Analysis of the debt status of households in poor areas based on economic capital using two-class boosted decision tree

Pita Jarupunphol and Wipawan Buathong

Department of Digital Technology,
Phuket Rajabhat University,
Phuket, Thailand
Email: p.jarupunphol@pkru.ac.th
Email: w.buathong@pkru.ac.th

Suthasinee Kuptabut*

Department of Computer,
Sakon Nakhon Rajabhat University,
Sakon Nakhon, Thailand
Email: suthasinee@snru.ac.th
*Corresponding author

Abstract: This study examines household debt determinants in Kut Bak district, Thailand, using a two-class boosted decision tree (TBDT) model to analyse 301 households across 30 financial, asset, and socio-economic variables. Compared with logistic regression, decision tree, random forest, and XGBoost, the model demonstrates superior performance, achieving an accuracy of 0.922, precision of 0.975, recall of 0.867, F1-score of 0.918, and AUC of 0.948. Key findings reveal that limited savings, minimal state assistance, and low ownership of productive assets significantly increase debt likelihood. Specific thresholds, such as savings below 4,500 units and cash reserves of 50 units or less, are strongly associated with indebtedness. The study highlights the model's effectiveness in predicting debt status and provides actionable insights for policymakers and organisations to enhance financial stability in rural communities. These results contribute to understanding socio-economic factors driving household debt in disadvantaged areas.

Keywords: data mining; household debt; machine learning; socio-economic factors; two-class boosted decision tree; TBDT.

Reference to this paper should be made as follows: Jarupunphol, P., Buathong, W. and Kuptabut, S. (2026) 'Analysis of the debt status of households in poor areas based on economic capital using two-class boosted decision tree', *Int. J. Data Mining, Modelling and Management*, Vol. 18, No. 2, pp.158–179.

Copyright © 2026 Inderscience Enterprises Ltd.

Biographical notes: Pita Jarupunphol earned his PhD in Computer Science from the University of Auckland, New Zealand. His research areas encompass data mining, machine learning, cybersecurity, human-computer interactions, and technology acceptance. Currently, he serves as a Lecturer in the Department of Digital Technology at Phuket Rajabhat University, Thailand.

Wipawan Buathong obtained her PhD in Information Technology from King Mongkut's University of Technology North Bangkok, Thailand. Her research applies data science, data mining, and machine learning across various fields. She is currently a Lecturer in the Department of Digital Technology at Phuket Rajabhat University, Thailand.

Suthasinee Kuptabut earned her PhD in Information Technology from King Mongkut's University of Technology Ladkrabang, Thailand. Her research focuses on the extensive applications of data analysis techniques in various disciplines. She currently serves as a Lecturer in the Department of Computer at Sakon Nakhon Rajabhat University, Thailand.

1 Introduction

In socio-economic research, understanding household debt dynamics in economically disadvantaged areas is crucial for developing effective policies and interventions. This study focuses on Kut Bak district, in Sakon Nakhon Province, Thailand, a region recognised for its economic challenges. The primary objective is to analyse the debt status of households within this area, which is pivotal in understanding these communities' broader economic health and resilience. Financial capital, in its various forms, plays a significant role in shaping the livelihoods of households in poor areas (Narongchai et al., 2016; Shah et al., 2021; Yalley et al., 2021). In rural economies like that of Kut Bak district, factors such as agricultural tools and vehicles, bank accounts, and savings patterns are integral to the economic well-being of families. This research aims to dissect these components to unveil a comprehensive picture of the economic status of these households, with a particular emphasis on their debt status.

To address the challenge of identifying household debt status in economically disadvantaged areas, various machine learning techniques can be employed to analyse complex relationships between financial capital and debt patterns (Alpaydin, 2020). Logistic regression offers a straightforward probabilistic approach (Ağlarcı and Bal, 2024), while decision tree provides an interpretable structure for capturing nonlinear relationships (Yuan and Tang, 2025). Random forest enhances decision tree by reducing overfitting through ensemble learning (Niu et al., 2020), and XGBoost further improves performance with its gradient-boosting framework and regularisation capabilities (Abbasimehr et al., 2023). Among these methods, two-class boosted decision tree (TBDT) also stands out as a robust choice due to its ability to sequentially refine predictions and uncover intricate patterns in the data. Given that the debt status of households in Kut Bak district is significantly shaped by a combination of financial capital indicators, including savings patterns, ownership of productive assets (e.g., agricultural tools and vehicles), and access to formal banking services, we posit that TBDT represents a compelling and suitable methodological choice for this research.

This article utilises TBDT, a robust data mining technique (Talia et al., 2016), to identify patterns within a dataset of 301 households. This technique is adept at uncovering hidden relationships between various economic factors and their influence on households' propensity to be in debt. By examining 30 distinct characteristics, ranging from banking habits to the possession of agricultural assets, the article intends to provide insights into the factors contributing to or alleviating the burden of debt in these households. Understanding the debt status of households in poor areas is not just an academic exercise, it has profound implications for policy-making, financial assistance programs, and overall economic development strategies. By identifying key factors that can influence household debt in the district, this research has the potential to contribute valuable knowledge to the field and support efforts to promote economic resilience in disadvantaged communities.

1.1 Objectives

The main objective of the research is to analyse households' debt status in the Kut Bak district, using the TBDT algorithm to identify and understand the factors influencing whether households are in debt. The study also aims to uncover patterns and determinants that contribute to or mitigate the likelihood of household debt, providing information that can inform policy making, financial advice, and community support initiatives to improve economic stability and reduce debt among vulnerable populations.

1.2 Hypotheses

- TBDT will be highly effective compared to other machine learning methods in identifying key predictors of debt and revealing nuanced relationships within the dataset, thereby offering deeper and more actionable insights for understanding household debt dynamics.
- Specific socio-economic factors, financial behaviours, and asset ownership patterns significantly influence the likelihood of households in Kut Bak district being in debt.
- Households with limited financial resources, minimal savings in agricultural and cooperative banks, lower levels of state-sponsored financial assistance, and minimal ownership of productive assets (such as small tractors and lawn mowers), coupled with low cash reserves, are more likely to be in debt.

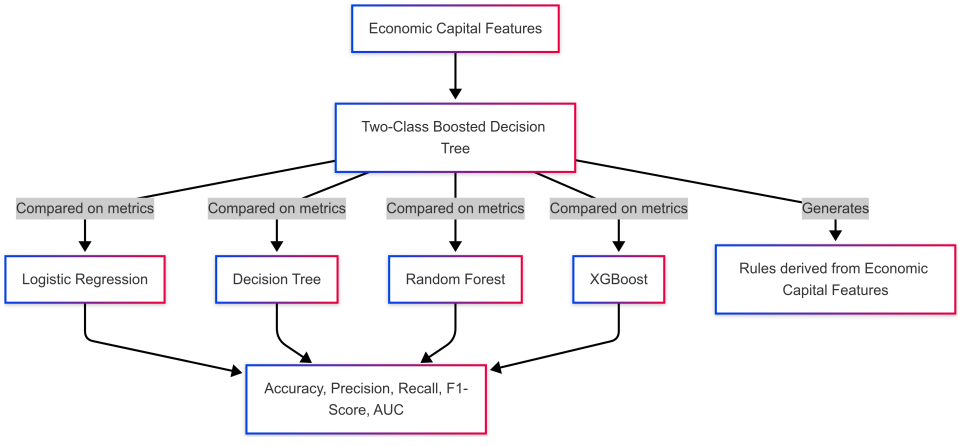
1.3 Conceptual framework

Figure 1 illustrates the conceptual framework of the research, providing a visual representation of the theoretical constructs, relationships, and variables under investigation. This framework outlines the characteristics of economic capital, such as socio-economic factors, financial behaviours, and asset ownership patterns, which are hypothesised to influence the debt status of households in Kut Bak district. It serves as a schematic guide to understand the study's approach to examining how these variables affect the likelihood of households being in debt. The framework also highlights the flow of analysis, where features of economic capital are input into the TBDT model

to identify significant predictors and generate interpretable rules derived from these features.

