkong, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3778450.3778458

1 Introduction

Predicting loan default is a critical task for financial institutions due to its direct impact on risk management, profitability, and overall financial stability. A loan default occurs when a borrower fails to meet repayment obligations, turning the loan into a non-performing asset. High default rates erode banks' earnings and capital reserves, can trigger liquidity crises, and constrain new lending. For example, in Malaysia, car loan defaults have been significant enough to result in about one-quarter of personal bankruptcies, prompting tighter underwriting and higher rejection rates for vehicle loans [1]. On a systemic level, widespread defaults can trigger financial distress. The 2008 subprime mortgage crisis illustrated how rising defaults among high-risk borrowers led to banking sector losses and a broader economic downturn [2]. Thus, robust default prediction models are needed to prevent unchecked lending to vulnerable borrowers.

The significance of default prediction is further evident in emerging markets where credit expansion has outpaced risk controls. In Indonesia's multifinance industry, non-performing financing (NPF) ratios spiked to 5.2% in 2020 amid the COVID-19 pandemic, reflecting borrowers' repayment challenges during economic shocks [3]. Although NPF levels have slightly improved post-pandemic, they remain above pre-2020 baselines. Certain segments like multipurpose and micro-loans exhibit particularly high default rates, indicating greater vulnerability in these portfolios. Such trends highlight that credit risk management remains a critical issue for lenders. Regulators and banks closely monitor default metrics as key indicators of financial health. Elevated default rates can necessitate higher loan-loss provisions and regulatory capital, and in severe cases may threaten the solvency of lending institutions or spark broader financial instability.

In sum, the ability to forecast loan default is foundational to both individual bank performance and macro-financial stability. This importance has driven this research and development of predictive models and risk scoring systems.

2 Literature Review

2.1 Machine Learning Applications in Loan Default Prediction

Recent studies confirm that machine learning (ML) models such as Random Forest, Gradient Boosting, XGBoost, and SVM outperform traditional logistic regression in loan default prediction, thanks to their ability to capture complex nonlinear relationships and handle high-dimensional data [2–5]. Ensemble models, particularly those combining Gradient Boosting, Random Forest, and XGBoost, yield even stronger performance in accuracy and robustness [3, 5]. For example, XGBoost demonstrated superior predictive performance, with higher AUC and precision than SVM or logistic regression [3]. ML models have been adopted in various markets, including Europe and Indonesia, and are widely recognized as the most effective for risk assessment [2, 3].

2.2 Handling Imbalanced Data in Default Prediction

A key challenge in default prediction is severe class imbalance—defaults are much rarer than non-defaults. This can cause models to miss actual defaults even when overall accuracy is high [3–5]. Techniques like SMOTE (Synthetic Minority Oversampling Technique) are commonly used to generate synthetic minority class

MLSC 2025, Hongkong, China

ACM ISBN 979-8-4007-2127-4/2025/10
https://doi.org/10.1145/3778450.3778458

samples and improve recall for default cases. Studies in Indonesia and large cross-country datasets confirm that using SMOTE or similar balancing methods significantly enhances model sensitivity to defaults, allowing for more effective risk management [3–6].

2.3 Model Interpretability and Explainability

While advanced ML models are powerful, their complexity raises concerns about interpretability. Feature importance analysis and post-hoc tools such as SHAP help reveal which variables drive model decisions, increasing transparency [2, 3, 5]. For instance, recent studies identify payment delay history as the top predictor, followed by financial ratios [3]. XGBoost's feature analysis and SHAP visualization make it possible for banks and regulators to understand and trust automated risk assessments [5]. The push for explainable AI (XAI) reflects both compliance demands and the practical need for interpretable, auditable models in financial services [2, 5].

2.4 Factors that Increase the Probability of Default

Prior research has extensively examined several borrower's socio-demographic and economic characteristics as significant predictors of default risk. High interest rates are consistently associated with increased likelihood of default [7, 10], indicating that greater borrowing costs can strain a borrower's repayment capacity and elevate credit risk. Among borrower demographics, marital status plays a substantial role: widowed borrowers have higher log-odds of default compared to single, married, or divorced individuals, while divorced borrowers also face a higher probability of default than married ones [7, 8]. Employment type is another relevant factor—self-employed individuals tend to have higher default rates, likely reflecting their more volatile income streams [7]. Education level, by contrast, exhibits an inverse relationship with default risk; borrowers with higher educational attainment are generally less likely to default, possibly due to enhanced financial literacy and more stable employment prospects [9]. Additionally, debt-to-income ratio, loan term, loan amount, and loan interest rate all positively correlate with the occurrence of loan default, emphasizing the importance of both borrower financial capacity and loan structure in credit risk assessment [10].

