



Predicting Informal Loans Using Machine Learning Algorithms

Pim Pinitjitsamut

Department of Economics, Rutgers University, USA

*Wisarut Suwanprasert**

Department of Economics and Finance, Middle Tennessee State University, USA.

Received 14 February 2025, Received in revised form 29 January 2026,

Accepted 23 February 2026, Available online 4 September 2026

Abstract

This paper applies machine learning methods to predict informal borrowing participation and informal loan amounts in Thailand, using individual-level survey data covering approximately 4,800 respondents across six regions. We evaluate the performance of K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting (XGBoost) models and benchmark their predictive accuracy against a standard logistic regression. XGBoost achieves the highest accuracy in predicting informal borrowing participation, while Random Forest performs best in predicting loan amounts. To move beyond predictive performance, we incorporate interpretable machine learning tools to examine how key socioeconomic characteristics contribute to model predictions. Correlation and interpretability analyses reveal clear segmentation between formal and informal credit markets: households with lower income and limited access to formal loans face substantially higher informal borrowing costs. Our results show that accurate prediction is possible using a relatively small set of observable socioeconomic indicators. The findings highlight how interpretable machine learning can complement traditional economic analysis and provide practical tools for identifying households vulnerable to high-cost informal credit in data-scarce environments.

Keywords: Informal Loans, Machine Learning, Random Forest, Predictive Modeling, Thailand.

JEL Classifications: E26, G51, O16, O17

* **Corresponding author:** Email: Wisarut.Suwanprasert@mtsu.edu

1. Introduction

Informal lending remains a central feature of financial systems in many developing economies, particularly where access to formal credit is limited. Informal loans often carry extremely high interest rates. In Thailand, these rates frequently exceed 60 percent annually, far above the legal limits applied to formal lending institutions (Siamwalla et al., 1990). Despite regulatory efforts, including interest rate caps on formal credit, informal lending persists and continues to expose borrowers to high financial costs and coercive recovery practices.

A key challenge in studying informal credit markets is the lack of reliable data. Because informal loans operate outside regulatory oversight, transactions are rarely recorded, and borrower behavior is difficult to observe using standard administrative sources. This limitation constrains empirical analysis and complicates efforts to identify households most vulnerable to high-cost informal borrowing. This paper addresses this challenge by applying machine learning techniques to predict informal borrowing participation and loan amounts using individual-level survey data from Thailand.

Using a dataset of approximately 4,800 individuals across 12 provinces and six regions, collected by the Department of Special Investigation under the Ministry of Justice, we apply three supervised machine learning algorithms: K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting (XGBoost). These methods are well-suited to settings characterized by complex interactions, nonlinear relationships, and a large set of observable household characteristics. We benchmark their performance against a standard logit model to situate the machine learning results within conventional econometric analysis.

Our results indicate that different algorithms perform best for different tasks. XGBoost achieves the highest accuracy in predicting whether households participate in informal borrowing, while Random Forest delivers the strongest performance in predicting informal loan amounts. This pattern highlights the importance of matching predictive methods to specific economic outcomes. While participation decisions are best captured by models that flexibly adjust decision boundaries, borrowing intensity appears more closely related to nonlinear interactions among household income, expenses, age, and loan terms.

Beyond predictive accuracy, this paper emphasizes interpretability and economic insight. We use feature importance measures and SHAP values to examine how key variables contribute to model predictions. The results align with established economic mechanisms. High household expenditures are indicative of liquidity constraints, occupations such as sellers and freelancers proxy for income instability, and regional indicators capture geographic disparities in access to formal credit markets. Importantly, these results are interpreted as predictive associations rather than causal effects.

The main contribution of this paper is to demonstrate how interpretable machine learning can be used to generate accurate and economically meaningful predictions of informal borrowing behavior in a data-scarce environment. We show that a relatively small set of observable socioeconomic characteristics is sufficient for effective risk screening. This makes machine learning a practical complement to traditional econometric approaches, particularly for policy applications aimed at identifying households vulnerable to reliance on high-cost informal credit.

This study is closely related to earlier work on informal lending in Thailand, including Pinitjitsamut & Suwanprasert (2022), Siamwalla et al. (1990), Tanomchat & Sampattavanija (2018), and Suwanprasert (2025). Pinitjitsamut & Suwanprasert (2022)

established key stylized facts regarding informal loans and employed econometric methods to analyze the determinants of loan decisions and amounts. Our paper extends their findings by applying machine learning techniques to predict these outcomes, providing a practical approach to infer informal loan conditions when traditional data sources are unavailable.

Siamwalla et al. (1990) highlighted the dominance of informal lenders during the 1984-1985 period, with 72.4 percent of borrowing households relying on them, attributing high interest rates to economic rents driven by information asymmetry. In contrast, Karaivanov & Kessler (2018) found lower interest rates on informal loans than on formal ones, while Tanomchat & Sampattavanija (2018) suggested that elevated interest rates reflect the high default risk of borrowers, with only powerful lenders willing to extend credit under such conditions. More recently, Suwanprasert (2025) highlighted persistent gender gaps in access to and use of informal loans.

Our work aligns with a broader body of literature that investigates the impacts of informal credit. Chemin (2008) used propensity score matching to examine how informal loans affect household expenditures, labor supply, and children's school enrollment in Bangladesh. Pitt & Khandker (1998) and Kaboski & Townsend (2005) explored the role of government-led lending programs in improving economic outcomes. Kislat (2015) used a difference-in-differences approach to assess the benefits of informal lending across income groups in rural Northeast Thailand.

Furthermore, this paper is connected to research examining the choice between formal and informal financial markets. For example, Guirking (2008) found that households in Peru rely on informal lenders when access to formal credit is limited, while Mohieldin & Wright (2000) suggested that borrowers in Egypt's informal sector face credit constraints that prevent them from engaging with formal lenders. Liu & Roth (2022) argue that informal borrowers may be trapped in cycles of debt, as lenders benefit from keeping them indebted over time.

The use of machine learning to improve predictions in data-limited settings is an emerging area of study. For example, Mathen et al. (2024) used nighttime light data to measure economic activity and inequality in India. Similarly, Fenz et al. (2024) applied a random forest algorithm to calculate population density in the Philippines and Thailand. Puttanapong et al. (2023) integrated satellite imagery with machine learning models to estimate real-time provincial GDP, while Chaengkham & Wianwiwat (2021) applied Support Vector Machine (SVM) models for monthly stock market trend forecasts. Kusunthum et al. (2022) demonstrated the effectiveness of neural networks, random forests, and decision trees in forecasting passenger demand for the Metropolitan Rapid Transit (MRT) system in Thailand. These examples illustrate the growing role of machine learning in enhancing predictive capabilities and informing decisions where traditional data are scarce.

The remainder of the paper is structured as follows. Section 2 describes our data and variables. Section 3 discusses the methodology, and Section 4 presents empirical findings using machine learning techniques. Section 5 discusses policy implications. Section 6 concludes.

2. Data and Variable Description

2.1 Data Source

The analysis in this paper is based on individual-level survey data provided by the Legal Aid Center for Debtors and Victims of Injustice (LADVIMOJ), under the Department of Special Investigation (DSI), Ministry of Justice. The data collection, conducted in 2014, covers 4,878 individuals across 12 provinces, representing all six regions of Thailand. The surveyed provinces include Bangkok and Pathum Thani (Bangkok Metropolitan); Saraburi, Ratchaburi, and Phitsanulok (Central); Chonburi (Eastern); Nakhon Si Thammarat and Songkhla (Southern); Chiang Rai (Northern); and Yasothon, Maha Sarakham, and Nong Khai (Northeastern), spanning 105 districts.

LADVIMOJ's mission is to provide legal assistance to borrowers burdened with informal loans bearing high interest rates, often a consequence of limited legal literacy and restricted access to formal financial systems. The dataset offers a unique view into the borrowing behaviors and socioeconomic profiles of individuals facing these challenges, capturing details on both informal and formal loans, interest rates, lender types, and loan purposes.

2.2 Variable Description and Cleaning Procedure

The dataset is cross-sectional and primarily denominated in Thai baht, with an approximate exchange rate of 33 baht to 1 USD at the time of data collection. Key variables include the amounts borrowed through formal and informal channels, their respective interest rates, and borrower characteristics such as age, gender, income, education, and expenditure patterns.

Formal loans are defined as credit obtained from regulated institutions, including banks and financial intermediaries, subject to oversight by the Bank of Thailand and the Ministry of Finance, with interest rate caps (e.g., 25% for personal loans and 16% for credit cards). In contrast, informal loans are sourced from unregulated entities such as local private lenders, loan sharks, and retailers, often with monthly interest rates that exceed legal thresholds and are not enforceable in court. Despite the risks, the flexibility and accessibility of informal loans make them appealing, particularly in the absence of viable formal credit options.

