Addis Ababa University

Addis Ababa Institute of Technology

College of Technology and Built Environment (CTBE)

Communication Engineering Graduate Program

Evaluating the Impact of Telecom Payment Behavior and Equb Participation on Credit Approval Decisions

By:

Selamawit Getasew

Advisor:

Dr. ROSA TSEGAYE

A Thesis Submitted to the School of Electrical and Computer Engineering Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecom Information System

February 2026

Addis Ababa, Ethiopia

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Communication Engineering Graduate Program

Evaluating the Impact of Telecom Payment Behavior and Equb Participation on Credit Approval Decisions

by:

Selamawit Getasew

Approval by:	Name	Signature	Date
Advisor:	Dr. Rosa Tsegaye	_____	_____
First Examiner:	Dr. Surafel Lemma	_____	_____
Second Examiner:	Dr. Sosina Mengistu	_____	_____

Declaration

I hereby declare that this thesis is my original work, conducted at Addis Ababa University under the supervision of Dr Rosa Tsegaye. To the best of my knowledge, all sources used have been properly cited. I also confirm that this thesis has not been submitted, either in whole or in part, for any other degree, diploma, or certification at this or any other institution.

_____

Signature

Selamawit Getasew

Date of Submission: February/2026

Abstract

Availability of credit helps in the development of small-scale businesses, and in most cases, traditional credit scoring is not effective for people without adequate financial history. This study aims to determine whether alternative data helps in improving the accuracy of microfinance institutions in predicting the success of loans.

This study aims to determine the effect of using alternative data in improving the accuracy of microfinance institutions in predicting the success of loans. This is done by comparing the performance of models using traditional data and models using traditional and alternative data. The data preprocessing steps involve handling missing values, limiting the number of outliers, scaling, and encoding. Stratified sampling and 5-fold cross-validation are used to perform the study. The performance is measured by accuracy, precision, recall, and F1 score.

The results obtained from the study show marginal improvements in the accuracy of microfinance institutions in predicting the success of loans. LightGBM secures the highest result, with an average F1 score of 0.857 that is realized when conventional features are employed. Other models' (for instance, Random Forest and XGBoost) effectiveness is at the par with LightGBM. Incorporating alternative data hardly changes the models' effectiveness.

The addition of alternative data to traditional data helps in enhancing the accuracy of microfinance institutions in predicting the success of loans, although the improvements are marginal and context-dependent.

Keywords: Artificial intelligence, alternative data, inclusive lending, credit evaluation, small business finance, predictive modelling, machine learning, economic empowerment, equitable finance.

Acknowledgment

First and foremost, all thanks and praise go to Almighty God, whose grace and mercy have been instrumental in guiding and supporting this study along its trajectory. I also thank the Blessed Virgin Mary for her company and support along the way. I would like to express my deepest gratitude to my supervisor, Dr. Rosa Tsegaye, for her support and guidance. Her support and feedback have greatly influenced and motivated me to write this thesis. I would like to thank my examiners, Dr. Surafel Lemma and Dr. Sosina Mengstu, for their valuable suggestions and feedback, which greatly contributed to the quality of my study. I would also like to thank Ethiotelecom and the School of Electrical and Computer Engineering, Addis Ababa University, for providing me the opportunity and resources to pursue my study. Most importantly, I would like to thank my family. To my mother, Firezewed Erkalem, thank you for all the love and support. To my husband, Ermias Wondafrash, thank you for all the patience and support. This achievement would not have been possible without the two of you.

Contents

Declaration 3

Abstract i

Acknowledgment ii

Table of Contents v

List of Table vi

List of Figures vi

Abbreviation vii

1 Background and Motivation 1

1.1 Introduction 1

1.2 Statement of the Problem 2

1.3 Research Question 3

1.4 Objectives 3

1.4.1 General Objective 3

1.4.2 Specific Objectives 3

1.5 Scopes and Limitations of the Research 4

1.6 Methodology 4

1.7 Significance of the study 5

1.8 Thesis organization 6

2 LITERATURE REVIEW 7

2.1 Loan predation 7

2.2 Traditional credit scoring 8

2.3 Traditional data 8

2.4 Alternative Data 9

2.5 Methods for Loan Prediction 10

2.5.1 Ensemble Models . . . 10
2.5.2 Feature Engineering . . . 11

3 PROPOSED SYSTEM MODEL 12

3.1 System model . . . 13
3.2 Data collection and preprocessing . . . 15
3.2.1 Data cleaning . . . 16
3.2.2 Feature preparation . . . 16
3.3 Model building . . . 17
3.3.1 Model Training Techniques . . . 17
3.3.2 XGBoost (Extreme Gradient Boosting) . . . 19
3.3.3 LightGBM (Light Gradient Boosting Machine) . . . 20
3.4 Evaluation . . . 21