3 Data and Methodology

3.1 Sources of Data

The dataset used for this study is searched from the website Kaggle and can be accessed with the link https://www.kaggle.com/datasets/saurabhbagchi/dish-network-hackathon/data, The dataset is from the State Bank of India and there are 121856 observations (samples) in it. Such a large sample implies that the data analytics result may tell a lot about practical significance, but some methods may lose its predictive power.

3.2 Variables

The dataset comprises a rich set of variables capturing demographic characteristics, financial status, and loan details of each client. For clarity, we categorize the key variables into numerical and categorical types, which are detailed in Table 1.

3.3 Data Cleaning

We observed that the dataset contains a large number of missing values as well as special symbols (such as $, #, @, etc.). To address this, all special symbols were first replaced with missing values (NaN). Next, all numerical variables were converted to the float data type, leaving only categorical variables as objects.

For missing values in numerical variables, we used the median of each column to fill them in, ensuring that outliers would not overly influence the imputed values. For categorical variables, we used a random imputation method based on the observed category probabilities: for each missing entry, a category is randomly sampled according to the existing distribution of values in that column. This approach preserves the original distribution of categorical features and avoids introducing bias by always filling with the mode.

3.4 Exploratory Data Analysis (EDA)

After the data cleaning, an exploratory data analysis is used to have a basic understanding of the dataset and give insights on modelling process. We began by visualizing the distributions of key variables such as age and users' credit scores reported by other financial institutions.

Figure 1.a shows the distribution of one of a credit score, with most values concentrated between 0.3 and 0.7. The distribution is right-skewed, indicating more borrowers have medium to high credit scores. Understanding this helps us decide if we need to normalize or transform the variable before modeling. Figure 1.b displays the distribution of age in days, roughly following a normal distribution centered around 36–41 years old. This shows most borrowers are middle-aged.

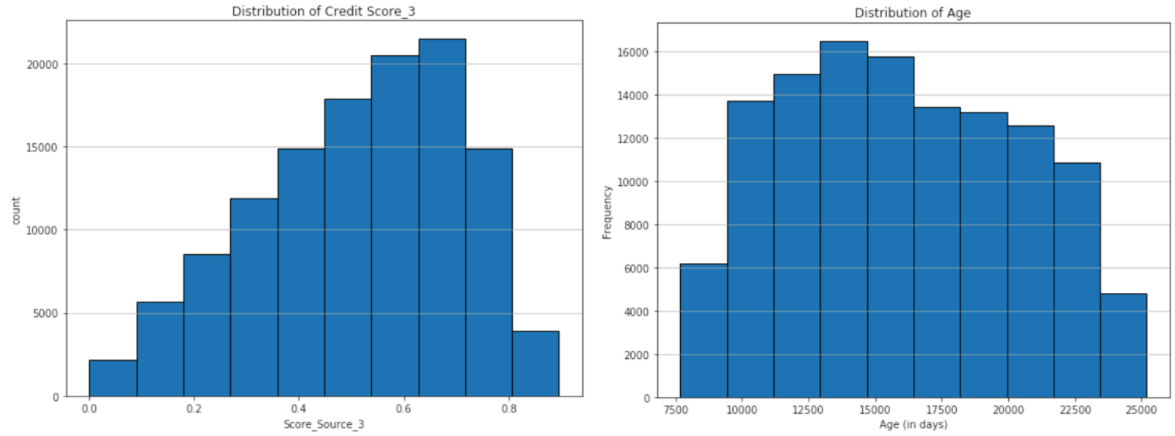
After selecting only, the data with default=1 (defaulters), we performed a bar plot and proportion analysis. For numerical variables, we divided them into 0.25, 0.5, 0.75, 0.9, and 0.99 quantiles, then analyzed the default proportion within each quantile. In Figure 2.a, it is found that approximately 10% of female borrowers defaulted (while only 7% of males did). Moreover, for borrowers with higher credit scores at other financial institutions, the likelihood of default is much lower, which is not surprising. The relatively high default rate among females is noteworthy. Studies based on a nationwide U.S. household survey indicate that women may experience more debt-related stress than men at similar debt levels [11]. Additionally, Chinese women are reported to have a higher abnormal default rate than men, which may be related to differences in financial literacy [12]. In Figure 2.b, an obvious decrease in the default rate can be seen as the credit score increases.