To evaluate the effectiveness of TBDT, it is compared with other widely-used machine learning techniques (Alpaydin, 2020) – logistic regression, decision trees, random forests, and XGBoost – across key performance metrics, including accuracy, precision, recall, F1-score, and area under the curve (AUC). This comparative analysis ensures the robustness of the chosen model and provides insights into its relative strengths in capturing complex relationships within the dataset. Furthermore, TBDT generates actionable rules based on the economic capital features, offering valuable interpretability and practical implications for policymakers and stakeholders aiming to address household debt challenges in economically disadvantaged areas. This analytical process sets the foundation for the empirical analysis conducted in this study.

Figure 1 Conceptual framework of the research (see online version for colours)



2 Literature review

The literature on household debt, particularly in poor areas, offers a rich array of insights and findings. This review aims to contextualise our study within the broader scholarly discourse, focusing on economic capital, debt patterns, and the use of TBDT in socio-economic research.

2.1 Household debt in rural economies

A considerable body of research has been dedicated to examining the determinants and implications of household debt within rural contexts, recognising its complex and multidimensional nature. The extant literature underscores the distinct structural and socio-economic factors that contribute to indebtedness in rural households, notably agricultural dependency, limited income diversification, and constrained access to formal financial institutions (Aldashev and Batkeyev, 2023; Shah et al., 2021; Xie et al., 2024). In agrarian economies, household financial stability is often highly volatile, as income sources fluctuate by seasonal agricultural cycles, leading to periodic liquidity

shortages and heightened vulnerability to external shocks. In addition, the predominance of informal credit mechanisms, including reliance on community lending, cooperatives, or high-interest private loans, exacerbates long-term financial instability for rural populations. Scholarly investigations have further demonstrated how external economic shocks, such as market fluctuations, extreme weather events, and policy changes, intensify financial distress and perpetuate cycles of debt, particularly in economically fragile regions. In particular, Aldashev and Batkeyev (2023) provide compelling evidence that the absence of financial buffers and exposure to unpredictable external forces significantly increases household debt burdens in rural economies. Similarly, Shah et al. (2021) highlight the role of poverty traps, wherein limited access to productive assets and financial resources restricts households' ability to escape recurring cycles of indebtedness.

Moreover, Xie et al. (2024) take a different approach by examining the role of agricultural insurance participation and its effectiveness in rural settings, questioning whether it alleviates or compounds the debt burden for rural households. Their findings suggest a complex relationship, where agricultural insurance can both provide the necessary capital for agricultural investments and yet, due to its high-interest rates and repayment structures, potentially lead to a cycle of indebtedness. In addition, Strzelecka et al. (2020) identified and evaluated socio-economic factors determining the debt of households in Central Pomerania using a logistic regression model. The data source was a survey among 1,000 households in Central Pomerania, Poland. The results showed that the following factors related to the socio-economic characteristics of households: economic education of the head of the household, developmental phase in the household, and socio-economic type of the household had a statistically significant positive impact on the probability that central Pomeranian households would use external sources of financing.

2.2 Economic capital and its influence

The role of economic capital in shaping household debt has been extensively studied (Evans and Gregson, 2023; Narongchai et al., 2016; Ono and Uchida, 2018). Evans and Gregson (2023) have shown how assets and income levels correlate with debt status. The authors analysed the reasons behind the historical exclusion of monetary considerations from the sociology of consumption studies. The authors re-examined data from prior research on UK household consumption through the perspective of monetary transactions and financial considerations. In addition, Narongchai et al. (2016) investigated the evolution of intergenerational transfers of economic capital in rural households of Northeastern Thailand, particularly in a community with a high elderly population. The research was setup in an agricultural area where rice and other crops are predominantly grown. The findings revealed that the social and economic landscape of the community had undergone significant changes due to economic development.

Moreover, Ono and Uchida (2018) considered the politics of public education policy in an overlapping-generations model with physical and human capital accumulation. The authors examined how debt and tax financing differ in growth and welfare across generations. The results showed that the growth rate in debt financing is lower than in tax financing and that debt financing creates a tradeoff between the present and future generations. These studies offer a foundational understanding of how various forms of

economic capital, from agricultural tools to bank accounts, impact the financial health of households.

2.3 Machine learning and economic capital

Machine learning models play a crucial role in analysing economic capital by uncovering complex relationships between financial behaviours, socio-economic variables, and household debt status. The selection of an appropriate model depends on the specific requirements of the analysis, such as the need for interpretability, robustness to nonlinearity, or the ability to handle high-dimensional data (Alpaydin, 2020). For instance, logistic regression offers a straightforward and interpretable approach, making it helpful in understanding the influence of individual economic factors (Ağlarcı and Bal, 2024; Strzelecka et al., 2020). Decision tree provides a simple, hierarchical structure for classification, but may lack generalisation when dealing with more complex patterns (Yuan and Tang, 2025). In contrast, ensemble methods such as random forest (Niu et al., 2020; Park et al., 2022), two-class boost decision tree, and XGBoost significantly improve robustness by aggregating multiple decision trees to enhance accuracy and reduce overfitting (Abbasimehr et al., 2023; An et al., 2021; Ríos Canales, 2016; Xu et al., 2024). These ensemble techniques are particularly effective in capturing nonlinear relationships and interactions among economic variables, making them valuable tools for evaluating financial stability and household debt risks. By leveraging these models, researchers can gain deeper insights into the economic factors influencing debt status, ultimately contributing to better financial decision-making and policy development. Research has demonstrated the efficacy of machine learning models in their transformative potential in multiple fields, allowing data-driven insights and informed decision-making (Abbasimehr et al., 2023; Ağlarcı and Bal, 2024; Niu et al., 2020; Yuan and Tang, 2025). The following equations illustrate the operational mechanisms of logistic regression, decision tree, random forest, and XGBoost.

2.3.1 Logistic regression

Logistic regression is a statistical method used for binary classification problems, where the goal is to predict the probability that a given input belongs to one of two classes. This technique models the relationship between the independent variables and the probability of the outcome using a logistic (sigmoid) function, which maps the input to a value between 0 and 1. The model estimates the parameters through the maximum likelihood estimation, allowing it to determine the most likely class label based on the input features $x = (x_1, x_2, ..., x_p)$ belonging to class 1 represented in Equation 1.

$$\hat{p}(y = 1 \mid x) = \frac{1}{1 + \exp\Big(-\Big(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\Big)\Big)}. \tag{1}$$

The *log-odds* (or logit) is given in equation (2), where a common decision rule is to classify an instance as class 1 if $\hat{p}(y = 1 \mid x) > 0.5$.

$$\log\left(\frac{\hat{p}(y = 1 \mid x)}{1 - \hat{p}(y = 1 \mid x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{2}$$

2.3.2 Decision tree

A decision tree is a supervised learning algorithm used for classification and regression tasks, where data are split into subsets based on specific characteristics, forming a tree-like structure. Each internal node represents a decision based on a feature, each branch represents the outcome of that decision, and each leaf node represents a final class label or output value. The model recursively partitions the data to maximise the homogeneity of the resulting subsets, making it interpretable and practical for capturing nonlinear relationships. Equation (3) explains the prediction function for a decision tree.

$$\hat{y} = \sum_{m=1}^{M} c_m \, \mathbb{I}(x \in R_m) \tag{3}$$

where

- $\{R_1, R_2, ..., R_M\}$ are the regions defined by the tree's splits
- c_m is the constant prediction (e.g., the mean response) in region R_m
- $\mathbb{I}(\cdot)$ is the indicator function.

