Determining the Effectiveness of Microfinance in Rural Areas of Bangladesh Using Machine Learning

by

Fardeen Rahman
20221032
Miftahul Jannah
21241025
Faiza Nooren Suhita
24341280
Sanjida Amin
22299246

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
February 2026.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

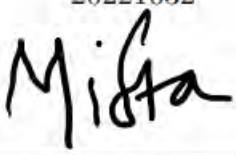
Student's Full Name & Signature:

Fardeen Rahman

20221032

Sanjida Amin

22299246



Miftahul Jannah

21241025

Faiza Nooren Suhita

24341280

Approval

The thesis titled "Determining Effectiveness of MicroFinance in Rural Areas of Bangladesh Using Machine Learning" submitted by

1. Fardeen Rahman (20221032)
2. Miftahul Jannah (21241025)
3. Faiza Nooren Suhita (24341280)
4. Sanjida Amin (22299246)

has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Spring 2026.

Examining Committee:

Supervisor:

(Member)

Dr.Chowdhury Mofizur Rahman

Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Rakin Bin Rabbani

Lecturer
Department of Computer Science and Engineering
Brac University

Abstract

Financial exclusion hampers economic growth in rural Bangladesh, as the conventional credit scoring model fails to identify the creditworthiness of the millions of unbanked citizens. Microfinance institutions aim to fill this gap; however, their impact assessments are unable to predict long-term borrower sustainability due to subjective judgment criteria. This thesis proposes a machine learning approach to evaluate the effectiveness of microfinance institutions based on non-traditional socioeconomic and behavioral indicators. A primary dataset of 1,001 rural borrowers with 85 engineered features was created, including digital literacy, household infrastructure, lifestyle factors, and psychometric indicators. Four supervised algorithms, LightGBM, XGBoost, Random Forest, and Support Vector Machine, were applied to predict five indicators of a successful loan: Income Generation, Standard of Living, Business Expansion, Ability to Save, and Asset Acquisition. Results show that the gradient boosting algorithms consistently outperform conventional baselines. LightGBM and XGBoost go hand in hand in successfully predicting the outcomes of microcredit. The findings demonstrate that alternative behavioral data can proxy formal credit histories. The framework provides MFIs as well as rural borrowers with insights on the possible outcome of their loans.

Keywords: Microfinance, machine learning, financial inclusion, rural lending, socioeconomic impact, borrower behavior, Bangladesh.

Dedication

This thesis is dedicated with love to our parents, whose unconditional love and support have been the pillars of our success. Their sacrifices have made us what we are today.

We also wish to give our deepest thanks to our respected faculty members. Besides teaching course contents, they have instilled curiosity and discipline that we need to be the engineers of tomorrow.

Acknowledgement

All praise is due to Almighty Allah for granting us the strength, knowledge, and perseverance to complete this project successfully. We are grateful to our families and friends for their unrelenting support in this journey.

Most importantly, we would like to thank our supervisors, Dr. Chowdhury Mofizur Rahman and Rakik Bin Rabbani, for their constant guidance throughout. We would also like to extend our gratitude to our research assistant, Shuchishmita Anwar.

Also, we thank Unaysa, Talim Chacha, Mahedi Bhai and Albab, who have helped us gather survey data required for this study. Lastly, we want to thank Sallu, Titly, Zayeed, Maharaj, and Tausif for their constant motivation and emotional encouragement, which kept us going.

Table of Contents

List of Tables

Chapter 1

Introduction

1.1 Background

Microfinance is the provision of financial services, such as small loans, insurance, savings accounts, and a payment system to low-income earners or households that cannot access traditional banking. Its main aim is to promote financial inclusion and empower marginalized communities, and reduce poverty [42].

Microfinance in Bangladesh began in the late 1970s with early pilot projects such as Professor Muhammad Yunus's "Jobra" experiment and the "Dheki Rin Prokolpa" which was first initiated by the Bangladesh Bank and Swanirvar Bangladesh [16]. In 1983, the Grameen Bank was formally established by Yunus. He introduced a group lending model where borrowers, primarily women, supported one another in repaying loans. This approach proved socially empowering and financially sustainable. As a result, achieving a remarkable repayment rate of 98%. BRAC adopted a more comprehensive strategy by integrating microfinance with legal aid, healthcare, and education at the same time. In 2002, its Ulta Poor Program was launched where 95% of participants overcame extreme poverty in the past two years [42].

These pioneering models laid the foundation for Bangladesh's leadership specially in the global microfinance innovation. On a larger scale, the microfinance movement in Bangladesh continued to evolve with time. Even though Grameen Bank achieved early success, the 1980s saw debate around the effectiveness of microfinance. This compares it with alternative development strategies, for example like conscientization. However, Grameen Bank continued to refine its model during this time by focusing on women-led group lending. Factors like weekly repayment schedules, and door-to-door banking,also proved particularly effective in rural contexts. The 16 Decisions framework was also introduced which encourages borrowers to adopt positive social behaviors such as sanitation, education for children, and tree planting. Thus, this demonstrated that financial inclusion could be integrated with social development. By 1990, Grameen had expanded to over 800 branches. They served more than a million clients across Bangladesh where a number of the population were women.

By the 1990s,in the field of poverty alleviation experimentation and data backed success solidified microfinance's place in poverty alleviation. In 1990, the government

established the Palli Karma Sahayak Foundation (PKSF) to coordinate and fund NGO MFI operations. As the years passed, by the 2000s, microfinance had moved from the periphery to the center of Bangladesh's financial landscape. This further proved that even the poorest populations could be viable clients. Grameen Bank's innovations, which included housing loans, education loans, and flexible repayment products, demonstrated its ability to adapt to the evolving needs of borrowers. The unique credit of this system without a collateral model inspired replication in over 100 countries. In the end, Professor Yunus and Grameen Bank won the Nobel Peace Prize in 2006 for their dedication to changing lives. The momentum for this work had been building since 1997, when the Microcredit Summit in Washington, D.C. set a high bar for what the world could achieve. The initiative aimed to provide microfinance services to 100 million of the most economically disadvantaged families globally by 2005. PKSF's role grew in parallel, with its contribution to MFIs' revolving loan funds increasing from 9% to 24% between 1996 and 2002. As a result, this was established as a central player in Bangladesh's microfinance ecosystem [16].

Microfinance has alleviated poverty for millions with great implications for the Sustainable Development Goals (SDGs), primarily SDG 1 (No Poverty) and SDG 5 (Gender Equality). Its impact varies regionally in the nation, with institutional strength and favorable socio-economic conditions amplifying success in some areas, while weak institutions and financial illiteracy limit its effectiveness in others [42]. However, long-term sustainability issues arise despite the prevailing short-term benefits, such as increased household income and access to healthcare and education. Debt traps are common among borrowers due to high rates of interest charged on loans and a variety of geographical disparities in access, which ultimately work to nullify the goals of reducing poverty as well as growing equality. Evidence from studies [35] suggests positive impacts in several areas. For instance, longer microfinance memberships in Bangladesh have negatively correlated with malnutrition. Moreover, with fewer sickness-related absences, health outcomes have generally improved. On the other hand, participants in savings and loan associations in Zanzibar have shown higher rates of home ownership and investment in housing improvements. Meanwhile, findings on education have been mixed, as only some invested in education, while others invested their microcredit elsewhere. In Bangladesh, microfinance is credited with lifting approximately 2.5 million people out of poverty over two decades, contributing to an annual reduction of 1% in moderate poverty and 1.3% in extreme poverty [35].

Nevertheless, some critics put their argument forward stating that microfinance only moderately reduces poverty. Also, how often fails to reach the poorest or most marginalized communities was a factor [42]. The economic impact is also mixed. On one side, while some recipients experience increased household income and improved business outcomes, others report overindebtedness or financial instability. The main reason for this was the rigid repayment schedules. Microfinance has had varying effects across different areas, for instance, some borrowers have been able to afford better education for their children or invest in income-generating assets. On the other hand, some even struggle to sustain long-term gains due to limited market access or even inadequate business training. Although women's empowerment remains a frequently cited benefit, often leading to greater participation in family

and community decisions. Factors such as education level, geographic location, and type of occupation also play crucial roles in determining microfinance's effectiveness. Hence, measuring the true impact of microfinance requires moving beyond its more visible short-term effects to assess its long-term sustainability and adaptability to diverse rural contexts. This calls for deeper, context-sensitive research to fully understand its potential and limitations.

1.2 Rationale of the Study or Motivation

While going through a stack of research papers, a similar yet unrecognised trend began to linger. While almost all papers boasted the usage of AI models, they had a blind spot. The persistent exclusion of rural borrowers' data was quite evident in most papers, as the data were limited to only high-end borrowers whose credit histories were available. However, are they the only representatives of a third-world country like Bangladesh? What about farmers, shopkeepers, and craftsmen who have minimal knowledge regarding financial transactions? What about the rickshaw pullers who always deal in cash and never leave a trail of digital receipts? Their records are missing. The negligence in financial studies conducted on the Bangladeshi population is primarily rooted in a lack of technological knowledge, and in some cases, concerns over privacy. Aiding such people are the MFIs, offering essential financial services such as: micro loans, insurance, and savings to people with low-income or small businesses [40]. Hence, specifically, the microfinance sector was chosen for the study, as it is made up of the farmers, shopkeepers, craftsmen, and rickshaw pullers, who make up the pillars of our economy, yet often remain financially deprived. These people's financial records are rarely documented and hence excluded when it comes to studies on financial analysis.

A notable study [41] has significant practical implications in emerging markets as they took the best advantage of oversampling by integrating SMOTE technique in ML which showed a significant improvement in result (overall accuracy of 91.46%) but the dataset was only limited to 56 Peruvian financial institutions which may not fully capture the diversity and since the paper was based on Peruvian institutions, algorithms such as:- logistic regression, gradient boosting applied in those regions can't handle the class imbalance of Bangladesh. Similarly, the paper [38] utilizes the high performance of machine learning, particularly DT, to predict the success of microcredit programs in empowering women, achieving an accuracy of 83.72%. While this paper intends to focus on the transformative role of microcredit in improving women's economic status, it did not verify the improvements in household income or security, which paper [16] successfully addressed. However, it lacked a detailed analysis of regional disparities in Bangladesh, which is vital to our research, as the variation in MFI effectiveness between urban and rural areas is different. On top of that, data was collected from only 43 respondents across 3 districts, which limits the generalization ability of the algorithm as ML requires highly diverse data to perform best [38].

Few papers used advanced algorithms such as LightGBM or Random Forest, which definitely enhanced the accuracy but pose a challenge when the resource setting is limited. Due to this ML models will lack scalability in demonstrating corner cases.

Although [40] [45] talked about working on it using SHAP and LIME to enhance model transparency and to make it more understandable for further reviewing, it's still underdeveloped. Making use of these algorithms can be difficult for MFIs with limited expertise and requires advanced IT infrastructure and skilled members which is quite impractical, given our main focus is on rural areas, and if these models are left uninterpreted they can probe the risk of becoming "BLACK BOX" which regular borrowers, lenders or even regulators cannot decipher [45].

Models such as logistic regression and gradient boosting, which are comparatively reliable, stumble in cases where there is a shortage of data. Due to such a huge gap in data records, the AI models had a limited interpretability scale. To get the best result we require more than just numbers; we need non-financial records such as family support that kept the business afloat, their day-to-day spending habits, years of experience and the hard work that paid off microloans, the portion of the income that went into raising their children and over healthcare, their mobile usage and social media activity [44]. These are the records that never made it to such algorithms.

Therefore, taking into account this scenario, the motivation of this thesis is to first collect a detailed dataset (consisting of both financial and behavioral data) from those who otherwise do not have their credit history documented and then integrate machine learning models on it to uncover complex relationships between borrower behavior and the overall impact of microfinance in Bangladesh. This research marks the first step in understanding the true weight of microcredit beyond numbers to see how it really shapes the lives of individuals.

1.3 Problem Statement

The rural populations are usually financially disadvantaged when it comes to accessing formal banking opportunities. Conventional credit models are very dependent on banking records and formal credit histories. Those who do not have collateral and recorded income, like farmers and small artisans, are normally excluded. Microfinance has been a significant instrument in this situation, particularly in poverty reduction and financial inclusion in rural areas. Nevertheless, the efficacy of microfinance schemes in enhancing economic and social well-being is controversial because of their high interest rates. Traditional credit assessment procedures premised on income and bank records are not always suitable in the rural setting, and this compels MFIs to resort to subjective judgment when deciding on who to lend to. Consequently, this exposes the lending institutions to the danger of lending to high-risk individuals. This paper seeks to investigate how non-traditional, alternative data can be utilized to determine the effects of microfinance on the financial well-being of rural borrowers in Bangladesh. Socio-economic features are combined with behavioral indicators (e.g., mobile phone use, duration of the call, and employment diversity). Several outcome variables are taken into account rather than paying attention to only loan repayment. The aim of the research is to determine patterns that can be used to identify sustainable borrowers by employing various machine learning models on each of the target variables. These target variables are carefully selected to capture both short-term and long-term effects of microfinance. This framework provides insights to both MFIs and borrowers on who is more likely to

benefit from the loan. Furthermore, it can identify key factors that lead to successful loan utilization.

1.4 Objective

Our study aims to:

- Examine the evolving impact of microfinance in Bangladesh
- Maximise financial inclusion of rural people
- Provide insights to both MFIs and borrowers about who is more likely to benefit from receiving microcredit, and how to optimise the usage of the loan
- Make microfinance more sustainable using machine learning

1.5 Methodology in Brief

The main motive of this study is to assess micrifinance in Bangladesh and dig deeper into how borrowers actually behave. This section provides the framework for how the research was carried out.
The novelty of our study is its primary dataset, which is obtained from door-to-door surveys in different districts of Bangladesh. The data is collected from people who have previously participated in microfinance throughout the country so that the results are not biased, since the effect of microfinance can vary from region to region. Moreover, we tried to interview people from different levels of poverty as well. The target size of this dataset, given certain time constraints, was set at a minimum of 1000 data points. Participation was completely voluntary. All participants signed a consent form prior to the interviews and no personal identifiers were used in our questionnaire to maintain anonymity.

After preprocessing the collected data, an unsupervised machine learning model, K-means clustering is used for data reduction. Finally, an encoded and more compact dataset is used for training four supervised machine learning models: Random Forest, SVM, LightGBM and XGboost. Stratified splitting is performed on the dataset, using 60% for training, 20% for testing, and a further 20% for independent validation. This is done to ensure our model performs well with unseen, real-world data. Five target variables were chosen, each acting as an indicator of microfinance impact. Both short-term and long-term effectiveness are measured to identify if and when microfinance can cause sustainable economic growth. Evaluation metrics like accuracy, precision, recall, F1 score, and AUC-ROC values are used to evaluate model performance. Finally, five-fold cross-validation was implemented to detect model bias.

1.6 Scopes and Challenges

Scope of the Study:
The novelty of this study is the collection of a primary dataset capturing a wide range

of behavioral and socio-economic factors. It directly takes into account the personal experiences of the otherwise financially excluded people of rural Bangladesh. Such a dataset opens door to several research directions.

- Substitutes behavioral data for those who lack formal credit histories.
- Incorporates detailed information on the utilization of the loan, the borrowing process, etc. Machine learning models can identify dominating factors determining a successful loan. Such insights can help borrowers make more informed decisions on how to use their microcredit to maximize its benefits.
- The multi-dimensionality of the dataset creates the scope to identify different borrower profiles, allowing loans to be more personalized for optimal outcomes.

Challenges Encountered:
Despite the study's ambitious goals, certain challenges were present:

- Contextual Variability: Economic realities shift from one district to another. Various cultural and regional nuances make it challenging to build a single model that performs well across the entire country.
- Lengthy Interviews: The data collection process was not a swift one. Each interview took almost 60 minutes. This was because of a number of reasons. While understanding local dialect was a challenge of its own, we also had to ensure participants truly understood the questions before answering to ensure accurate data.
- Small Size of Dataset: Even though we could successfully gather 1,001 data points, it is not large enough for deep learning models like neural networks.

1.7 Research Gap

Although microfinance and machine learning are not novel subjects of research, a specific focus on their intersection in the specifics of the socio-economic environment of rural Bangladesh is a relatively under-researched area. An overview analysis of existing literature shows that there are several serious gaps that this study will fill:

- Dependence on Conventional Credit Data: The majority of the current credit scoring predictive models continue to rely on formal banking history and internet footprints. This is a huge challenge to the millions of rural borrowers lacking a formal credit history.
- Geographic Homogeneity: Numerous existing research has been done with data from countries like Peru or Kenya. The same models would not perform as well with rural Bangladeshi data because of regional differences and cultural specifics. Socio-economic factors not only differ between countries, but also can differ between districts.
- Discontinuous and Little Data: In Bangladesh, prior research using ML has usually been limited in the sample size (small) or dimension (women empowerment). There is a strong requirement of multidimensional data to be collected on a grand level and capture the loan repayment and behavioral measures.

- Analysis of the short-term and long-term effect: The available literature often concentrates on short-term effects such as the returns of loans.

By introducing a primary dataset of 1,001 rural borrowers and integrating behavioral data with machine learning algorithms, this thesis seeks to form a bridge between these gaps.

Chapter 2

Literature Review

2.1 Preliminaries

The microfinance industry is increasingly adopting innovative technology to deal with economic uncertainty and adopt financial inclusion. This is due to the scarcity in data and multi-dimensional nature of borrower profiles. To better understand how microfinance affects rural communities, researchers are turning to machine learning. By applying various computational models, they can predict defaulters and evaluate borrower risk more accurately. Moreover, both the efficiency of the lending institutions and the overall economic results of the programs can be measured [15] [18].

In microfinance research, logistic regression is most commonly used to classify borrowers based on binary outcomes. It works best on limited data due to its interpretability and ease of use [8]. Decision trees are also frequently used for segmenting borrowers based on features such as income level, repayment history, or household size [4]. Decision trees are easy to interpret, however, they can sometimes overfit the data [9].By combining results from multiple decision trees, random forest improves accuracy and reduces overfitting. Thus, this makes it a strong choice for predicting loan outcomes in rural microfinance settings.

Gradient boosting methods, including XGBoost and LightGBM, have gained significant interest due to their high prediction accuracy in complex datasets [21] [23]. These models are particularly useful in analyzing non-linear relationships and learning subtle patterns in borrower behavior. The model Support Vector Machine is effective in situations where data is in high dimension and the separation between borrower categories are not clearly defined [5]. They are usually used to categorize clients into high and low-risk groups accordingly.

[3]. When working with rural population, the KNN algorithm turned out to be very useful as it classifies borrowers based on similarities to others and for more complex task such as analysing mobile, money records or sequential repayment behaviours researchers have used ANN. This includes convolutional and recurrent neural networks [20] [7]. Even though they can be difficult to interpret, these models can efficiently catch and learn deep patterns from large datasets.

SHAP and LIME artificial intelligence tools, which are used to address these issues

of model transparency [24] [22]. They help identify the factors that most influence a machine learning model's decision. They are essential for ensuring fairness and accountability in rural microfinance programs. In preliminary classification tasks or when working with small datasets, simpler models like Naive Bayes and perceptron are used [2].

Regression techniques are also common in impact evaluation. Linear regression, ordinary least squares and linear probability models are used to measure continuous outcomes, for example, income change or loan size [11]. Advanced regression techniques such as elastic net, step-wise regression and Bayesian ridge are applied to handle multi-collinearity and improve model performance [13] [6]. Quantile random forests are useful for estimating ranges of outcomes instead of just a single prediction [14].

Partial least squares and principal component analysis are used in research where there are highly correlated variables or require dimensionality reduction[10]. For multiclass classification tasks, linear discriminant analysis and multinomial regression are often employed. They help to categorize borrowers based on risk level or type of loan received [1].

Emerging approaches include control learning algorithm and Gaussian score models to adapt borrower behaviour over time and contribute to developing personalized credit scoring system [29]. Fintech platforms using blockchain are also being integrated with machine learning to improve the security and transparency of rural lending [30].

Additionally, generalised random forests and policy trees are used to find out how microfinance outcomes are different from one group to another, this helps use to understand who is being benefitted the most from microfinance programs [28].

These machine learning models, when put together, can form a more inclusive, efficient and fair microfinance system that can be used by MFIs and practitioners to better serve those in need.