3.5 Data Preprocessing

In this study, we adopted differentiated preprocessing strategies. Missing values were filled using median imputation and random imputation other than in modelling. For OLS regression, categorical variables were first converted into dummy variables (one-hot encoding) to meet the linear model's assumption of numerical input and to allow for straightforward coefficient interpretation.

Table 1: List of Variables

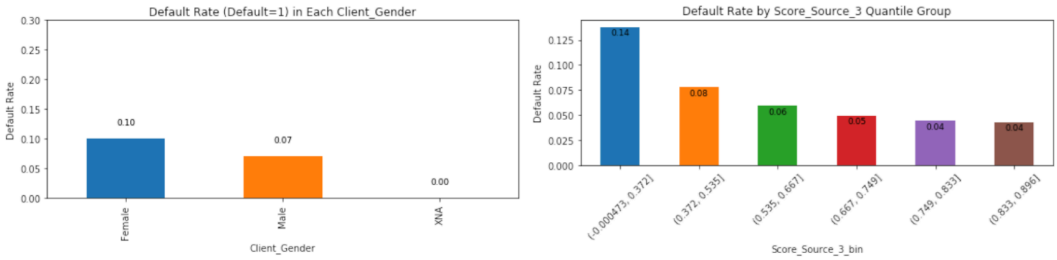
Class	Variables	Definition
Numerical Variables	Age_Days	Client's age in days at time of loan application.
	Employed_Days	Days since client started working.
	Registration_Days	Days since client changed registration details.
	ID_Days	Days since client updated identity documents.
	Own_House_Age	Age of client's house in years.
	Credit_Amount	Total amount of credit requested.
	Loan_Annuity	Monthly loan annuity amount.
	Child_Count	Number of children client has.
	Score_Source	External credit score of the client.
	Social_Circle_Default	Number of social connections who defaulted recently.
	Phone_Change	Days since client changed phone number.
	Credit_Bureau	Number of loan inquiries in the past year.
Categorical Variables	Client_Gender	Gender of the client (Male/Female).
	Car_Owned/Bike_Owned/House_Own	Whether client owns a car/bike/house (0 = No, 1 = Yes).
	Active_Loan	Whether client has another active loan (0 = No, 1 = Yes).
	Accompany_Client	Who accompanied the client when applying for the loan.
	Client_Income_Type	Type of client's income source (e.g., Salaried, Pensioner, etc.).
	Client_Education	Highest education level attained by client.
	Mobile_Tag/Homephone_Tag	Whether mobile/home phone was provided.
	Workphone_Working	Whether work phone was reachable (0 = No, 1 = Yes).
	Client_Occupation	Client's occupation type.
	Client_City_Rating	Rating of the city the client resides in (1–3).
	Application_Process_Day	Day of the week the loan application was submitted (0–6).
	Client_Permanent_Match_Tag	Whether contact and permanent address mismatch
	Client_Contact_Work_Tag	Whether work and contact address mismatch.
	Type_Organization	Type of organization where client is employed.



a. Distribution of credit scores b. Distribution of Ages

Figure 1: Distribution of key factors

In contrast, for tree-based models such as XGBoost, Gradient Boosting, and Random Forest, we used the Iterative Imputer to estimate missing values. This technique is especially effective for predictive modeling where accuracy is prioritized over interpretability, particularly in high-dimensional datasets [16]. Categorical variables were encoded using label encoding because tree models use value-based splits rather than assuming linear effects., which can naturally handle ordinal and nominal variables without requiring dummy expansion. Moreover, dataset is split into training set and



a. Default rate in each gender b. Default rate by score source quantiles

Figure 2: Default rates by segment across classification metrics

testing set to facilitate the training and evaluation. As the severe class imbalance found between default and non-default, SMOTE oversampling method is used to increase the amount of minority class samples (default) to boost the recall of minority, which is the key in this research. The ultimate proportion reaches 1:1 after the processing. By tailoring preprocessing methods to the characteristics of each modeling technique, we aimed to preserve the interpretability of OLS regression while maximizing the predictive performance of non-linear machine learning models.