For classification, a standard formulation is illustrated in equation (4), where p_m is typically the proportion of training samples in region R_m that belong to class 1.

$$\hat{p}(y = 1 \mid x) = \sum_{m=1}^{M} p_m \, \mathbb{I}(x \in R_m) \tag{4}$$

2.3.3 Random forest

Random forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks or the mean prediction for regression tasks. Each tree is built using a random subset of the data (bootstrap sample) and a random subset of features at each split, reducing overfitting and increasing model robustness. A random forest aggregates the predictions of multiple decision trees. For regression, equation (5) describes that the prediction is the average of the predictions from B trees, where $T_b(x)$ is the prediction from the b^{th} tree.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{5}$$

In many cases, the forest computes the average probability, and the final class is determined by thresholding or majority vote as explained in equation (6).

$$\hat{p}(y = 1 \mid x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{6}$$

2.3.4 XGBoost

Extreme gradient boosting (XGBoost) is an advanced ensemble learning algorithm that builds decision trees sequentially, where each subsequent tree corrects the errors of the previous ones by minimising a regularised objective function. It employs gradient boosting with enhancements such as handling missing data, optimising split points, and incorporating L1/L2 regularisation to improve accuracy and reduce overfitting. Known for its efficiency, scalability, and high performance, XGBoost is widely used in machine learning competitions and real-world applications for classification and regression tasks. XGBoost builds an ensemble of trees in an additive manner. For an instance x_i, the prediction is represented in equation (7), where $\mathcal{F}$ is the space of regression trees and K is the number of boosting rounds.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \;\; f_k \in \mathcal{F} \tag{7}$$

The objective function minimised during training is described in equation (8).

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l\big(y_i, \hat{y}_i\big) + \sum_{k=1}^{K} \Omega(f_k) \tag{8}$$

XGBoost incorporates regularisation to control the complexity of the model and prevent overfitting, ensuring better generalisation on unseen data. This is achieved by adding a penalty term to the objective function, discouraging overly complex trees. Equation (9) illustrates a typical regularisation term.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{9}$$

where

- T is the number of leaves in the tree
- w is the vector of leaf scores
- γ and λ are regularisation parameters.

For binary classification, the score $\hat{y}_i$ is converted to a probability using the sigmoid function as illustrated in equation (10).

$$\hat{p}(y = 1 \mid x_i) = \frac{1}{1 + \exp\Big(-\hat{y}_i\Big)} \tag{10}$$

2.3.5 Two-class boosted decision tree

TBDT is an ensemble learning method that combines multiple weak decision trees to improve predictive accuracy for binary classification tasks. Each tree is sequentially trained to correct the errors made by the previous ones, with a focus on misclassified instances, using a boosting algorithm such as AdaBoost or gradient boosting. By

aggregating the predictions of these trees, the model achieves higher performance and robustness compared to individual trees, effectively capturing complex patterns in the data while maintaining interpretability (An et al., 2021; Hassan and Saeed, 2023; Ríos Canales, 2016). Equation (11) shows the predicted score for an input x is given by:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{11}$$

where

- M is the total number of decision trees
- $h_m(x)$ is the prediction of the m^{th} weak decision tree
- γ_m is the weight (learning rate-adjusted contribution) of tree h_m.

To convert this score into a probability for class $y \in \{0, 1\}$, the sigmoid transformation is applied in equation (12).

$$P(y = 1|x) = \frac{1}{1 + e^{-F(x)}} \tag{12}$$

Equation (13) presents that the loss function typically used in binary classification with boosting is the log loss, where N is the number of training examples.

$$L = -\sum_{i=1}^{N} [y_i \log P(y_i|x_i) + (1 - y_i) \log(1 - P(y_i|x_i))] \tag{13}$$

This technique is widely used in applications requiring binary classification, such as spam detection, customer churn prediction, fraud detection, and medical diagnoses where decisions are typically binary (e.g., disease/no disease). It combines multiple decision trees through boosting, an ensemble learning method. The algorithm starts with a single base decision tree to make initial predictions. These predictions are used to evaluate the accuracy and identify instances in which the models predictions were incorrect. In subsequent iterations, new decision trees are created, each focusing on correcting the errors made by the previous trees in the ensemble. Each tree is trained on the dataset, with increased emphasis on the instances that were misclassified by previous trees.

After each tree is added, the algorithm adjusts the weights of the instances based on the accuracy of the prediction. Misclassified instances are given higher weights, making them more likely to be correctly predicted in the next round. The process continues until a predetermined number of trees have been added or no further improvements can be made to predict accuracy. By aggregating the predictions of multiple trees, the TBDT model often achieves higher accuracy than individual decision trees. This model can handle various data types without extensive pre-processing, including numerical and categorical variables. In addition, it is less prone to overfitting than individual decision trees, especially when the number of trees is correctly tuned.

3 Methodology

This research used a mixed methods approach to analyse the debt status of households in the Kut Bak district, focusing on the relationship between economic capital and debt. The study combined quantitative data analysis using TBDT to provide a holistic understanding of the factors influencing household debt.

3.1 Data collection

The study concentrated on a selected sample of 301 households in Kut Bak district, a region characterised by its unique socio-economic dynamics. To gather comprehensive data, researchers employed a structured questionnaire designed to delve into a broad spectrum of economic attributes. The dataset comprises a total of 30 columns. Among these, 14 columns provide detailed information about individuals' agricultural and work-related equipment or vehicles. This includes tractors, farming tools, and other machinery crucial for their vocational activities. The remaining 16 columns focus extensively on financial aspects, encompassing variables such as available cash, bank deposits, and additional financial assistance sources. These financial columns offer a comprehensive overview of the individual's economic status and resources. The primary focus of this research is encapsulated in one key column labelled 'in debt', which categorises respondents based on their debt status with a simple 'yes' or 'no' response. This column serves as the target variable for the study. The aim is to analyse and understand how the various equipment, vehicle ownership, and diverse financial metrics correlate with and potentially influence an individual's likelihood of debt.

3.2 Ethical considerations

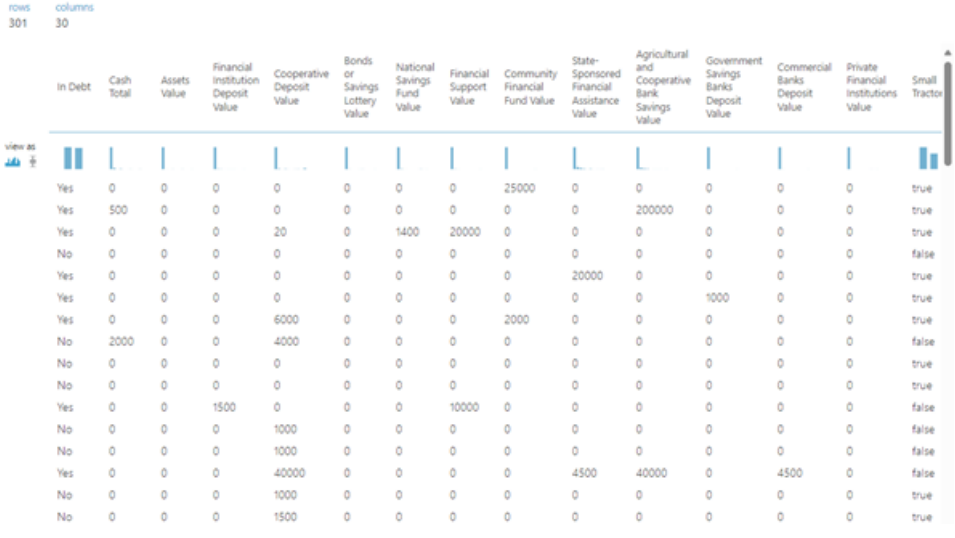
Data was gathered through field surveys and interviews, conducted by a team trained in data collection techniques, ethical practices, and ensuring consistency and reliability in their approach. These team members, fully briefed on the study's goals, personally visited each participant, carefully observing their living environments to collect detailed information. The research strictly adhered to ethical standards, having been granted clearance by the ethics committee of Sakon Nakhon Rajabhat University (full board review) under the approval number HE 65-099, which was valid from 31 August 2022 to 31 August 2023. Throughout the interview process, it was paramount to maintaining the participants' confidentiality and being sensitive to their situations.