4 EXPERIMENTS 24

4.0.1 Tool . . . 24
4.1 Dataset Description . . . 24
4.2 Data visualization . . . 26
4.3 Data Cleaning . . . 32
4.3.1 Normalization . . . 32
4.4 Feature Engineering . . . 33
4.4.1 Features Removed or Transformed . . . 33
4.4.2 Features Added through One-Hot Encoding . . . 34
4.4.3 Feature Grouping . . . 34
4.5 Handling Missing Values . . . 35
4.6 Cross-Validation . . . 36
4.7 Imbalanced Data and Dataset Splitting . . . 37

5 Experiment Setup 39

6 RESULT AND DISCUSSION 41

6.1 Experimental Results 41

7 Conclusion and Recommendations **45**

7.1 Future Research Directions 46

List of Tables

1 Steps in Ensemble Methods 18

2 Dataset Column Description 25

3 Dataset Preparation Steps and Number of Rows 26

4 Descriptive Statistics of Features Before Normalization 32

5 Single Train–Test Split Results (Evaluation on 20% Test Set) 42

6 Model Performance Comparison 43

List of Figures

3.1 Proposed system model architecture 14

4.2 Scatter Plot of Income vs. Previous Loan Amount 29

4.3 Count plots of categorical features vs. loan status. 31

4.4 Distribution of loan status 38

6.5 F1-score by Model and Feature Set 44

NFIS	National Financial Inclusion Strategy
FICO	Fair Isaac Corporation
ANN	Artificial Neural Network
XGBoost	Extreme Gradient Boosting
IQR	Interquartile Range Method
LightGBM	Light Gradient Boosting Machine
DL	Deep Learning
RNN	Recurrent Neural Networks
CNNs	Convolutional Neural Networks
GPU	Graphical Processing Unit
KNN	K-Nearest Neighbor
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAXENT	Maximum Entropy
MEF	Mobile Ecosystem Forum
ML	Machine Learning
NB	Naive Bayes
NN	Neural Network
OC-SVM	One-Class Support Vector Machine Classifier
PCs	Personal Computers
RBS	Rule-Based System
RF	Random Forest
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
MSE	Mean Squared Error

1 Background and Motivation

1.1 Introduction

Financial loan prediction is a key aspect for banks and financial institutions as it is oneof the main means they use to check one's creditworthiness. On the other hand, theseinstitution face the challenge of how to lower their loan risk while at the same timeencouraging financial inclusion [1]. Financial institutions may decide on not serving a big part of the adult population ifthey impose very strict credit assessment procedures that only look into the traditionalfinancial data, which is contrary to the idea of financial inclusion. However, thesustainability of financial institutions might be compromised by very loose financialmechanisms that do not properly assess the risk [2]. Ethiopia has comparatively lower formal financial inclusion than other Eastern African countries and is almost reaching a significant change in its financial sector. The country's revised National Financial Inclusion Strategy (NFIS 2021, 2025) targets a rise in financial inclusion from the existing 46% to 70% of the adult population by 2025. The document also highlights the role of digitized mobile money services as the main driver of financial inclusion in the country [1]. Ensemble learning represents a powerful concept in the field of machine learning wherein the prediction capability of a model is enhanced through bringing together various models (also known as base learners) to collectively perform at a level which surpasses that of individual models. Essentially, this concept merges different algorithms and harnesses them to work at a level better than any single algorithm alone. Take Random Forest for example, it was found to be a more effective tool for handling financial data and combating overfitting issues through the use of ensemble learning. The significance of incorporating alternative data sources like social media and transactions in ensemble learning to give a well-rounded evaluation of the creditworthiness of applicants was another part of the research made public recently. It also showed that using a broad set of loan feature variables with Random Forest methods is a winning strategy for loan approvals. Besides that, paper [5] came up with an ensemble learning based loan prediction system.

The increasing accessibility and use of mobile money services in developing countries

have played a major role in promoting financial inclusion by making it possible for unbanked population segments to get micro-loans [1]. On the other hand, this progress raises a fundamental question: how to widen credit availability and still have a good control over risks.

Including alternative data is a good way to enhance the precision and reliability of loan prediction models since such data can supply a more detailed and changing picture of a borrower's financial health. More specifically, alternative data captures behavior patterns over time like transaction regularity, mobile payment frequency, and utility payment discipline, which can be overlooked by traditional credit records. Studies using combination methods, e.g., XGBoost and Random Forest, have revealed that these extra indicators can lead financial institutions to more accurately predict default risk, in particular, for individuals with a small amount of formal credit history [2] [3].