4 Machine Learning

4.1 OLS Regression

To explore more features that may influence the default rate, the study utilizes OLS regression as preliminary method aiming to discover the predictive factors of default rate, where we have the following OLS regression function:

$$\text{Default}_i = \beta_0 + \beta_1 X_{\{1i\}} + \beta_2 X_{\{2i\}} + \ldots + \beta_k X_{\{ki\}} + \epsilon_i \quad (1)$$

Here, Default_i is the binary outcome variable indicating whether individual i defaults, $X_{\{1i\}}, X_{\{2i\}}, \ldots, X_{\{ki\}}$ represent the variables other than ID and default in the dataset, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients to be estimated, and ϵ_i is the error term. The regression result is as following:

Among all variables in Table 2, the most significant predictors of default are the external credit score features, which have strong negative coefficients (especially 0.217 for score source 3). This indicates higher credit scores lead to lower default probability, holding all other variables constant. Additionally, basic demographic and economic factors such as car ownership, age, gender, and education level play an important role. Car owners, older individuals, males, and those with higher education generally being less likely to default. Workphone_working has the significantly negative coefficient, meaning that if the work phones of the loan borrowers are on, the default rate tends to decline. However, many categorical variables, particularly those related to organization type and clients' occupations, show limited or insignificant impact in this linear model.

The overall fit of the OLS regression model can be assessed using several key parameters. With an R-squared value of 0.047, the model explains only a small portion of the variance in the default outcome. The model is statistically significant as indicated by the F-statistic of 45.37 and a Prob (F-statistic) of 0.00, confirming that at least some predictors are related to the outcome variable. The analysis is based on a large dataset with 121,856 observations, which increases the reliability and precision of the results. However, low R-squared indicates that the relationship between features and default is non-linear. Overall, while the model is statistically robust thanks to its large sample size, its practical predictive ability is limited. There's a need for further model techniques.

4.2 Modeling

We adopt four mainstream machine learning algorithms in this study to compare their predictive performance and choose the best one, including Logistic Regression, Random Forest, Gradient Boosting and XGBoost.

4.2.1 Logistic Regression. The idea of logistic regression is to map input variables into a probability space between 0 and 1 through a linearly weighted combination of features, and use this mapping for classification prediction. The formula of the logistic regression model is shown in Equation 2):

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \quad (2)$$

Where $x_1, x_2, \ldots, x_n$ are input features, $\beta_1, \beta_2, \ldots, \beta_n$ are model parameters.

4.2.2 Random Forest. Specifically, random forest performs bootstrap sampling with replacement on the training set to generate multiple distinct subsets of data, each of which is used to train a decision tree. At each node split, a random subset of features is selected from all available features to determine the optimal split, thereby reducing the risk of overfitting. The final prediction of the random forest is determined by aggregating the predictions of all decision trees, as shown in Equation 3):

$$H(x) = \frac{1}{T} \sum_{i=1}^{T} h_i(x) \quad (3)$$

where T is the number of decision trees, and $h_i(x)$ is the prediction from the i-th tree.

4.2.3 Gradient Boosting. Gradient Boosting builds multiple decision trees sequentially, where each new model fits the residuals