3.3 Data preparation

We first addressed missing data by evaluating the nature of the missingness and applying appropriate techniques such as imputation for continuous variables and mode substitution for categorical variables, ensuring no loss of vital information (Kalita et al., 2024). Continuous variables were normalised to ensure they were on a similar scale, facilitating better model convergence. In preparing the dataset, we addressed missing values using imputation methods suited to variable type, aiming to minimise potential bias in the analysis. For numerical variables, mean or median imputation was applied based on the variable's distribution, ensuring that the central tendency of the data

was preserved without significantly altering its variability. We used mode imputation for categorical variables, replacing missing values with the most frequently occurring category to maintain category representation and mitigate the risk of introducing bias.

Figure 2 The dataset after the preparation phase (see online version for colours)



rows 301 columns 30

In Debt	Cash Total	Assets Value	Financial Institution Deposit Value	Cooperative Deposit Value	Bonds or Savings Lottery Value	National Savings Fund Value	Financial Support Value	Community Financial Fund Value	State-Sponsored Financial Assistance Value	Agricultural and Cooperative Bank Savings Value	Government Savings Banks Deposit Value	Commercial Banks Deposit Value	Private Financial Institutions Value	Small Tractor
Yes	0	0	0	0	0	0	0	25000	0	0	0	0	0	true
Yes	500	0	0	0	0	0	0	0	0	200000	0	0	0	true
Yes	0	0	0	20	0	1400	20000	0	0	0	0	0	0	true
No	0	0	0	0	0	0	0	0	0	0	0	0	0	false
Yes	0	0	0	0	0	0	0	0	20000	0	0	0	0	true
Yes	0	0	0	0	0	0	0	0	0	0	1000	0	0	true
Yes	0	0	0	6000	0	0	0	2000	0	0	0	0	0	true
No	2000	0	0	4000	0	0	0	0	0	0	0	0	0	false
No	0	0	0	0	0	0	0	0	0	0	0	0	0	true
No	0	0	0	0	0	0	0	0	0	0	0	0	0	true
Yes	0	0	1500	0	0	0	10000	0	0	0	0	0	0	false
No	0	0	0	1000	0	0	0	0	0	0	0	0	0	false
No	0	0	0	1000	0	0	0	0	0	0	0	0	0	false
Yes	0	0	0	40000	0	0	0	0	4500	40000	0	4500	0	false
No	0	0	0	1000	0	0	0	0	0	0	0	0	0	true
No	0	0	0	1500	0	0	0	0	0	0	0	0	0	true

In addition, we have attached a comprehensive list of all variables used in this study, detailing the meaning and relevance of each variable to the analysis of household debt. This list includes variables related to financial status, asset ownership, and socio-economic factors, which are critical in understanding debt dynamics in rural households. By providing this appendix, we aim to enhance the transparency and interpretability of our model and findings. In this study, we employed the TBDT model, selected for its proven strengths in managing complex, nonlinear relationships and effectively handling imbalanced class distributions. These features make it particularly well-suited to the socio-economic dataset under analysis. The model's ensemble learning approach iteratively refines predictions, enabling it to uncover intricate patterns and interactions among variables, which are critical for a comprehensive understanding of household debt dynamics.

To broaden the perspective, we intend to compare the performance of the TBDT model against other widely used classification algorithms, including logistic regression, support vector machines, and random forests. This comparative analysis will assess key performance metrics such as accuracy, precision, recall, and F1-score to contextualise the results and highlight the model's relative effectiveness. The study used Microsoft Machine Learning Studio on a personal computer with a Core i5-13400F processor operating at 2.50 GHz and 16 GB of RAM. Providing this information ensures transparency regarding the computational resources utilised, which is essential for reproducibility and evaluating the scalability of the methodology. Figure 2 depicts the dataset in its prepared state, ready for training.

3.4 Model training

Following the comprehensive data preparation process, the model training phase is critical in applying the TBDT algorithm to analyse household debt status in Kut Bak district. The dataset contains 301 entries, each with 30 unique attributes. These entries are divided into two categories based on their debt status: 'yes' signifies individuals in debt, while 'no' indicates individuals not in debt. In this case, the TBDT algorithm is selected from the available machine learning models list. This algorithm is chosen for its ability to handle binary classification tasks, such as distinguishing between households in debt (yes) and those not (no). The dataset was first split into training and test sets, typically following a 70:30 ratio. The training set was then used for parameter optimisation of the TBDT model to enhance predictive accuracy. The test set remained separate and was only used to evaluate the final model's performance, preventing any overfitting from exposure to test data during the training phase.

3.5 Model evaluation

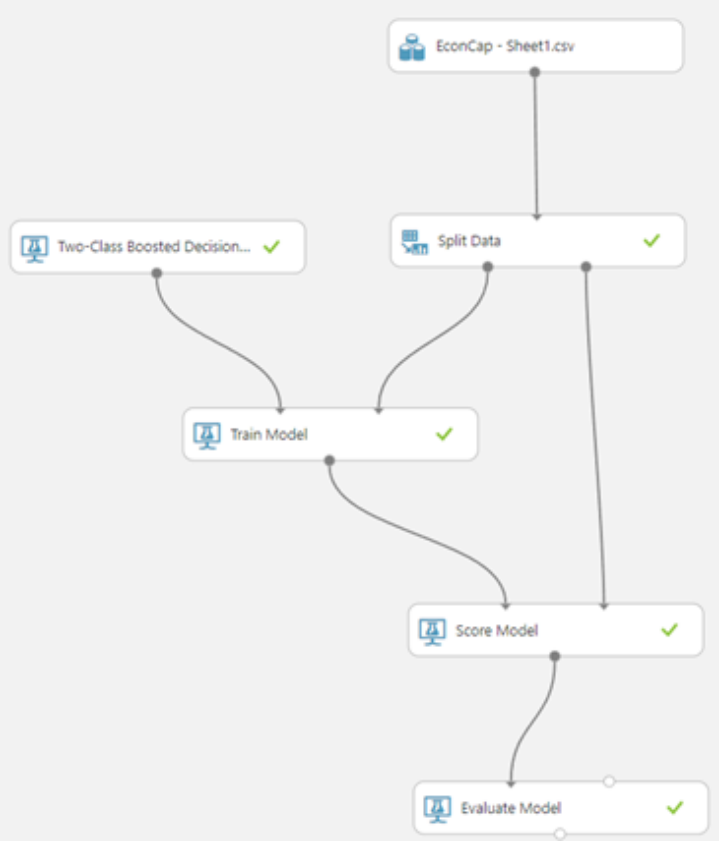
Once training is complete, the model performance is evaluated using evaluation metrics (Dalianis, 2018). Key metrics such as accuracy, precision, recall, and the area under the ROC curve are examined to assess how well the model can predict the status of household debt. Based on the performance metrics, the model parameters may be adjusted and the model re-trained to further enhance its predictive accuracy. This iterative process continues until the desired performance level is achieved. The best-performing model configuration is selected as the final model. This model embodies the optimal balance between accuracy and generalisability to predict the status of household debt in the Kut Bak district. In addition, the effectiveness of the TBDT model will be evaluated by comparing its performance with other competitive models, including logistic regression, decision tree, random forest, and XGBoost, to determine its efficiency in predicting household debt status. This comparative analysis ensures that the TBDT model is assessed against widely used classification techniques, highlighting its strengths and limitations in capturing complex socio-economic patterns influencing household indebtedness.