Socioeconomic variables were further refined to facilitate analysis. Occupations are grouped into nine categories: sellers, business owners, contract-based workers, farmers, freelancers, private sector employees, government-owned corporation employees, government employees, and the unemployed. These categories were aggregated into salaried and non-salaried groups based on income stability, reflecting differences in financial security and access to credit.

Loan motivations were classified into four categories: (i) necessary expenses (e.g., medical bills, tuition); (ii) unnecessary expenses (e.g., luxury items); (iii) investments; and (iv) debt refinancing. Necessary expenses include essential costs such as medical bills, tuition, and household necessities, as well as traditional family-related expenditures. Unnecessary expenses, on the other hand, cover non-essential purchases such as luxury goods and electronic devices. Debt refinancing involves consolidating or repaying existing formal or informal loans, highlighting the interplay among different credit sources.

Total personal expenses are calculated by aggregating expenditures across categories such as housing, food, utilities, communication, education, transportation, medical costs, insurance, investments, and vehicle payments. This aggregation provides

a comprehensive view of household financial obligations, an essential factor in predicting borrowing behavior.

Following data cleaning to address inconsistencies and remove outliers, the final dataset consists of 4,628 observations.

Table 1 reports descriptive statistics for all variables used in the analysis, including means, medians, and standard deviations, for the full sample and for borrowers. For the modeling stage, the classification task uses the full analytic sample (N = 4,628) to predict informal borrowing participation. Loan amount prediction is estimated conditional on borrowing. In addition, we construct a two-part expected loan measure that combines predicted borrowing probabilities with predicted conditional loan amounts. As a robustness check, we also estimate an unconditional specification in which non-borrowers' loan amounts are set to zero. We discuss the interpretation and policy relevance of each approach.

Table 1: Summary Statistics.

Variable	Observations	Mean	Median	Std. Dev.
Amount of all loans (total)	4,628	193,142.30	40,000	547,941.8
Amount of all loans (conditional on having a loan)	3,357	227,764.00	50,000	607,536.3
Amount of informal loans (total)	4,628	22,961.70	0	105,505.9
Amount of informal loans (conditional on having an informal loan)	1,957	54,300.90	20,000	156,937.6
Interest rate on informal loans (% per month) (conditional on having an informal loan)	1,957	16.50	10	63.4
Amount of formal loans (total)	4,628	142,250.80	0	498,423.0
Amount of formal loans (conditional on having a formal loan)	2,216	297,083.40	80,000	687,696.7
Interest rate on formal loans (%) (conditional on having a formal loan)	2,216	14.10	6	26.1
Total family expenses	4,628	17,716.80	13,300	22,450.1
Total personal expenses	4,628	7,238.50	2,895	14,331.3
Total personal income	4,628	15,085.70	11,900	17,408.1
Personal savings	4,628	749.70	0	1,966.0

Source: Authors' calculations.

3. Methodology

In this section, we apply machine learning techniques to predict the likelihood that individuals will take out informal loans and, if applicable, estimate the loan amount. By combining numerical and categorical variables, we aim to optimize predictive accuracy and identify key socioeconomic and financial characteristics associated with borrowing behavior.

3.1 Data Preparation

The dataset includes both numerical and categorical variables. Numerical variables capture household composition (e.g., number of household members, income earners, college students, unemployed members, and stay-at-home members) as well as individual and financial characteristics, including age, second-job status, total household income, expenditures, savings, loan amounts (formal and informal), outstanding formal loan balances, formal loan interest rates, informal loan terms, and informal loan interest rates. Categorical variables include gender, household role, province of residence, education level, occupation, and stated reasons for borrowing.

Several preprocessing steps were applied prior to estimation. First, key financial variables—including total income, savings, outstanding formal loan balances, and total personal and family expenditures—were log-transformed to reduce skewness. Second, all numerical variables were standardized by subtracting their means and dividing by their standard deviations, producing z-scores. This normalization ensures that variables are on a comparable scale, which is important for models sensitive to the magnitude of inputs. Finally, categorical variables were transformed using one-hot encoding, with each category represented by a binary indicator. This encoding allows the models to incorporate categorical information without imposing an ordinal structure on the categories.

3.2 Machine Learning Models

We apply three supervised machine learning algorithms—K-Nearest Neighbors (KNN), Random Forests, and Extreme Gradient Boosting (XGBoost)—to classify informal loan behaviors. We rank features, both numerical and categorical, using Random Forest feature-importance selection. This selection sorts variables based on the magnitude of their effects on the target variable. In decision trees, every node is a condition of how to split values in a single feature so that similar values of the dependent variable end up in the same set after the split. For classification problems, the condition is based on Gini impurity, while for regression trees, it is based on variance. Using the training set, we compute the contribution of each feature to the average decrease in impurity across trees.

We construct two datasets. Dataset 1 includes all variables, including the borrower's reported loan purpose. Dataset 2 excludes the loan-purpose variable to simulate conditions in which policymakers or survey designers cannot observe loan purposes, as is often the case in administrative data. This design allows us to assess whether accurate prediction is feasible in such restricted-data settings.

We use Random Forest feature importance to rank predictors. It is important to note that feature importance reflects correlations with the prediction outcome, not causal effects. To better interpret these results, we further employ SHAP values to visualize how changes in individual features influence predicted probabilities. These tools improve transparency while making clear that our analysis remains predictive rather than causal.

The dataset is split into training and test sets at 60:40, ensuring sufficient training data while retaining a substantial portion for out-of-sample testing. To enhance reproducibility, we set the random state to 23.

The first algorithm we applied was K-Nearest Neighbors (KNN), a simple yet effective method that classifies new data points based on the majority class of their nearest neighbors in the training set. Distances between the new sample and all points in the dataset were calculated, and the sample's classification was determined by its closest neighbors. Although KNN is easy to implement and interpret, it can be computationally expensive for large datasets and is sensitive to the choice of distance metrics.

Next, we used Random Forest, an ensemble-based algorithm that constructs multiple decision trees from random subsets of the data. Each tree is built using a random selection of features and samples, and the final classification is determined by majority voting among the trees. One of the key advantages of Random Forest is its ability to reduce overfitting by averaging predictions across multiple trees, resulting in a more generalized model. In this study, the Random Forest model consisted of 150 trees, with the same 60:40 training-to-testing split to ensure ample training data.

Finally, we applied XGBoost, another ensemble learning method based on gradient boosting. Unlike Random Forest, XGBoost builds trees sequentially, with each tree trained to correct the errors of its predecessor, thereby minimizing overall error across iterations. For the XGBoost models, we tested several configurations. The first model used default parameters, while the second was fine-tuned with a maximum depth of 8, a learning rate of 0.1, and a subsample ratio of 0.5. In the third model, we performed a grid search to optimize hyperparameters and improve performance. The final model included Gamma XGBoost tuning, which penalizes overfitting by introducing a regularization term to control the complexity of the trees.

4. Results

4.1 Correlation analysis and feature ranking

A correlation analysis was run on 16 of the most important features identified through a Random Forest feature importance ranking to understand their relationships better and identify potential multicollinearity issues. Table 2 lists these variables, which capture household demographics, financial circumstances, and borrowing characteristics central to informal credit decisions.