Table 2: OLS Regression Result

Variable	Coef	t-Statistic	P-Value	Significance
const	0.117	0.304	0.761	
Client_Income	0.000	-2.824	0.005	***
Car_Owned	-0.029	-13.025	0.000	***
Bike_Owned	-0.001	-0.420	0.674	
Active_Loan	0.001	0.541	0.588	
House_Own	0.007	4.250	0.000	***
Child_Count	0.000	-0.054	0.957	
Credit_Amount	0.000	-4.549	0.000	***
Loan_Annuity	0.000	6.299	0.000	***
Population_Region_Relative	0.000	-0.242	0.809	
Age_Days	0.000	-4.682	0.000	***
Employed_Days	0.000	-0.852	0.394	
Registration_Days	0.000	-2.623	0.009	***
ID_Days	0.000	-7.821	0.000	***
Own_House_Age	0.001	5.188	0.000	***
Mobile_Tag	0.094	0.351	0.725	
Homephone_Tag	0.006	2.913	0.004	***
Workphone_Working	-0.007	-3.885	0.000	***
Client_Family_Members	0.002	0.640	0.522	
Cleint_City_Rating	0.020	12.266	0.000	***
Application_Process_Day	0.001	1.882	0.060	*
Application_Process_Hour	-0.001	-3.198	0.001	***
Score_Source_1	-0.082	-13.466	0.000	***
Score_Source_2	-0.008	-7.926	0.000	***
Score_Source_3	-0.217	-47.256	0.000	***
Social_Circle_Default	-0.047	-5.587	0.000	***
Phone_Change	0.000	-11.074	0.000	***
Credit_Bureau	0.001	1.130	0.258	
Client_Education_Junior secondary	0.033	4.604	0.000	***
Client_Education_Post Grad	-0.024	-0.756	0.450	
Client_Education_Secondary	0.026	13.067	0.000	***
Client_Gender_Male	-0.022	-11.537	0.000	***
Client_Gender_XNA	-0.142	-0.921	0.357	
Loan_Contract_Type_RL	-0.024	-8.858	0.000	***
Client_Housing_Type_Home	-0.004	-1.117	0.264	
Client_Housing_Type_Municipal	0.013	2.319	0.020	**
Client_Housing_Type_Office	-0.012	-1.324	0.185	
Client_Housing_Type_Rental	0.006	0.791	0.429	
Client_Housing_Type_Shared	0.016	1.209	0.227	
Client_Occupation_Core	-0.002	-0.357	0.721	
Client_Occupation_Drivers	0.009	1.905	0.057	*
Client_Occupation_Laborers	0.005	1.254	0.210	
Variable	**Coef**	**t-Statistic**	**P-Value**	**Significance**
Client_Occupation_Managers	-0.004	-0.917	0.359	
Client_Occupation_Sales	0.003	0.701	0.483	
Type_Organization_Business Entity Type 3	0.012	0.541	0.588	
Type_Organization_Medicine	0.003	0.160	0.873	
Type_Organization_Other	0.007	0.308	0.758	
Type_Organization_Self-employed	0.018	0.840	0.401	
Type_Organization_XNA	0.009	0.401	0.688	
No. Observations	121856.000			
F-statistic	45.37***			
R-squared	0.047			

Notes. *** p < .01, ** p < .05, * p < .10

Table 3: Comparison of Model Performance on the Testing Set

Models	Weighted			Class 1(default)			Accuracy
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Logistic	0.56	0.69	0.54	0.17	0.67	0.27	0.70
RandomForest	0.60	0.65	0.62	0.68	0.22	0.34	0.86
GBoosting	0.62	0.60	0.61	0.31	0.25	0.27	0.90
XGBoost	0.79	0.57	0.61	0.65	0.15	0.25	0.93

of the previous model to gradually reduce overall prediction error. Assuming F_m is a weak model, gradient boosting adds a new estimator that learns to predict the residuals of F_m , as shown in Equation 4):

$$F_{m+1}(x) = F_m(x) + h(x) \tag{4}$$

4.2.4 XGBoost. XGBoost is an enhanced version of Gradient Boosting. Its core idea is to iteratively train a series of decision trees, where each new tree fits the residuals of the previous model to continuously improve overall predictive performance. Additionally, XGBoost applies second-order Taylor expansion to optimize the objective function during training, which accelerates convergence and improves both accuracy and robustness.

By comparing multiple models, we can more comprehensively evaluate the performance of different algorithms on the dataset, avoid biases introduced by a single model, and improve the reliability of results. During the hyperparameter tuning process, we mostly use HalvingGridSearchCV to efficiently identify optimal parameters.

4.3 Results

4.3.1 Model Prediction Comparison. Table 3 presents the performance of each model on the test set. We selected weighted metrics, Class 1-specific scores, and overall accuracy for horizontal comparison.

According to the results in Table 3, when focusing solely on overall accuracy and weighted F1-score, XGBoost appears to be the top performer (Accuracy: 0.93, Weighted F1: 0.61), followed by Gradient Boosting. However, when we shift our attention to the minority class—i.e., loan default cases—logistic regression stands out with a recall rate of 0.67, significantly higher than other models. This indicates that logistic regression is much more effective at capturing default samples. In contrast, tree-based models (Random Forest, Gradient Boosting, XGBoost) suppress recall rates for defaults to the 0.15–0.25 range. In other words, although these models excel in overall performance, they sacrifice the ability to identify actual defaulters. Logistic regression, on the other hand, achieves the highest recall (0.67) for defaults, but with a precision of only 0.17, indicating a high false-positive rate. Still, it successfully identifies 67% of true defaulters.