2.2 Review of Existing Research

This section provides an overview of existing research exploring the intersection between the fields of ML and microfinance. Also, we saw it fit to look into some studies on microfinance in Bangladesh that use more traditional methods instead of ML algorithms, to have a better idea of the field. We will first look into the variety of ways in which researchers have implemented machine learning to assist the process of lending microcredit and later dive deeper into existing studies specifically in the context of Bangladesh. Once we find an evident research gap, we can work on overcoming that through this thesis.

First, it is seen that ML algorithms have been a great tool that helps to enhance credit management and risk mitigation. A study [36] evaluates machine learning models like SVM, decision trees, gradient boosting (XGBoost, LightGBM), KNN

and Neural Network for credit scoring, using case studies and model performance tests rather than a specific dataset. It compares traditional methods like FICO and logistic regression with supervised, unsupervised and reinforcement learning. Metrics for comparison include ROC-AUC, precision, recall and F1-score. Explainability tools like SHAP and LIME were used in order to meet fairness and transparency requirements. The study uses information on credit bureau reports and transaction records, mobile payment logs, utility bills, e-commerce and social media behavior, internal banking and fintech case studies. The two case studies that are used are that of UK High Street Bank, which used decision trees and gradient boosting to detect 83 risks missed by traditional models and WeBank (China), which adopted ML with alternative data like mobile payments, social media, etc., to serve clients with no credit history. In this case, their model achieved a land driving financial inclusion. Overall, this research concluded that ML models outperform traditional credit scoring by improving accuracy, lowering non-performing loans and widening access to credit, since the case studies confirm a high predictive success of 91.2%.

UK banks Decision trees and random forests offer good accuracy and interpretability, while XGBoost and LightGBM provide maximum precision. Neural networks and deep learning excel with large datasets but lack transparency, as there is little interpretability, making them less suitable for regulated environments. SVMs and KNN work for classification but struggle with complex, high-dimensional financial data. This study also acknowledges its fair share of barriers, mainly regarding regulation (data privacy, fairness and explainability). There is always the possibility of hidden bias, whether it be from the data obtained or those within AI algorithms. Overall, the study enlightens us on the different ML algorithms and how they perform with data from microfinance, as well as the limitations they come with.

The role of fintech in modernizing credit systems and reducing risk for small-scale lenders and enterprises is explored in an article [37]. It highlights how outdated assessment methods fall short. Also, it suggests that creditworthiness be measured through alternative data points like mobile records and social media footprints. To further protect these processes, the research recommends using blockchain technology. This is to maintain transparent and secure records. The main idea here was to bridge the gap between technology and finance. Overall, the paper creates room for more research, especially involving blockchain techniques along with ML.

Another study [40] focuses on how machine learning (ML) can help MFIs to be more financially inclusive by improving credit scoring, fraud detection and customer segmentation. Yet again, ML leverages alternative data (e.g., mobile transactions, social media, utility payments), which is critical for serving unbanked populations, outperforming traditional credit scoring. The research used both primary and secondary data with both qualitative and quantitative components. They conducted 18 expert interviews with MFI stakeholders such as fintech firms, regulators and beneficiaries and surveyed 320 microfinance clients on loan accessibility and service delivery using a mix of closed and open-ended questions to achieve data saturation. The secondary data that was collected includes anonymized MFI records (loan approvals, repayments, demographics) and World Bank/IMF datasets, as well as fintech platform data. Then, the data was preprocessed (missing data handled via

imputation; outliers removed using z-scores/IQR; normalized features and applied one-hot encoding for categorical variables (e.g., occupation, loan purpose); engineered new features (e.g., loan-to-income ratio) and used PCA for dimensionality reduction). For credit Scoring, decision trees, random forests, logistic regression and XGBoost/LightGBM were used, while random forests, SVM (risk classification) and autoencoders/isolation forests (anomaly detection) were used for fraud detection and K-means and hierarchical clustering were used for customer segmentation. All models were evaluated on precision, accuracy, recall, F1 Score and AUC. Finally, LightGBM performed best, whereas autoencoders and isolation forests, despite showing potential, needed refinement. Lastly, K-means Clustering was effective (silhouette score 0.72) for customer segregation, enabling more personalized services based on client demographics. The paper also highlights ethical considerations and bias mitigation throughout. The research finally concludes that machine learning boosts financial inclusion by 13–17% compared to traditional methods.

Furthermore, two more studies were reviewed that included the involvement of microfinance data from Peru. First, a research [34] explores how different ML algorithms can reduce credit risk and enhance decision-making. The study focuses on Puno, a highly rural region with clients from a microfinance institution. This data is then cleaned and preprocessed (missing values were dealt with, one-hot encoding was applied to categorical values and empirical rural-specific variables were defined, such as client financial history, payment capacity, along with theoretical ones. Financial parameters (e.g., capital disbursement, agreed interest rate and savings accounts), demographic parameters (e.g., age, civil status and level of education) and economic parameters (e.g., primary/secondary activities, occupation, economic sector, etc.) were included. Moreover, unique variables specific to rural microfinance were also added, which took into account factors like the type of primary activity and the number of borrowers. ML models, including Logistic Regression, Decision Tree, Random Forest, SVM, ANN and KNN, were used on the dataset that was split into 75yet again evaluated using metrics like precision, accuracy, recall, F1 Score and AUC-ROC and a confusion matrix was also used as a validation tool. The results revealed that the ANN model surpassed traditional methods (dominating AUC-ROC, the most comprehensive metric), improving default prediction accuracy from 76.81%.

Nevertheless, limited data coverage, bias in data and limited interpretability are potential limitations of the study. Besides using ML algorithms for enhanced credit scoring while lending microcredit, such algorithms can be integrated with other aspects of microfinance too. For instance, the second paper dealing with Peruvian data [41] aims to predict the failure rate of MFIs in emerging countries and introduces the Adjusted Gross Granular Model (ARGM), a hybrid machine learning tool combining granular computing and advanced analytics designed to process messy, imbalanced datasets, which are typical of MFIs. Data was collected from 56 Peruvian financial institutions (banks, MFIs, rural lenders) from 2014 to 2023, a period marked by economic crises, policy shifts and COVID-19. However, there was a data imbalance since only 11.4418.67). To mitigate this class imbalance, SMOTE oversampling and ensemble methods (Random Forest, Gradient Boosting) were used. Also, metrics like AUC, MCC and F1-Score were chosen over accuracy due to asym-

metric misclassification costs when concluding.

Similar to the last one, a study [43] also uses ML techniques to forecast the performance of MFIs and uncover key drivers of financial sustainability. A global dataset of 9,059 firm-year observations (2003–2018), obtained from the World Bank's Mix Market, included parameters like operational self-sufficiency, portfolio yield, administrative cost ratio and outreach indicators. Linear models like Linear Regression, Linear Regression with Stepwise Selection and Elastic Net and Partial Least Squares (PLS) were used. Moreover, two ensemble learning models- random forest and quantile random forest were also used along with Bayesian Ridge, KNN and Support Vector Regression (SVR). Three key regression metrics: Root Mean Squared Error (RMSE), R-squared and Mean Absolute Error (MAE), were used to assess model accuracy. It was found that Random Forest and Quantile Random Forest had stable and consistent residuals across training and testing samples, suggesting a well-balanced bias– variance trade-off, while Elastic Net and SVR showed signs of overfitting or underfitting. Random Forest achieved high R-squared values across all test splits and is considered the best model in predicting MFIs' financial performance. Operational self-sufficiency was found to be the critical variable that influences profitability and institutional sustainability, along with financial revenue to assets and operating expense to total assets, which also play significant roles in influencing financial outcomes.

We reviewed a paper [45] that focused on the expanding integration of ML technologies with the microfinance sector. It focuses mainly on enhancing credit scoring and financial inclusion by exploring alternative data sources like social media activity, mobile phone usage and utility bill payments. It also emphasized how e-commerce behavior is being leveraged to improve loan accessibility and risk assessment, especially for borrowers lacking formal credit histories. Besides, the ML algorithms like decision trees, random forest, deep learning, logistic regression SVM and CNNs are highlighted. They can better handle large quantities of unstructured data and uncover intricate patterns that traditional models often miss. These models can detect non-linear relationships between the data types and repayment behaviors, thus improving predictive performance. Additionally, the use of NLP, blockchain and XAI is explored to enhance transparency, traceability and the interpretability of AI decisions. The incorporation of biometric data and psychometric data is also elaborated in this paper. This offers innovative methods for identity verification and credit evaluation. However, there are certain challenges that have been mentioned. These include data privacy, algorithmic bias, regulatory compliance and technical limitations. Despite these concerns, the literature concludes that AI presents a promising future for improving financial access, decision-making and sustainability in the microfinance sector.

We have also reviewed studies on microfinance in Bangladesh to better grasp the sector's current landscape. The first paper [38] is the closest to what we aim to build upon, even though it was conducted on a relatively small scale. This study explores the use of ML algorithms to evaluate the impact of microcredit. It pinpoints specifically on empowering rural women in Bangladesh. A survey containing 21 categorical and demographic parameters was administered to 43 underprivileged

female microcredit recipients (ages 18–50) from Narayanganj, Kishoreganj and Jessore. Algorithms such as Naive Bayes, SMO, k-NN, decision tree and random forest were applied to the dataset. The dataset was split into 20% and 80% for testing and training, respectively. Accuracy, precision, recall, F1-score, MCC, ROC-AUC and PR curves were used as evaluation metrics. It was seen that the decision tree model performed best, achieving an accuracy of 83.72%. This study concludes that ML has the potential to identify beneficiaries who are more likely to succeed with microcredit. Despite this promise, research on ML applications in microfinance remains limited in Bangladesh because of one major reason, which is the scarcity and fragmentation of structured, high-quality data. Many microfinance institutions (MFIs) operate at local or regional levels. They either lack strong digital infrastructures or are hesitant to share client data due to privacy concerns and regulatory uncertainties. Rural borrowers often lack digital receipts, which makes it difficult to build reliable datasets and another factor that proves to be a challenge is the technical barrier. Since ML requires expertise in data science, this may not be readily available within traditional development or finance-focused research teams. To add more to this, funding and institutional support for interdisciplinary work combining social development and advanced computation are limited. Lastly, in a sensitive sector like microfinance, concerns around interpretability and fairness of ML models may discourage adoption. This is critical in this context because decisions have real human consequences. As a result, ML holds significant potential to optimize microfinance targeting, risk prediction and impact evaluation. Its integration in the Bangladeshi context is still at a growing stage. Hence, it calls for further exploration with proper infrastructure, ethical safeguards and stakeholder collaboration. Then, we looked into research solely on the microfinance sector of Bangladesh, which used more traditional methods instead of ML methods.

Moving on to a study [33] which conducted interviews on 350 households, from September to December in 2020 in three villages: Khalishakundi, Malipara and Silimpur in Daulatpur Upazila, Kushtia, Bangladesh had used Monthly household income, Age of respondent, Microfinance loan participation, Education (years), Household size, Land holding (hectares), Spouse's income, Loan availability, etc as their key variables. On the other hand, the Logit Model, which was used to analyze the probability of loan adoption, found a negative impact of age, average income and household size on the likelihood of borrowing. It suggests that older, richer, or larger households are less likely to adopt microfinance. It also emphasized how credit availability positively influences adoption, meaning respondents prefer MFIs over banks for easier access. While most research focuses on microfinance's impact on individual borrowers, a study [25] estimates its macro-level impact. This was specifically done on Bangladesh's GDP using a Computable General Equilibrium (CGE) model. The model incorporates microfinance as part of productivity improvements, capital collection and reallocation of labor and capital. Furthermore, simulations compare actual GDP with a counterfactual case where microfinance is absent. From here, we gain insights into the extent to which microfinance impacts the economy of Bangladesh. It seems that MFIs can greatly boost the country's GDP while alleviating poverty.

Next, a study [44] is designed to determine key socio-demographic and programmatic

factors that impact the effectiveness of microfinance loans in Bangladesh. The data used was self-collected via surveys of 1,000 microentrepreneurs across all 8 divisions of Bangladesh using stratified random sampling. Standard logistic regression was used to evaluate how loan amount, business type, experience and support factors affect borrowers' evaluation of MFI loans in promoting entrepreneurship. The study concludes that microfinance is solely inadequate for successful entrepreneurship. Larger loan amounts, business type, experience, access to other funding and family support significantly influence positive attitudes toward MFI loans. MFIs are viewed more favourably by experienced and multiple-loan clients. On the other hand, those with training or family support are less likely to credit MFIs for business success.

Furthermore, a research [35] examines the effects of government (BRDB) and NGO (AIDComilla) microcredit programs on rural livelihoods and empowerment. Using data fromm 300 microfinance participants and 200 control respondents, researchers used propensity score matching (PSM), rosenbaum bounds, multi regression and fixed effects instrumental variable (FE-IV) techniques to address potential bias. The study evaluates impacts across basic needs, overall quality of life and empowerment and categorizes them into whether that be economic, social, or political. The findings indicate that microfinance programs have a positive influence on both livelihood and empowerment. This research also focused on how women benefited more than men in income generation. Despite having such positive results, the study shows methodological limitations. This is mainly due to the potential hidden biases, political interference in government lending and inconsistent training and monitoring practices.

Finally, we looked into studies on the performance of MFIs in Bangladesh. First, a study [19] evaluates the efficiency of Bangladesh's ten major MFIs from 2003 to 2011. This was done using a stochastic frontier output distance function with a Cobb-Douglas specification. It found that total factor productivity (TFP) grew. Inefficiency is significantly linked to cost per borrower and operational self-sufficiency. It also found that larger MFIs improved faster compared to the smaller MFIs.

Another study [17] investigates the implementation of Point of Sale (POS) technology by the microfinance institution Come to Save (CTS) in rural Bangladesh. This demonstrates its effectiveness in streamlining operations and improving efficiency. Using qualitative methods, the study found that POS use significantly enhanced staff productivity, reduced human errors and fraud and increased client satisfaction. The paper highlights the potential and challenges of scaling technology in small rural MFIs. These findings reflect the current state of microfinance in Bangladesh, which is intricate information for our research since we aim to introduce machine learning to a sector associated with the poorest of the poor in the country, lacking access to advanced technologies. We have also looked into studies on the Bangladeshi microfinance sector to get ideas about how it impacts the rural people, how the MFIs perform, what kind of data is typically taken, specific to these regions, etc. Notably, there has been little research specifically on using machine learning algorithms when it comes to this sector, using data from Bangladesh. This is obvious, given how new ML is to the emerging economy of Bangladesh. We believe that our research, in-

tegrating ML technologies with our highly saturated dataset involving information from all around the country, can be groundbreaking and in turn, can take this sector onto the next level.

2.3 Summary of Key Findings

Below is the summary of the most relevant studies from the literature reviewed, represented in tabular format.

Table 2.1: Key findings from the literature review

Study	**Aim of the research**	**Performance/Key findings**	**Research gap**
[36]	Explore how a range of machine learning can enhance traditional credit scoring methods and help make financial services more accessible to everyone	Decision trees and random forests offer good accuracy with interpretability, XGBoost and LightGBM provide higher precision. Neural networks and deep learning excel with large datasets but lack transparency, SVMs and KNN work for classification but struggle with complex, high-dimensional financial data.	Data quality and missing values in emerging markets; Regulatory/privacy constraints on alternative data; Deep-learning "black-box" interpretability; Limited real-time digital histories in some regions; Need for robust X frameworks

Study	Aim of the research	Performance/Key findings	Research gap
[40]	Explore how machine learning can support MFIs in promoting financial inclusion by improving credit scoring,fraud detection and customer segmentation.	LightGBM: Highest performance in credit scoring (Accuracy: 89.6%, AUC: 0.92): Random Forests: Strong results in loan approval and fraud detection (Loan approval accuracy: 86.7%, Fraud detection accuracy: 87.6%, AUC: 0.88); SVM: Competitive performance across tasks: Autoencoders and Isolation Forests: Showed potential for anomaly detection, but require further refinement; K-means Clustering: Effective for customer segmentation (Silhouette score: 0.72)	Interpretability, Data Quality, Deployment Challenges
[34]	The main goal is to build a more accurate and reliable model for evaluating rural clients looking for microcredits by using ML techniques to assess creditworthiness, enhance decision making accuracy and lower default rates.	Decision Tree performed the best in terms of accuracy, precision, recall and F1 score (accuracy of 0.9230 and AUC ROC of 0.8880); ANN had the highest AUC ROC (0.9372), making it strong in distinguishing between classes; kNN had the lowest AUC ROC and accuracy, indicating weaker overall performance.	Limited Data Coverage, Bias in Data, Limited interpretability and external validation

Study	Aim of the research	Performance/Key findings	Research gap
[41]	Proposes an Adjusted Rough-Granular Model (ARGM) that combinies machine learning and granular calculating techniques to predict MFI failures.	ARGM achieved AUC of 0.925, F1-Score of 0.9231, Accuracy of 92.54%, Sensitivity of 89.55%, MCC of 0.8520; Six institutions out of the 30 that were left after preprocessing the data were flagged as high risk.	False negatives persist (missed failures). Class imbalance, Dependency on Peruvian data; generalizability untested.
[43]	Apply a range of ML regression methods to forecast financial sustainability of MFIs and identify key factors of performance, guiding practitioners and policymakers in risk management.	Random Forest delivered the lowest RMSE and highest R^2; operational self-sufficiency emerged as the top predictor.	Doesn't capture rural, borrower-level impacts; regional heterogeneity masked by global sample; limited interpretability and external validation; focus on financial performance only.
[45]	Reviews advancements in AI-based credit scoring models for microfinance institutions, exploring how alternative data sources and AI can improve loan accessibility, risk assessment and financial inclusion	Case studies report improvements like 15–30% better default prediction, 27–40% higher loan approvals and 40–65% faster loan processing compared to traditional methods.	Concerns on data privacy, algorithmic bias, explainability.
[38]	Suggests using ML algorithms to predict the impact of microcredit on women's empowerment based on socio-economic and demographic data collected in rural Bangladesh.	Best performance: Decision Tree, Accuracy: 83.72%, MCC: 0.723, ROC: 0.836.	Existing methods often focus on specific aspects of PolSAR image classification and may not provide comprehensive solutions. The proposed method addresses this by combining spatial information from multiple scales.

A critical blind spot in the application of modern computational models to financial inclusion is addressed in this research. This is highlighted mainly in the persistent exclusion of rural borrowers who lack formal banking records. As mentioned in our literature review, while many studies report high model accuracy, they often rely on datasets drawn from urban or well-documented borrowers. These do not reflect the financial realities of rural populations. As a result, this creates a significant gap in ensuring inclusive, fair and scalable microfinance solutions.

By grounding our study in the specific context of Bangladesh and focusing on rural communities, we aim to bridge this gap through the thoughtful integration of non-traditional, non-financial data. These include daily expenditures, mobile usage and social behaviors. Our methodology is designed to evaluate both the technical effectiveness and the practical interpretability of various predictive models. This will have special attention to those that can operate in low-resource settings.

Chapter 3

Requirements, Impacts and Constraints

3.1 Societal Impact

Microfinance is a big part of the lives of people living in rural Bangladesh. This is particularly true for the informal economy workers who earn their income on a day-to-day basis. Lacking traditional credit history causes them to be excluded from the traditional financial system of Bangladesh as well. We used socioeconomic and behavioral data as alternative checks for creditworthiness. Our study looks into unconventional banking methods and aims to help the rural and minor communities, such as farmers, small traders, and daily wage earners, by making loans more accessible to them. As they are the majority of this country, including them has the potential to increase economic participation in small-scale entrepreneurship, agriculture, and the local market. This will also help to reduce long-standing inequalities between urban and rural populations. Traditional methods are often biased and subjective, due to which many potential borrowers are being denied. By understanding how microfinance loans are actually used and identifying the factors that are most affected by such loans, we tried to provide insights that help both MFIs and borrowers. The former will find it easier to identify which kind of borrowers are more likely to benefit from the loans, and the latter will have more knowledge on the optimum way to use the loan, and what part of their finances will benefit the most.

Ultimately, MFIs can offer loans appropriately, design better repayment plans, and avoid issuing unsuitable loans. This ensures that the MFIs continue to be financially viable while also serving vulnerable rural populations. To prevent rural people from falling into debt or being financially exploited, there are multiple other suggested alternatives that are backed by our research. Besides, if the results of our research reach the borrowers, it will help them gain awareness of how microfinance statistically impacts such groups instead of blindly trusting the promises made to them. They can, in turn, make better judgments on where to invest the received microcredit to maximize its benefits.