Table 2: Variables Included in Correlation Analysis

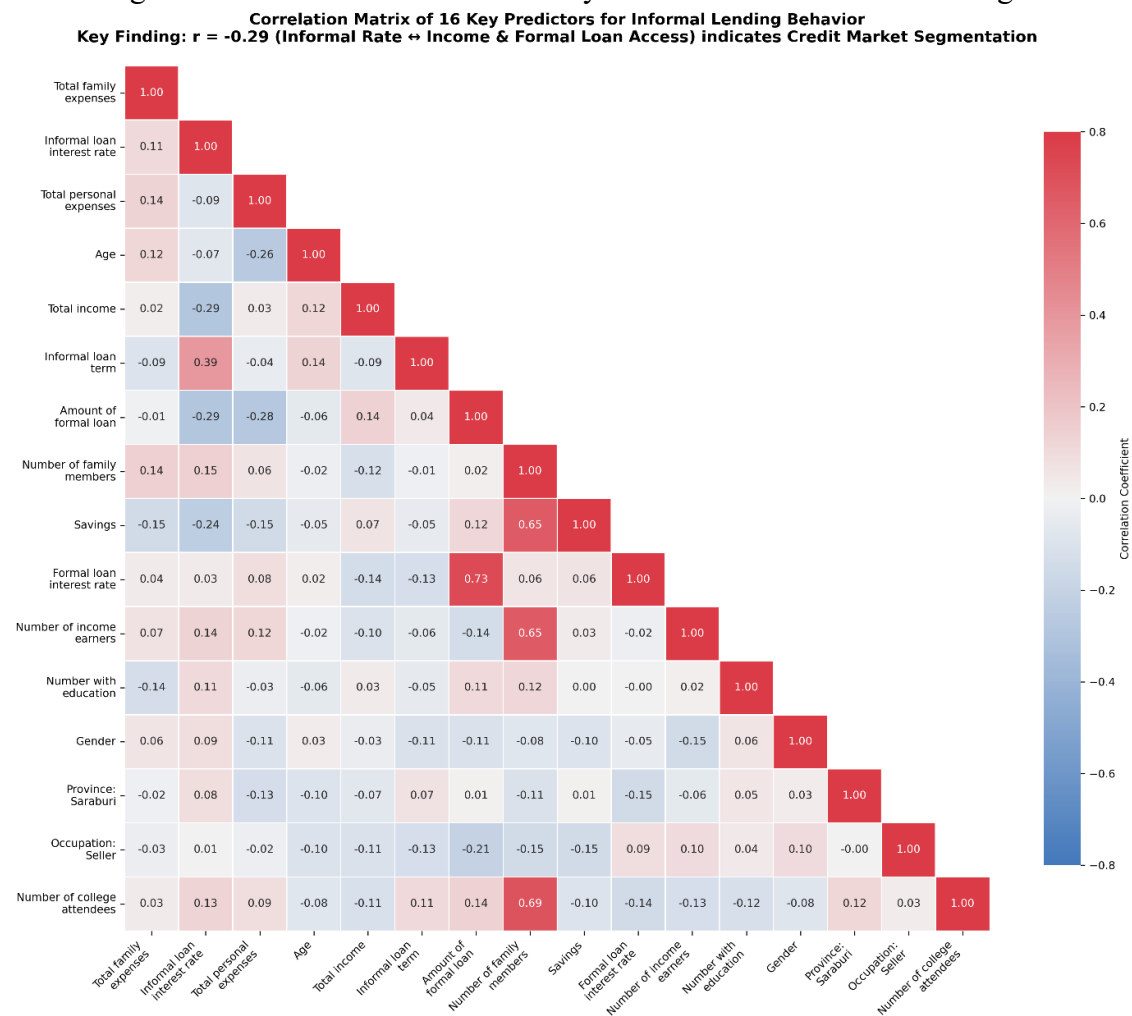
No.	Variable Name	Type
1	Total family expenses	Continuous
2	Total personal expenses	Continuous
3	Age	Continuous
4	Total income	Continuous
5	Informal loan interest rate	Continuous
6	Informal loan term	Continuous
7	Amount of formal loan	Continuous
8	Formal loan interest rate	Continuous
9	Number of family members	Count
10	Number of income earners	Count
11	Number of members with education	Count

No.	Variable Name	Type
12	Number of college attendees	Count
13	Savings	Continuous
14	Gender	Binary
15	Province: Saraburi	Binary
16	Occupation: Seller	Binary

Source: Authors' calculations.

Figure 1 presents the complete correlation matrix for all 16 variables, visualizing the pairwise relationships among household demographics, financial variables, and loan characteristics. The heatmap format reveals several noteworthy patterns that align with theoretical expectations, with color intensity indicating correlation strength and direction.

Figure 1. Correlation Matrix of 16 Key Predictors for Informal Lending Behavior



Source: Authors' calculations.

4.1.1 Positive Correlations

Figure 1 reveals multiple notable patterns that align with theoretical expectations. There are strong positive relationships between household structure variables. For example, family size is linked to the number of income earners ($r = 0.65$) and college students ($r = 0.69$). This means that larger families usually have more people working and going to college. Savings show a moderate positive association with household size

($r = 0.65$), suggesting that larger households might save more overall, even though their per-person costs could be higher.

The total amount of outstanding formal loans is strongly correlated with formal loan interest rates ($r = 0.73$). This aligns with risk-based pricing systems that charge higher rates for larger loans or higher-risk customers. This correlation likely shows a direct relationship between loan size and interest costs, as well as between creditworthiness characteristics and the borrower's ability to borrow and the terms that formal lenders offer.

Additionally, a positive correlation between informal loan interest rate and informal loan term ($r = 0.39$) suggests that borrowers, when faced with a higher interest rate, may negotiate longer repayment periods to ease immediate cash flow pressures. Some downsides of this strategy are that borrowers extend their indebtedness and increase their total interest burden over time.

4.1.2 Negative Correlations

The darker blue-shaded cells in Figure 1 indicate systematic negative correlations among several key variables. Most notably, informal loan interest rates are negatively correlated with both total income ($r = -0.29$) and the amount of formal loans ($r = -0.29$). Although these correlations are moderate in magnitude, their consistency points to a clear pattern: households with lower incomes and more limited access to formal credit tend to face substantially higher interest rates when borrowing from informal lenders. The modest size of these correlations likely reflects heterogeneity within informal lending markets. Informal interest rates vary not only with income and access to formal loans but also with factors such as social networks, regional market conditions, and individual bargaining dynamics that are not observed in the data.

This pattern extends to other indicators of financial vulnerability, although the correlations remain in the weak-to-moderate range. Informal loan interest rates are negatively correlated with savings ($r = -0.24$), indicating that households with limited financial buffers face higher borrowing costs precisely when they are least able to absorb them. Similarly, the negative correlation between formal loan amounts and personal expenses ($r = -0.28$) suggests that individuals experiencing greater financial strain may be excluded from formal credit markets and, as a result, rely on more costly informal alternatives.

Although the magnitudes of these correlations are modest (all $|r| < 0.30$), they indicate that informal credit terms are shaped by multiple factors beyond those observed in the data. Importantly, the fact that these indicators consistently point in the same direction reinforces the presence of systematic disadvantage in informal lending markets.

Occupational and demographic factors further illuminate these stratification patterns, albeit with relatively weak correlations that nonetheless point in theoretically meaningful directions. The seller occupation category is negatively correlated with both formal loan access ($r = -0.21$) and savings ($r = -0.15$), indicating structural disadvantages across both formal and informal credit channels. In addition, the negative correlation between personal expenses and age ($r = -0.26$) suggests a life-cycle pattern. Younger borrowers tend to face higher personal expenses relative to income, which can place them under greater financial pressure and increase reliance on informal borrowing, even under unfavorable terms. Although these correlations do not represent strong bivariate relationships, they suggest underlying dynamics that may operate through interactions and nonlinearities—patterns that are difficult to detect using correlation analysis alone but can be captured by machine learning methods.

4.1.3 Implications for Multicollinearity and Model Specification

The correlation analysis in Figure 1 indicates no severe multicollinearity concerns that would impede model estimation. The only pairwise correlation approaching conventional thresholds is between formal loan amounts and formal interest rates ($r = 0.73$), which remains within acceptable bounds. Overall, the 16 features provide largely complementary rather than redundant information. Most correlations are weak to moderate (with the majority below $|r| = 0.30$), a pattern that is well-suited to machine learning applications. This dispersion suggests that the features capture distinct dimensions of borrowing behavior, allowing the algorithms to combine information effectively without the instability typically associated with highly collinear predictors.

The moderate correlation magnitudes further indicate that these variables operate through separate pathways, consistent with the multidimensional nature of informal credit decisions. Household structure variables—such as family size, number of income earners, and number of college attendees—exhibit some clustering ($r = 0.65$ – 0.69) but remain sufficiently distinct from financial status indicators and loan characteristics to justify their separate inclusion. This empirical structure supports a modeling framework that jointly incorporates demographic, financial, and institutional factors. Moreover, the modest bivariate correlations suggest that strong predictive performance, when achieved, is likely driven by interactions and nonlinear relationships rather than simple additive effects—precisely the types of patterns that ensemble methods are designed to capture.

4.1.4 Feature Importance Rankings and Model Input Selection

Beyond correlation patterns, Random Forest feature importance rankings provide an additional perspective on predictor relevance. Table 3 reports the ten most important features, ranked by their contribution to predictive accuracy rather than by simple bivariate associations with the outcome. This ranking offers a data-driven hierarchy of variables that matter most for classification performance.

Table 3: Top 10 Features by Random Forest Importance (Classification Task)

Rank	Feature	Importance (%)
1	Informal loan interest rate	20.86
2	Total family expenses	11.11
3	Age	10.17
4	Informal loan term	10.08
5	Total personal expenses	9.21
6	Total income	9.14
7	Formal loan interest rate	4.74
8	Savings	4.73
9	Amount of formal loan	4.49
10	Number of family members	4.28

Source: Authors' calculations.

Informal loan interest rates emerge as the dominant predictor by a wide margin, accounting for nearly 21 percent of the model's discriminatory power in the classification task. This prominence likely reflects both their direct signaling role—where higher rates may indicate lender risk assessment or borrower urgency—and their role as a summary statistic that captures unobserved aspects of creditworthiness and local market conditions.

Family and personal expenses form the second tier of importance, jointly contributing more than 20 percent to overall model performance. This pattern underscores the role of household financial strain in shaping informal borrowing decisions. Households facing high expenditure burdens relative to income are

substantially more likely to rely on informal credit, even when formal alternatives may exist in principle.