From the perspective of credit risk screening in financial institutions, such as banks, which often prioritize identifying high-risk borrowers, logistic regression may be more appropriate. However, its high false-positive rate implies a need for manual post-review to filter out incorrect alerts. Although tree-based models perform well on the majority class and thus boost overall accuracy, the extremely low proportion of default cases causes the weighted metrics to be diluted by the majority class. Their ability to detect defaulters is, therefore, quite limited. On the other hand, if institutional resources are constrained and only a small set of warning cases can be reviewed, a model with higher precision, such as XGBoost with threshold tuning, may be preferred.

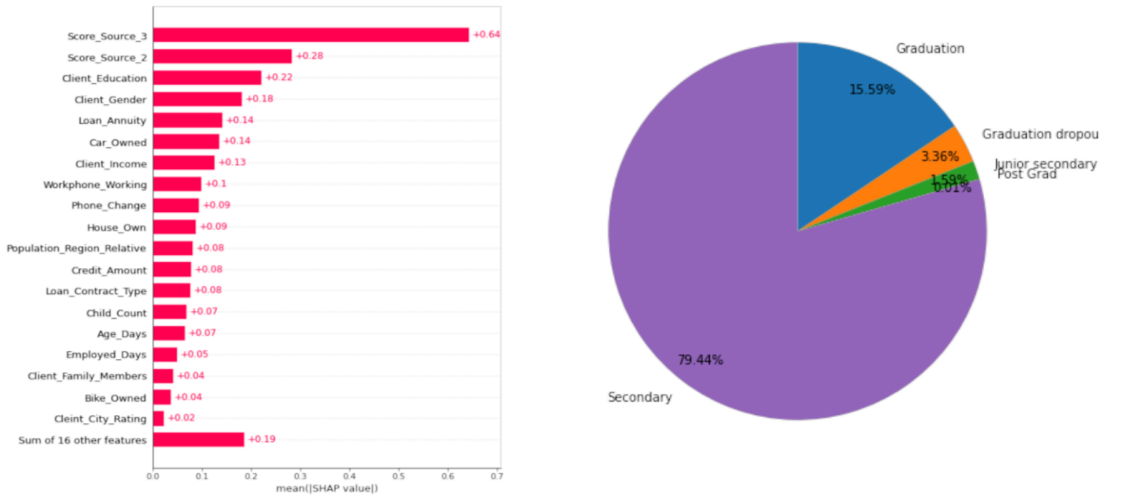
4.3.2 Model Interpretability Analysis. In this study, the SHAP method was employed to interpret the predictions of different machine learning models. As shown in Figure 3.a (SHAP bar plot), which ranks features by their mean absolute SHAP values, the most influential predictors for default classification were credit rating and education level. Gender, loan annuity, and car ownership status were also among the top-ranked features. Other variables such as population density, number of children, household size, and urban classification level had relatively lower importance.

The high ranking of education level among defaulters led to further statistical analysis. As education is a categorical variable, we examined its distribution using bar plots (Figure 3.b). The results show that borrowers with secondary education make up over 80% of the default group, while those with university degrees account for approximately 16%.

Indian high school graduates are often concentrated in informal employment (such as street vending and daily wage labor), leading to highly volatile incomes. Over 50% of workforce