3.2 Ethical Issues

Before collecting data, we ensured that all participants signed a consent form shown in figure 3.1 and the whole process was completely voluntary. They were informed about the usage of their data. Furthermore, any kind of personal information that can be used to identify the personnel was removed from the dataset. Thus, the data remained completely anonymous throughout the research.

Due to the fact that we are dealing with sensitive data, we gathered the data by visiting the participants in person to ensure accuracy and privacy of the data. Talking to them face-to-face helped us clarify the questions and better understand their perspectives.

It was ensured that only the information related to the study was recorded, as they would often share their personal experiences as well. This made the research more ethical, as no unnecessary or personal data were recorded in the dataset.

Finally, data were collected from multiple districts of Bangladesh to ensure fairness and transparency with the dataset. This step also reduced the risk of overgeneralization and strengthened the whole research even more.

3.3 Constraints

We required a large dataset, a minimum of 1000 responses to ensure reliable model training and performance. Collecting such a huge volume of data from the rural population was extremely time-consuming and logically challenging, as our participants had to be diverse.

On top of that, our initial questionnaire was prepared in English; however, it had to be translated into Bangla to ensure better understanding among respondents. While initial data collection was done on pen and paper, the responses were put into a Google Form, which was then converted to an Excel sheet (csv). These column labels had to be translated back into English so that the images generated from our code show the appropriate names of features. Also, considering how the participants come from intense poverty, it was safe to assume that they lacked enough digital literacy for online data collection. This is why a door-to-door approach, albeit time-consuming, was our only option.

অংশগ্রহণের মধ্যে কী কী বিষয় জড়িত:

- আপনাকে একটি প্রশ্নাবলীর উত্তর দিতে এবং একটি সাক্ষাৎকারে অংশ নিতে বলা হবে।
- আপনার অনুমতিক্রমে, সাক্ষাৎকারটি নির্ভুলতার জন্য রেকর্ড করা যেতে পারে।
- কোন সঠিক বা ভুল উত্তর নেই - আপনার সৎ অভিজ্ঞতা মূল্যবান।

গোপনীয়তা:

- আপনার ব্যক্তিগত পরিচয় গোপন রাখা হবে এবং কোনও প্রতিবেদনে প্রদর্শিত হবে না।
- সংগৃহীত তথ্য নিরাপদে সংরক্ষণ করা হবে এবং শুধুমাত্র গবেষণার উদ্দেশ্যে ব্যবহার করা হবে।

সুবিধা এবং ঝুঁকি:

- কোনও সরাসরি আর্থিক সুবিধা নেই, তবে আপনার ইনপুট আরও ভাল প্রোগ্রাম ডিজাইনে অবদান রাখবে।
- এই গবেষণায় অংশগ্রহণের সাথে কোনও ঝুঁকি জড়িত নেই।

সম্মতি বিবৃতি:

দয়া করে নীচের বিবৃতিটি পড়ুন এবং অংশগ্রহণ করতে সম্মত হলে স্বাক্ষর করুন:

আমি উপরের তথ্যটি পড়েছি বা আমাকে পড়ে শোনানো হয়েছে। আমি প্রশ্ন জিজ্ঞাসা করার সুযোগ পেয়েছি এবং সন্তোষজনক উত্তর পেয়েছি। আমি স্বেচ্ছায় এই গবেষণায় অংশগ্রহণ করতে সম্মত। আমি বুঝতে পারি যে আমি যেকোনো সময় প্রত্যাহার করতে পারি।

স্বাক্ষর / থাম্বপ্রিন্ট: _____

তারিখ: _____

সাক্ষাৎকারগ্রহীতার নাম: _____

স্বাক্ষর: _____

তারিখ: _____

Figure 3.1: Consent Form

Chapter 4

Proposed Methodology

4.1 Methodology Overview

After reviewing the literature on existing research, we moved forward to create our own primary dataset. This would be done using a carefully constructed questionnaire. The goal was to capture diverse information about each participant, including demographics, contextual information and behavioral patterns, as well as their attitudes and opinions, like beliefs, satisfaction levels and expectations in the context of their financial situation. Obtaining this information helps ensure the dataset was rich, multidimensional and balanced.

The target size of the dataset was set at 1000, as it was attainable given the time constraints, while also being large enough to capture diversity and be suitable for supervised model training. The dataset can then provide a strong statistical foundation for both unsupervised learning (clustering) and supervised learning. Door-to-door surveys were conducted throughout the eight divisions of Bangladesh.

The collected data was preprocessed by cleaning it and handling any missing values. All categorical data were converted to numeric data by one-hot encoding, which was then scaled. The dataset initially contained 100 rows, or features. To further simplify this, related features were grouped into specific categories using K-means clustering. Finally, four supervised machine learning algorithms: LightGBM, XGBoost, Random Forest and Support Vector Machine, were implemented. These models were trained across a set of target variables, each acting as an indicator of how effective the microcredit was. Finally, they were validated using 5-fold cross-validation and evaluated using metrics such as accuracy, precision, AUC-ROC and F1-score to ensure robust and interpretable results.

Ultimately, the aim is to generate data based insights that can improve decision-making and promote sustainable growth within Bangladesh's microfinance sector. The framework of our methodology is shown in figure 4.1.

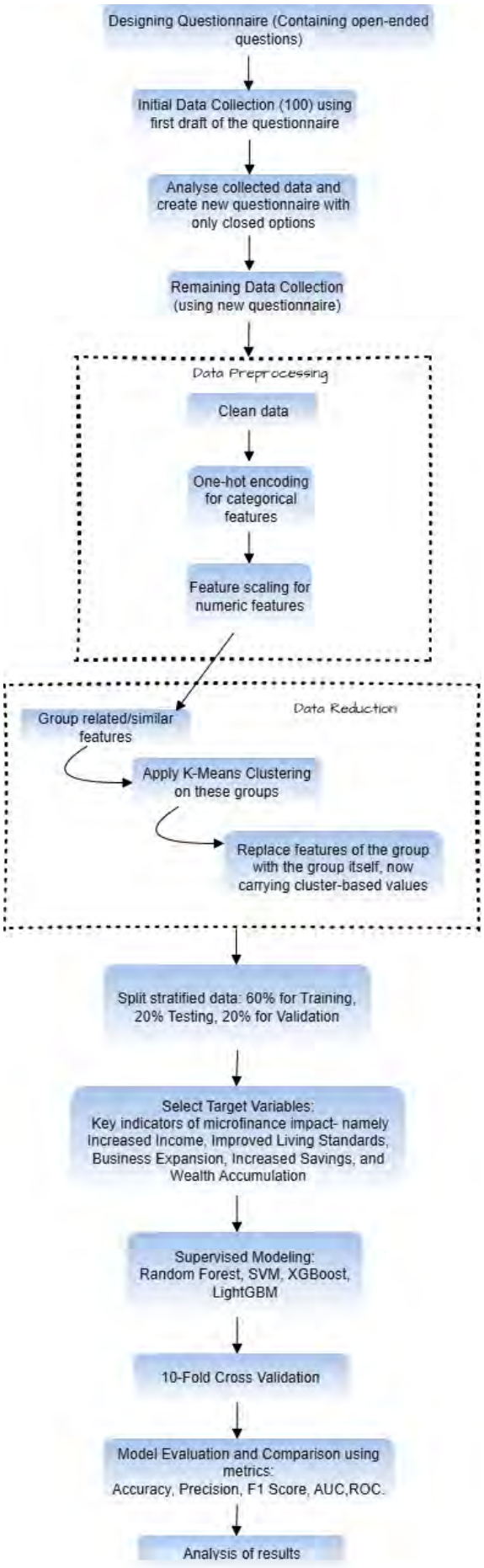


Figure 4.1: Methodology Flowchart

4.2 Model Specification

K-means clustering
K-means clustering is an unsupervised clustering algorithm that partitions a dataset into k groups/clusters so that the points within the same clusters are more similar to each other than to those in other clusters[32]. It is widely used when datasets are fairly simple and clusters are expected to be roughly spherical and similar in size. Because our dataset's features lack clear separation, k-means is required to identify all the underlying clusters to run further predictive algorithms.

The algorithm works with unlabeled data for groups that follow similar patterns. This allows it to detect patterns that may not be obvious to supervised methods. To calculate similarity, it typically relies on Euclidean distance.

The algorithm works in 4 steps:

1. It will randomly assign k centroids according to the number of clusters
2. It will work in a loop to assign every data point to its nearest centroid using the Euclidean distance
3. After every point has been assigned, we re-calculate the centroid by taking the average of the points in each cluster
4. We stop the iteration when we start getting the same centroid values again and again.

K-means aims to minimise the Within Cluster Sum of Squared distances (WCSS) using the formula [27] :-

$$\text{WCSS} = \sum \_k = 1^K \sum \_i \in C\_k \|x\_i - \mu\_k\|^2 \tag{4.1}$$

Where Sk is the set of observations in the kth cluster and (X)'kj is the jth variable of the cluster center for the kth cluster.

To choose the best K in k-means, two of the most commonly used methods are[31]:-

1. **Elbow method**- It involves calculating the WCSS for all the values of k (k1 $\sim$ k max) and plotting the graph of WCSS vs K. The point where the WCSS starts to flatten, that is, when it resembles an elbow shape, is the corresponding k value to consider it to be the optimal fit. Since the method is fairly simple, it is sometimes ambiguous. It works best when there is a clear change in the marginal gain from adding clusters.
2. **Silhouette method**- It measures the mean distance to other points in the same clusters, commonly known as cohesion and mean distance to other neighbouring clusters called separation. We determine how well the dataset has been clustered from the value of the silhouette method. If s $<$ 0, they are highly misaligned and s=1 refers to datasets being well clustered. S=0 indicates that the clusters are on the boundary.

K-means works best when clusters are roughly even-sized and properly scaled. We need to ensure that our dataset is preprocessed and features are properly scaled while keeping in mind that outliers can pull the centroids further away. To deal with that, we can remove the outliers while preprocessing. Dimensionality reduction methods like PCA are most commonly used before applying k-means to a dataset to get a proper 2D visualisation. In high-dimensional data, distances are hard to calculate and as k-means greatly relies on Euclidean distance, it tends to be a weakness if not done beforehand. PCA also removes less important data, which might deflect the output [12].

Random Forest
A random forest algorithm builds a "forest" combining multiple decision trees to achieve better predictive performance. It trains the tree on different subsets of data and features and then forms a prediction based on majority vote for classification or averaging for regression [39]. In real-world scenarios, it has a massive impact as working with the results of multiple decision trees can help us to handle non-linear and complex relationships. When a single decision tree might draw spiked boundaries, RF would draw a much smoother and natural curve by combining the results of all the decision trees. As opposed to decision trees, it reduces overfitting as no particular tree is "too well trained" to learn the patterns. Hence, it performs better on unseen data along with the trained dataset. This makes RF more accurate and generalisable.

Our dataset contains both categorical and numerical features which have to be dealt separately so naturally random forest was the best choice as it can inherently handle mixed feature types and doesn't require transformation. It can also tell us how much each input contributes in our model's prediction given our dataset has many variables initially.

The process begins by creating multiple bootstrap samples from the original dataset. If we have n number of data points, each tree will randomly be assigned with random data points (some may be repeated). Data points not included in a given sample are known as out-of-bag samples, which can be used to estimate errors. This ensures that all the trees are working with different data points, introducing variance among them. As opposed to a decision tree, RF allows all the branches to grow to their full depth without pruning, which makes each tree individually stronger and at each split within a tree, RF randomly selects a feature instead of considering all the features. This randomness creates diversity among trees, reducing correlation and improving performance [26].

RF works best under large and complicated datasets as it handles many irrelevant features. While deep learning models require heavy tuning, RF acquires high accuracy with minimal parameter tuning.

Despite it being one of the most widely used algorithms, it slacks off when the datasets are highly imbalanced unless resampling or class weighting is applied. it's also highly time-consuming, given that we have a large dataset. Additionally, working with large forests can also be computationally expensive.

Support Vector Machine
It is also a supervised machine learning algorithm that is primarily used for classification and sometimes for regression analysis as well.

Its main goal is to find the best decision boundary to separate different classes in our dataset with the maximum margin possible. In the case of higher dimensions, we find the optimal plane called a "hyperplane". The bigger the margin, the better it performs on unseen input data points. The data points that are the closest to the hyperplane and affect the margin are known as support vectors and the one that maximises the distance between them is called "hard margin".

The general equation for hyperplane is :-

$$w^T x + b = 0 \tag{4.2}$$

Where:

- w is the normal vector to the hyperplane (the direction perpendicular to it).
- b is the offset or bias term representing the distance of the hyperplane from the origin along the normal vector w.

When clusters are spread out evenly, it's relatively straightforward to determine the best-fit line. However, when the data points of different classes are placed on a single line, it gets quite difficult to compute the best-fit line. Hence, we take the help of a technique called kernel to map data points into higher dimensions to make it easier to visualize and separate classes. This helps the algorithm to handle non-linearly separable data more effectively.

The distance from the support vectors to the plane can be calculated by the distance between a data point x_i and the decision boundary can be calculated as:

$$d\_i = \frac{w^T x\_i + b}{\|w\|} \tag{4.3}$$

where $\|w\|$ represents the Euclidean norm of the weight vector w.
Since the purpose of this hyperplane is to maximise the margin between to classes, it leads us with the minimising equation:-

$$\min\_w, b \frac{1}{2}\|w\|^2 \tag{4.4}$$

Subject to the constraint:

$$y\_i(w^T x\_i + b) \geq 1 \quad \text{for } i = 1, 2, 3, \cdots, m \tag{4.5}$$

Where:

- y_i is the class label ($+1$ or -1) for each training instance.
- x_i is the feature vector for the i-th training instance.

- m is the total number of training instances.

The condition $y\_i(w^T x\_i+b) \geq 1$ ensures that each data point is correctly classified and lies outside the margin.

In the real-world dataset, it's still not enough to get the best plane due to outliers. In that case, we introduce slack variables (ζ) that measure how much a data point is away from the margin. So we modify our minimizing equation:-

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\sum\_i = 1^m \zeta\_i \\ \text{subject to} \quad & y\_i\left(w^T x\_i + b\right) \geq 1 - \zeta\_i, \quad i = 1, 2, \ldots, m, \\ & \zeta\_i \geq 0, \quad i = 1, 2, \ldots, m. \end{aligned} \tag{4.6}$$

The larger the C, the classification is done almost correctly and the smaller the C, the more margin violations allowed, meaning it is more generalized.

Extreme Gradient Boosting (XGBoost)
Extreme Gradient Boosting (XGBoost) is a poweful supervised machine learning algorithm that belongs to the family of ensemble methods. It is an advanced and optimised implementation of gradient boosting that uses decision tree as its base or weak learners. Unlike simple ensemble approaches that train models independently, XGBoost follows a boosting strategy in which models are added one at a time and each addition improves the overall performance. This iterative error-correction mechanism allows the model to capture complex non-linear relationships and uncover subtle patterns in data more effectively [21].

XGBoost optimizes a regularized objective function, ringing a balance between model accuracy and complexity. The objective function can be written as:

$$\mathcal{L} = \sum\_i = 1^m l(y\_i, \hat{y}i) + \sum k = 1^K \Omega(f\_k) \tag{4.7}$$

where $l(y\_i, \hat{y}\_i)$ represents the loss function. This loss function measures the difference between the predicted value $\hat{y}\_i$ and the true label y_i. Here, $\Omega(f\_k)$ is a regularization term that is responsible for reducing overfitting.

Our literature review repeatedly mentions XGBoost and how consistently well it has performed in financial and microfinance related prediction tasks. It is particularly suitable for a dataset like ours, since it has intricate socio-economic relationships, because of its capacity to handle class imbalance through parameter tuning and represent non-linear interactions.

XGBoost requires careful hyperparameter tuning. For example, the learning rate, maximum tree depth and regularization coefficients. When tuned properly, though, it offers a great trade-off between accuracy and interpretability. Moreover, since it is a tree-based model, it can also generate feature importance scores, which further help to understand variable contributions to model predictions.

Light Gradient Boosting Machine (LightGBM)
It also belongs to the family of Gradient boosting decision tree methods. It essentially works similarly as XGBoost but it introduces several critical algorithms that makes it significantly faster and more memory efficient especially for large datasets. Similar to XGBoost, LightGBM also constructs decision trees one after another to reduce prediction errors. Nonetheless, it presents two significant innovations. Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) notably decrease computational expenses while maintaining model effectiveness. [23].

LightGBM builds trees leaf-wise, not level-wise. This allows the algorithm to focus on leaves with higher loss and produce deeper trees, leading to faster convergence and improved accuracy. The objective function of LightGBM follows the same gradient boosting formulation:

$$\mathcal{L} = \sum \_i = 1^m l(y\_i, \hat{y}i) + \sum k = 1^K \Omega(f\_k) \tag{4.8}$$

but with additional efficiency mechanisms that enable scalable learning on high-dimensional data.
According to findings from the literature review [36][40], LightGBM has demonstrated excellent performance with microfinance datasets. Its ability to process large volumes of data with mixed feature types makes it well-suited for socio-economic survey data.

The literature review consistently reports superior performance of the two gradient boosting models over classical machine learning techniques when applied to microfinance and financial behavior datasets. Thus, XGBoost and LightGBM were selected along with Random Forest and SVM for the purpose of this research.

4.3 Data Collection

The surveys were conducted across different rural regions in Bangladesh, ensuring there was no geographic bias. Face-to-face interviews were conducted, which not only helped participants feel comfortable but also ensured authentic and more accurate data. A consent form was signed by all participants and no personal identifiers were recorded to maintain anonymity. A challenge that we faced was how time-consuming each interview was. Instead of giving concise and to-the-point answers, many participants provided extensive additional information. Moreover, many participants lacked literacy, so they were not asked to fill out the survey questionnaire; instead, it was completed by us. Understanding local dialects was hard on our part, but what was even harder was explaining each question thoroughly to them. Ensuring they truly understood our questions is key to accurate data collection. In the end, 1,001 data points were gathered.

The questionnaire for the interviews was carefully constructed to include diverse information about each participant. This included questions on demographics like age, gender, education, occupation and geographical location. Then we asked ques-

tions about their current quality of life, including if they have access to clean water, electricity, internet, etc. and their current health status. Data on behavioral patterns such as preferences and lifestyle choices, as well as their attitudes, opinions and beliefs on certain financial notions, what they expect from the microfinance sector were collected. We also gathered information about the participants' families (their marital status, number of children, financially dependent members and breadwinners, along with their occupations). A section of the survey consisted of details on the received microcredit: its amount and interest on it, how it was used, repayment status, etc. Other essential information, like the reason why the microcredit was taken and exactly how it was used were asked about. Additionally, some other financial information, such as whether they save and if so, what portion of their income they save were taken. Non-financial information, such as their level of digital literacy, familiarity with online banking and overall experience with microcredit, were obtained. As a result, the dataset was comprehensive, capturing a wide range of personal characteristics and external factors. Figure 4.2 shows a section of the raw data that was collected in the first stage of this research.

For the first 100 surveys, the questionnaire included an "Others" option for several questions. This allowed respondents to provide additional answers to questions that better reflected their views (given none of the options already provided were suitable). An example is, when asked about which social media platform they used, many respondents mentioned "Imo". This is a platform that was not originally included as an option in the questionnaire but received a surprisingly large number of responses. These open responses offered valuable insights. After carefully reviewing these answers, relevant suggestions were later incorporated as new answer options where appropriate. The initial survey also included open-ended questions. The answers to these were then converted into a limited set of specific options to improve clarity and consistency in the results. This was mainly done because data collection and coding had to be done simultaneously. Allowing open-ended questions would have required repeated preprocessing, as unique responses needed to be handled individually. This would not be feasible. Moreover, it appeared that many answers to certain questions were essentially similar but worded differently. For example, in a question about what the participants expect from MFIs in the future, some people wrote "lower interest rate" while many others wrote "reduced interest rates". During preprocessing, these variations would be mistakenly assigned different numeric values, causing logical inconsistencies. To address this, again, each response was carefully reviewed and new standardized options were introduced to group similar answers appropriately. Finally, we observed that many respondents selected multiple answers for certain questions, as multiple scenarios applied to their situations. This was especially seen in questions regarding what changes microcredit had brought about. Most people chose multiple options, such as income generation, increased savings, better living standards, etc. This was addressed by splitting the responses into separate columns with binary indicators for each option. At last, the questionnaire included mostly closed-ended questions, with a few exceptions for numeric values (such as the amount of microcredit taken). Figure 4.3 shows the updated questionnaire.