3.6 Data analysis

The model's findings are explored to uncover patterns and relationships within the data. This involves examining the rules and conditions under which households are classified as being in debt. Such analysis might reveal, for example, that households owning certain types of agricultural machinery are more likely to be in debt, or that those with specific banking habits have a lower incidence of debt. The analysis generates insights into the socio-economic dynamics influencing debt status in Kut Bak district. These insights lead to the formulation of hypotheses regarding the causative factors of debt, which can be tested in future research or used to inform policy-making and intervention strategies. The findings from the data analysis phase are visualised using charts, graphs, and tables to effectively communicate the relationships and patterns identified. Figure 3 illustrates the data mining process, encompassing the stages from data splitting through to model evaluation and utilising the TBDT algorithm. This process includes dividing

the dataset into two portions: 70% for training and 30% for testing. The trained model was then applied to the remaining 30% of the data, designated as the testing set, to generate performance scores. The evaluation phase followed, where the predictive accuracy and effectiveness of the model were systematically assessed.

Figure 3 The data mining process using the TBDT algorithm (see online version for colours)



4 Results and discussion

Figure 4 represents the evaluation of the model's performance, presenting metrics such as accuracy, precision, recall, F1-score, and AUC. The metrics indicate the model's effectiveness in classifying households within Kut Bak district into 'in debt = yes' and 'in debt = no' categories, based on the TBDT algorithm. The key performance indicators are detailed as follows:

1 Accuracy (0.922): This metric, standing at 92.2%, signifies the model's overall correctness in identifying both 'in debt = yes' and 'in debt = no' statuses across all predictions. It highlights the model's strong generalisation ability, indicating that it can correctly predict the debt status of households in most cases.

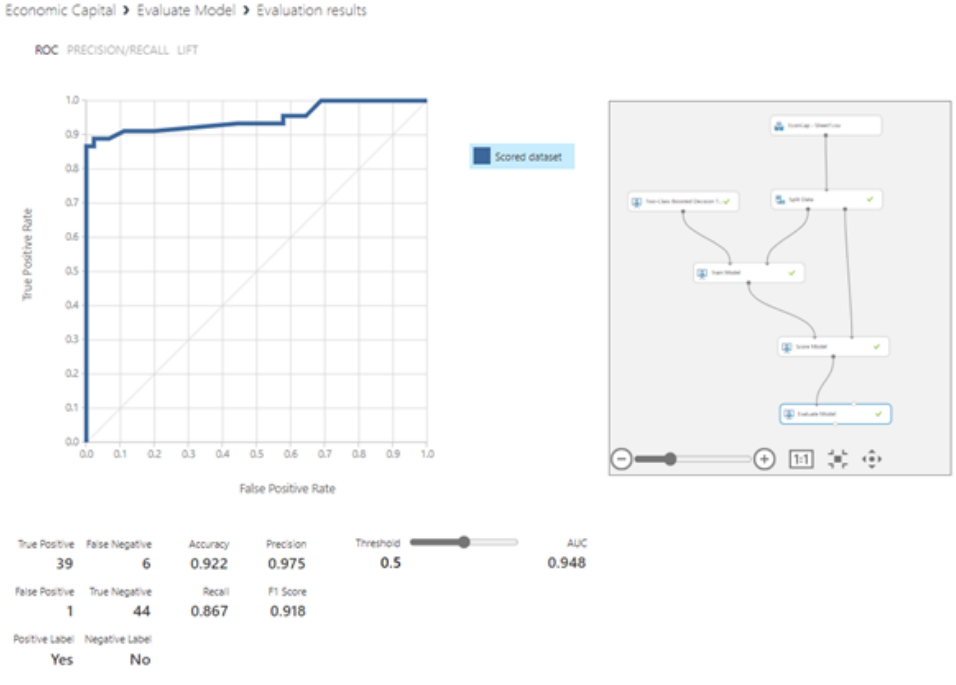
2 Precision (0.975): The precision rate of 97.5% reflects the model's reliability in predicting households that are in debt. This high precision indicates that when the model predicts a household is in debt, it is correct 97.5% of the time, showcasing a high level of trustworthiness in its positive classifications.

3 Recall (0.867): With a recall rate of 86.7%, this metric shows the model's capacity to identify all relevant instances of households in debt. It implies that the model successfully captures most of the 'in debt' cases, although there is a margin for missing some cases.

4 F1-score (0.918): The F1-score, combining precision and recall in a single measure, stands at 91.8%. This score is particularly useful when an equilibrium between precision and recall is essential. It suggests the model maintains an excellent balance between accurately predicting 'in debt' households and minimising false positives.

5 AUC (0.948): An AUC of 94.8% indicates an excellent model performance distinguishing between the 'in debt = yes' and 'in debt = no' classes. This high value suggests that the model has a high probability of ranking a randomly chosen 'in debt = yes' instance more highly than an 'in debt = no' instance.

Figure 4 The overall performance metrics of the model (see online version for colours)



These validation metrics demonstrate the robust predictive power and practical applicability of the model in identifying and understanding the debt status among households in the Kut Bak district. The high accuracy and precision indicate a reliable model for stakeholders to derive insights into the community's financial health. While slightly lower than precision, the substantial recall rate still ensures that a significant

proportion of indebted households are correctly identified, offering a solid basis for targeted interventions. Furthermore, the exceptional AUC value reinforces the model's discriminative capacity, highlighting its utility in nuanced scenarios where the distinction between varying degrees of financial stability is crucial. This becomes particularly relevant in socio-economic research and policy-making, where identifying households at risk of financial distress is crucial.

4.1 Comparison of classification model performance

To assess the effectiveness of the TBDT model in analysing the status of household debt, its performance must be compared with other competitive models. This study evaluates TBDT alongside XGBoost, random forest, logistic regression, and a standard decision tree to determine its efficiency. The implementation is carried out using Python in Google Colab, incorporating a data split of 70:30 for training and testing, along with a confusion matrix analysis for performance validation. The results demonstrate the efficiency of the two-class enhanced decision tree, providing insight into its classification capabilities. Figure 5 illustrates a comparative analysis of the classification model results, highlighting the performance differences between metrics.