Demographic characteristics, such as age and household structure, exhibit moderate importance (approximately 10 percent and 4 percent, respectively), indicating that they meaningfully—but secondarily—influence informal borrowing behavior. Their role likely reflects life-cycle factors, job stability, and family responsibilities, which shape credit needs and access but operate less directly than immediate financial constraints.

In contrast, formal credit variables display relatively low importance despite their conceptual relevance. Both formal loan amounts and formal interest rates contribute less than 5 percent to overall classification accuracy. This likely reflects the fact that their effects are mediated through other variables or that they exhibit limited variation within the sample of informal borrowers. Rather than diminishing their theoretical importance, this pattern suggests that, conditional on other predictors, formal credit variables provide relatively little additional information for classification purposes.

4.1.5 Bridging Correlation Analysis and Machine Learning Approaches

Figure 1 illustrates that informal credit behavior is complex and multidimensional, motivating the use of machine learning methods. While bivariate correlations offer useful initial insights into pairwise relationships, they are limited in their ability to capture nonlinearities, threshold effects, and conditional dependencies that are likely to characterize actual borrowing decisions. The predominance of weak-to-moderate correlations (with most $|r| < 0.30$) suggests that informal borrowing is shaped by a combination of factors rather than a single dominant predictor. Ensemble methods such as Random Forest and XGBoost are well-suited to this setting, as they model hierarchical decision rules and uncover complex patterns that linear approaches may overlook.

The correlation of $r = -0.29$ between informal interest rates, income, and access to formal credit underscores this point. Although these linear associations are moderate in magnitude, they point to underlying dynamics that raise important questions about functional form and interaction effects. For example, disadvantages faced by low-income households may be amplified when access to formal credit is limited, resulting in nonlinear or multiplicative effects. Similarly, informal interest rates may respond sharply once income falls below certain thresholds or may vary systematically across regions or occupations. Such questions cannot be addressed using correlation analysis alone. By contrast, machine learning methods can explore these complexities through recursive partitioning of the feature space and the identification of localized patterns, providing a richer characterization of informal credit behavior.

In addition, feature importance rankings complement correlation matrices by identifying predictors whose effects may be obscured in bivariate comparisons. Variables with weak pairwise correlations can still play an important role when interacting with other features or when their influence is concentrated within specific subpopulations. This distinction between correlation-based and importance-based measures underscores the value of a modeling approach that integrates both perspectives: correlations are useful for assessing multicollinearity and broad associations, while feature importance scores guide variable selection and interpretation in predictive models.

4.2 Classification Results

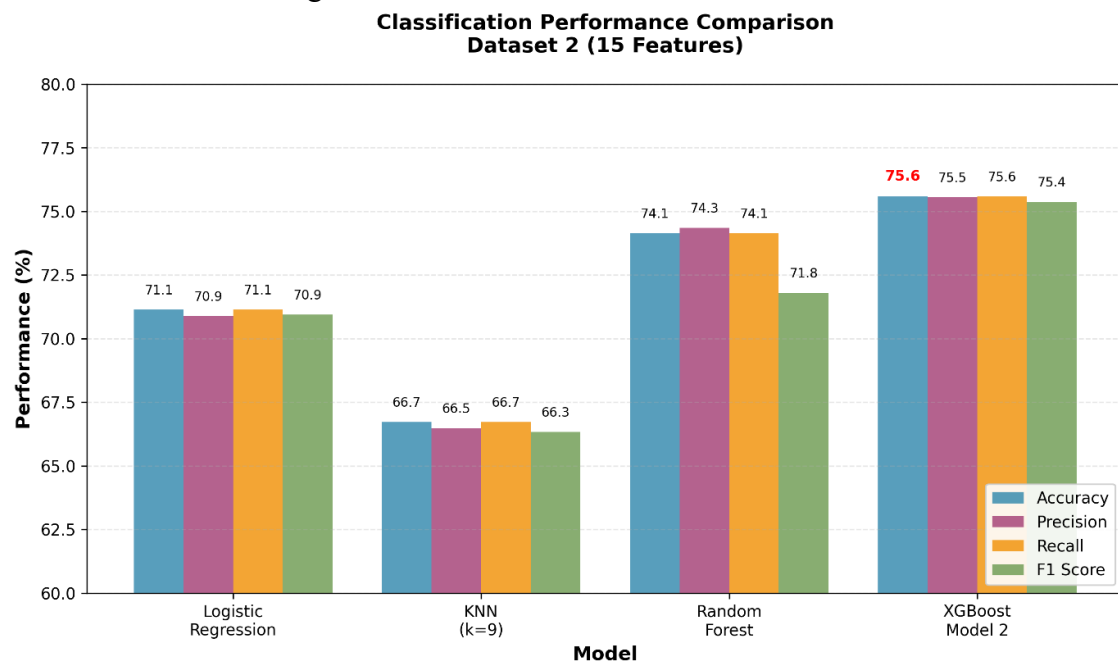
We conduct four experiments for each machine learning technique by combining two datasets with two feature sets. Dataset 1 includes all available variables, while Dataset 2 excludes information on the reason for borrowing. For each dataset, models are

estimated using either the full set of features or the top 16 features identified by Random Forest importance rankings. This design allows us to evaluate classification performance across varying data configurations. Classification accuracy measures the extent to which each model correctly predicts whether an individual engages in informal borrowing. Accuracy results for all 12 experiments are reported in Figure 2.

The classification analysis reveals clear performance differences across modeling approaches. Figure 2 presents a comprehensive comparison of four methods: logistic regression as a baseline, K-Nearest Neighbors (KNN), Random Forest, and XGBoost.

XGBoost delivers the strongest performance, achieving an accuracy of 75.58 percent, a precision of 75.55 percent, a recall of 75.58 percent, and an F1 score of 75.36 percent. The close alignment across these metrics indicates balanced classification performance, with the model effectively identifying both borrowers and non-borrowers. XGBoost's advantage likely reflects its gradient-boosting structure, which iteratively corrects classification errors and adapts flexibly to complex decision boundaries.

Figure 2: Classification Performance Metrics



Source: Authors' calculations.

Random Forest performs slightly less well, with an accuracy of 74.14 percent. Its strength lies in aggregating predictions from many decision trees, each trained on a different bootstrap sample and a random subset of features. This ensemble approach reduces variance and improves generalization but lacks the sequential error-correction mechanism that characterizes boosting methods.

Logistic regression, which serves as the econometric benchmark, achieves an accuracy of approximately 72 percent. While this performance is notable, it falls short of that of the ensemble methods. The gap suggests that informal borrowing behavior involves nonlinearities or interactions that are not well captured by the linear index structure of logistic regression. Although logistic regression remains attractive for its simplicity and interpretability, the 3–4 percentage point improvement in accuracy delivered by ensemble methods is meaningful for practical classification tasks.

K-Nearest Neighbors (KNN) exhibits the weakest performance, with an accuracy of 66.72 percent—only marginally better than random classification in a roughly balanced sample. This relatively poor performance likely reflects KNN's sensitivity to

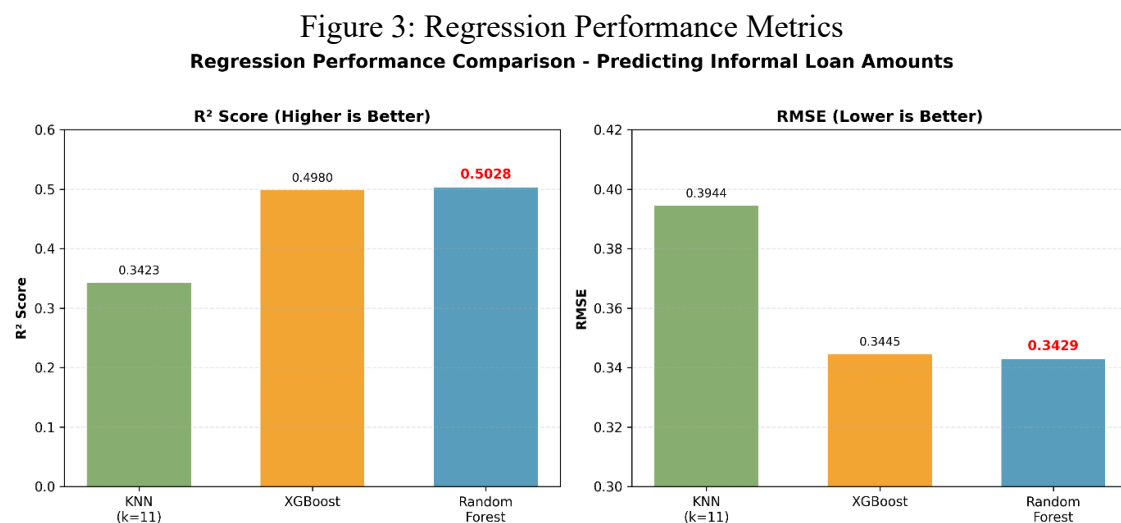
irrelevant features and its vulnerability to the curse of dimensionality. With 16 predictors, the feature space becomes sparse, reducing the reliability of distance-based similarity measures. The limited improvement over naive classification suggests that KNN's notion of local similarity does not align well with the structure of informal borrowing decisions.