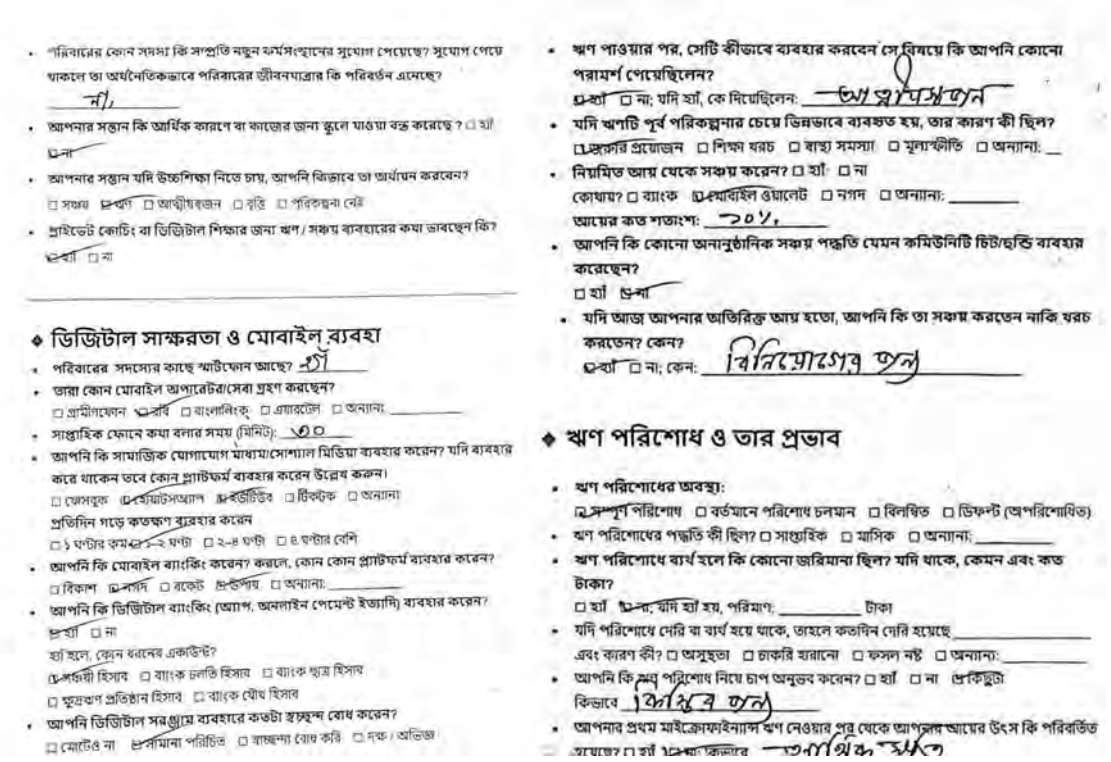


Figure 4.2: Raw Data

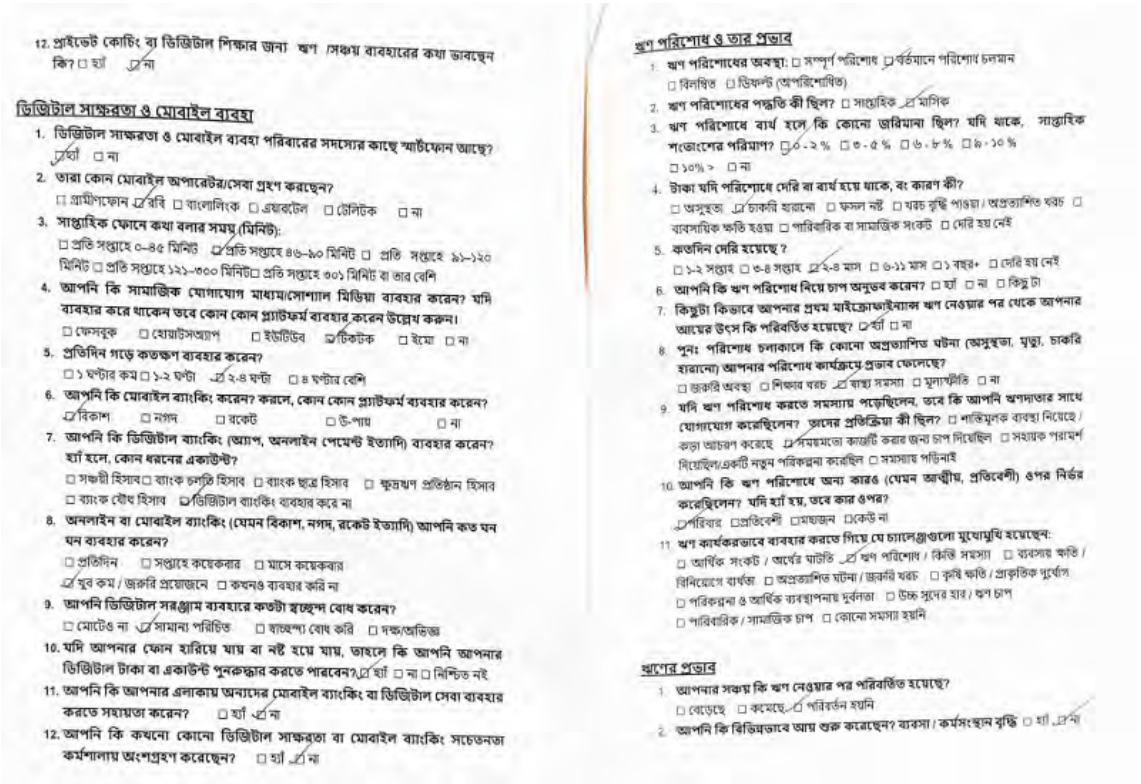


Figure 4.3: Updated Raw Data

4.3.1 Data Analysis

The dataset used in this study was collected door to door through surveys among rural households, resulting in a total of 1,001 observations. Each participant represents an individual or household that has participated in microfinance services. The dataset contains 85 features representing a wide range of socio-economic, behavioral and lifestyle indicators. These features initially had both numerical variables, such as age, amount of loan taken, mobile phone usage duration, etc., as well as categorical variables, such as occupation type, education level, household characteristics and financial behavior indicators. Later they were encoded into numerical representation to feed into machine learning models.

Prior to model development, the dataset was examined for data quality issues. We made sure that there were minimal missing values and when present, median imputation was used to fix them. This helped us to have complete information about the participants and have consistency in data collection.

We focused on 5 binary target features such as:

1. Increase in Disposable Income
2. Improvement in the Standard of living
3. Business Expansion
4. Ability to Save
5. Asset Acquisition

These collectively help us to determine the effectiveness of microfinance. Each target variable is treated as an independent binary classification problem where 1 indicates a positive outcome and 0 indicates a negative outcome.

The class distribution reveals varying degrees of imbalance across targets, shown in figure 4.4.

- For the target variable, Increase in Disposable Income, the dataset contains 533 negative outcomes and 468 positive outcomes, indicating a relatively balanced distribution
- For Improvement in Living Standards, we had 593 negative outcomes and 408 positive outcomes, showing a moderate imbalance
- Business Expansion follows a similar pattern with 605 negative outcomes and 396 positive outcomes
- For Ability to Save, we had a much profound imbalance with 651 negative outcomes and 350 positive outcomes
- For asset acquisition, we had 780 negative outcomes and 221 positive outcomes, which shows the most severe imbalance

Model performance analysis accounted for the observed class imbalance, especially for targets related to asset acquisition and savings. Accuracy alone was insufficient to fully reflect the model's effectiveness, so greater emphasis was put on class-sensitive metrics such as precision, recall and F1-score.

Overall, the analysis indicates that alternative socio-economic and behavioral features can be effectively used in supervised learning frameworks to assess our possible desired outcome.

Figure 4.4: Class balance distribution across all five target variables.

4.3.2 Data Cleaning

Data cleaning formed the foundation of the preprocessing stage, as the performance of the models and the reliability of the results depend heavily on the quality of the input data. Data was cleaned thoroughly and irrelevant columns were dropped, for example, the unique IDs of each participant (which helped to ensure no duplicate data was present). Additionally, we introduced value ranges for certain questions to simplify data interpretation, such as age, average social media usage time, weekly average phone call duration, etc.

One-hot encoding was applied, resulting in a fully numeric dataset. The data was then scaled to ensure that no single variable dominated others due to differences in magnitude. Finally, stratification was done where applicable, i.e., where the target variable contained imbalanced data.

This cleaning process ensured consistency, removed noise and prepared the dataset for effective dimensionality reduction and clustering.

4.3.3 Data Reduction

Our data was still dense and complex after data cleaning and preprocessing. To further reduce the dataset size and decrease the overall number of features, the K-means clustering algorithm was used to cluster sets of similar variables together to form higher-level composite features. The related variables were initially organized manually into conceptual categories and clustering was then carried out at each category to produce interpretable segmentations. For example, information on the accessibility of smartphones, SIM providers, use of social media, time spent on smartphones on average and familiarity with online banking was combined into one feature called Digital Literacy. K-means clustering then divided them into optimal clusters, with the resulting cluster values representing the new feature. The same measure was applied with other groups of variables, including access to clean water, electricity and up-to-date health status, which were all converted into a feature named Lifestyle. Different features about the composition of the participants' families were also simplified with the help of the same approach.

The algorithm selected the best number of clusters (k) using the Elbow Method and Silhouette Score. For each category, K-Means assigned each data point to one of the clusters based on feature similarity. The resulting clusters represented distinct profiles within the dataset. PCA was then used to reduce dimensions and visualize cluster separations.

For the digital literacy category, features including smartphone access, SIM provider, social media usage, average phone usage time, online banking familiarity and comfort with digital tools were used. This segmented borrowers into 2 groups, the former consisting of 35.8% (lower digital engagement) and the latter consisting of 64.2% (higher digital engagement). PCA was applied to visualize cluster separation in two dimensions. Feature discrimination using ANOVA revealed that the most influential variables included comfort with digital tools, mobile banking usage frequency, smartphone accessibility, weekly phone usage and the ability to recover digital money or accounts. Overall, one cluster represents a comparatively lower digitally literate group and the other has higher digital engagement.

Similarly, family-related variables were clustered to simplify information regarding household structure and composition. This category included features such as marital status, number of family members, number of children, whether children live with parents and recent changes in family size. The clustering procedure selected k = 7 clusters, achieving an average silhouette score of 0.406, which indicates moderately strong separation. The cluster distribution was as follows: Cluster 0 (27.0%), Cluster 1 (31.7%), Cluster 2 (5.4%), Cluster 3 (5.1%), Cluster 4 (0.2%), Cluster 5 (6.2%) and Cluster 6 (24.4%). The most discriminative features were marital status and the number of family members. These clusters reflect different household profiles ranging from small nuclear families to large extended households.

The lifestyle category consisted of questions that reflect the current living standards of the borrowers. For instance, road conditions, availability of jobs in the area, whether or not they have access to clean water, electricity, waste management, internet connectivity, etc. It also took into account the borrower's current health

status, whether healthcare services are accessible, etc. The algorithm divided the data points into two clusters with a silhouette score of approximately 0.303, again indicating moderate separation. Cluster 0 (85%) represented households with comparatively limited access to infrastructure, while Cluster 1 (15%) represented households with improved living conditions.

Figure 4.5 shows the results of the clustering. Through this clustering-based reduction process, the dataset was transformed into a more compact one, while still preserving meaningful behavioral and socioeconomic patterns.

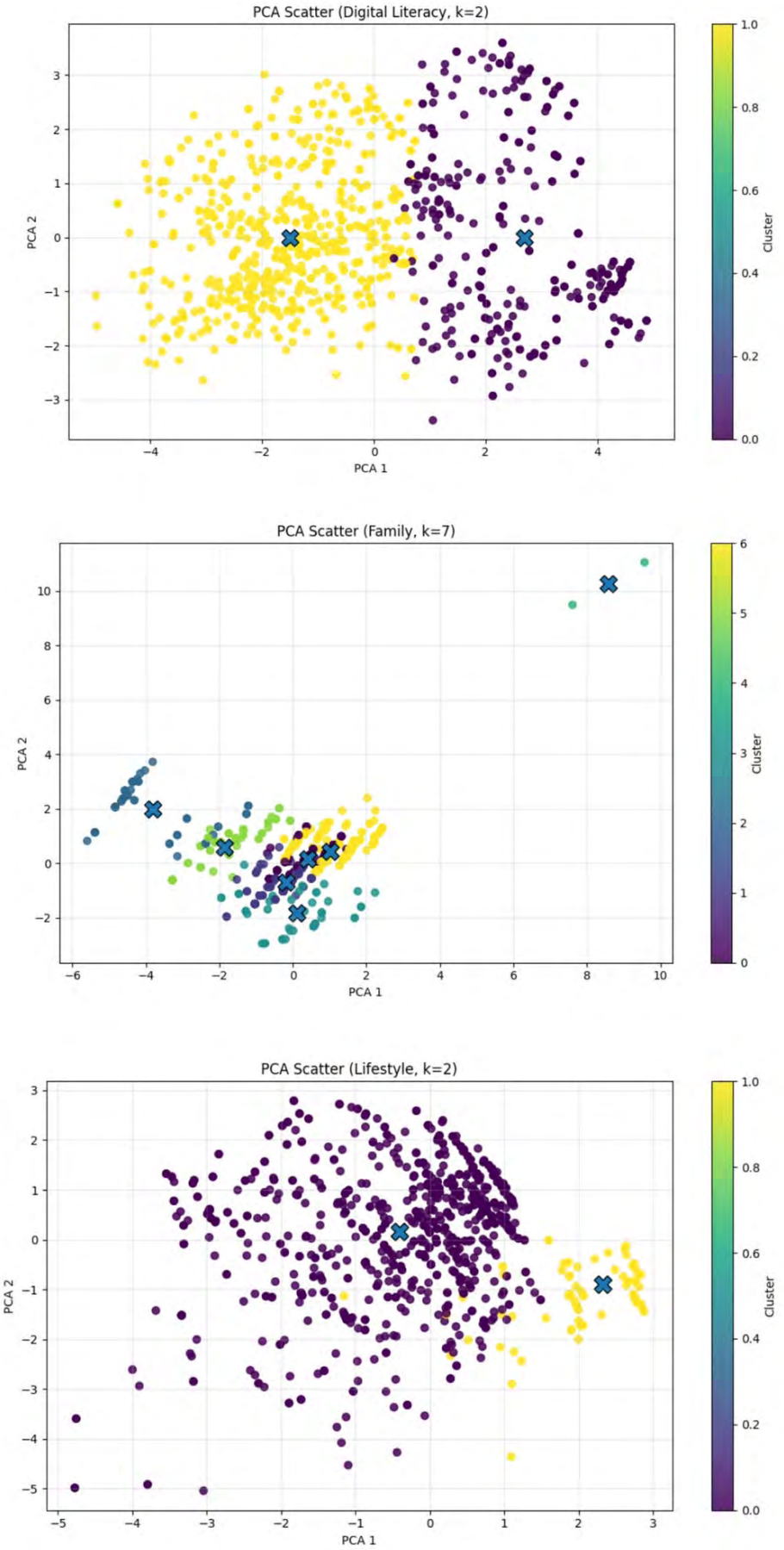


Figure 4.5: PCA Scatter Plots for Digital Literacy, Family and Lifestyle Variables

4.3.4 Summary of Preprocessed Data

After thorough data cleaning and feature engineering, the dataset was still quite dense. The clustering process then reduced dataset complexity substantially, grouping correlated variables into meaningful profiles and improving interpretability. Overall, this procedure reduced the original feature space from 100 variables to 85. The final dataset therefore had 1,001 rows and 85 columns with only numeric and scaled values between 0 and 1, forming a robust foundation for subsequent analysis and modeling.

4.4 Implementation of Selected Models

Each of the four supervised machine learning models, namely Random Forest, SVM, XGBoost and LightGBM was trained and tested on five different target variables, each acting as an indicator of microfinance effectiveness. These target variables consisted of binary, yes or no questions of whether the microfinance received has resulted in an increase in disposable income, improved living standards, caused an expansion in business, increased savings, or wealth accumulation. These five target variables were selected as measurable indicators of economic empowerment. We chose multiple target variables since microcredit can have varying impacts on different people, depending on how it is utilized. We tried to look into it from different perspectives. For example, people view success differently; while some believe increased income is the best output one can have from microfinance, others may believe better savings lead to a more secure future, making that the best indicator of a successful loan. Moreover, some of the target variables, like increased disposable income and increased savings, showcase short-term effects of the microcredit, whereas others, like expansion of business or wealth accumulation, point towards more long-term effects. Also, instead of looking into it only from a monetary point of view, we tried to take into account other perspectives as well. For instance, improved living standards refer to an overall improvement in the quality of life, going beyond the basic financial status of the participants. Together, they provide a comprehensive view of how microfinance participation influences both short-term financial behavior and long-term wealth outcomes. For each target variable, a separate classification task was formulated.

The remaining variables in the dataset, including demographic attributes and the constructed composite features such as digital literacy, family structure and lifestyle, were used as input predictors. The final design for this study consisted of five independent experimental setups per model. To ensure fair model comparison, the same training and testing procedure was applied across all four algorithms. Stratified sampling was performed to preserve the original class distribution of each target variable, particularly in cases where class imbalance was present. This approach minimized bias and ensured that minority classes were adequately represented during training and evaluation.

For each model, hyperparameter tuning was performed. For example, for the model Random Forest, parameters such as the number of trees and maximum tree depth were adjusted. For SVM, kernel type and regularization parameters were tuned.

XGBoost and LightGBM were optimized using 3 factors, which included learning rate, maximum depth and number of estimators. This tuning process helped prevent overfitting and improved the generalization on unseen data.

Random Forest Classifier Configuration:

For the Random Forest ensemble, hyperparameter tuning was strictly focused on variance reduction. This was to mitigate the risk of overfitting inherent in high-dimensional survey data. We utilized 100 estimators. This quantity was sufficient to stabilize the error rate via the Law of Large Numbers without incurring the diminishing returns and computational latency associated with larger ensembles.

Crucially, the tree structure was constrained using a maximum depth of 8 and a minimum number of samples per leaf of 5. This pruning strategy was chosen over deeper trees because unconstrained random forests tend to memorize specific noise in the training set. Thus, this created rules for individual respondents rather than general socio-economic trends. A depth of 8 allows the model to capture complex interactions, but also prevents it from getting too noisy. On the other hand, we ensured that every final classification decision is supported by at least 5 distinct data points, enforcing the statistical significance. Finally, to address class imbalance, we employed class_weight='balanced_subsample'. This reweighed the classes based on the bootstrap sample for each specific tree and also ensured that the model does not bias its predictions toward the majority class simply due to frequency.

Support Vector Machine (SVM) Configuration

As the baseline classifier, the Support Vector Machine was configured to handle the non-linear boundaries of socio-economic classification. Since the relationship between variables like Digital Literacy and Loan Repayment is rarely linear, the Radial Basis Function was utilized. The regularization parameter C was set to 1.0. A higher value of C would have forced the optimizer to strictly classify every training point correctly. This would lead to a hard margin that overfits outliers.

On the contrary, a lower value of C would create a margin too wide, causing it to lose discriminative power. The value of 1.0 represents an optimal soft margin compromise, which allows for some misclassification in the training phase. This is to preserve generalization for unseen data. Furthermore, we enabled probability=True using the Platt scaling to generate probability scores required for ROC-AUC analysis. Additionally, class_weight='balanced' was applied, which automatically adjusts the margin penalties. This is inversely proportional to class frequencies, which prevents the decision boundary from being pushed too far into the minority class region. 5-fold cross-validation was used on all the models to detect possible overfitting. The model performance was evaluated using standard classification metrics, which included accuracy, precision, recall, F1-score and confusion matrices. These metrics were computed separately for each of the five target variables. In addition, feature importance scores were extracted from the tree-based models to recognize the most influential predictors contributing to each outcome variable. The experimental design allowed both vertical and horizontal analysis. The former compared

the performance of the different models on the same target variable, while the latter compared how a single model performed across different impact indicators. This dual perspective provides a more comprehensive understanding of which algorithms were most suitable for modeling specific aspects of microfinance outcomes.