1.2 Statement of the Problem

With increasing availability and use of mobile money services in developing countries, these financially excluded people got access to the micro-loans becoming one of the main drivers of financial inclusion [1]. But this progress also brings a big problem: to increase credit access while still managing risk well. Conventional credit scoring systems are heavily dependent on formal financial data like income statements, credit history, and employment records, these are the data that many people in low-income or informal sectors do not have [4]. This dependence makes financial access impossible for millions of people and, in the result, keeps the circle of their exclusion from the most needed financial services.Without a good understanding of a borrower's financial condition, defaults may soar and the microcredit business model may become unsustainable. For this reason, credit institutions are increasingly exploring alternative data sources such as mobile phone usage patterns, utility payment records, and mobile transaction histories as a potential solution to this credit assessment gap [5].In this study, alternative data refers specifically to borrower-level behavioral and transaction variables extracted from digital financial activity. The key variables include mobile money transaction frequency, average

transaction amount, cash-in/cash-out regularity, airtime top-up frequency, utility bill payment timeliness, month-to-month account activity consistency, and community-based savings participation indicators, including Equb contribution behavior. These variables are used as proxies for financial discipline, liquidity behavior, and repayment capacity, especially for applicants with limited formal credit records [6].Given this backdrop, this research systematically compares two feature sets: (i) traditional variables such as income, employment status, and prior credit information, and (ii) a combined set that adds the above alternative-data variables. The objective is to determine whether adding these alternative features improves loan prediction performance and expands access to credit for borrowers with limited formal credit histories.

1.3 Research Question

RQ1: What are the differences in model performance when the model is trained on traditional data versus when it is trained on a mixture of both alternative and traditional data?

RQ2: Which ensemble models (Random Forest, XGBoost, LightGBM) provide the highest performance in each data scenario?

1.4 Objectives

1.4.1 General Objective

Developing and evaluating the performance of machine learning models that combine alternative data with conventional financial metrics for loan approval prediction to be able to enhance accuracy and financial inclusion through credit scoring.

1.4.2 Specific Objectives

To achieve the general objective, we have to formulate these specific objectives:

- To identify and categorize relevant alternative data sources for loan approval prediction.

- To prepare and preprocess the alternative and traditional data.
- To build and compare loan approval predictive models, such as Random Forest, XGBoost, and LightGBM, using alternative and traditional data.
- To evaluate model performance under different data scenarios.
- To provide recommendations for financial institutions and policymakers on the responsible use of alternative data in credit decision-making.

1.5 Scopes and Limitations of the Research

This research aims to identify the impact of alternative data such as mobile phone usage, utility payment records, and informal financial behaviors in enhancing loan approval prediction models for Tele birr users in Ethiopia. The work proposes and tests machine-learning-based predictive models that are primarily based on traditional financial data, alternative data or a combination of both. Ensemble learning methods are also employed while the effectiveness of the models is gauged with the help of standard predictive accuracy measures so as to identify the role of alternative data. Results will hopefully help financial institutions as well as policymakers in devising data-driven strategies for inclusive credit assessment. Attention is given only to recent datasets and the research has no bearing on the creation of commercial credit systems or the legal/regulatory frameworks for implementation.

1.6 Methodology

These are the phases of research methodology that have been used in order to achieve the goals of this research. Literature review: Literature review: Related literature has been reviewed to understand the existing state of arts and to identify the gaps in current loan prediction techniques. **Research problem identified:**The problem statement was identified based on the gap in the reviewed literature. **Data collection:** The inputs data i.e. the traditional and alternative data have been collected from Ethio Telecom's Tele birr platform, specifically, from the customers' financial histories. **Data analysis:**

The data has been examined using data visualization techniques in order to understand the dataset content and characteristics. **Data preprocessing:** Data preprocessing has been carried out to clean the data by eliminating inconsistencies and handling missing values. In addition, feature engineering has been performed to derive new features from the existing data. Finally, the data has been standardized as per the requirements of the selected ensemble learning method, and thus the dataset has been prepared for the model training. **Training and testing:**Ensemble-based machine learning models were trained using traditional data, alternative data, and combined datasets. The ensemble techniques were applied to integrate multiple base learners to enhance loan prediction performance. Cross-validation was used for both model training and testing to ensure robustness, minimize overfitting, and provide reliable performance evaluation. Result and discussion: In the final step of the study, the models have been evaluated using performance evaluation metrics such as precision, recall, and f1-score. Subsequently, the result of the experiment is discussed.

1.7 Significance of the study

This research will greatly help banks, financial regulators, digital service providers and citizens in the underserved areas of emerging countries. Since traditional credit scoring methods keep on excluding large parts of the population due to the absence of formal financial data, this work points out the power of alternative data to connect the underserved groups with the financial system. It is an evidence-based study that supports inclusive financing by looking at mobile phone usage, utility payments, and digital transaction histories to design better credit scores. In the case of banks and digital lenders, results are a source of knowledge on the role alternative data plays in improving the precision and dependability of credit prediction models, as well as in lowering credit defaults, and at the same time, reaching even unbanked people. For authorities and lawmakers, research makes a case for well-thought data governance policies that help strike a balance between innovation and protecting personal data, fairness and openness to the public. Furthermore, the paper is a great contribution to the debate on credit

scoring based on data at hand by presenting a side-by-side comparison of traditional and alternative data sources in the case of poor countries. In the end, this research is a step toward financial inclusion by real responsibility as it provides proof that the use of innovative data can lead to easier credit access for marginalized people without a sacrifice in risk management.