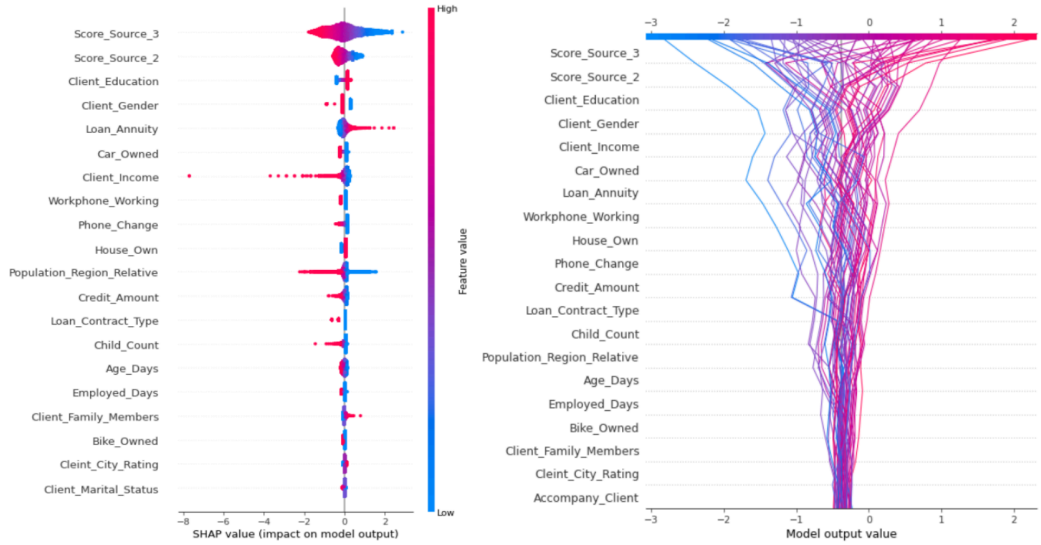
According to statistics [13], among the 34.4 million workers in India with only a high school diploma, about 50% are self-employed, and only one-third hold regular salaried positions. Around 4 million are engaged in temporary wage jobs. In contrast, half of those college and graduate degree holders are more likely to have a job where salaries are more predictable. Educational attainment is strongly correlated with income stability, which in turn significantly affects repayment ability.

Figure 4.a is a beeswarm SHAP plot. The x-axis shows the mean absolute SHAP values (i.e., the average impact on default probability), and features are ranked on the y-axis by importance. We observe that higher credit scores are strongly associated with lower default risk (a negative relationship). Interestingly, first, the probability of default is higher for female borrowers compared to males. This may reflect the socio-economic conditions faced by women in India. For example, female labor force participation is only 25%, and women are concentrated in low-paying, informal jobs with greater income volatility. Additionally, although many loans are registered in the names of women, the vehicles are often used by husbands or fathers, meaning women do not control the repayment cash flow [14]. Second, a higher number of house ownerships is associated with greater default risk. Although counterintuitive, a



a. SHAP Bar Plot b. Proportion of defaulters' education levels

Figure 3: SHAP model interpretability plot



a.SHAP Beeswarm Plot b. SHAP Decision Plot

Figure 4: Analysis of Factors Influencing the Final Decision

study using property-level data suggests that homeowners who extract equity face increased leverage and mortgage payments so that they are more likely to default [15].

Furthermore, as Loan_Annuity increases, the probability of default also rises—likely linked to tighter cash flow among borrowers. Lower Client_Income is associated with negative SHAP values (i.e., suppressing default), while Population_Region_relative may reflect access to more stable income sources. For Age_Days, younger borrowers have more positive SHAP values, indicating that younger individuals are more prone to default. Similarly, borrowers with shorter Employed_Days (e.g., newly hired employees) show higher risk as reflected in positive SHAP values.

Other features positively associated with default include: Credit_Amount (existing debt), Family members (greater household burden), whereas negatively associated features include Car_Owned, suggesting that clients who already possess fixed assets like a car are less likely to default.

Figure 4.b Decision Plot illustrates the cumulative SHAP values of all features for a specific instance, showing the predicted "default probability trajectory." The plot should be read from top to bottom, starting from the base value. Movement to the left (–SHAP) indicates a decrease in default probability, while movement to the right (+SHAP) indicates an increase. From the plot, it is evident that Client_Education and Loan_Annuity are major contributors pushing the default probability upward. In contrast, factors such as Client_Income and Car_Owned help suppress the risk by pulling the probability downward. From a combined feature perspective, even though the client shows relatively stable income and owns assets, the co-occurrence of a low external credit score, high loan amount, and high monthly annuity still raises the default risk. Therefore, we recommend further manual review or increasing the loan interest rate for such cases.

5 Conclusion

This study compares multiple predictive models for loan default classification. Among the models tested, Logistic Regression offered the best overall performance in correctly identifying defaulters, balancing the accuracy with the best recall rate of default. The challenge of class imbalance was a central concern in this research. By applying oversampling techniques SMOTE, the study was able to improve the models' ability to recognize defaulters more effectively. This adjustment helped prevent the underestimation of risk and ensured a more balanced and fair evaluation of borrower profiles.