Figure 2 summarizes these performance differences, showing that XGBoost consistently outperforms the other models across all four evaluation metrics. More broadly, the results indicate that ensemble methods—particularly Random Forest and XGBoost—deliver substantial gains over traditional approaches in predicting informal borrowing behavior. These findings highlight the value of ensemble methods for research on credit access and suggest their potential usefulness in applications that seek to identify households at elevated risk of informal borrowing.

4.3 Regression Results

Regression Performance: Predicting Loan Amounts

Following the classification analysis, we evaluate the models' ability to predict the amounts households borrow from informal sources. In contrast to the classification results, the regression analysis reveals a different performance ranking across models. Figure 3 reports the coefficient of determination (R^2) and root mean squared error (RMSE) for the three machine learning approaches.



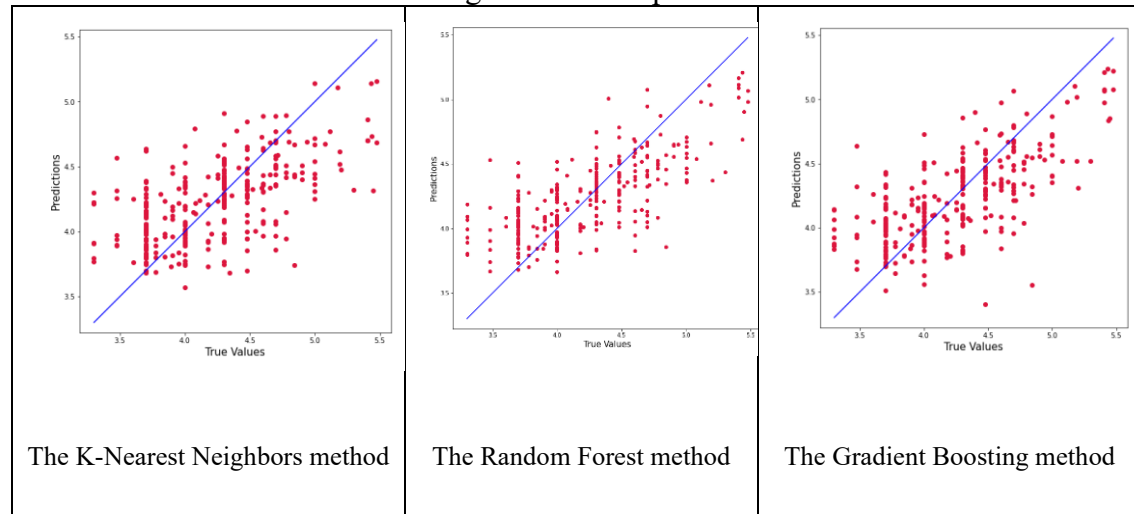
Source: Authors' calculations.

Random Forest performs best in the regression task, achieving an R^2 of 0.5028 and an RMSE of 0.3429, indicating that it explains slightly more than half of the variation in informal loan amounts. This level of predictive accuracy is notable given the inherent unpredictability of borrowing behavior, which depends not only on observable household characteristics but also on unobserved factors such as urgent liquidity needs, access to social networks, and lender availability. Random Forest's strength likely stems from its ability to capture complex, nonlinear relationships between predictors and outcomes without imposing restrictive functional form assumptions.

XGBoost performs similarly but slightly less well, with an R^2 of 0.4980—representing a difference of less than one percent of explained variance—and an RMSE of 0.3445, which closely matches the Random Forest error rate. While both methods capture comparable underlying patterns, Random Forest's aggregation of fully grown trees may allow it to better represent variation in loan amounts than XGBoost's sequential error-correction approach, which is often optimized for classification boundaries.

Overall, the results suggest a modest but consistent advantage for Random Forest in predicting borrowing intensity.

Figure 4: Scatterplots



Source: Authors' calculations.

Consistent with its weaker classification performance, K-Nearest Neighbors (KNN) attains an R^2 of only 0.3423 in the regression task. This substantially lower explanatory power reinforces the view that distance-based methods struggle with the dimensionality and feature heterogeneity present in the data. The roughly 16 percentage point gap in explained variance between KNN and the ensemble approaches underscores the superior performance of Random Forest and Gradient Boosting in both classification and regression settings.

Overall, the results confirm the suitability of ensemble methods for predicting informal loan amounts, with Random Forest consistently delivering the strongest performance across experimental configurations.

Figure 4 presents scatterplots of actual versus predicted informal loan amounts for each machine learning method. Across all experiments, the models capture a substantial share of the variation in loan amounts, with R^2 values ranging from 0.35 to 0.51. The strongest performance is observed when using all features from Dataset 1, where R^2 values between 0.48 and 0.51 indicate that nearly half of the variation in borrowing amounts is explained by the input variables.

Predictive accuracy is further evaluated using the root mean squared error (RMSE), which ranges from 0.3391 to 0.3944 across specifications. Lower RMSE values correspond to better predictive performance, with the lowest RMSE of 0.3391 achieved when all features from Dataset 1 are included. This result highlights the benefit of a more comprehensive information set for predicting loan amounts. At the same time, restricting the models to the top features identified by importance rankings leads to only modest increases in RMSE, indicating that prediction remains robust even with a reduced feature set.

The scatterplots visually reinforce these findings, showing close alignment between actual and predicted loan amounts in the best-performing models. The ability to maintain strong predictive performance with fewer variables suggests that a small set of key socioeconomic factors—such as income, expenses, and loan terms—accounts for much of the variation in informal loan sizes. These results underscore the practical applicability of the models and highlight the central role of these variables in shaping informal borrowing intensity.

4.4 Feature Importance Analysis

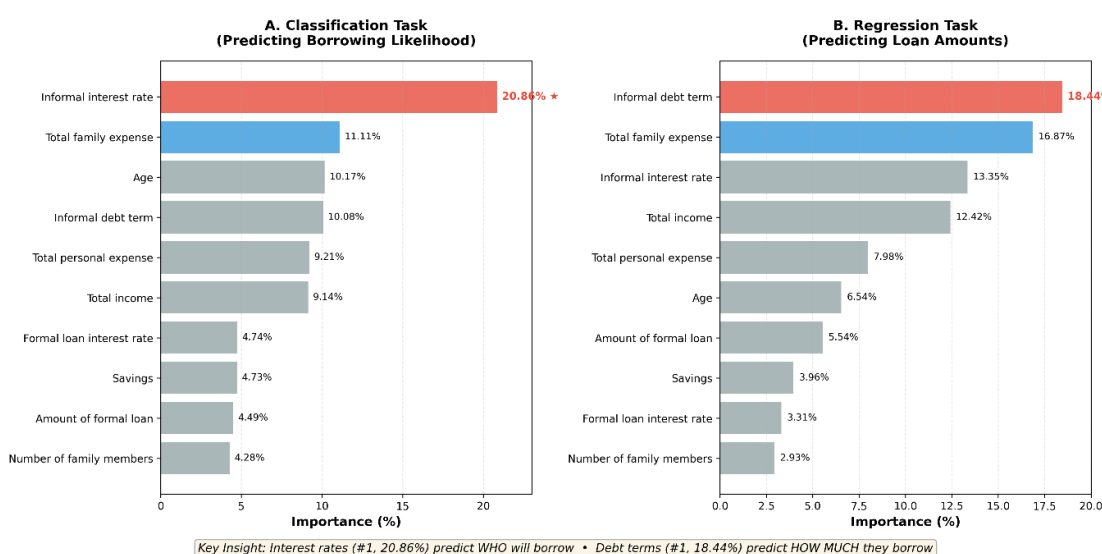
4.4.1 Task-Specific Patterns in Informal Credit Markets

The preceding sections establish that different machine learning models excel at different predictive tasks, with some performing better in classification and others in regression. This raises a natural set of questions: Do the same features drive both informal borrowing participation and borrowing intensity? And does the relative importance of these features change depending on whether the objective is to identify borrowers or to predict loan amounts?

Addressing these questions provides insight into the distinct mechanisms operating at the extensive and intensive margins of informal borrowing and helps clarify how household characteristics shape different dimensions of credit behavior.

Figure 5: Feature Importance Comparison

Feature Importance: Task-Specific Rankings Reveal Distinct Lending Dynamics



Source: Authors' calculations.

4.4.2 Classification: Interest Rates Signal Credit Access

In the classification task, informal loan interest rates emerge as the most important predictor, accounting for 20.86 percent of total feature importance—nearly twice the contribution of the second-ranked variable. This prominence indicates that interest rates serve as a key signal of both credit market conditions and borrower circumstances. Elevated informal interest rates likely reflect lenders' assessments of borrower risk while also capturing borrowers' willingness to accept unfavorable terms when alternative sources of credit are limited.