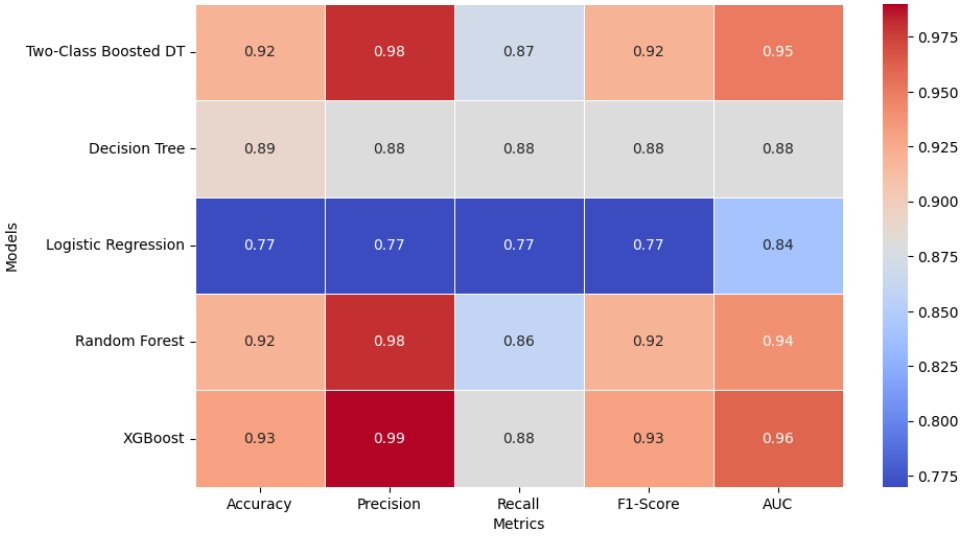
- The accuracy of the models varies significantly, with XGBoost achieving the highest accuracy at 0.93, followed closely by TBDT and random forest, both at 0.92. These results indicate that these three models demonstrate strong reliability in making general predictions. However, the decision tree model achieves an accuracy of 0.89, placing it in the middle range of performance. Logistic regression performs the lowest, with an accuracy of only 0.77, suggesting that it faces challenges in correctly classifying household debt status compared to the other models.
- The precision of the models varies significantly, with XGBoost achieving the highest precision at 0.99, followed closely by TBDT and random forest, both at 0.98. These models exhibit exceptional performance by minimising false positives, ensuring that non-debt households are rarely misclassified as debt-ridden. The decision tree also performs well, with a precision of 0.88. On the other hand, logistic regression demonstrates the lowest precision at 0.77, indicating a higher tendency to incorrectly classify non-debt households as struggling with debt.
- Recall is highest for the decision tree and XGBoost (0.88), followed closely by TBDT (0.87) and random forest (0.86). These models effectively identify many debt-ridden households, making them reliable for scenarios where missing a positive case is costly. In contrast, logistic regression performs the lowest with a recall of 0.77, implying it is more likely to overlook actual debt-ridden households.
- The F1-score, which balances precision and recall, is particularly valuable when minimising false positives and false negatives is critical. Among the models evaluated, XGBoost (0.93), TBDT (0.92), and random forest (0.92) demonstrate the best performance in achieving this balance. These models excel in accurate classification while effectively controlling errors, making them highly reliable choices. Additionally, the decision tree model achieves an F1-score of 0.88, indicating strong performance, though it falls slightly behind the top-performing

models. In contrast, logistic regression, with a lower F1-score of 0.77, struggles to maintain an optimal balance between precision and recall.

- XGBoost, with an AUC of 0.96, exhibits the highest capability in distinguishing between debt-ridden and non-debt households, establishing itself as the most effective classification model. It is followed closely by TBDT (0.95) and random forest (0.94), demonstrating strong performance in classification tasks. In contrast, logistic regression, with a lower AUC of 0.84, shows the weakest ability to separate the two groups, making it less reliable for accurately classifying debt-ridden versus non-debt households. The decision tree performs moderately with an AUC of 0.88 but still falls short compared to ensemble methods like XGBoost and random forest.

The comparative results show that XGBoost outperforms all other models, achieving the highest rank across all performance metrics. Its accuracy of 0.93 demonstrates superior predictive ability, ensuring many correct classifications. Its precision of 0.99 suggests that it effectively minimises false positives, which is crucial when misidentifying households struggling with debt could have profound implications. Additionally, XGBoost maintains a strong recall of 0.88, successfully identifying many debt-ridden households. The F1-score of 0.93 reflects a well-balanced performance between precision and recall, making it a highly reliable model. Furthermore, its AUC of 0.96 indicates an excellent ability to differentiate between debt and non-debt households, reinforcing its robustness.

Figure 5 Performance comparison of different models (see online version for colours)



The TBDT follows closely behind XGBoost in overall ranking. With an accuracy of 0.92, it performs nearly as well as XGBoost in terms of correct classifications. Its precision of 0.98 means it also makes very few false positive errors. However, its recall is slightly lower at 0.87, indicating that it may miss some actual cases of debt-ridden households compared to models with higher recall. The F1-score of 0.92 ensures a good

balance, making it a strong alternative to XGBoost. Its AUC of 0.95 also confirms that it effectively distinguishes between the two classes. This model is a solid second choice, especially in scenarios where interpretability and computational efficiency matter. The random forest performs similarly to TBDT in accuracy and F1-score, at 0.92. However, its recall is slightly lower at 0.86, which means it may miss a few more actual debt-ridden households compared to XGBoost or TBDT. Its precision remains high at 0.98, ensuring that most positive classifications are correct. The AUC of 0.94 suggests that it still has strong discrimination ability, though not as powerful as XGBoost. This model remains a good choice for classification but does not surpass the two leading models.

The decision tree model has an accuracy of 0.89, which is lower than TBDT but still acceptable. Its precision and recall are 0.88, which shows balanced performance, but does not excel in either category. The F1-score follows the same trend at 0.88, maintaining consistency. With an AUC of 0.88, it is evident that while this model performs decently, it lacks the refined learning power of boosted models. The decision tree tends to be more interpretable than the ensemble models. Nevertheless, the trade-off between simplicity and predictive power suggests that TBDT would be preferable. Logistic regression ranks the lowest for all performance metrics, with an accuracy of 0.77, significantly lower than the other models. Its precision, recall, and F1-score are all 0.77, indicating it struggles to make the correct classifications. The AUC of 0.84 suggests it has difficulty distinguishing between the two classes, making it the weakest performer. Given its lower complexity, logistic regression might still be helpful in cases where a simple and interpretable model is needed, but it does not perform well for this particular problem.

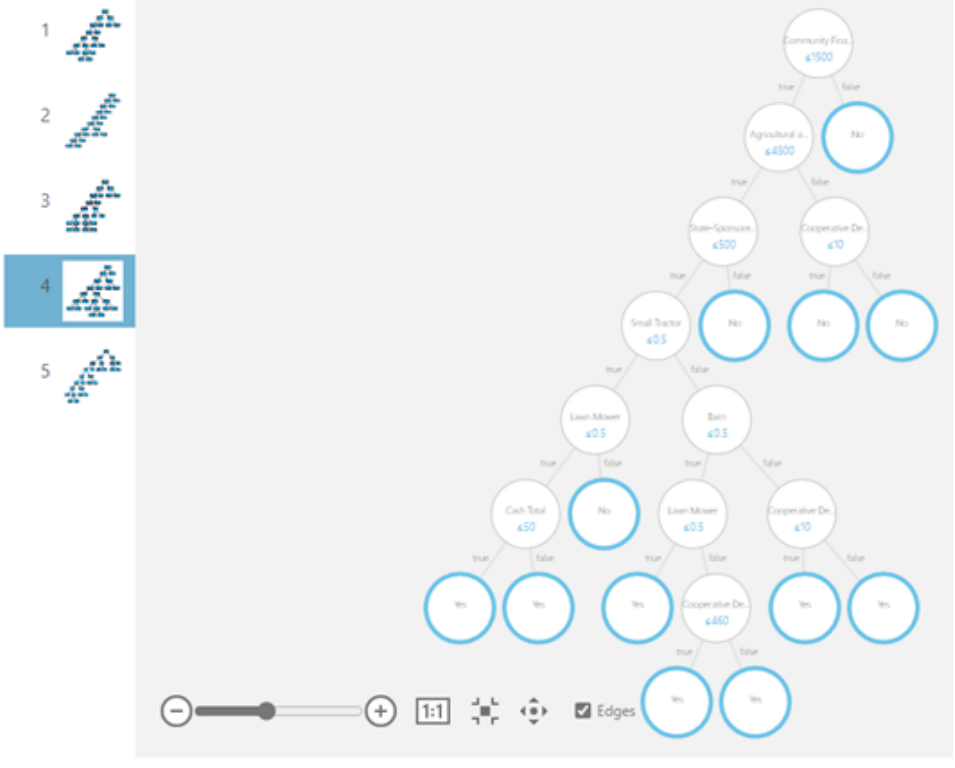
The results confirm the hypothesis that TBDT demonstrates excellent performance across all critical metrics, except recall, where its performance is nearly equivalent to that of the traditional decision tree. Notably, the model achieves higher accuracy, precision, F1-score, and AUC compared to other methods, establishing its efficiency and suitability for classification tasks within this context. Furthermore, the findings highlight the superior performance of ensemble-based decision tree methods, such as random forest and XGBoost, over traditional classification approaches like logistic regression and standalone decision trees. This underscores the advantages of ensemble techniques in capturing complex patterns and relationships within the dataset. Collectively, these results validate TBDT as a robust and reliable model for analysing household debt status in Kut Bak district (An et al., 2021; Ríos Canales, 2016).