To sum up, this implementation framework provided a reproducible approach for assessing the predictive capability of multiple machine learning models across diverse socio-economic targets. The study ensures a robust evaluation of model performance and generates meaningful insights into the impact of microcredit borrowers' financial and social well-being.

XGBoost (Extreme Gradient Boosting) Configuration

A slow learning strategy was prioritized for the configuration of XGBoost to maximize generalization capability. We implemented a combination of a high number of boosting rounds, where we had 600 estimators and a low learning rate of 0.05. This approach ensures that the model learns incremental and subtle patterns in the residuals instead of making aggressive corrections that could overshoot the optimal minimum. However, this often occurs with higher rates of 0.1 or 0.2. To control model complexity, the maximum depth of each tree was restricted to 4. Unlike Random Forest, boosting relies on combining many weak learners. Here, a shallow depth of 4 prevents individual trees from becoming too dominant, hence forcing the algorithm to rely on the ensemble effect. Furthermore, we set the minimum child width to 3, which acts as a regularizer. This works by requiring a minimum sum of instance weights in a child node and working with values lower than this would allow the model to isolate outliers. On the other hand, using higher values might underfit the model. Additionally, a class weight parameter is used for handling class imbalance. Furthermore, stochasticity was introduced via subsampling with the value 0.85 and column sampling with the value 0.85. This means each tree is trained on a random 85% subset of data and features. This decorrelates the trees and thus prevents the model from relying too heavily on any single feature and making it robust to noise.

LightGBM (Light Gradient Boosting Machine) Configuration

LightGBM relies on a leaf-wise growth strategy and can be prone to overfitting on smaller datasets. Subsequently, hyperparameter selection focused heavily on capacity control. We utilized 800 estimators with an even more conservative learning rate of 0.03. The critical parameter, number of leaves, was set to 31. Since LightGBM grows complex trees, restricting leaves to 31, which is roughly equivalent to a depth of 5 in a balanced tree ($2^5 \approx 32$), prevents the model from creating overly specific branches. These branches capture noise rather than signal. We explicitly set minimun child samples to 20, ensuring that no leaf node is created unless it represents at least 20 borrowers. This is a vital constraint for rural survey data since allowing smaller leaf sizes would result in the model learning the idiosyncrasies of specific villages or families rather than generalizable economic behaviors. Similar to XGBoost, we employed a class weight parameter for handling class imbalance.

Chapter 5

Result Analysis

This chapter presents the experimental results and subsequent technical analysis of the machine learning framework designed to assess the impact of microfinance.

5.1 Performance Evaluation

The performance of four supervised machine learning models: LightGBM, XGBoost, Random Forest, and SVM, is evaluated across five distinct socio-economic target variables. The dataset, consisting of 1,001 datapoints, was divided into 600 training samples, 200 test samples, and 201 validation samples to ensure robust model training, evaluation, and generalization.

5.1.1 Increase in Disposable Income

The first target variable is an obvious one, Increase in Disposable Income, representing whether borrowers were able to generate surplus income after meeting essential household expenses. This serves as a direct indicator of short-term financial improvement resulting from the microfinance taken. The dataset consists of 468 positive cases (46.8%) and 533 negative cases (53.2%). A total of 84 numeric features were used against this target variable.

LightGBM achieved comparatively strong generalization on unseen data. On the test set, it recorded an accuracy of 0.7800, precision of 0.7579, recall of 0.7742, F1-score of 0.7660, and ROC-AUC of 0.8633. On the independent validation set, LightGBM outperformed other models with an accuracy of 0.8955, precision of 0.8687, recall of 0.9149, F1-score of 0.8912, and ROC-AUC of 0.9318. These results indicate that LightGBM generalized better than the other models and achieved the most balanced performance across precision and recall. The model achieved the highest correct classifications with 88 true negatives and 73 true positives, while producing only 19 false positives and 20 false negatives.

Next, the XGBoost model also demonstrated strong predictive capability. During cross-validation, it achieved an accuracy of 0.8841 ± 0.0259, precision of 0.8757 ± 0.0402, recall of 0.8802 ± 0.0678, F1-score of 0.8756 ± 0.0325, and ROC-AUC of 0.9451 ± 0.0168. Although the train-test accuracy gap of 0.1145 suggested some

overfitting, XGBoost consistently produced the highest discriminative power among all models during cross-validation, indicating strong ability to separate borrowers who experienced income growth from those who did not. The model correctly identified 84 non-income-increase cases and 72 income-increase cases, with 23 false positives and 21 false negatives.

Random Forest achieved a cross-validation accuracy of 0.8002 ± 0.0356, precision of 0.7425 ± 0.0405, recall of 0.8824 ± 0.0602, F1-score of 0.8049 ± 0.0352, and ROC-AUC of 0.8954 ± 0.0234. Overfitting analysis showed a train-test accuracy gap of 0.1060, indicating mild overfitting. On the independent test set, Random Forest achieved an accuracy of 0.7700, precision of 0.7043, recall of 0.8710, F1-score of 0.7788, and ROC-AUC of 0.8524. Validation performance remained stable with an accuracy of 0.7910 and ROC-AUC of 0.8900, confirming reasonable generalization. The model correctly classified 73 non-income-increase cases and 81 income-increase cases, while misclassifying 34 false positives and 12 false negatives.

Lastly, SVM was applied with a radial basis function (RBF) kernel with parameter C=1.0. During 5-fold stratified cross-validation, SVM achieved an accuracy of 0.8112 ± 0.0416, precision of 0.7707 ± 0.0494, recall of 0.8546 ± 0.0709, F1-score of 0.8084 ± 0.0437, and ROC-AUC of 0.9030 ± 0.0325. The corresponding training scores (accuracy = 0.9348, ROC-AUC = 0.9885) indicate mild overfitting, which is expected due to the high-dimensional feature space and non-linear kernel mapping. On the independent test set, the SVM model achieved an accuracy of 0.7900, precision of 0.7383, recall of 0.8495, F1-score of 0.7900, and ROC-AUC of 0.8570. The confusion matrix showed that the model correctly classified 79 borrowers who experienced income growth and 79 borrowers who did not, while misclassifying 42 cases in total. The relatively high recall demonstrates that SVM was particularly effective at identifying borrowers who benefited from increased disposable income, although its lower precision indicates a higher rate of false positives compared to gradient boosting models. Overall, the SVM provided a strong geometric baseline, confirming that income-related socio-economic patterns are separable in high-dimensional feature space and supporting the robustness of the results obtained from tree-based ensemble models. Table 5.1 shows the summarized table and figure 5.1 shows all the Confusion Matrix and graphs for the target variable "Increase in Disposable Income".

Despite moderate overfitting observed in the boosting models, the use of 5-fold stratified cross-validation and independent test and validation sets confirmed strong generalization performance. All models achieved ROC-AUC values above 0.85, demonstrating that alternative socio-economic and behavioral data can reliably predict income improvement outcomes among rural microfinance borrowers.

The feature importance analysis across Random Forest, XGBoost, and LightGBM models, shown in figure 5.2, reveals a consistent emphasis on loan structure, borrower financial behavior, and experiential feedback variables as the strongest predictors of increased disposable income. In the Random Forest model, the most influential features include loan outcome perceptions (no change), desired changes in interest rate and loan conditions, penalties for repayment failure, and loan size variables such as amount of the first and most recent loan, indicating that borrowers' financial

stress and loan terms strongly shape income outcomes. XGBoost further highlights borrowers' expectations from loan experience, unexpected repayment shocks such as illness or job loss, and loan purpose (health/medical and daily household expenses) as key drivers, suggesting that external vulnerability and consumption priorities significantly affect surplus income generation. LightGBM places dominant weight on agreed weekly interest rate, year and amount of first loan, financial behavior, occupation, family characteristics, and education, reflecting a broader socio-economic perspective that integrates both structural and behavioral dimensions. Collectively, these results show that 'Increase in disposable income' is not driven by a single factor but emerges from the interaction between loan design (interest rate and size), borrower behavior, occupational context, and household responsibilities, confirming the importance of incorporating alternative socio-economic and behavioral data when assessing microfinance impact.

Overall, LightGBM emerged as the best model for the particular target variable due to its highest validation accuracy (0.8955) and ROC-AUC (0.9318), along with balanced precision and recall. XGBoost was a close second, as it had the strongest cross-validation ROC-AUC (0.9451). However, it showed slightly weaker generalization. These findings support that gradient boosting methods are suitable for predicting whether or not a microcredit can lead to income generation, with the help of several behavioral and demographic data.

Table 5.1: Performance Comparison for Target Variable: Increase in Disposable Income

Model	**Split**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	5-Fold CV	0.9062	0.8531	0.9658	0.9059	0.9778
	Validation	0.7910	0.7321	0.8723	0.7961	0.8900
	Test	0.7700	0.7043	0.8710	0.7788	0.8524
XGBoost	5-Fold CV	0.9986	0.9983	0.9986	0.9985	1.0000
	Validation	0.8955	0.8687	0.9149	0.8912	0.9318
	Test	0.7800	0.7579	0.7742	0.7660	0.8633
LightGBM	5-Fold CV	1.0000	1.0000	1.0000	1.0000	1.0000
	Validation	0.8955	0.8687	0.9149	0.8912	0.9467
	Test	0.8050	0.7935	0.7849	0.7892	0.8899
SVM	5-Fold CV	0.9348	0.8940	0.9765	0.9334	0.9885
	Validation	0.7811	0.7232	0.8617	0.7864	0.8707
	Test	0.7900	0.7383	0.8495	0.7900	0.8570

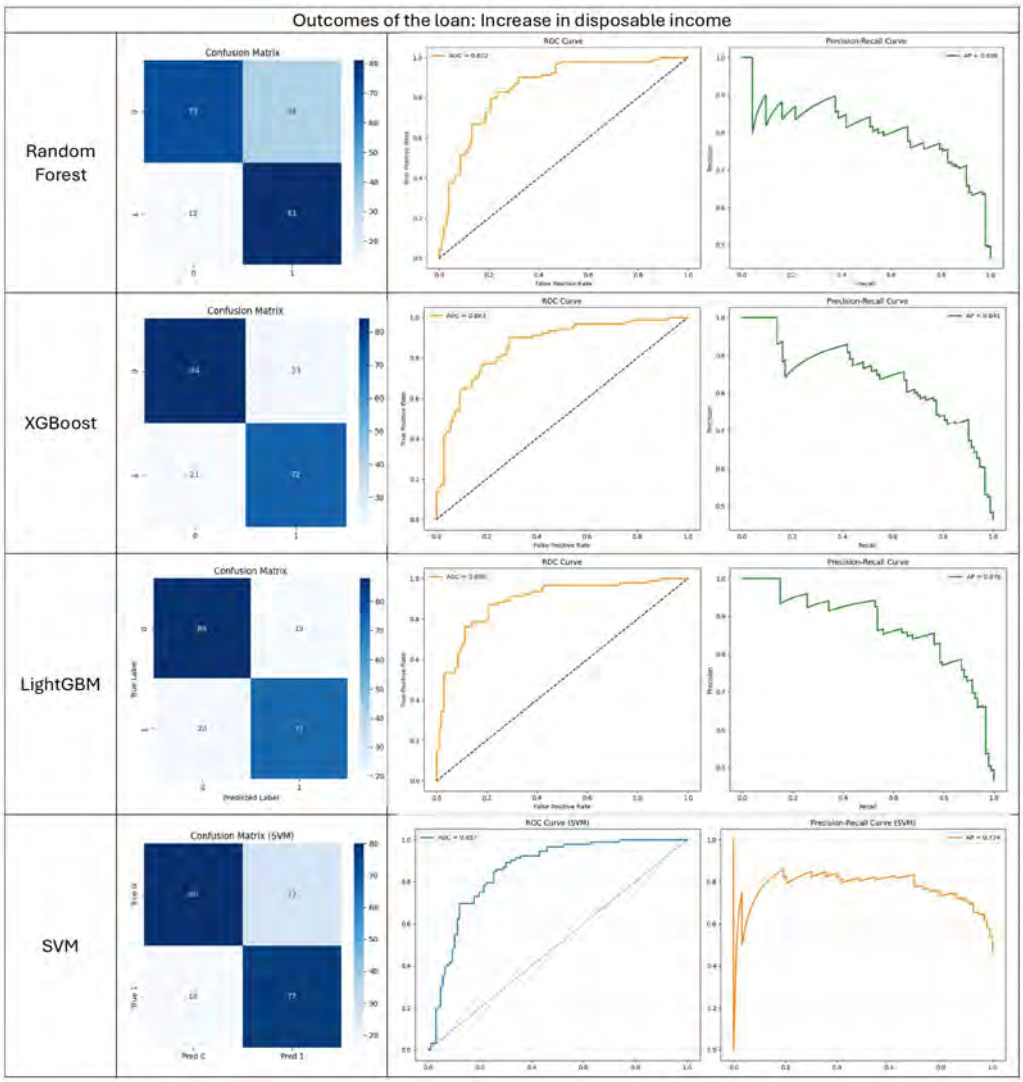


Figure 5.1: Results of Classification of Increase in Disposable Income

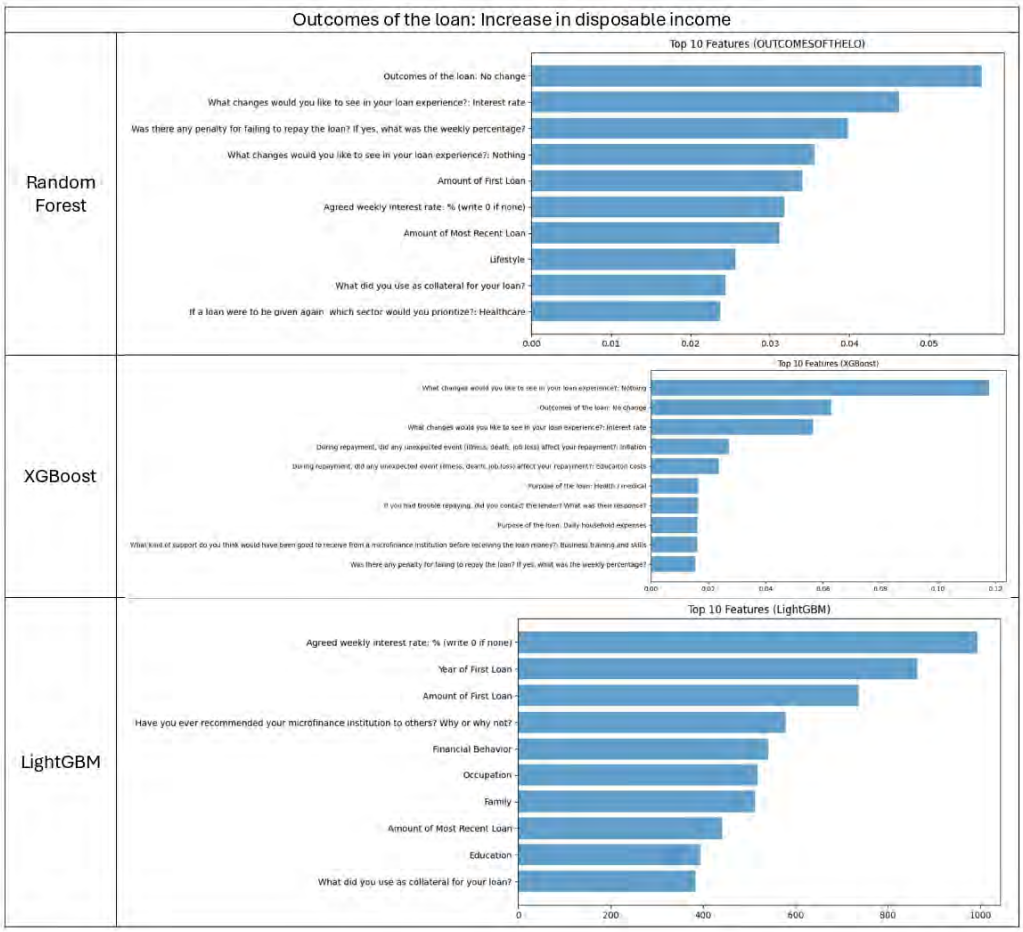


Figure 5.2: Feature Importance determining Increase in Disposable Income

5.1.2 Improvement in living standards

The second target variable, Improved Standard of Living, is more of a non-monetary perspective to look from. It reflects whether microfinance participation contributed to measurable improvements in borrowers' living conditions, including better access to basic needs, enhanced household facilities, and overall quality of life. This goes beyond income generation and dives into the social impact of such microcredit. This variable contained 408 positive cases (40.8%) and 593 negative cases (59.2%), forming a moderately imbalanced but suitable binary classification task.

Here, the LightGBM model showed competitive performance on the validation set. It achieved an accuracy of 0.8159, precision of 0.8000, recall of 0.7317, F1-score of 0.7643, and ROC-AUC of 0.8841. The model maintained balanced precision and recall and demonstrated stable discrimination between improved and non-improved living standards. The confusion matrix reveals that the best balance was achieved here with 107 true negatives and 64 true positives, while yielding only 11 false positives and 18 false negatives.

The XGBoost model demonstrated stronger performance. During cross-validation, XGBoost achieved an accuracy of 0.8701 ± 0.0306, precision of 0.8494 ± 0.0474, recall of 0.8307 ± 0.0392, F1-score of 0.8393 ± 0.0365, and ROC-AUC of 0.9305 ± 0.0241. On the test set, XGBoost reached an accuracy of 0.8300, precision of 0.7857, recall of 0.8049, F1-score of 0.7952, and ROC-AUC of 0.8982. The model correctly identified 100 negative cases and 66 positive cases, with 18 false positives and 16 false negatives.

As for Random Forest, cross-validation produced an accuracy of 0.8143 ± 0.0441, precision of 0.7744 ± 0.0615, recall of 0.7720 ± 0.0788, F1-score of 0.7712 ± 0.0575, and ROC-AUC of 0.8985 ± 0.0317. On the independent test set, it achieved an accuracy of 0.8350, precision of 0.8025, recall of 0.7927, F1-score of 0.7975, and ROC-AUC of 0.9043. Validation performance remained consistent with an accuracy of 0.7811, F1-score of 0.7179, and ROC-AUC of 0.8812, indicating stable generalization. The relatively high recall demonstrates that Random Forest was effective in identifying households that experienced improvements in living standards, although its precision suggests some false positives. The confusion matrix shows that model correctly classified 102 negative cases and 65 positive cases, while producing 16 false positives and 17 false negatives.

Finally, for SVM, 5-fold cross-validation yielded an accuracy of 0.8192 ± 0.0445, precision of 0.7644 ± 0.0653, recall of 0.8113 ± 0.0489, F1-score of 0.7859 ± 0.0497, and ROC-AUC of 0.9017 ± 0.0307. On the independent test set, SVM achieved an accuracy of 0.8050, precision of 0.7654, recall of 0.7561, F1-score of 0.7607, and ROC-AUC of 0.8860. The confusion matrix indicated that the model correctly classified 99 households without improvement and 62 households with improvement, while misclassifying 19 false positives and 20 false negatives, demonstrating a balanced trade-off between sensitivity and specificity. Table 5.2 shows the summarized table and figure 5.3 shows all the Confusion Matrix and graphs for the target variable "Improvements in Living Standards".

Validation analysis revealed mild overfitting in the gradient boosting models, as indicated by train–test accuracy gaps exceeding 0.10; however, the use of 5-fold stratified cross-validation and independent validation sets confirmed strong generalization performance. All four models achieved ROC-AUC values above 0.88, demonstrating that non-financial socio-economic and behavioral data can reliably predict improvements in living standards among rural microfinance borrowers.