The correlation analysis in Figure 1 helps contextualize this result. Informal interest rates are negatively correlated with both income ($r = -0.29$) and access to formal credit ($r = -0.29$), suggesting segmentation between formal and informal credit markets. Low-income households not only face higher borrowing costs but also constitute a disproportionate share of informal credit users. In this sense, interest rates function as a summary statistic that captures unobserved aspects of financial vulnerability and market power, making them particularly informative for classification.

The second tier of predictors consists of family and personal expenses, which contribute 11.11 percent and 9.21 percent, respectively, and together account for more than 20 percent of predictive power. These variables highlight the role of immediate

financial pressure in driving informal borrowing decisions. Households facing expenditure burdens that exceed their current income are substantially more likely to seek informal credit, even when borrowing costs are high.

Demographic factors also play a meaningful role. Age ranks third, contributing 10.17 percent, consistent with life-cycle patterns in credit access and financial stability. Younger households often have limited savings and less stable employment, increasing their reliance on informal borrowing despite potentially longer repayment horizons. Informal loan term length (10.08 percent) and total income (9.14 percent) complete the top five predictors, indicating that both the structure of existing debt and current earning capacity influence the likelihood of informal borrowing.

4.4.3 Regression: Debt Terms Determine Borrowing Scale

The regression results reveal a distinct pattern relative to the classification task. Informal interest rates fall to third place (13.35 percent), while informal loan term length becomes the most important predictor, accounting for 18.44 percent of total importance. This shift highlights a key distinction between the extensive and intensive margins of informal borrowing: interest rates are more informative for identifying who borrows, whereas loan terms are more informative for determining how much is borrowed. Longer repayment periods reduce the required payment per period, enabling borrowers to take on larger loan amounts even when total interest costs are high.

Household spending remains highly influential in the regression setting. Family expenses account for 16.87 percent of predictive importance, closely mirroring their role in the classification task (11.11 percent). This consistency indicates that financial pressure affects both margins of borrowing. While high expenses increase the likelihood of entering the informal credit market, they also raise the amount borrowed once households decide to borrow, reflecting the need to finance ongoing obligations.

Total income ranks fourth in the regression analysis (12.42 percent), representing a modest increase relative to its classification importance (9.14 percent). This pattern suggests that income plays a larger role in determining borrowing capacity than in determining access. Lenders may implicitly use income as a benchmark for repayment ability when setting loan amounts, while borrowers may also self-select loan sizes that align with their perceived repayment capacity.

4.4.4 Interpreting Task-Specific Rankings: Extensive vs. Intensive Margins

The divergence between classification and regression importance rankings highlights distinct dynamics at the extensive and intensive margins of informal borrowing. At the extensive margin—whether a household borrows—informal interest rates are the most influential factor. Interest rates reflect both supply-side considerations, such as lenders' risk assessments and market power, and demand-side conditions, including borrower urgency and limited alternatives. The negative correlations between informal interest rates, income, and access to formal credit reinforce this interpretation: interest rates effectively summarize the disadvantage faced by financially vulnerable households.

In contrast, debt terms play a more prominent role at the intensive margin—how much a household borrows. Longer repayment periods directly expand borrowing capacity by lowering required payments per period, making larger loans feasible even when interest rates are high. Household expenses remain highly influential at both margins, underscoring the central role of financial stress in informal credit decisions. While expenses increase the likelihood of entering the informal credit market, they also raise the amount borrowed once participation occurs. The mechanisms differ, however: access depends largely on willingness or ability to accept prevailing loan terms, whereas

borrowing intensity depends on how repayment can be structured to match cash-flow constraints.

These task-specific patterns have direct implications for policy design. Policies aimed at improving credit access may focus on reducing interest rate disparities, for example, by expanding affordable formal credit options or mitigating excessive pricing in informal markets. By contrast, policies targeting over-indebtedness may be more effective if they address loan structure, such as regulating repayment periods or imposing income-based borrowing limits. In this sense, machine learning feature importance rankings do more than enhance predictive accuracy; they help identify which policy levers are most relevant for different objectives.

Figure 5 summarizes these findings: informal interest rates dominate classification performance, while debt terms are more important for regression outcomes. The consistently moderate-to-strong importance of household expenses across both tasks reinforces their central role in informal borrowing decisions. At the same time, the relatively low importance of formal credit variables—formal loan amounts and interest rates consistently rank below 6 percent—suggests that formal and informal credit markets operate largely in parallel. Borrowers appear to view them as imperfect substitutes rather than direct alternatives.

4.5 SHAP Analysis

4.5.1 SHAP Methodology

SHAP (Shapley Additive Explanations) decomposes individual model predictions into feature-specific contributions, providing both the magnitude and direction of each feature's effect. Unlike traditional Random Forest feature importance measures, which summarize average contributions across the sample, SHAP offers a unified framework that supports feature ranking, directional interpretation, individual-level explanations, and the identification of nonlinear relationships between predictors and outcomes.

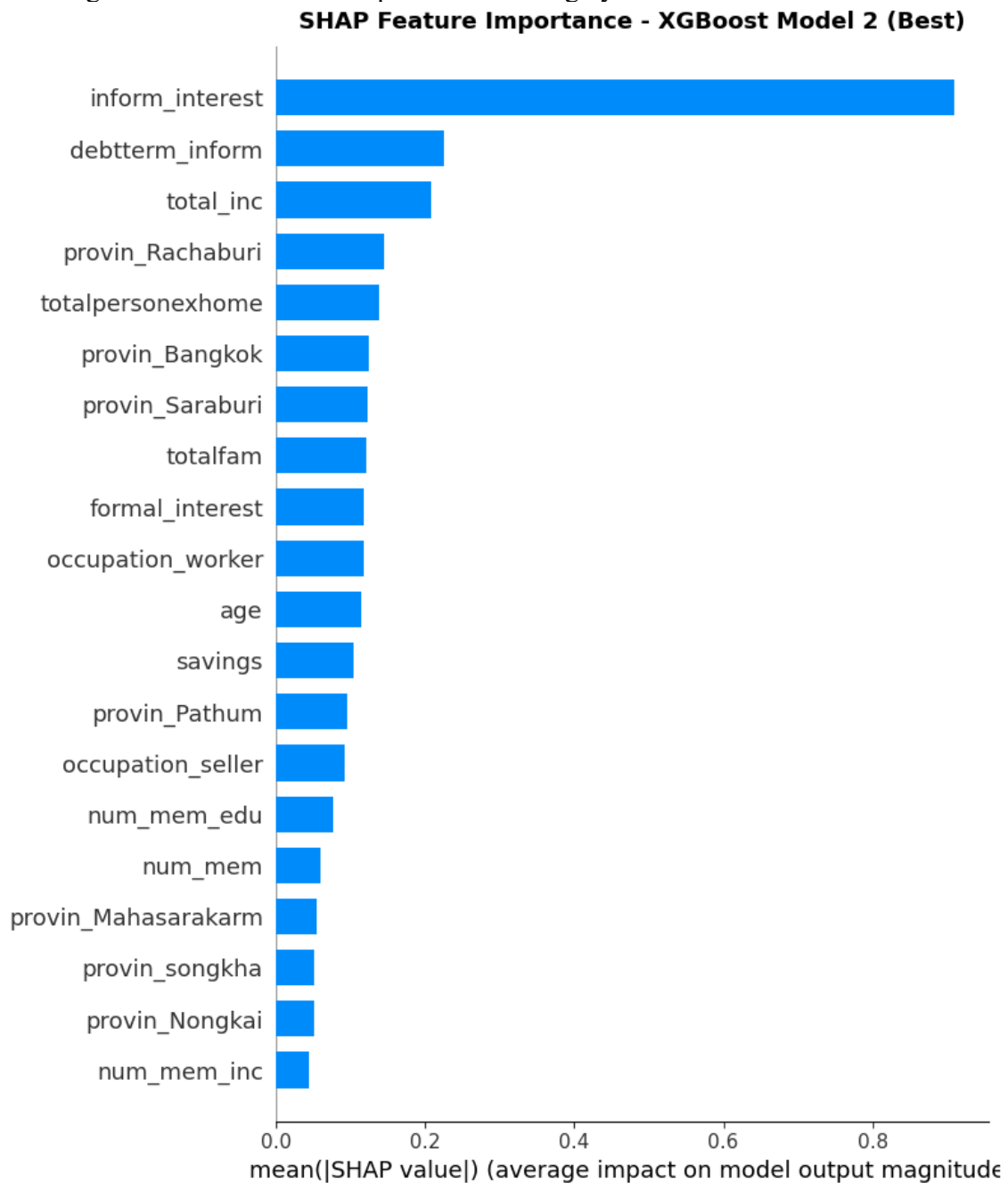
In this study, we apply SHAP to our best-performing classification model, XGBoost Model 2, which achieves an accuracy of 75.58 percent. SHAP values are computed using the TreeExplainer algorithm, which is specifically designed for tree-based models and efficiently computes exact Shapley values. We calculate SHAP values for 100 randomly selected observations from the test set (X_{test}), corresponding to approximately 20 percent of the test sample. The XGBoost classifier is trained with the following hyperparameters: learning rate of 0.1, 140 estimators, maximum depth of 5, minimum child weight of 2, subsample ratio of 0.8, and column subsample ratio of 0.8.