1.8 Thesis organization

This thesis structure is such that the content of the remaining chapters is organized as follows. Chapter 2 is dedicated to a thorough literature review which, among other things, provides a basis for the current work by delving into financial inclusion and credit scoring and discussing the source of data, i.e., traditional versus alternative, in loan prediction. Chapter 3 is the system model where at the beginning the framework is laid out, the description of data, and the tools of machine learning that the study will use to measure how accurate, in terms of loan prediction, the different models are. In Chapter 4, the experimental environment is described including details about cleaning of data, model building, and testing, and will emphasize the model working with the combination of traditional and alternative data features. Chapter 5 depict the findings and results, their meaning and importance are discussed, the performances of the model over different data are compared, and possibilities of applying the findings are elaborated. Ultimately, in Chapter 6, the major lessons are recapped and directions are suggested for further work, policy making, and responsible utilization of alternative data in credit scoring systems.

2 LITERATURE REVIEW

This review of the literature explores the expanding influence of alternative data in transforming loan prediction techniques. In view of the fact that traditional credit scoring models impose restrictions on many people, banks are turning to the use of new types of data in order to have a better understanding of customers' creditworthiness. According to the Alliance for Financial Inclusion, alternative data "refers to non-traditional information or unconventional data sources that are not usually employed in conventional credit scoring models" and therefore offer additional insights that cannot be obtained from the standard credit reports [7]. This review examines the ways in which such alternative data can improve the preciseness of loan prediction and simultaneously mitigate the weaknesses of the conventional models.

2.1 Loan predation

Loan prediction is one of the essential activities in the financial sector as it requires using historical data to forecast whether a loan applicant will repay the loan or not. This forecasting allows the financial entities to understand the risk involved in providing the money. By examining different variables such as credit history, income, employment status, and the amount of the loan, machine learning models can assess the chances of repayment and thus, lenders will be able to make better decisions [8]. From the borrower's point of view accurate loan prediction is of great help providing one with the necessary financial understanding. Research points out that "a well-forecasted loan repayment schedule helps them manage cash flow, allocate funds strategically, and avoid financial distress" [9]. This is even more so true for small business owners and individual borrowers who depend on predictable financial plans for their business continuity and personal financial well-being. Borrowers who are aware of their repayment capacity can not only seek better loan terms but can also make well-informed financial decisions [9].Some studies reveal that incorporating data from sources other than credit bureau can significantly improve the prediction of loans. The article [10] highlights that such alternative data as

cell phone usage, mobility, social media activity, and online shopping habits are even more predictive than the traditional data alone [10].

2.2 Traditional credit scoring

2.3 Traditional data

Loan prediction with traditional data means using structured and tabular collections of historical data from financial institutions or banks. This information includes aspects like the applicant's demographics, financial condition, credit history, loan amount, employment, and the final loan status (approved or denied). Such datasets serve as the basis for the training and testing of machine learning models that are designed to predict loan approvals or defaults. Diverse data on loan approvals have recently been used in many research papers. It is a common practice to utilize the Kaggle Bank Loan Prediction dataset that has 148, 670 records and 37 attributes. Among various characteristics, the dataset embodies aspects like income, credit history, loan amount, and property details which are indispensable for loan decisions. Using this dataset and through machine-learning approaches, including AdaBoost, Random Forest, Decision Tree, SVM, and GaussianNB, Haque et al. (2024) compare the respective performances and reveal that ensemble approaches, e.g., AdaBoost, are capable of the highest accuracy (99.99%). This is an indication that such datasets can be effectively used for developing powerful models [11]. Researchers have also relied on smaller institutional datasets. For instance, Anand et al. (2022) utilized 850 records, to predict loan behavior employing similar machine learning techniques [12]. In brief, these traditional datasets, with their precisely characterized variables and binary target classes (approved or denied), form a sound platform for financial predictive analysis and have gained the status of benchmarks for comparing various machine learning methods in the recent scholarly work [13]. Nevertheless, the conventional credit scoring methods have their fair share of disadvantages. By their very nature, they leave out individuals who do not have credit footprints, e.g., young people, immigrants, or those working in an informal sector. As a result, they contribute to

financial exclusion [14] [15]. Besides, these methods can also keep alive the biases of the past that have been imprinted in historical data and therefore, result in unfair treatment of certain demographic groups. Moreover, the traditional credit scoring is incapable of exploiting other possible data sources, such as utility payment records or mobile money transactions which can paint a fuller picture of the financial behavior of a person [16]. Therefore, a call for credit assessment methods that are more inclusive and capable of addressing the issues of the existing ones is gaining ground.