The feature importance analysis for Improved Standard of Living across Random Forest, XGBoost, and LightGBM, shown in figure 5.4, highlights the dominant role of loan characteristics, household context, and borrower behavior in shaping quality-of-life outcomes. In the Random Forest model, the most influential predictors include asset acquisition, loan purpose (particularly agriculture), penalties for repayment failure, loan size (amount of first loan), and agreed weekly interest rate, indicating that both productive investment and repayment burden strongly affect living standards. XGBoost places high importance on loan outcome perceptions (no change), loan purpose such as foreign travel or agriculture, asset acquisition, and delay in loan disbursement, suggesting that how loans are used and external constraints during repayment significantly influence household welfare. LightGBM emphasizes more structural and socio-demographic variables, ranking weekly interest rate, year and amount of first loan, family characteristics, financial behavior, education, and occupation among the top predictors, reflecting a broader interaction between financial terms and household capacity to translate credit into improved living conditions. Collectively, these models show that improvements in standard of living are driven not only by access to credit itself, but by the combination of loan design, usage purpose, repayment experience, and the borrower's socio-economic environment, underscoring the importance of incorporating alternative behavioral and contextual data when evaluating the social impact of microfinance.

To conclude, XGBoost was the best-performing model for this target variable, achieving the highest cross-validation ROC-AUC of 0.9305 and a strong test F1-score of 0.7952. It also achieved a balanced precision and recall. LightGBM ranked closely behind. These results reinforce what we had already seen for the first target variable. Gradient boosting is particularly suitable for a dataset like the one in this study.

Table 5.2: Performance Comparison for Target Variable: Improved Standard of Living (BG3)

Model	**Evaluation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
LightGBM	5-Fold Cross Validation	0.8851	0.8710	0.8453	0.8567	0.9475
	Independent Validation	0.8209	0.8026	0.7439	0.7722	0.8959
	Test Set	0.8550	0.8533	0.7805	0.8153	0.9256
Random Forest	5-Fold Cross Validation	0.9174	0.8906	0.9090	0.8997	0.9801
	Independent Validation	0.7811	0.7568	0.6829	0.7179	0.8812
	Test Set	0.8350	0.8025	0.7927	0.7975	0.9043
XGBoost	5-Fold Cross Validation	0.9982	0.9978	0.9978	0.9978	0.9998
	Independent Validation	0.8159	0.8000	0.7317	0.7643	0.8841
	Test Set	0.8300	0.7857	0.8049	0.7952	0.8982
SVM	5-Fold Cross Validation	0.9408	0.9138	0.9439	0.9286	0.9882
	Independent Validation	0.8192	0.7644	0.8113	0.7859	0.9017
	Test Set	0.8050	0.7654	0.7561	0.7607	0.8860

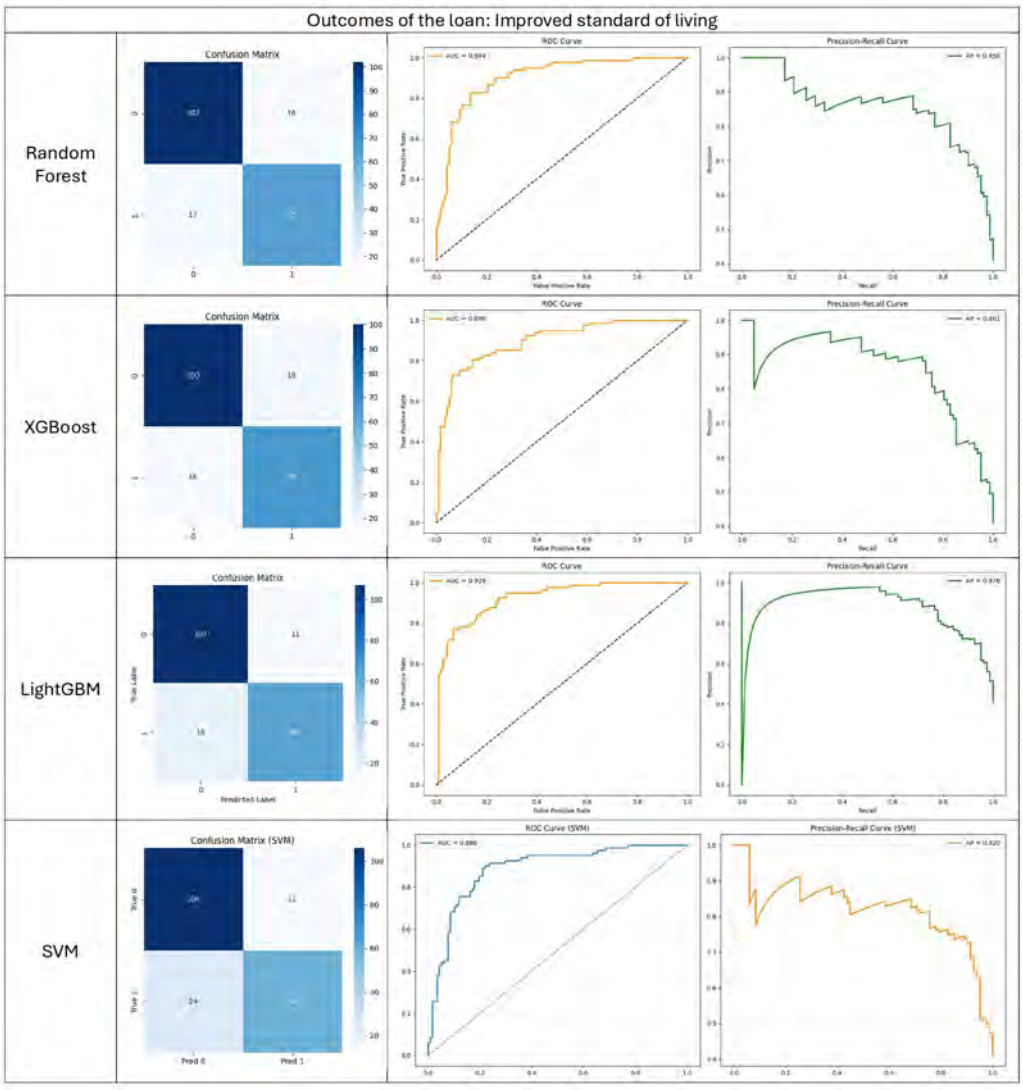


Figure 5.3: Results of Classification of Improved Standard of Living

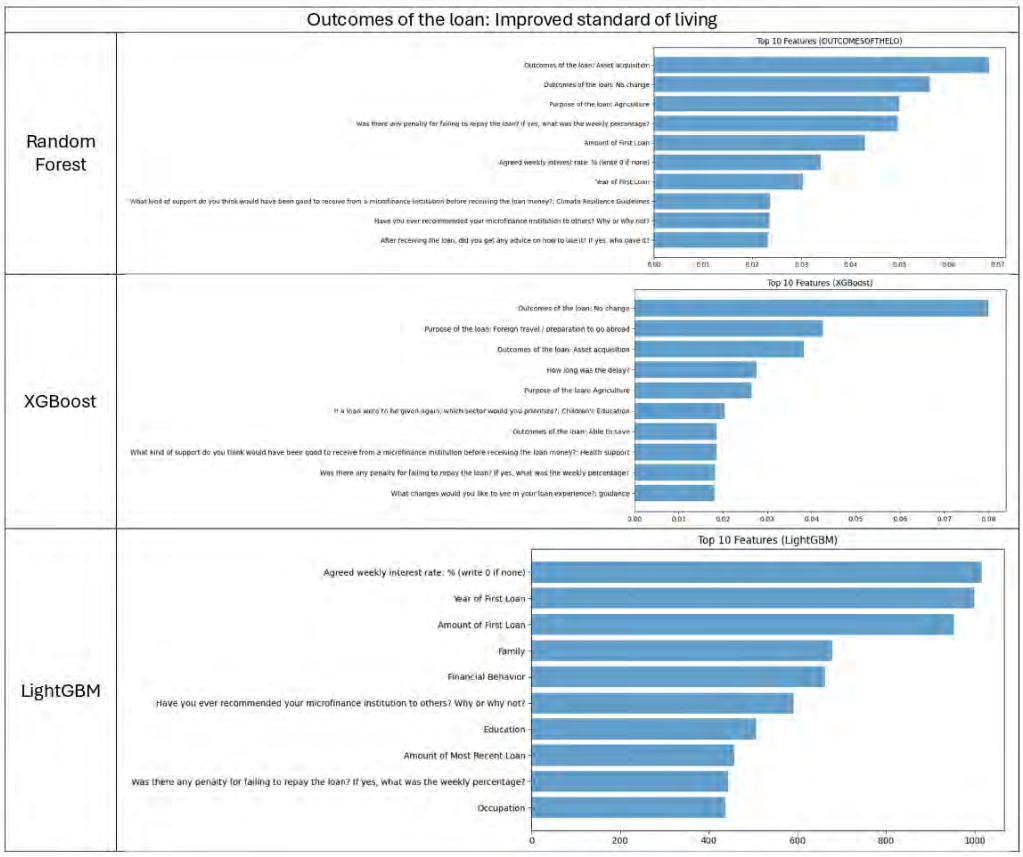


Figure 5.4: Feature Importance determining Improved Standard of Living

5.1.3 Expansion in business

The third target variable, Business Expansion, measures whether borrowers were able to expand their income-generating activities after receiving microfinance loans. This indicator reflects productive utilization of credit rather than short-term consumption. It is a crucial measure of sustainable microfinance impact, as business expansion suggests a generation of long-term income and employment opportunities. The dataset contained 396 positive cases (39.6%) and 605 negative cases (60.4%) for this variable, resulting in a moderately imbalanced binary classification problem.

Here, LightGBM again showed strong generalization performance on the validation set with an accuracy of 0.8706, F1-score of 0.8354, and ROC-AUC of 0.9241. These results indicate that LightGBM maintained balanced precision and recall. The model showed the best overall balance, correctly identifying 114 non-expansion and 65 expansion cases, with the lowest false positives (7) and only 14 false negatives.

Similar to the previous target variables, XGBoost demonstrated stronger predictive capability, achieving cross-validation accuracy of 0.8951 ± 0.0294, F1-score of 0.8689 ± 0.0358, and ROC-AUC of 0.9513 ± 0.0231. On the test set, XGBoost recorded an accuracy of 0.8800, F1-score of 0.8400, and ROC-AUC of 0.9400, confirming its ability to model non-linear socio-economic relationships related to entrepreneurial growth. The model achieved the highest correct negative classifications with 113 true negatives and 63 true positives, producing only 8 false positives and 16 false negatives.

Random Forest, in this case, achieved cross-validation accuracy of 0.8252 ± 0.0229 with an F1-score of 0.7930 ± 0.0254 and ROC-AUC of 0.9112 ± 0.0247. On the test set, it achieved an accuracy of 0.8500 and ROC-AUC of 0.9205, while validation performance remained stable with an F1-score of 0.7630 and ROC-AUC of 0.8892. This indicates that Random Forest was effective at identifying borrowers who expanded their businesses, though with moderate false positives. The model correctly classified 104 non-expansion cases and 66 expansion cases, while misclassifying 17 false positives and 13 false negatives.

Additionally, SVM, with an RBF kernel achieved cross-validation accuracy of 0.8344 ± 0.0277, precision of 0.7500 ± 0.0385, recall of 0.8764 ± 0.0432, F1-score of 0.8075 ± 0.0315, and ROC-AUC of 0.9239 ± 0.0326. On the test set, SVM achieved an accuracy of 0.8010, precision of 0.7128, recall of 0.8375, F1-score of 0.7701, and ROC-AUC of 0.8769. The high recall indicates that SVM was particularly effective at identifying borrowers who experienced business expansion, making it a strong geometric baseline model. The model correctly classified 103 non-expansion and 60 expansion cases, but produced higher misclassifications with 18 false positives and 19 false negatives compared to the boosting models. Table 5.3 shows the summarized table and figure 5.5 shows all the Confusion Matrix and graphs for the target variable "Expansion in Buisness".

The feature importance analysis for Business Expansion, shown in figure 5.6, demonstrates that entrepreneurial growth is primarily influenced by loan purpose, financial structure, and borrower socio-economic behavior across all three tree-based models.

In the Random Forest model, the most dominant predictor is purpose of the loan (business), followed by no change in loan outcome, ability to save, and year of first loan, indicating that business-oriented credit use and prior financial trajectory strongly determine expansion outcomes. XGBoost further emphasizes loan purpose such as foreign travel or business, unexpected repayment shocks (illness, job loss), and lack of institutional support before receiving the loan, highlighting the role of external vulnerability and institutional guidance in shaping entrepreneurial success. LightGBM places the greatest weight on agreed weekly interest rate, year and amount of first loan, and borrower characteristics such as family structure, financial behavior, occupation, and age, reflecting the interaction between financial burden and household capacity to reinvest capital productively. Across all models, business expansion emerges as a multidimensional outcome driven not only by the amount of credit received but also by how the loan is used, the borrower's financial discipline, occupational stability, and resilience to repayment shocks, underscoring the importance of integrating behavioral and contextual variables when evaluating sustainable microfinance impact.

Overall, XGBoost was the best-performing model for this target variable, achieving the highest ROC-AUC (0.9513) and a strong test F1-score (0.8400). LightGBM followed closely with excellent validation performance, while SVM provided a robust baseline with high recall, and Random Forest contributed interpretability through feature importance analysis.

Table 5.3: Performance Comparison for Target Variable: Business Expansion

Model	**Split (Size)**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	5-fold CV	0.9156	0.8477	0.9593	0.9000	0.9772
	Test (20%)	0.8500	0.7952	0.8354	0.8148	0.9205
	Independent Validation	0.7960	0.7097	0.8250	0.7630	0.8892
XGBoost	5-fold CV	0.9991	0.9978	1.0000	0.9989	1.0000
	Test (20%)	0.8800	0.8873	0.7975	0.8400	0.9400
	Independent Validation	0.8706	0.8462	0.8250	0.8354	0.9241
LightGBM	5-fold CV	0.9041	0.8761	0.8840	0.8794	0.9586
	Test (20%)	0.8950	0.9028	0.8228	0.8609	0.9538
	Independent Validation	0.8607	0.8421	0.8000	0.8205	0.9219
SVM	5-fold CV	0.9338	0.8672	0.9834	0.9216	0.9913
	Test (20%)	0.8150	0.7692	0.7595	0.7643	0.9135
	Independent Validation	0.7910	0.7021	0.8250	0.7586	0.8775

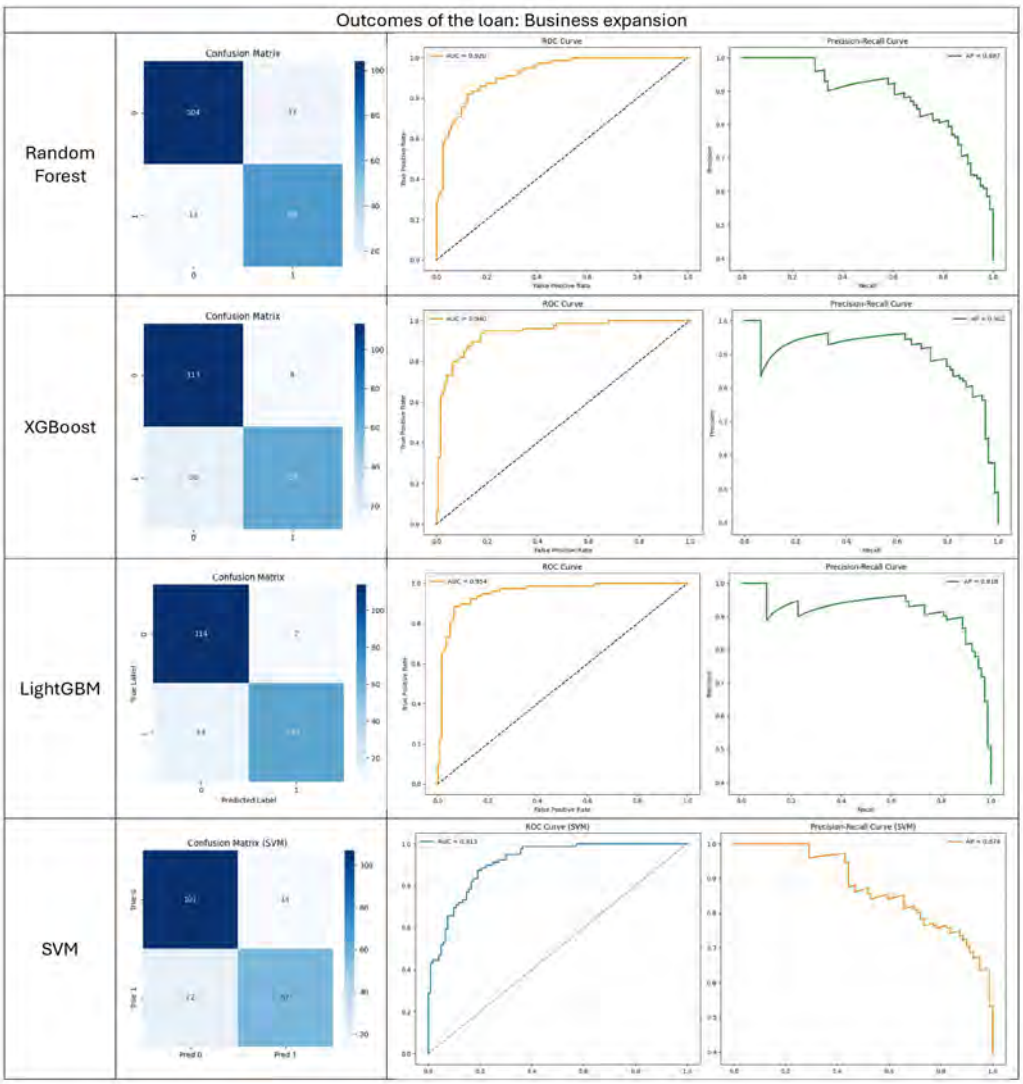


Figure 5.5: Results of Classification of Business Expansion

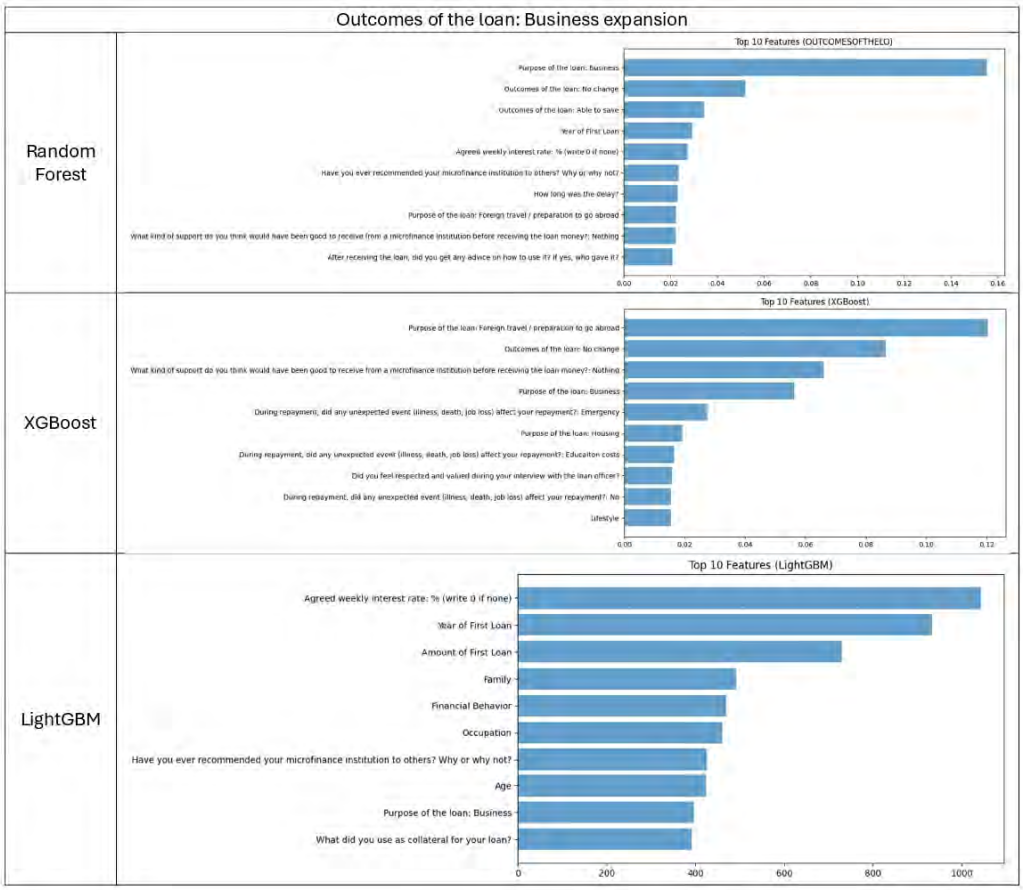


Figure 5.6: Feature importance determining Business Expansion

5.1.4 Ability to Save

The fourth target variable, Able to Save, represents whether borrowers were able to accumulate savings as a result of the microfinance intervention. This variable also serves as a key indicator of financial stability and long-term asset building. The dataset contained 350 positive cases (35.0%) and 651 negative cases (65.0%) for this variable, resulting in a moderately imbalanced dataset.