Looking ahead, more accurate prediction of loan default will likely require the inclusion of additional contextual and behavioral data. While traditional credit scores and demographic variables offer valuable insights, they may not fully capture the volatility in a borrower's financial condition. Unexpected life events, such as sudden job loss, health emergencies, or family disruptions can significantly affect repayment ability but are often missed in statistic datasets. Incorporating real-time or alternative data sources that reflect changes in income, spending behavior, or social context could greatly enhance the responsiveness and accuracy of credit risk models. Based on this, this paper recommends lenders improve their risk assessment processes by providing the following specific strategies:

1. Beyond existing credit and income verification variables, establish a real-time behavioral data access layer, which can monitor borrowers' bank statements, for example, extracting income fluctuation coefficients and unusual spending indicators over the past 90 days.

2. Lenders can access LinkedIn or social security data through compliance partnerships to capture sudden unemployment or career gaps. Once data is accessed, it should be uniformly monitored to ensure variable stability.

3. Regularly run group fairness checks. If any adverse impact is detected on specific groups, immediately trace the source of the features and adjust the weights.

4. With the threshold optimization goal of "maximizing recall while ensuring an overall accuracy rate of $\geq 80\%$," dynamic threshold ranges should be set in the decision engine. Borrowers falling within this range will automatically trigger a second-level manual review to reduce false positives.

References

[1] Lim, H. E., & Goh, S. Y. (2017). Estimating the determinants of vehicle loan default in Malaysia: An exploratory study. International Journal of Management Studies (IJMS), 24(1), 73–90.

[2] Barbaglia, L., Manzan, S., & Tosetti, E. (2023). Forecasting loan default in Europe with machine learning. Journal of Financial Econometrics, 21(2), 569–596. https://doi.org/10.1093/jjfinec/nbab010

[3] Fawwaz, M. J., & Zulkarnain. (2025). Development of loan default prediction models in Indonesia's multifinance industry. Enrichment: Journal of Multidisciplinary Research and Development, 3(2), 195–210. https://doi.org/10.55324/enrichment.v3i2.357

[4] Kobone, B. T., & Montshiwa, T. V. (2025). Impact of sample size on the robustness of machine learning algorithms for detecting loan defaults using imbalanced data. Journal of Applied Data Sciences, 6(3), 1830–1849.

[5] Zhang, X., Zhang, T., Hou, L., Liu, X., Guo, Z., Tian, Y., & Liu, Y. (2025). Data-driven loan default prediction: A machine learning approach for enhancing business process management. Systems, 13(7), 581. https://doi.org/10.3390/systems13070581

[6] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[7] Antar, A. (2024). Evaluating the impact of borrower characteristics, loan specific parameters, and property conditions on mortgage default risk. Theoretical and Practical Research in Economic Fields, 15(2), 481–501. https://doi.org/10.14505/tpref.v15.2(30).24

[8] Huang, W., Chen, A., & Qian, Y. (2022). The role of marital status in the online lending market: Evidence from the Renrendai platform. Asia Pacific Journal of Financial Studies, 51(4), 594–617. https://doi.org/10.1111/ajfs.12389

[9] Aliano, M., Alnabulsi, K., Cestari, G., & Ragni, S. (2023). The role of gender and education in peer-to-peer lending activities: Evidence from a European cross-country study. European Scientific Journal, 14, 95. https://eujournal.org/index.php/esj/article/view/16433

[10] Xu, J. (2025). Factors influencing loan default: An empirical analysis based on microscopic evidence. Journal of Economics, Business and Management, 13(1), 14–22. https://www.joebm.com/vol13/JOEBM-V13N1-841.pdf

[11] Dunn, L. F., & Mirzaie, I. A. (2022). Gender differences in consumer debt stress: Impacts on job performance, family life and health. Journal of Family and Economic Issues, 43(4), 897–914.

[12] Chen, J., Jiang, J., & Liu, Y. (2018). Financial literacy and gender difference in loan performance. Journal of Empirical Finance, 48, 307–320.

[13] Mehrotra, S. (2019). Informal employment trends in the Indian economy: Persistent informality, but growing positive development (Employment Working Paper No. 254, pp. 4–5). International Labour Organization.

[14] Reserve Bank Innovation Hub (RBIH). (2023). Financial inclusion of women: Current evidence from India. https://www.orfonline.org/research/financial-inclusion-of-women-current-evidence-from-india

[15] Laufer, S. (2018). Equity Extraction and Mortgage Default. Review of Economic Dynamics, 28(April), 1–33.

[16] Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 34(9), 6390–6404.