2.4 Alternative Data

Alternative data for loan prediction is a term that denotes non-conventional data sources used in identifying the creditworthiness of possible borrowers. Through these sources, lenders can get supplementary insights that go beyond the traditional credit reports and financial statements. For instance, social media activity, utility payment history, rental payment history, mobile phone usage data, and e-commerce transaction data are some of the examples of non-traditional data. This data, when combined with the traditional credit information, can further help lenders in making well-informed decisions while assessing loan applications [2]. Leveraging alternative data for loan prediction brings more than a few advantages to the table. For one, it can make predictions more accurate by incorporating additional behavioral and transactional signals which can lead to an overall increase in model performance [2]. Secondly, it can open up credit markets to a wider range of people, as lenders will now be capable of credit-testing individuals who have no formal credit histories. Thirdly, it can lower risk at a large scale, since increased credit behaviors and financial patterns can provide insights into a borrower's reliability. Lastly, it can facilitate personalized lending, as lending institutions can better understand their borrowers and offer them loan products accordingly. Besides helping in making better loan decisions, alternative data also lends itself to loan approval processes that are quicker and backed by a wider set of indicators [2]. The initial results from the field also highlight that the incorporation of alternative data has the potential to not only improve credit risk prediction but also widen credit availability to underbanked segments, thereby

suggesting its tangible contribution to the cause of financial inclusion [2].

2.5 Methods for Loan Prediction

One of the ways that lending institutions in the financial sector have been able to make better, faster, and more reliable decisions on which loans to give and which ones to deny is through the use of advanced machine learning and data analytic techniques. These aims to completely improve the level of getting the right answer while at the same time becoming faster and less prone to errors in the whole loan issuing and risk evaluation procedures. machine learning models which have been able to analyze loans that have already been given and get great relationships between borrower characteristics and loan outcomes are the topic of quite a few papers. To name a few, Random Forest, Gradient Boosting, Support Vector Machines (SVM), Decision Trees, and Logistic Regression. These data modeling types have been demonstrated to be quite successful at locating very intricate patterns that can then be used to predict loan approvals and defaults with a high degree of accuracy [17] [18].The main idea of ensemble learning is to combine the advantages of different machine learning models so as to yield better adoption results, in this way the overall efficiency is increased. Some specific methods like AdaBoost, Stacking, and Voting-based ensembles have shown better accuracy and stability than the single models. Ensemble models by integrating different algorithms are capable of lowering the fluctuation and reduce the possibility of overfitting, thus they are very powerful for loan prediction problems.

2.5.1 Ensemble Models

Studies demonstrate that ensemble models like AdaBoost and XGBoost achieve exceptional accuracy (up to 99.99%) when trained on datasets combining traditional financial metrics with alternative data such as mobile transactions, utility payments, and digital footprints [19], [20]. For instance, AI-driven models leveraging alternative data sources can assess borrowers' risk profiles through spending patterns and bill payment consistency, enabling real-time credit decisioning and dynamic loan adjustments [19].

This approach is particularly impactful in developing economies, where traditional credit histories are scarce, and alternative data provides critical insights into informal financial behaviors [21]. The machine learning technique for non-traditional data not only enhances predictive performance but also expands access to credit for underserved populations, as fintech lenders increasingly adopt these methods to reduce bias and improve risk assessment [22] [23]. Research highlights that model incorporating alternative data-such as psychometrics, social media activity, and digital transaction histories-outperform traditional logistic regression and linear discriminant analysis, particularly for high-risk or unbanked borrowers [19] [23]. For example, ensemble frameworks like Random Forest and XGBoost, when applied to datasets enriched with alternative data, improve AUC scores and reduce false positives by capturing complex, nonlinear relationships in borrower behavior [24]. This dual emphasis on advanced algorithms and diverse data sources enables lenders to achieve greater accuracy while mitigating ethical risks associated with biased or opaque scoring systems. As a result, ensemble learning and alternative data are reshaping credit markets, promoting equitable access to financial services globally [24] [25].

2.5.2 Feature Engineering

Proper feature engineering is a very important step to get the most out of interpretable as well as powerful loan prediction models. The process normally involves changing raw data into useful features that explain the characteristics of the borrower and are also good predictors of loan performance. Some examples could be credit score, loan amount, loan purpose, customer income, type of employment, etc. Besides, transforming the input features is always helpful in significantly changing model accuracy and the reliability of loan prediction results. Tuning features thoughtfully can signal the underlying data patterns better to the machine learning models, which in turn results in performance improvement [26]. In a nutshell, current loan prediction methods rely on various combinations of traditional as well as modern machine learning algorithms, deep neural networks, ensemble methods, and well-thought-out feature engineering to produce in-depth, fast, and accurate predictions in the financial sector. For this paper, we will go with ensemble-based