During 5-fold stratified cross-validation, LightGBM achieved an accuracy of 0.8851 $\pm$ 0.0277, precision of 0.8393 $\pm$ 0.0393, recall of 0.8314 $\pm$ 0.0645, F1-score of 0.8342 $\pm$ 0.0433, and ROC-AUC of 0.9507 $\pm$ 0.0294, indicating strong discriminative power. Although the training performance reached near-perfect values, a train–test accuracy gap of 0.1149 suggested mild overfitting, which was mitigated through regularization and independent validation. On the independent test set, LightGBM achieved an accuracy of 0.8300, precision of 0.7500, recall of 0.7714, F1-score of 0.7606, and ROC-AUC of 0.8914, demonstrating stable generalization. The confusion matrix showed that the model correctly classified 112 non-saving borrowers (true negatives) and 54 saving borrowers (true positives), while producing 18 false positives and 16 false negatives. Overall, these results indicate that LightGBM effectively captures non-linear relationships between loan structure, financial behavior, and socio-economic variables associated with borrowers' ability to accumulate savings, making it a strong predictive model for assessing financial resilience among rural microfinance participants.

XGBoost exhibited superior predictive power, achieving cross-validation accuracy of 0.8831 $\pm$ 0.0311, F1-score of 0.8319 $\pm$ 0.0482, and ROC-AUC of 0.9408 $\pm$ 0.0320. On the test set, XGBoost obtained an accuracy of 0.8000, F1-score of 0.7333, and ROC-AUC of 0.9003, confirming its effectiveness in modeling non-linear socio-economic patterns associated with saving behavior. The model achieved high correct classifications with 105 true negatives and 55 true positives, while generating 25 false positives and 15 false negatives.

Random Forest achieved moderate but consistent performance, with cross-validation accuracy of 0.8241 $\pm$ 0.0348 and an F1-score of 0.7385 $\pm$ 0.0590. On the test set, it reached an accuracy of 0.7750 and ROC-AUC of 0.8622, while validation performance remained stable with an F1-score of 0.7000 and ROC-AUC of 0.8485. The model correctly identified 103 non-saving and 52 saving cases but produced relatively higher misclassifications, including 27 false positives and 18 false negatives, indicating weaker discrimination compared to boosting-based models.

During 5-fold stratified cross-validation, SVM achieved an accuracy of 0.8132 $\pm$ 0.0362, precision of 0.7125 $\pm$ 0.0448, recall of 0.7829 $\pm$ 0.0878, F1-score of 0.7440 $\pm$ 0.0561, and ROC-AUC of 0.8874 $\pm$ 0.0354, indicating stable discriminative performance with balanced sensitivity and specificity. On the independent test set, the SVM model obtained an accuracy of 0.7700, precision of 0.6395, recall of 0.7857, F1-score of 0.7051, and ROC-AUC of 0.8671. The confusion matrix showed that the model correctly classified 99 non-saving borrowers (true negatives) and 55 saving borrowers (true positives), while producing 31 false positives and 15 false negatives. The relatively high recall demonstrates that SVM was effective at identifying bor-

rowers who developed saving behavior, although its lower precision indicates a higher rate of false positives compared to gradient boosting models. Overall, the SVM provided a strong geometric baseline, confirming that saving behavior can be separated in high-dimensional socio-economic feature space using non-linear decision boundaries, even though ensemble-based methods achieved superior overall performance. Table 5.4 shows the summarized table and figure 5.7 shows all the Confusion Matrix and graphs for the target variable "Ability to save".

The feature importance analysis for Ability to Save, shown in figure 5.8, indicates that savings behavior is primarily shaped by loan structure, financial discipline, and broader lifestyle conditions across all three tree-based models. In the Random Forest model, the most influential predictors include loan outcome perceptions (no change), agreed weekly interest rate, amount of first loan, asset acquisition, and business expansion, suggesting that both loan burden and prior financial progress strongly determine whether borrowers can generate surplus income for savings. XGBoost highlights loan outcome (no change), asset acquisition, borrower expectations from loan experience, and lifestyle factors, along with education costs during repayment and digital literacy, reflecting the role of financial shocks and financial awareness in shaping saving capacity. LightGBM places dominant weight on year and amount of first loan, weekly interest rate, and financial behavior, followed by occupation, family structure, education, and delay in loan disbursement, showing that long-term engagement with microfinance and household responsibilities significantly influence saving outcomes. Collectively, these findings demonstrate that the ability to save emerges from the interaction of loan size and interest conditions, behavioral discipline, socio-economic stability, and prior improvement in living standards, reinforcing the importance of incorporating behavioral and contextual variables when evaluating financial resilience among rural microfinance borrowers.

Overall, XGBoost was the best-performing model for this target variable, achieving the highest ROC-AUC (0.9408) and strong F1-scores across cross-validation and test sets. LightGBM followed closely with robust validation performance, while SVM offered high recall as a reliable baseline model, and Random Forest contributed interpretability through feature importance analysis. These findings demonstrate that boosting models are particularly suitable for capturing the complex socio-economic factors underlying saving behavior among microfinance borrowers.

Table 5.4: Performance Comparison for Target Variable: Ability to Save

Model	**Split**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	5-Fold CV	0.9477	0.9232	0.9279	0.9255	0.9881
	Validation	0.7910	0.7000	0.7000	0.7000	0.8485
	Test	0.7750	0.6582	0.7429	0.6980	0.8622
XGBoost	5-Fold CV	0.9996	0.9987	1.0000	0.9994	1.0000
	Validation	0.8010	0.7027	0.7429	0.7222	0.8885
	Test	0.8000	0.6875	0.7857	0.7333	0.9003
LightGBM	5-Fold CV	1.0000	1.0000	1.0000	1.0000	1.0000
	Validation	0.8358	0.7606	0.7714	0.7660	0.9013
	Test	0.8300	0.7500	0.7714	0.7606	0.8914
SVM	5-Fold CV	0.8132	0.7125	0.7829	0.7440	0.8874
	Validation	0.7463	0.6118	0.7429	0.6710	0.8399
	Test	0.7700	0.6395	0.7857	0.7051	0.8671

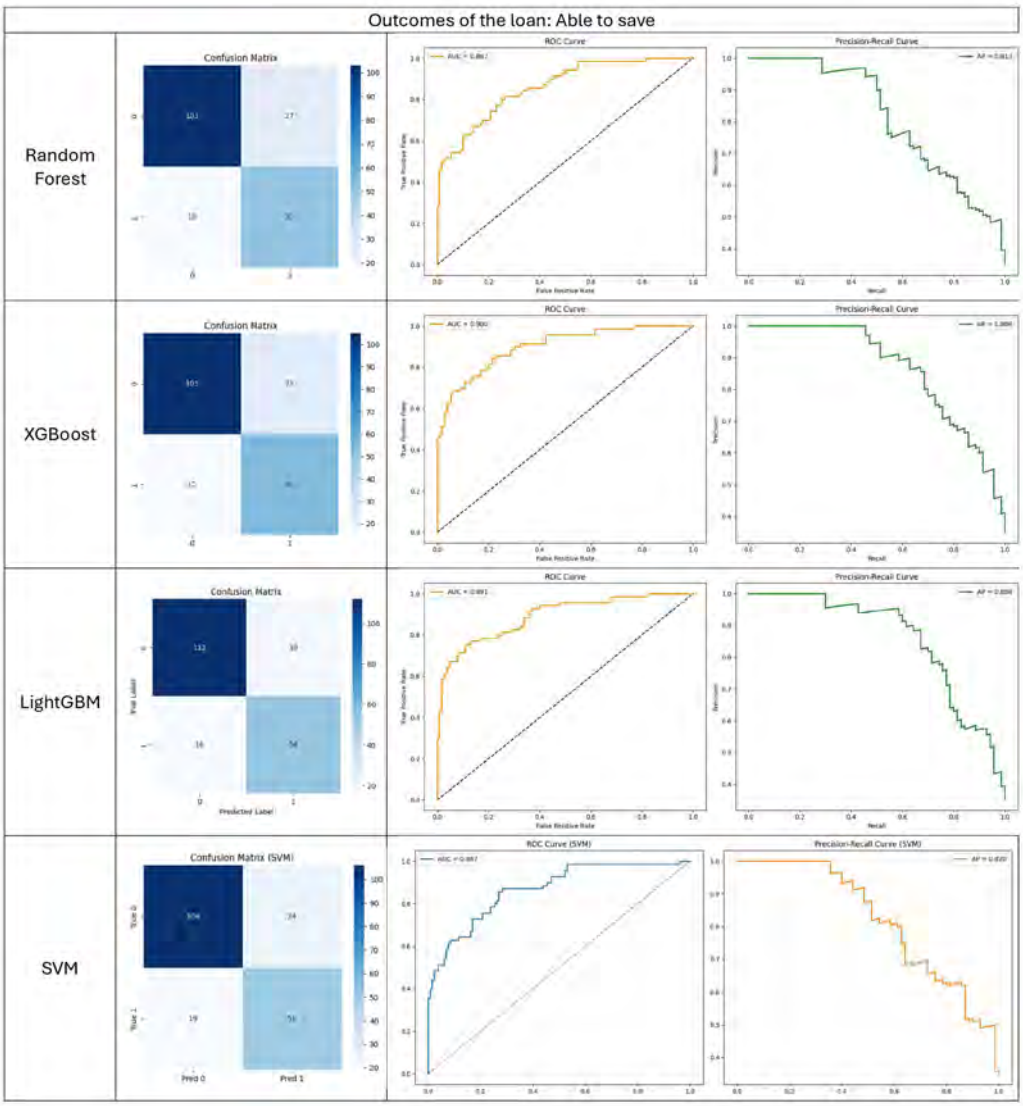


Figure 5.7: Results of Classification of Ability to Save

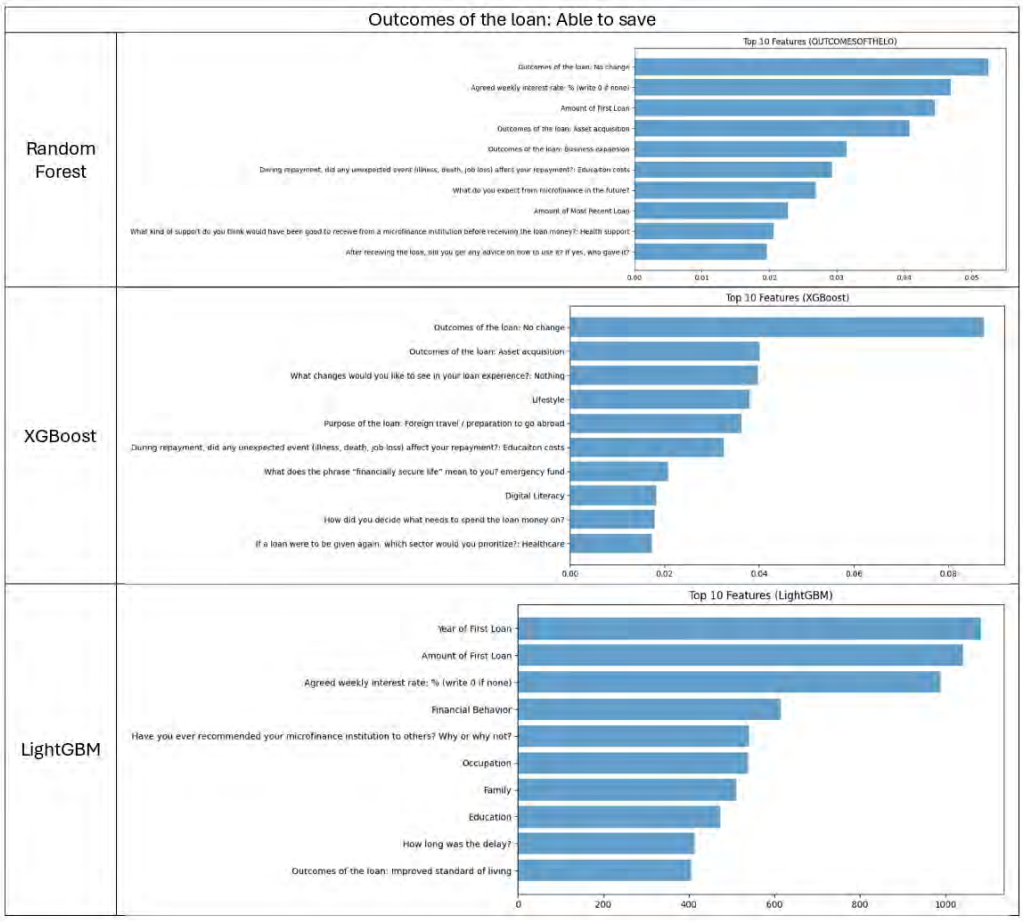


Figure 5.8: Feature importance determining Ability to Save

5.1.5 Asset Acquisition

LightGBM demonstrated strong predictive performance for the Asset Acquisition target. During 5-fold stratified cross-validation, it achieved a mean accuracy of 0.9994, precision of 0.9975, recall of 1.0000, F1-score of 0.9987, and a perfect ROC–AUC of 1.0000. On the independent validation set, LightGBM achieved an accuracy of 0.8507 and a precision of 0.6522. On the test set, the model maintained stable generalization performance with an accuracy of 0.8600, precision of 0.7105, recall of 0.6136, F1-score of 0.6585, and ROC–AUC of 0.8864. Although near-perfect cross-validation performance resulted in a noticeable train–test accuracy gap, the consistently high ROC–AUC values indicate that LightGBM effectively captured nonlinear relationships between loan structure and socio-economic capacity in predicting asset acquisition outcomes.

XGBoost exhibited strong overall performance for this target variable. During 5-fold cross-validation, it achieved an accuracy of 0.9990, precision of 0.9955, recall of 1.0000, F1-score of 0.9977, and ROC–AUC of 0.9995, indicating excellent discriminative power. On the independent validation set, XGBoost recorded an accuracy of 0.8557 and a precision of 0.6596. On the test set, it achieved an accuracy of 0.8750, precision of 0.7436, recall of 0.6591, F1-score of 0.6988, and ROC–AUC of 0.8888, confirming robust generalization and reliable identification of borrowers capable of long-term capital investment. These results highlight XGBoost's effectiveness in modeling complex nonlinear socio-economic patterns related to asset formation.

Random Forest also demonstrated strong performance while providing interpretability through feature importance analysis. During 5-fold cross-validation, it achieved an accuracy of 0.9544 and ROC–AUC of 0.9896. On the validation set, Random Forest recorded an accuracy of 0.8607, while on the test set it achieved an accuracy of 0.8550, precision of 0.7273, recall of 0.5455, F1-score of 0.6234, and ROC–AUC of 0.9021. Although its recall was lower than that of the boosting models, Random Forest maintained stable performance under class imbalance and effectively identified borrowers who converted loans into tangible assets.

SVM with an RBF kernel served as a geometric baseline model for comparison. During 5-fold cross-validation, it achieved an accuracy of 0.9661, precision of 0.8837, recall of 0.9754, F1-score of 0.9272, and ROC–AUC of 0.9936. On the validation set, SVM achieved an accuracy of 0.8607. On the test set, it recorded an accuracy of 0.8850, precision of 0.7561, recall of 0.7045, F1-score of 0.7294, and ROC–AUC of 0.9151, demonstrating strong discriminative capability between borrowers who acquired assets and those who did not. The high ROC–AUC confirms that asset acquisition patterns are separable in high-dimensional feature space and are not artifacts of tree-based models alone. Table 5.5 summarizes the results, while Figure 5.9 presents the corresponding confusion matrices and ROC curves.

Across all models, mild overfitting was observed in boosting-based approaches due to near-perfect cross-validation metrics. However, consistently high test ROC–AUC values (ranging from 0.8864 to 0.9151) and stable test accuracies (0.8550–0.8850) confirm strong generalization performance. The use of 5-fold stratified cross-validation alongside independent validation and test sets ensures that these

results reflect genuine underlying patterns rather than artifacts of data splitting.

Overall, XGBoost achieved the strongest balance between predictive accuracy and generalization for Asset Acquisition, with the highest test accuracy (0.8750) and competitive ROC–AUC (0.8888). LightGBM followed closely with comparable test performance, while SVM provided a strong geometric baseline and Random Forest contributed interpretability through feature importance analysis. These findings confirm the effectiveness of ensemble and boosting frameworks in predicting long-term capital investment outcomes in microfinance contexts.

Table 5.5: Performance Comparison for Target Variable: Asset Acquisition

Model	**Split**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	5-Fold CV	0.9544	0.8861	0.9105	0.8980	0.9896
	Validation	0.8607	0.7000	0.6364	0.6667	0.8800
	Test	0.8550	0.7273	0.5455	0.6234	0.9021
XGBoost	5-Fold CV	0.9990	0.9955	1.0000	0.9977	0.9995
	Validation	0.8557	0.6596	0.7045	0.6813	0.8690
	Test	0.8750	0.7436	0.6591	0.6988	0.8888
LightGBM	5-Fold CV	0.9994	0.9975	1.0000	0.9987	1.0000
	Validation	0.8507	0.6522	0.6818	0.6667	0.8631
	Test	0.8600	0.7105	0.6136	0.6585	0.8864
SVM	5-Fold CV	0.9661	0.8837	0.9754	0.9272	0.9936
	Validation	0.8607	0.6739	0.7045	0.6889	0.8898
	Test	0.8850	0.7561	0.7045	0.7294	0.9151