4.2 Decision rules from the TBDT model

Figure 6 illustrates a series of five TBDT models, each developed to analyse and predict the debt status of households within Kut Bak district. Model 4 has been selected for a detailed discussion due to its exemplary performance in identifying the nuanced rules that determine whether households are categorised as in debt or not in debt. The rules extracted from model 4 offer crucial patterns and associations in understanding the financial behaviours and conditions that lead to household debt. While the model has generated many rules that dictate the debt status outcome, we will focus on a curated selection of these rules to elucidate and interpret their significance. This selective approach highlights the most impactful and representative rules, providing a clear and concise overview of the key factors contributing to or mitigating the likelihood of

household debt within the community. For example, rules that involve variables such as the value of community financial fund contributions, agricultural and cooperative bank savings, state-sponsored financial assistance, and the ownership of essential assets such as small tractors and lawn mowers have been particularly telling. These rules not only reflect the economic resilience or vulnerability of households, but also paint a broader picture of the socio-economic landscape of the Kut Bak district.

Figure 6 Examples of rules extracted from model 4 (see online version for colours)



Model 4 reveals that a combination of rules leads to classifications of households, with five rules indicating a status of not being in debt (in debt = no) and seven rules suggesting a status of being in debt (in debt = yes), culminating in a total of 12 distinct outcomes. In this scenario, the rules presented in Table 1 are derived from the TBDT model, which identifies splits in the data based on their ability to minimise classification errors. Unlike association rule mining (Ledmi et al., 2023), which uses metrics such as support and confidence to evaluate rule quality, TBDT prioritises variables and thresholds that contribute to the overall model's performance metrics, such as accuracy, precision, recall, and F1-score. The rules were selected to demonstrate the model's interpretability and practical application in analysing household debt status. Table 1 elaborates on examples of rules from Model 4, providing the rules and their corresponding descriptions.

Table 1 provides a selection of representative rules extracted from model 4, emphasising complex and impactful interactions among variables. Although Figure 6

illustrates the tree structure with more straightforward rules, the rules in Table 1 incorporate multiple variable thresholds, offering more profound insight into the relationships identified by the model. The redundancy observed in some rules highlights the complexity of real-world data and the prioritisation of more predictive conditions by machine learning models. For example, the cooperative deposit value appears to be a less significant variable compared to the community financial fund and agricultural/cooperative bank savings. As a result, the cooperative deposit value variations do not affect the predicted outcome, as the model relies primarily on the more influential variables to drive its predictions.

Table 1 Examples of rules and descriptions from model 4

No.	*Rules*	*Description*
1	If (community financial fund value <= 1,500) = true, (agricultural and cooperative bank savings value <= 4,500) = false, (cooperative deposit value <= 10) = true, then in dept = no	This rule reflects that households with a strong financial foundation, as demonstrated by their savings levels, are less likely to be indebted, even if they have small investments or deposits in community or cooperative schemes.
2	If (community financial fund value <= 1,500) = true, (agricultural and cooperative bank savings value <= 4,500) = true, (state-sponsored financial assistance value <= 500) = false, then in dept = no	The rule highlights an interesting insight from the data: households that may not have high values in community or cooperative savings but receive a considerable amount of state-sponsored financial assistance are predicted to be financially stable enough to not be in debt. This suggests that government assistance plays a critical role in supporting household financial health, potentially offsetting limited personal savings or investments.
3	If (community financial fund value <= 1,500) = true, (agricultural and cooperative bank savings value <= 4,500) = true, (state-sponsored financial assistance value <= 500) = true, (small tractor <= 0.5) = true, (lawn mower <= 0.5) = false, then in dept = no	This rule underscores a nuanced view of household financial health, suggesting that not being in debt is associated with a careful balance of savings, receipt of financial assistance, and practical asset management. Households that maintain moderate savings, receive some level of government assistance, and make selective decisions about the assets they own are seen as less likely to be in debt.
4	If (community financial fund value <= 1,500) = true, (agricultural and cooperative bank savings value <= 4,500) = true, (state-sponsored financial assistance value <= 500) = true, (small tractor <= 0.5) = true, (lawn mower <= 0.5) = true, (cash total <= 50) = true, then in dept = yes	This rule provides critical insight into how a confluence of financial constraints and minimal asset ownership correlates with a higher likelihood of being in debt. It underscores the importance of one or two financial indicators and a broader assessment of a household's financial health, including liquid assets and tangible property, in understanding debt dynamics.

Given these results, future work could explore the refinement of the model to improve recall without significantly compromising precision. This could involve additional feature engineering, experimenting with different model parameters, or incorporating more nuanced socio-economic variables. Additionally, the insights derived from this model can inform targeted financial literacy programs, debt relief initiatives, and economic development policies tailored to the specific needs and characteristics of the Kut Bak district, ultimately contributing to a more nuanced understanding and mitigation of household debt challenges.

5 Conclusions

In conclusion, this research has comprehensively analysed household debt status in the Kut Bak district, employing the sophisticated TBDT model within Microsoft Machine Learning Studio. The study examined a dataset comprising 302 households characterised by 30 distinct attributes, aiming to identify patterns and determinants influencing the likelihood of household indebtedness. The TBDT model was compared against logistic regression, decision tree, random forest, and XGBoost to assess its predictive performance. The validation phase demonstrated that TBDT achieved high accuracy (0.922), precision (0.975), recall (0.867), F1-score (0.918), and AUC (0.948), outperforming logistic regression, decision tree, and random forest, while closely competing with XGBoost (An et al., 2021; Hassan and Saeed, 2023). These performance metrics validate the efficacy of TBDT as a robust machine learning approach, reinforcing its potential for predictive analytics in socio-economic research.

In this study, the model automatically selected variable thresholds and the order of processed variables within each decision tree in the boosted ensemble, rather than relying on predefined ones. This adaptive approach enabled each tree to refine predictions iteratively, improving accuracy and robustness. The decision trees within the TBDT ensemble, particularly Model 4, contributed unique corrective adjustments, leading to improved classification accuracy. Despite minor variations in rule sets, this boosting methodology ensured consistency and a holistic understanding of household debt factors.