machine learning methods to emphasise these benefits. To be more specific, we use ensemble methods like Random Forest and Gradient Boosting to build predictive models of Tele Birr consumers in Ethiopia using both conventional and non-traditional forms of data. Ensemble learning enables our predictive models to better understand the complex interactions that lenders and borrowers have in order to accurately predict a person's credit risk. Ensemble learning, along with alternative data, is a powerful combination and continues to be a strong focus of research. Ensemble methods utilise a variety of models where each makes predictions based on a different type of data. For instance, one model may analyse a user's mobile phone behaviour while another may be based on social media information. The merger of these models yields a comprehensive picture of a borrower's creditworthiness [27]. Scientific studies reveal that the performance of credit scoring systems is significantly enhanced by ensemble learning, even when only limited data sets like airtime recharge records are available [26]. By mixing different models, lenders can not only address each individual approach's shortcomings but also make strong and robust predictions. However, one must be careful when selecting and validating the models in order to avoid overfitting and make sure that the predictions are still generalizable [28]. In a larger frame, this paper brings out that alternative data, in combination with ensemble learning, has immense potential for developing credit assessment tools that are not only more accurate but also fair. When taken together, these two elements open doors for the growth of financial inclusion while keeping a close eye on risk management in lending.

3 PROPOSED SYSTEM MODEL

This chapter outlines the research design and methodology adopted for this study. A well-structured methodology provides the foundation for conducting rigorous and credible research, guiding the process from problem identification to data analysis. By following a systematic approach, the study ensures that its findings are both valid and reliable. In this study, we utilize ensemble learning techniques, a strong category of machine learning methods that improve predictive accuracy and reliability by integrating multiple

base models. We specifically implement the Random Forest, XGBoost, and LightGBM algorithms, which are known for their effectiveness in processing both conventional structured data and diverse alternative data sources. Ensemble strategies such as bagging and boosting improve loan prediction performance by reducing overfitting and enhancing generalization.

3.1 System model

The present system model indicates a methodical and logical approach for loan prediction improvement by breaking down the whole process into three large steps: Data Preparation, Model Development, and Model Evaluation & Deployment as mentioned in the Figure 3.1. It has been scrupulously planned stage by stage to provide the maximum resultant accuracy in a brief and lucid manner for the loan approval procedure. In an attempt to discover possible ways of credit scoring improvement, the proposed system focuses on alternative data exploration and using ensemble learning methods for further extraction of features. The experimental process is initiated by gathering both typical data (e.g., credit history, salary level, workplace, etc.) and alternative data (participation in Equb, utility payments, mobile money usage, etc.) After data collection, a preparation phase is conducted including data cleaning, transformation, and feature engineering to ensure consistency and quality. The datasets are then split into training and testing sets. Ensemble models have been trained on Random Forest, XGBoost, and LightGBM algorithms using traditional and alternative data as an input. The performance metrics such as accuracy, precision, recall, F1-score, and AUC have been used for the evaluation of each model. The final scenes would be presented as charts and a table showing the results of traditional and alternative data models that are used to measure the value of predictive power and contribution of alternative data. This system model is an aid to the main objective of the thesis which is to examine and prove the function of alternative data in helping loan prediction and making financial services available to the less privileged groups.

The proposed model adheres to a methodical machine learning workflow that is purpose-

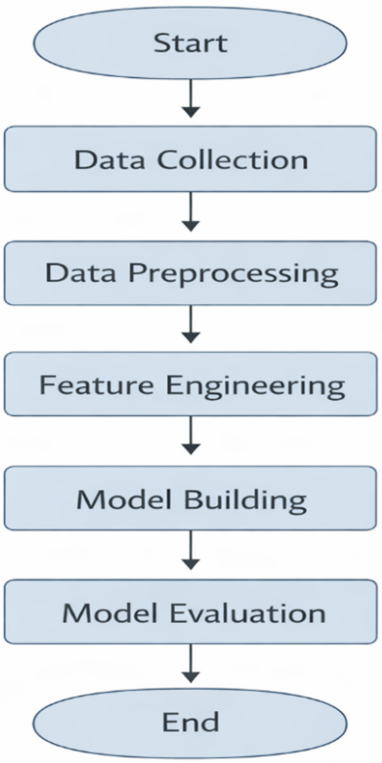

Figure 3.1: Proposed system model architecture

fully laid out to take raw data from unprocessed to a stage where an optimal prediction model can be identified. The first step in the sequence is data collection, which entails gathering of sets of data that relate to the problem domain. The gathered data is then prepared for analysis through preprocessing, which refers to activities such as dealing with missing data, normalization, scaling, and encoding for the purpose of making data consistent and high in quality. Besides, during this step, feature engineering is carried out to fabricate, pick, and alter features that improve a machine's behavior to account for data's underlying patterns. Afterwards, the stage of model creation-led by model building includes selecting suitable algorithms and determining hyperparameters for the multiple potential models candidates gotten to be developed. Evaluation of a model refers

the process of determining its merit by the use of relevant the statistical measures such as accuracy, precision, recall, F1-score, or root mean square error (RMSE). In the end, among the available models, the one that best satisfies the comparative evaluation criteria is picked as the final solution.