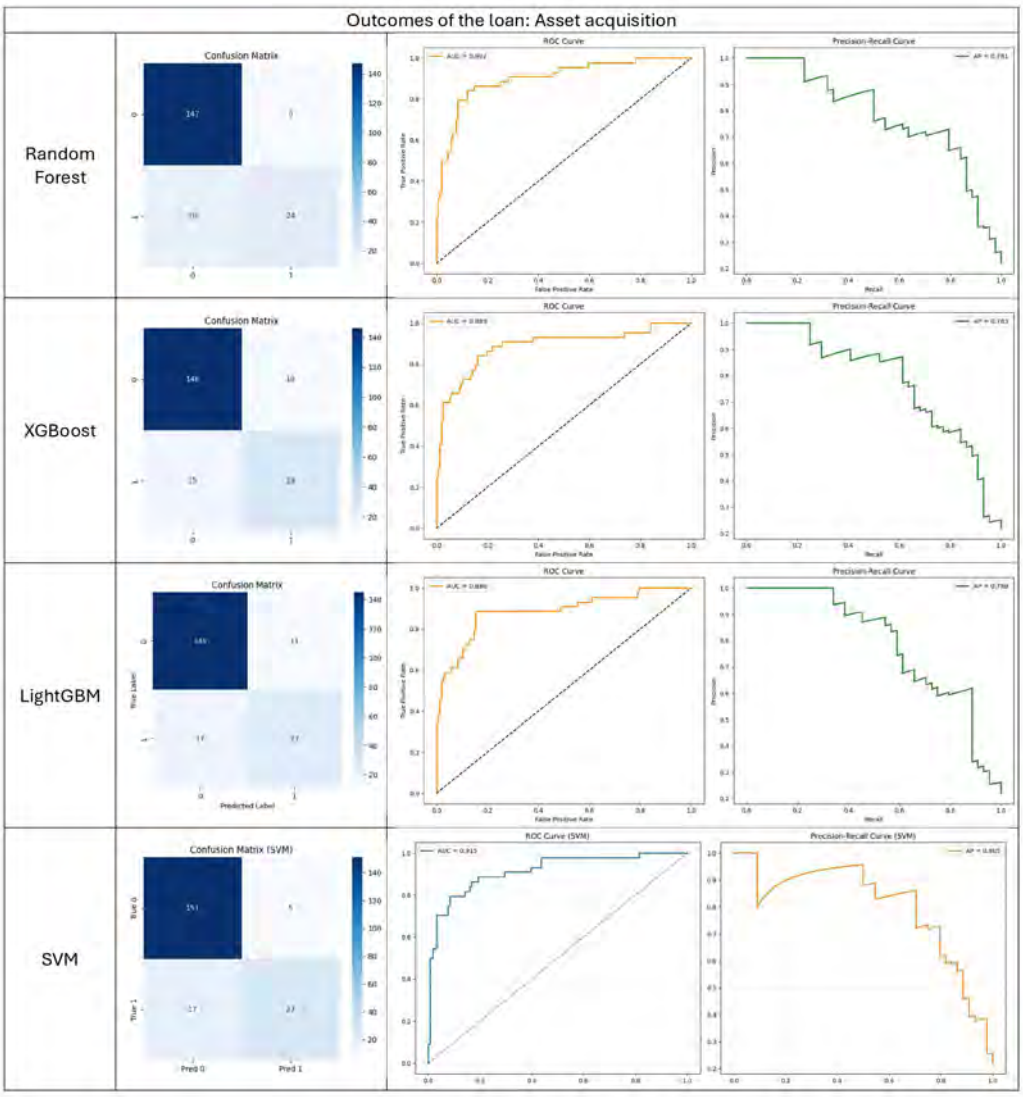


Figure 5.9: Results of Classification of Asset Acquisition

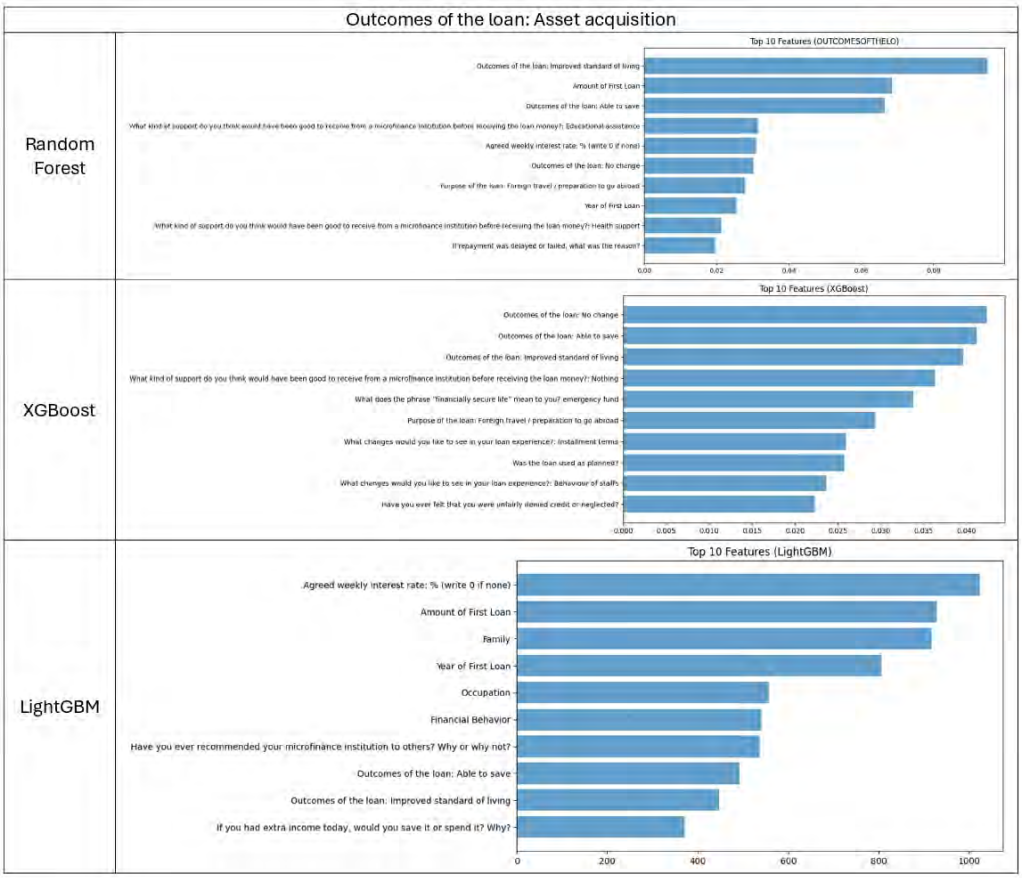


Figure 5.10: Feature importance determining Asset Acquisition

5.2 Analysis of Design Solutions

The high-tier performance metrics provided a preliminary validation of the proposed structure. However, a critical examination of model behavior is necessary to ensure these results represent true generalization. This is to ensure that it does not represent an artifact of the training process. A primary concern in analysis is overfitting, which occurs when models learn the training data so well that they start capturing random noise and idiosyncratic outliers rather than the underlying socioeconomic patterns. This results in high accuracy on training data but fails to perform well when presented with new and unseen data.

Performance analysis across the 3 models, LightGBM, XGBoost and Random forest identifies the presence of marginal overfitting. This is mainly shown by a distinct train-test accuracy gap. The LightGBM model achieved a high accuracy in the training set of our primary target variable, "Increase in Available Income" but when evaluated on test data, it had an accuracy gap of 0.1048. Similarly, XGBoost had a gap of 0.1147 when evaluated on test data. These results suggested that gradient boosting algorithms are more sensitive to specific patterns within training samples as they utilize an repetitive boosting process where each tree tries to correct the previous errors. Furthermore, they inherently work to seek out patterns that capture data noise.

In contrast to these models, SVM and Random Forest have helped to achieve an architectural baseline. The SVM utilizes an RBF kernel to provide a geometric baseline that maps non-linear relationships into higher dimensions to find a continuous, curved decision boundary. On the other hand, these tree-based models, like LightGBM, partition data into discrete decision regions. For the target "Business Expansion," the same model achieved a high ROC-AUC of 0.9239, and the fact that SVM consistently achieved high scores shows that the dataset has robust socioeconomic patterns independent of the specific model architecture.

To mitigate these methodological challenges, the study utilizes 5-fold Stratified Cross-Validation which partitioned the 1001 collected data points into 5 class-balanced groups. By training and testing the models 5 times across each fold, this approach verifies that performance metrics result from the consistent identification of underlying signals. Finally, all reporting prioritized performance on a completely independent Validation Set. This ensures the finalized results represent the models' objective ability to analyze the lived financial realities of the broader rural population.

Across the comprehensive evaluation of the five target variables, XGBoost emerged as the overall superior model for analyzing microfinance impact, demonstrating the highest predictive performance in four out of five key indicators . For the primary indicator, Increase in Income, XGBoost achieved a mean test accuracy of 0.8841 and a ROC-AUC of 0.9451, showcasing exceptional discriminative power. It maintained this dominance for Business Expansion and Living Standard, reaching a peak accuracy of 0.9171 for Asset Acquisition. While LightGBM outperformed XGBoost specifically for Ability to Save with a ROC-AUC of 0.9507, XGBoost's consistent

ability minimized the train-test accuracy gap averaging between 0.0819 and 0.1165. This suggests it provides a more robust and stable generalization across the diverse socioeconomic dataset. Consequently, XGBoost is recommended as the foundational architectural choice for this microfinance impact framework.

Table 5.6: Model Performance Comparison Across All Target Variables

Target Variable	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Increase in Income	XGBoost	0.8000	0.6875	0.7857	0.7333	0.9003
	LightGBM	0.8300	0.7500	0.7714	0.7606	0.8914
	Random Forest	0.7750	0.6582	0.7429	0.6980	0.8622
	SVM	0.7900	0.7383	0.8495	0.7900	0.8570
Improved Standard of Living	XGBoost	0.8300	0.7857	0.8049	0.7952	0.8982
	LightGBM	0.8550	0.8533	0.7805	0.8153	0.9256
	Random Forest	0.8350	0.8025	0.7927	0.7975	0.9043
	SVM	0.8050	0.7654	0.7561	0.7607	0.8860
Business Expansion	XGBoost	0.8800	0.8873	0.7975	0.8400	0.9400
	LightGBM	0.8950	0.9028	0.8228	0.8609	0.9538
	Random Forest	0.8500	0.7952	0.8354	0.8148	0.9205
	SVM	0.8150	0.7692	0.7595	0.7643	0.9135
Ability to Save	XGBoost	0.8000	0.6875	0.7857	0.7333	0.9003
	LightGBM	0.8300	0.7500	0.7714	0.7606	0.8914
	Random Forest	0.7750	0.6582	0.7429	0.6980	0.8622
	SVM	0.7700	0.6395	0.7857	0.7051	0.8671
Asset Acquisition	XGBoost	0.8750	0.7436	0.6591	0.6988	0.8888
	LightGBM	0.8600	0.7105	0.6136	0.6585	0.8864
	Random Forest	0.8550	0.7273	0.5455	0.6234	0.9021
	SVM	0.8850	0.7561	0.7045	0.7294	0.9151

5.3 Comparisons and Relationships

Table 5.6 compares the evaluation on the test split across all models. The relationship between the four implemented models provides a comprehensive predictive power. While models like LightGBM, XGBoost, and Random Forest utilize hierarchical data splitting to identify patterns, SVM utilizes RBF kernels to identify an optimal decision boundary in higher dimensions.

There is a trade-off between the precision of the boosting models and the stable recall of the SVM baselines. Even though LightGBM achieved the highest accuracy for immediate financial indicators like "Increase in Availability Income", the SVM model maintained a better recall of 0.8764 for "Business Expansion". This shows that SVM can be used as a safety net that captures potential beneficiaries that the boosting model might not capture. It is critical to understand that even if the boosting models are optimal for maximizing specific performance metrics, the inclusion of SVM is necessary to generalize and resist the inherent noise of the data.

The Random Forest analysis identified a set of 15 features that were listed as the findings of feature importance. The features such as occupation, financial behavior, and lifestyle had a great impact on the outcome. The fact that these same features consistently appear as high-weight variables in both LightGBM and XGBoost models proves that the boosting models are optimising their predictions based on the same core drivers of impact instead of just chasing random noise. Despite using a fundamentally different distance-based logic, the SVM's high performance confirms that these features possess universal predictive power across fundamentally different algorithmic families. This multi-model agreement proves that indicators like digital literacy and occupation are robust, real-world predictors of a borrower's financial future rather than artifacts of any specific algorithm.

5.4 Discussions

The detailed analysis of the tree-based architectures LightGBM, XGBoost, and Random Forest shows a major convergence of predictive signals, where a core set of unique features dominates the impact assessment through five target variables consistently.

The most common predictors that appeared at the top of the feature importance rankings for all the models were variables such as occupation, digital literacy, and financial behavior. This is very important as this proves that LightGBM and XGBoost prioritized these features to minimize gradient loss for high-precision targets, such as an increase in disposable income. The Random Forest uses them to ensure stability throughout highly varying indicators, such as asset acquisition. The presence of occupation in almost all the different tree splitting logics proves that these are the main socio-economic factors of financial accessibility in rural Bangladesh. This architectural consensus does a good job in filtering out the random noise present in human survey data, confirming that a borrower's vocational stability is an impor-

tant factor that determines their microfinance success. By using this framework in these common features, the system provides a much clearer roadmap for institutional decision making processs.

Chapter 6

Conclusion

6.1 Summary of Findings

The experimental results demonstrate that machine learning models can effectively utilize non-financial socio-economic data to predict the impacts of microfinance in Bangladesh. Among the models that were evaluated, LightGBM and XGBoost achieved the highest accuracy, and SVM helped to confirm the validity of the underlying socio-economic patterns by providing a stable and reliable baseline. Despite mild overfitting in gradient boosting models, independent validation and stratified cross-validation confirmed strong generalized performance throughout all targets. We can conclude that applying machine learning techniques to alternative behavioral and demographic data in fields like microfinance can result in accurate, fair, and sustainable microfinance decision-making.

6.2 Contributions to the Field

This thesis makes several significant contributions to the microfinance research landscape, particularly in Bangladesh. One of the most important contributions is the creation of a large, authentic primary dataset, at a scale and level of detail that is rarely found. During the literature review, it was seen that very limited research has been conducted worldwide, with almost none in Bangladesh, in terms of using machine learning in determining the efficacy of microfinance. This is mainly due to the scarcity of publicly available data. Financial institutions typically consider borrower information confidential, and access restrictions make it nearly impossible for researchers to obtain comprehensive datasets for analytical or predictive modelling.

This research uses a relatively large-scale dataset. Consequently, the research gives one of the most detailed, multidimensional datasets of microfinance clients in rural Bangladesh, which allows more intense comprehension of individual, social, and financial determinants of microfinance effectiveness. Therefore, the most significant value of the study is the dataset that has been created, making it one of the few primary datasets from this sector that is compatible with machine learning models.

Next, by using machine learning models to forecast the effects of microfinance on the borrowers, it can provide insights that can be used by the MFIs and policymakers

to enhance the financial inclusion strategies. Furthermore, such insights also help rural borrowers, as they gain awareness on how to utilize microcredits optimally to ensure sustainable economic improvement.

This thesis is an initial step towards using advanced data-driven methods to improve the microfinance sector of Bangladesh.

6.3 Recommendations for Future Work

Future work can be done from this research in a number of ways. First, the questionnaire can be utilized as a foundation to expand the dataset. More advanced models, like the neural networks, will be effective with a larger data set since they will need more data to discover more complex relationships, which can then provide much better predictive performance. This takes this study on to the next level by implementing deep-learning techniques on our data.

Besides, the high dimensionality of the dataset can be used for borrower segmentation using unsupervised models. This can help create distinct borrower profiles and allow loans to be more personalized. Such a method will take MFIs to the next level as their efficiency will increase significantly, leading to further economic growth.

Lastly, one more direction that can be taken would be to include longitudinal data. This implies that the participants will be tracked in various loan cycles or years in order to determine the long-term, enduring effect of microfinance. It may also reveal how borrower behavior changes over time. Such data can be combined with time series or sequential models.

Overall, the dataset developed in this thesis has a lot of potential to be explored. Future studies can gain deeper insights into various directions in the microfinance sector by increasing the size of the sample and using richer modeling methods.

Bibliography

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[2] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735

[8] M. Zeller, "Towards enhancing the role of microfinance for safety nets of the poor," Zentrum für Entwicklungsforschung (ZEF), Bonn, Germany, Tech. Rep. 19, 1999. [Online]. Available: http://ageconsearch.umn.edu/record/210627

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[10] S. Wold, M. Sjöström, and L. Eriksson, "Pls-regression: A basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001. DOI: 10.1016/S0169-7439(01)00155-1 [Online]. Available: https://doi.org/10.1016/S0169-7439(01)00155-1

[11] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press, 2002.

[12] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.

[13] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Mar. 2005, ISSN: 1369-7412. DOI: 10.1111/j.1467-9868.2005.00503.x eprint: https://academic.oup.com/jrsssb/article-pdf/67/2/301/49795094/jrsssb_67_2_301.pdf. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2005.00503.x

[14] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 983–999, 2006.

[15] B. Armendáriz and J. Morduch, *The Economics of Microfinance*, revised paperback. Cambridge, MA: MIT Press, 2007, ISBN: 978-0262512015.

[16] S. Ahmed, "Microfinance institutions in bangladesh: Achievements and challenges," *Managerial Finance*, vol. 35, no. 12, pp. 999–1010, 2009.

[17] A. S. M. Musa and M. S. R. Khan, "Benefits and limitations of technology in mfis: Come to save (cts) experience from rural bangladesh," *Journal of Electronic Commerce in Organizations (JECO)*, vol. 8, no. 2, pp. 54–65, 2010.

[18] D. Karlan and J. Zinman, "Microcredit in theory and practice: Using randomized credit scoring for impact evaluation," *Science*, vol. 332, no. 6035, pp. 1278–1284, 2011.

[19] S. Bairagi, "Productivity and efficiency analysis of microfinance institutions (mfis) in bangladesh," 2014.

[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[23] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[25] S. Raihan, S. Osmani, and M. B. Khalily, "The macro impact of microfinance in bangladesh: A cge analysis," *Economic Modelling*, vol. 62, pp. 1–15, 2017.

[26] S. J. Rigatti, "Random forest," *Journal of insurance medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[27] A. Kuraria, N. Jharbade, and M. Soni, "Centroid selection process using wcss and elbow method for k-mean clustering algorithm in data mining," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 4, no. 11, pp. 190–195, 2018.

[28] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019. DOI: 10.1214/18-AOS1709

[29] S. Athey and S. Wager, "Estimating treatment effects with causal forests: An application," *Observational Studies*, vol. 5, no. 2, pp. 37–51, 2019. DOI: https://doi.org/10.1353/obs.2019.0001

[30] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Blockchain technology in financial services: A comprehensive review of the literature," *IEEE Access*, vol. 7, pp. 124 058–124 076, 2019.

[31] H. Humaira and R. Rasyidah, "Determining the appropiate cluster number using elbow method for k-means algorithm," in *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*, 2020, pp. 1–8.

[32] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020. DOI: 10.1109/ACCESS.2020.2988796

[33] B. Banu, M. M. Hossain, M. S. Haque, and B. Ahmad, "Effect of microfinance adoption on rural household income in selected upazila of kushtia district of bangladesh," *Bangladesh Journal of Multidisciplinary Scientific Research*, vol. 3, no. 1, pp. 24–32, 2021.

[34] H. I. Condori-Alejo, M. R. Aceituno-Rojo, and G. S. Alzamora, "Rural micro credit assessment using machine learning in a peruvian microfinance institution," *Procedia Computer Science*, vol. 187, pp. 408–413, 2021, 2020 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI2020, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.04.117 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921009297

[35] M. S. U. Mazumder, "The effects of microfinance programs on recipients' livelihoods in rural bangladesh," *The European Journal of Development Research*, vol. 34, no. 3, pp. 1383–1418, 2022.

[36] G. D. Onipe, "The role of machine learning in enhancing credit scoring models for financial inclusion," *World Journal of Advanced Research and Reviews*, vol. 17, no. 3, pp. 1095–1106, 2023. DOI: 10.30574/wjarr.2023.17.3.0400 [Online]. Available: https://wjarr.com/content/role-machine-learning-enhancing-credit-scoring-models-financial-inclusion

[37] H. E. Omokhoa, C. S. Odionu, C. Azubuike, and A. K. Sule, "Innovative credit management and risk reduction strategies: Ai and fintech approaches for microfinance and smes," *IRE Journals*, vol. 8, no. 6, p. 686, 2024.

[38] J. A. Polin, M. F. H. Sarker, M. D. K. Dolon, N. Hasan, M. M. Rahman, and Z. N. Vasha, "Predicting the effects of microcredit on women's empowerment in rural bangladesh: Using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 4, pp. 2699–2706, 2024.

[39] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024.

[40] M. S. Shak et al., "Innovative machine learning approaches to foster financial inclusion in microfinance," *International Interdisciplinary Business Economics Advancement Journal*, vol. 5, no. 11, pp. 6–20, Jun. 2024, ISSN: 2375-9615 (eISSN), 2375-9607 (pISSN). DOI: 10.55640/business/volume05issue11-02 [Online]. Available: https://www.iibajournal.org/index.php/iibeaj/article/view/50

[41] Y. J. Garcia-Lopez, P. H. Marquez, and N. N. Morales, "Microfinance institutions failure prediction in emerging countries, a machine learning approach," *PLoS One*, vol. 20, no. 4, e0321989, 2025.

[42] M. Mahedi, A. Saha, A. Pervez, and S. J. Shaili, "Micro-credit in bangladesh: A comprehensive review of its evolution, impact, and challenges using quantitative and qualitative evidences," *Asian Journal of Economics, Business and Accounting*, vol. 25, no. 5, pp. 1–18, 2025.

[43] T. Ting, M. A. Mia, M. I. Hossain, and K. K. Wah, "Predicting the financial performance of microfinance institutions with machine learning techniques," *Journal of Modelling in Management*, vol. 20, no. 2, pp. 322–347, 2025. DOI: 10.1108/JM2-10-2023-0254 [Online]. Available: https://doi.org/10.1108/JM2-10-2023-0254

[44] M. A. H. Choudhury and S. Mandal, "Important determinants of mfi loan effectiveness for successful micro entrepreneurship in bangladesh,"

[45] M. A. Rehman, M. Ahmed, and S. Sethi, "Ai-based credit scoring models in microfinance: Improving loan accessibility, risk assessment, and financial inclusion," *The Critical Review of Social Sciences Studies*, vol. 3, no. 1,