SHAP values measure how each feature shifts a prediction from the model's base value—the average predicted outcome—toward the final predicted value for a given observation. In the binary classification setting, SHAP values are interpreted as contributions to the log-odds of informal loan participation, where positive values increase the predicted probability of borrowing and negative values decrease it.

4.5.2 SHAP Feature Importance Ranking

Figure 6 presents the mean absolute SHAP values, which provide a clear ranking of features based on their overall contribution to the model's predictions. A striking pattern emerges: informal loan interest rate (`inform_interest`) is by far the most important feature, with a mean $|\text{SHAP}|$ value of 0.910—more than 3.6 times larger than that of the second-ranked variable, informal loan term (`debtterm_inform`, 0.245). This substantial gap indicates that, when observed, informal interest rate information is the single most informative indicator of informal loan participation in the classification model.

Figure 6: SHAP Feature Importance Ranking by Mean Absolute SHAP Value



Source: Authors' calculations.

Table 4: Top 10 Features by Mean Absolute SHAP Value in XGBoost Classifier (Model 2)

Rank	Feature	Mean SHAP	Variable Description
1	inform_interest	0.91	Informal loan interest rate (%)
2	debtterm_inform	0.245	Informal loan term (months)
3	total_inc	0.208	Total household income (Baht)
4	provin_Ratchaburi	0.145	Residence in Ratchaburi province (binary)
5	totalpersonexhome	0.138	Number of household members

Rank	Feature	Mean SHAP	Variable Description
6	provin_Bangkok	0.125	Residence in Bangkok (binary)
7	provin_Saraburi	0.124	Residence in Saraburi province (binary)
8	totalfam	0.121	Total family size
9	formal_interest	0.118	Formal loan interest rate (%)
10	occupation_worker	0.117	Worker occupation category (binary)

Source: Authors' calculations.

Table 4 shows the top ten features, ranked by mean absolute SHAP value. Beyond `inform_interest`, the remaining features exhibit more comparable levels of importance and span distinct categories of borrower characteristics. The next three features reflect different dimensions of borrowing behavior: `debtterm_inform` captures loan contract characteristics, `total_inc` reflects household economic capacity, and `provin_Ratchaburi` represents geographic variation. In addition, `formal_interest` appears among the top ten features, indicating that information from formal credit markets also contributes to classification performance.

Overall, these rankings suggest that households' borrowing behavior is shaped by a combination of informal loan conditions, household financial capacity, and geographic context. The presence of both formal and informal credit variables among the most important features further suggests that borrowers' financial portfolios often span multiple credit sources, a pattern reflected in the model's classification decisions.

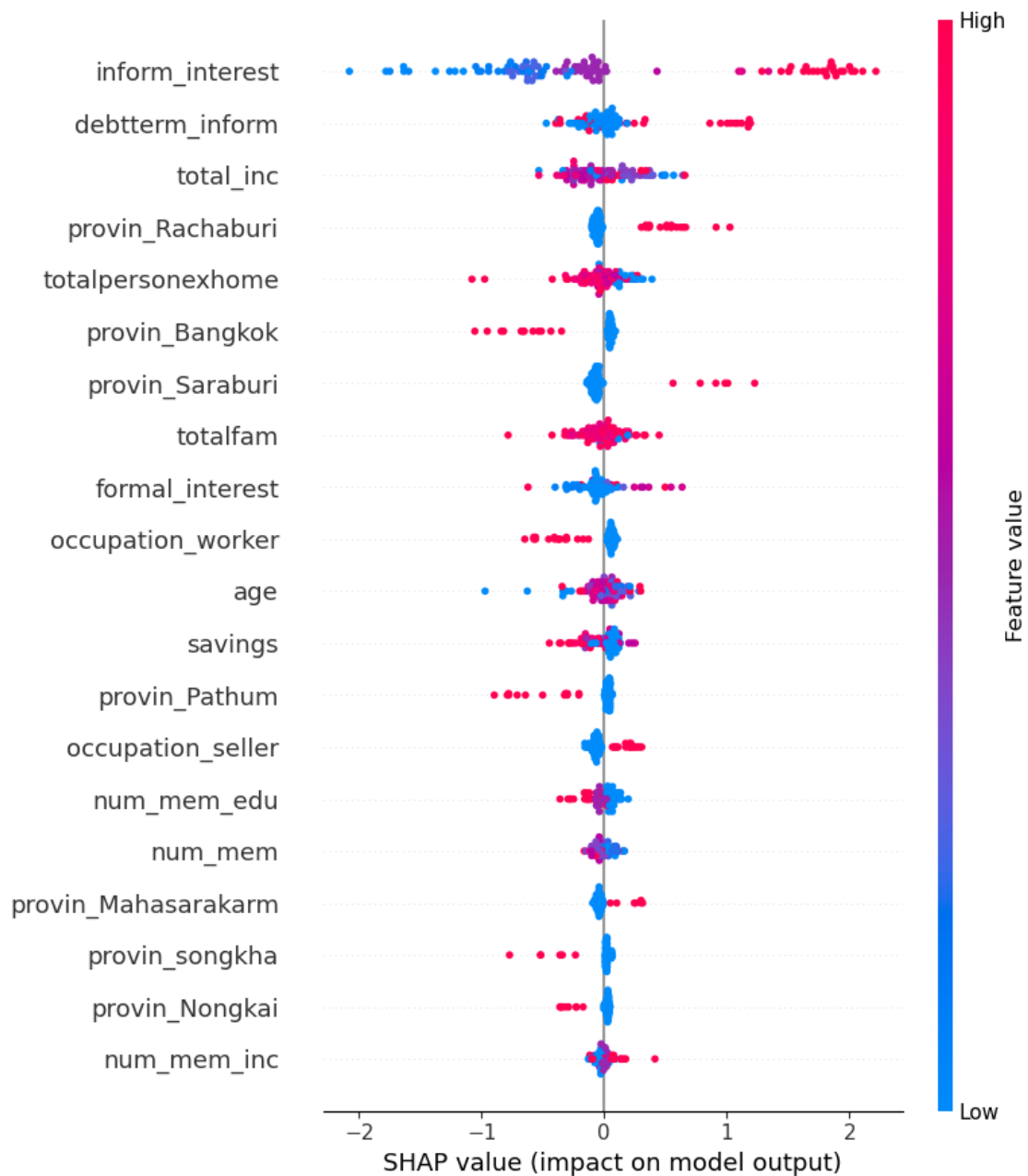
4.5.3 Feature Impact Patterns

Figure 7 presents a SHAP summary plot that combines feature importance rankings with information on the direction of each feature's effect on model predictions. Features are ordered by their mean absolute SHAP values, reflecting their overall influence on predicted outcomes. Each point corresponds to an individual observation, with colors indicating feature values (red for high values and blue for low values). The horizontal axis shows the direction and magnitude of each SHAP value, where positive values increase the predicted probability of informal borrowing and negative values decrease it.

The most pronounced pattern is observed for `inform_interest`. High values of informal loan interest rates (red points) are concentrated on the positive side of the distribution ($\text{SHAP} > 0$), while low or zero values (blue points) are concentrated on the negative side ($\text{SHAP} < 0$). This pattern has a straightforward interpretation: `inform_interest` functions as a diagnostic indicator, as only households with informal loans report paying informal interest rates. Observations with `inform_interest` equal to zero therefore correspond to non-borrowers, while positive values strongly signal participation in informal credit, resulting in positive SHAP contributions.

Other features exhibit more nuanced patterns. For `total_inc` (household income), higher values are associated with negative SHAP values, indicating that wealthier households are less likely to rely on informal credit markets. This finding is consistent with economic intuition, as households with higher income typically have better access to formal credit. Geographic variables display heterogeneous effects. For example, residence in Ratchaburi is associated with positive SHAP contributions, potentially reflecting regional differences in informal credit market conditions or local borrowing norms.

Figure 7: SHAP Feature Summary Plot
SHAP Summary Plot - XGBoost Model 2 (Best)
(Feature Impact on Informal Loan Prediction)



Source: Authors' calculations.

5. Policy Implications

This paper provides several policy-relevant insights into informal lending in Thailand by identifying the factors most strongly associated with both informal borrowing participation and borrowing intensity. By focusing on predictive patterns rather than causal claims, the analysis offers practical guidance for designing targeted interventions aimed at reducing reliance on high-cost informal credit and improving financial inclusion.