The analysis revealed strong relationships between household financial characteristics, asset ownership, and debt status. Factors such as contributions to community financial funds, savings in agricultural and cooperative banks, state-sponsored financial assistance, and ownership of productive assets (e.g., small tractors and lawnmowers) significantly influenced household debt likelihood. Specifically, the model's decision rules suggested that households with greater financial stability – characterised by higher savings, financial aid, and prudent asset management – were less likely to be in debt. The findings provide valuable insights into socio-economic determinants of household debt in rural Thailand, offering a data-driven foundation for policymaking, financial planning, and targeted interventions (Evans and Gregson, 2023; Shah et al., 2021; Yalley et al., 2021). The study underscores the importance of comprehensive financial support systems, including state-sponsored assistance and community financial initiatives, to strengthen household financial resilience.

Future research could expand on this work by incorporating additional socio-economic variables, leveraging longitudinal data to track debt trends over time,

and applying the model to diverse regional contexts to validate its generalisability (Xie et al., 2024). Furthermore, further exploration of external economic factors such as market dynamics and policy shifts could refine insights into household debt risks. Ultimately, this study not only enhances our understanding of household debt dynamics in the Kut Bak district but also demonstrates the power of advanced machine learning models in uncovering complex socio-economic patterns, providing a roadmap for future research and policy development in financial stability and debt mitigation.

Acknowledgements

The project is funded by the Program Management Unit on Area-Based Development (PMU A), Thailand, under Grant No. A14F650075.

References

Abbasimehr, H., Paki, R. and Bahrini, A. (2023) 'A novel XGBoost-based featurisation approach to forecast renewable energy consumption with deep learning models', *Sustainable Computing: Informatics and Systems*, Vol. 38, p.100863, https://doi.org/10.1016/j.suscom.2023.100863.

Ağlarcı, A.V. and Bal, C. (2024) 'Effect of various factors on classification performance of ordinal logistic regression', *International Journal of Data Mining, Modelling and Management*, Vol. 16, No. 2, pp.196–208, https://doi.org/10.1504/ijdmmm.2024.138813

Aldashev, A. and Batkeyev, B. (2023) 'Household debt, heterogeneity and financial stability: evidence from Kazakhstan', *Central Bank Review*, Vol. 23, No. 2, pp.100119, DOI: 10.1016/j.cbrev.2023.100119.

Alpaydin, E. (2020) *Introduction to Machine Learning*, MIT Press, https://doi.org/10.7551/mitpress/13811.001.0001.

An, S.H., Yeo, S.H. and Kang, M. (2021) 'A study on a car insurance purchase prediction using two-class logistic regression and two-class boosted decision tree', *Korean Journal of Artificial Intelligence*, Vol. 9, No. 1, pp.9–14, DOI: 10.24225/kjai.2021.9.1.9.

Dalianis, H. (2018) 'Evaluation metrics and evaluation', in Dalianis, H. (Ed.): *Clinical Text Mining: Secondary Use of Electronic Patient Records*, pp.45–53, Springer International Publishing, Cham, DOI: 10.1007/978-3-319-78503-5_6.

Evans, D.M. and Gregson, N. (2023) 'Money, debt and finance: reclaiming the conditions of possibility in consumption research', *Sociology*, Vol. 57, No. 6, pp.1491–1506, SAGE Publications Ltd., DOI: 10.1177/00380385231156339.

Hassan, S. and Saeed, M.S. (2023) 'A comparative study evaluated the performance of two-class classification algorithms in machine learning', *Kurdistan Journal of Applied Research*, Vol. 8, No. 2, pp.43–50, DOI: 10.24017/science.2023.2.5.

Kalita, J.K., Bhattacharyya, D.K. and Roy, S. (2024) '3 – data preparation', in Kalita, J.K., Bhattacharyya, D.K. and Roy, S. (Eds.): *Fundamentals of Data Science*, pp.31–46, Academic Press, DOI: 10.1016/B978-0-32-391778-0.00010-7.

Ledmi, M., Souidi, M.E.H., Hahsler, M., Ledmi, A. and Mohamed, C.K. (2023) 'Mining association rules for classification using frequent generator itemsets in arules package', *International Journal of Data Mining, Modelling and Management*, Vol. 15, No. 2, pp.203–221, https://doi.org/10.1504/ijdmmm.2023.131399.

Narongchai, W., Ayuwat, D. and Chinnasri, O. (2016) 'The changing of intergenerational transfers of economic capital in rural households in Northeastern, Thailand', *Kasetsart Journal of Social Sciences*, Vol. 37, No. 1, pp.46–52, DOI: 10.1016/j.kjss.2016.01.005.

Niu, T., Chen, Y. and Yuan, Y. (2020) 'Measuring urban poverty using multi-source data and a random forest algorithm: a case study in guangzhou', *Sustainable Cities and Society*, Vol. 54, p.102014, https://doi.org/10.1016/j.scs.2020.102014

Ono, T. and Uchida, Y. (2018) 'Human capital, public debt, and economic growth: a political economy analysis', *Journal of Macroeconomics*, Vol. 57, pp.1–14, DOI: 10.1016/j.jmacro.2018.03.003.

Park, H.J., Kim, Y. and Kim, H.Y. (2022) 'Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework', *Applied Soft Computing*, Vol. 114, p.108106, https://doi.org/10.1016/j.asoc.2021.108106.

Ríos Canales, V. (2016) *Using a Supervised Learning Model: Two-Class Boosted Decision Tree Algorithm for Income Prediction*, https://prcrepository.org:443/xmlui/handle/20.500.12475/557.

Shah, S., Chaudhry, I. and Farooq, F. (2021) 'Poverty human capital and economic development nexus: a case study of multan division', *Sustainable Business and Society in Emerging Economies*, Vol. 3, No. 4, pp.461–469, https://doi.org/10.26710/sbsee.v3i4.1924.

Strzelecka, A., Kurdyś-Kujawska, A. and Zawadzka, D. (2020) 'Application of logistic regression models to assess household financial decisions regarding debt', *Procedia Computer Science*, Vol. 176, pp.3418–3427, DOI: 10.1016/j.procs.2020.09.055.

Talia, D., Trunfio, P. and Marozzo, F. (2016) 'Chapter 1 – introduction to data mining', in Talia, D., Trunfio, P. and Marozzo, F. (Eds.): *Data Analysis in the Cloud*, Computer Science Reviews and Trends, pp.1–25, Elsevier, Boston, DOI: 10.1016/B978-0-12-802881-0.00001-9.

Xie, S., Zhang, J., Li, X., Xia, X. and Chen, Z. (2024) 'The effect of agricultural insurance participation on rural households' economic resilience to natural disasters: evidence from China', *Journal of Cleaner Production*, Vol. 434, p.140123, DOI: 10.1016/j.jclepro.2023.140123.

Xu, X-M., Liu, Y.S., Hong, S., Liu, C., Cao, J., Chen, X-R., Lv, Z., Cao, B., Wang, H-G., Wang, W., Ai, M. and Kuang, L. (2024) 'The prediction of self-harm behaviours in young adults with multi-modal data: an XGBoost approach', *Journal of Affective Disorders Reports*, Vol. 16, p.100723, https://doi.org/10.1016/j.jadr.2024.100723.

Yalley, J., Francis, O., Anlimachie, M. and Avoada, C. (2021) 'Human capital, economic growth and poverty reduction nexus: why investment in free compulsory universal education matters for Africa', *International Journal of Humanities and Social Sciences*, Vol. 13, No. 2, pp.50–60, https://doi.org/10.26803/ijhss.13.2.3.

Yuan, J. and Tang, S. (2025) 'Approximating decision trees with priority hypotheses', *Theoretical Computer Science*, Vol. 1023, p.114896, https://doi.org/10.1016/j.tcs.2024.114896.