3.2 Data collection and preprocessing

Keeping in line with the company policies and the ethical standards of research at Ethio Telecom, the current study has completely avoided using any private or confidential customer information. The dataset was free of any personally identifiable details such as names, phone numbers and exact addresses. Anonymity and confidentiality were maintained throughout the research process. Formal administrative approvals for data access were first obtained from the respective department heads and the Information Technology unit. This project uses a dataset of 50, 000 records which come from the users of Ethio Telecom mobile money super app. This data provides deep demographic, financial, and behavioral profiling of digital loan users. The socio-economic attributes considered include age, gender, marital status, education level, employment type and monthly income. On the other hand, the dataset also contains loan-related information such as the history of amounts of past and current loans, credit history and loan repayment status. In addition, by including users' behavioral patterns in the mobile ecosystem like telecom and utility bill payments, total top-up amounts, and participation in informal savings groups (Equb), the dataset offers a very comprehensive profiling of individuals. After data was collected, it was preprocessed to ensure that the data is of good quality and suitable for the purposes of machine learning. Some of the tasks at this stage are data cleansing, changes of data forms, and data integration from multiple sources if applicable. Missing values were addressed cautiously so as not to introduce any bias, and the final dataset was obtained, which would be the basis for model development and evaluation.

3.2.1 Data cleaning

Data cleaning was done to enhance data quality and make sure data is reliable before training the models. As the e ducation Level attribute is categorical, the most frequent value was used to fill the missing values in it. Rows having missing values in Loan Status attribute were simply removed as it is the target variable and having unlabeled cases would confuse the model. Column names were harmonized to grant smooth data access and prevent errors during processing. For instance, to remove the leading space, the colu mn originally called "Loan Status" was simply renamed as "Loan Status". Using the Interquartile Range (IQR) technique, out liers (extreme values) in the numerical features such as Income, Previous Loan Amount, Total Top-up Amount and Equb Contribution Amount were identified and treated appropriately. To avoid an unbalanced effect on model training, the extreme values were limited to reasonable levels without distu rb ing the overall d ata distribution. Besides that, numeric columns were explicitly changed into numeric data types. When converted, non-numeric values were tra nformed into missing values which were then handled during the preprocessing steps. The above operations resulted in a dataset that was clean, consistent, and ready for machine learning.

3.2.2 Feature preparation

Feature preparation was done to convert the cleaned dataset into a numerical representation that is feasible for the machine learning models. Several categorical variables, namely Gender, Marital Status, Education Level, Employment Type, Telecom Bill Payment Pattern, Equb Membership (Yes/No), and Region were handled with one-hot encoding. This change turns categorical properties into numeric features represented as binary arrays. Because different numerical features had different scales, they were all normalized so that the model learning would be balanced. As a matter of fact, Income and Equb Contribution Amount were scaled by Min, Max normalization, which changes their values to lie between 0 and 1. This is done so that the features with bigger numerical ranges don't overshadow others during learning. For the sake of comparison, the developed features were split up into three different feature groups:

- **Traditional Features**, packaging borrower demographics (such as age, gender, education) and financial characteristics in the conventional sense.
- **Alternative Features**, describing mobile and behavioral ecosystem-related indicators
- **Combined Features**, featuring both traditional and alternative attributes

For each feature group, individual datasets were created. Loan Status was still kept as the target variable. This method of grouping will allow the researchers for a thorough assessment of the model's performance using different feature sets and it will help to achieve quite directly the objectives of the study's experiment.

3.3 Model building

As we discussed in Chapter 2, constructing a Machine learning (ML) model refers to the task of developing a model that has the capability to learn from a dataset and then to classify new, previously unseen examples. The model learning and testing require datasets for training and testing, and performance metrics.

3.3.1 Model Training Techniques

Loan prediction usually involves many different variables such as income, credit history, type of employment, etc., and their interaction can be very complicated. Random Forest can recognize these complex patterns and relationships between input variables even if we haven't created specific features for each possible interaction.

Minimizing Overfitting: Decision trees are quite prone to overfitting, particularly when they are very deep. Random Forest method helps in overfitting by averaging the outputs of multiple trees where each tree is built on a different random subset of data and features. This method of ensemble or bagging not only cuts down the variance of the model but also the process of generalizing that model to new data is improved.

Feature Importance: Random Forest can show which features are more important by calculating their relative influence on the model's output. In predicting loans, this

information can be a great help not only to pinpoint the main reasons behind loan defaults but also to understand the overall lending process.

Handling Various Data Types: Random Forest is capable of dealing with both types of features numerical and categorical even though one-hot encoding of these features is mostly beneficial. It does not require any complex preprocessing like scaling.

Robustness to Outliers and Noise: Random Forest, due to its ensemble nature and random sampling, does not get easily affected by the presence of outliers and noise points in the training data.