It is important to emphasize that the results are predictive rather than causal. The variables identified as influential—such as occupation, expenditures, savings, and loan

terms—should be interpreted as risk indicators useful for screening households that are more likely to rely on informal credit, not as direct causes of borrowing behavior. Accordingly, the policy implications focus on how these predictive signals can be translated into operational tools for targeting and early intervention.

A central finding is that models using a reduced set of variables—primarily income, expenditures, savings, and loan terms—achieve predictive performance comparable to models using the full set of covariates. This result is particularly relevant in data-scarce environments. Governments, financial institutions, and development agencies could design low-cost risk-screening tools based on a small number of survey questions or digital-finance indicators. Such tools could be used to flag households with high predicted probabilities of informal borrowing, enabling early intervention through targeted financial literacy programs, access to more flexible formal credit products, or the expansion of social safety nets. Importantly, this approach avoids overstating causality by treating observed characteristics as signals of vulnerable groups.

The findings also highlight the potential role of machine learning in the development of data-driven credit scoring systems. Traditional credit scoring models often exclude individuals in the informal sector due to the absence of collateral or formal financial records. By incorporating non-traditional indicators—such as household expenditures, income volatility, and occupation—predictive models can help financial institutions responsibly extend credit to underserved populations. For example, such models could be integrated into digital lending platforms to assess borrowing risk using inexpensive and readily available inputs.

Geographic heterogeneity in predicted informal borrowing further suggests scope for regionally targeted interventions. Certain provinces exhibit higher predicted probabilities of informal borrowing even after controlling for household characteristics. Policymakers can use these geographic signals to allocate resources more efficiently, for instance, by expanding microfinance initiatives or strengthening local financial support mechanisms in higher-risk areas.

The interpretability results also inform the design of financial literacy and awareness programs. Households with persistently high expenditures relative to income may benefit from programs emphasizing budgeting, savings accumulation, and awareness of the long-term costs of informal borrowing. Similarly, groups with unstable income sources—such as sellers and freelancers—may require tailored interventions that focus on income smoothing and risk management strategies appropriate to their employment conditions.

Finally, the predictive framework developed in this paper can complement regulatory efforts aimed at monitoring informal lending practices. By identifying the household-level conditions most strongly associated with informal borrowing, policymakers can better anticipate demand-side pressures that sustain high-cost informal credit markets. This demand-side perspective can enhance existing supply-side enforcement efforts and help prioritize protective measures for the most vulnerable populations.

Overall, the results demonstrate how interpretable machine learning models can serve as practical tools for policy targeting, enabling more efficient identification of households at risk of relying on informal credit and supporting the design of focused, data-informed interventions.

6. Conclusion

This study examines informal lending in Thailand using machine learning methods to shed light on a relatively underexplored segment of the credit market. Using individual-level survey data from approximately 4,800 respondents across 12 provinces and six regions, we develop predictive models to classify informal borrowing participation and to estimate informal loan amounts.

We compare the predictive performance of three machine learning approaches—K-Nearest Neighbors (KNN), Random Forest, and Gradient Boosting—across both tasks. The results indicate that ensemble methods outperform simpler alternatives. Random Forest delivers the most consistent performance across classification and regression tasks, reflecting its ability to accommodate nonlinear relationships and interactions among household characteristics. Gradient Boosting also performs well, particularly in classification settings with reduced feature sets, underscoring the robustness of ensemble approaches in data-constrained environments. Together, these findings highlight the value of machine learning not only for improving predictive accuracy but also for uncovering economically meaningful patterns in informal borrowing behavior.

Beyond performance comparisons, the analysis emphasizes interpretability. Feature importance rankings and SHAP-based explanations reveal systematic differences between the extensive margin of borrowing (whether households borrow informally) and the intensive margin (how much they borrow). Interest rates are most informative for identifying participation, while loan terms and household expenditures play a central role in determining borrowing intensity. These results illustrate how predictive models can provide insights into the structure of informal credit markets while remaining firmly within a non-causal framework.

The study has several limitations. First, the cross-sectional nature of the data restricts the analysis to a single point in time, preventing an examination of dynamic borrowing behavior. Second, reliance on self-reported survey responses may introduce measurement error. Future research could address these limitations by incorporating longitudinal data, linking survey responses to administrative records, or exploring additional behavioral and institutional factors that shape informal lending decisions.

Overall, this paper demonstrates how interpretable machine learning methods can complement traditional economic analysis in settings where data are limited and informal activity is difficult to observe. By providing accurate and transparent predictions of informal borrowing behavior, the approach offers a practical framework for researchers and policymakers seeking to better understand and address the persistence of informal credit markets.

Acknowledgments

We thank the editor, Juthathip Jongwanich, and the anonymous referees for their valuable comments and suggestions. We are grateful to the Department of Special Investigation (DSI), Ministry of Justice, Thailand, for generously providing the data used in this study. We also thank participants at the 18th Thailand's National Conference for Economists for helpful comments and discussions. This research was supported by the Puey Ungphakorn Institute for Economic Research. All remaining errors are our own.

References

- Chaengkham, S., & Wianwiwat, S. (2021). Stock market index prediction using machine learning: Evidence from leading Southeast Asian countries. *Thailand and the World Economy*, 39(2), 56–64.
- Chemin, M. (2008). The benefits and costs of microfinance: Evidence from Bangladesh. *Journal of Development Studies*, 44(4), 463–484.
- Dutt, P., & Tsetlin, I. (2020). Income distribution and economic development: Insights from machine learning. *Economics & Politics*, 33(1), 1–36.
- Fenz, K., Mitterling, T., Martinez, A. M. Jr., Bulan, J. A. N. M., Durante, R. L. S., Martillan, M. A., Addawe, M. B., & Roitner-Fransecky, I. (2024). Compiling granular population data using geospatial information. *Asian Development Review*, 41(1), 263–300.
- Guirkinger, C. (2008). Understanding the coexistence of formal and informal credit markets in Piura, Peru. *World Development*, 36(8), 1436–1452.
- Jeromi, P. D. (2007). Regulation of informal financial institutions: A study of money lenders in Kerala. Reserve Bank of India Occasional Papers, 28(1), 1–32.
- Kaboski, J., & Townsend, R. (2005). Policies and impact: An analysis of village-level microfinance institutions. *Journal of the European Economic Association*, 3(1), 1–50.
- Karaivanov, A., & Kessler, A. (2018). (Dis)advantages of informal loans – Theory and evidence. *European Economic Review*, 102, 100–128.
- Kislat, C. (2015). Why are informal loans still a big deal? Evidence from Northeast Thailand. *The Journal of Development Studies*, 51(5), 569–585.
- Kislat, C., Menkhoff, L., & Neuberger, D. (2017). Credit market structure and collateral in rural Thailand. *Economic Notes*, 46, 587–632.
- Klühs, T., Koch, M., & Stein, W. (2019). Don't expect too much: High income expectations and over-indebtedness. *Rationality and Competition Discussion Paper Series*, 200, CRC TRR 190 Rationality and Competition.
- Kusonkhum, W., Srinavin, K., Leungbootnak, N., & Chaitongrat, T. (2022). Using a machine learning approach to predict the Thailand underground train's passenger. *Journal of Advanced Transportation*, 2022.
- Liu, E., & Roth, B. (2022). Contractual restrictions and debt traps. *Review of Financial Studies*, 35(3), 1141–1182.
- Mathen, C. K., Chattopadhyay, S., Sahu, S., & Mukherjee, A. (2024). Which nighttime lights data better represent India's economic activities and regional inequality? *Asian Development Review*, 41(2), 193–217.
- Mohieldin, M., & Wright, P. W. (2000). Formal and informal credit markets in Egypt. *Economic Development and Cultural Change*, 48(3), 657–670.
- Pinitjitsamut, P., & Suwanprasert, W. (2022). *Informal loans in Thailand: Stylized facts and empirical analysis* (Discussion Paper No. 173). Puey Ungphakorn Institute for Economic Research. Retrieved from <https://www.pier.or.th/dp/173/>
- Pitt, M., & Khandker, S. (1998). The impact of group-based credit programs on poor households in Bangladesh: Does the gender of participants matter? *Journal of Political Economy*, 106(5), 958–996.
- Puttanapong, N., Prasertsoong, N., & Peechapat, W. (2023). Predicting provincial gross domestic product using satellite data and machine learning methods: A case study of Thailand. *Asian Development Review*, 40(2), 39–85.

- Siamwalla, A., Pinthong, C., Poapongsakorn, N., Satsanguan, P., Nettayarak, P., Mingmaneeenakin, W., & Tubpun, Y. (1990). The Thai rural credit system: Public subsidies, private information, and segmented markets. *The World Bank Economic Review*, 4(3), 271–295.
- Suwanprasert, W. (2025). The gender gaps in informal loan markets. *Applied Economics*, 1-18.
- Tanomchat, W., & Sampattavanija, S. (2018). Dependence of informal interest rates and level of lenders' influence in the informal loan market in Thailand. *International Advances in Economic Research*, 24(1), 47–63.