Handling Missing Values (Indirectly): Random Forest, in itself, doesn't explicitly handle missing values but the ways in which individual decision trees are constructed can indirectly manage them. For instance, some versions use surrogate splits or allocate missing values to the major class during the tree building phase. However, the preferred approach is to deal with missing values during the data cleaning stage before the model is trained. **Key Formula:** For regression, the Random Forest prediction for input x is:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \tag{3.4a}$$

where N is the number of trees and $T_i(x)$ is the prediction from the i-th tree [29]. For classification, the final class is determined by majority vote among all trees. [29]. Firstly,

Table 1: Steps in Ensemble Methods

Step	Description
Bootstrapping	Random samples are drawn from the original dataset with replacement to construct each individual decision tree.
Feature Bagging	At each split in the tree, a random subset of features is selected to determine the best split.
Aggregation	Final predictions are obtained by majority voting for classification tasks or by averaging for regression tasks.

by using the randomness of data and feature, Random Forest becomes a strong and powerful method for various prediction challenges [29]. Talking about this project: We developed a Random Forest model by using traditional and combined features to identify the status of loans ('Paid' or 'Defaulted'). The model analyzed feature patterns to decide

whether a loan would be paid back or defaulted. By looking at its score against models like XGBoost and LightGBM the loan prediction capability of Random Forest is measured and the role of different features in its success becomes clear.

3.3.2 XGBoost (Extreme Gradient Boosting)

XGBoost (short for Extreme Gradient Boosting) is a very fast and strong gradient boosting algorithm which recently gained popularity as a tool for most machine learning problems, including loan prediction. It is a kind of ensemble learning method that creates decision trees one by one and each subsequent tree tries to fix the errors of the previous ones [30]. Here's how XGBoost applies to loan prediction:

Very Good Performance: XGBoost has a track record of excelling in a wide array of tabular datasets. The boosting strategy enables it to unearth intricate patterns and interactions in the data, hence it is capable of producing highly precise predictions of the risk of loan default.

Regularization: XGBoost implements several regularization methods (L1 and L2 regularization) to control overfitting. This aspect is critical in the realm of loan prediction, where a model should continue to perform well even when it is exposed to new loan applications that it has never seen before during training.

Missing Values Handling: One of the greatest features of XGBoost is its ability to automatically deal with missing data. It determines the most appropriate direction to take (left or right in a tree split) when a value is missing, from the analysis of the training data. Thus, the manual data cleaning step of imputing missing values can be avoided.

Very Fast and Scalable: The XGBoost tool was made with the goal of high performance and can process massive datasets without any trouble. Its ability to perform computations in parallel is one of the reasons why it is capable of handling training on huge amounts of loan data.

Feature Importance: Similar to Random Forest, XGBoost gives a measure of feature importance, i.e. it tells how each feature has contributed to the final prediction. This can be quite helpful in identifying most of the important factors that have an influence

on whether a loan will be defaulted or not. On top of these, it allows user to do both classification (e.g. Paid or Default as loan status) and regression (e.g. loan amount, interest rate, etc. are continuous variables) depending on their requirement. As far as this work is concerned, we used traditional and combined feature sets separately to train two different XGBoost models for loan status prediction. Usage of XGBoost boosting and regularization features enabled us to develop an accurate model capable of pinpointing potential loan default candidates. Subsequently, we evaluated XGBoost along with Random Forest and LightGBM models for their efficacy as powerful ensemble methods [30] [31].

3.3.3 LightGBM (Light Gradient Boosting Machine)

LightGBM is a highly popular and exceptionally efficient gradient boosting framework. It is famous for its quicker speed and great performance especially when dealing with big data. Like XGBoost, it also builds trees sequentially, each time fixing the errors of the previous one. In this way LightGBM is implemented in loan prediction [32]:

Speed and Efficiency: LightGBM is remarkably faster compared to normal gradient boosting methods, thus it is perfectly able to handle huge loan datasets. It employs two ingenious methods namely Gradient-based One-Side Sampling (GOSS) to select data points and Exclusive Feature Bundling (EFB) to bundle mutually exclusive features, hence speeding up the training process without compromise on accuracy.

High Accuracy: Along with speed, LightGBM is able to keep high accuracy, in fact it is usually on par with or even surpasses XGBoost. Due to its boosting technique, it is capable of finding complex structures and interactions in the data.

Handling Large Datasets: LightGBM is very good at processing huge data composed of many features, which is a typical scenario for financial applications such as loan prediction.

Regularization: There are regularization parameters in LightGBM that help in minimizing overfitting, thus enabling the model to give good results even on new loan applications. Feature Importance: LightGBM allows one to visualize the importance of features so that a better understanding can be obtained about which features have the greatest impact on the prediction of loan default.

Handling Categorical Features: LightGBM is able to handle categorical features by itself, which means that one does not have to do this manually all the time. Nevertheless, here, we used one-hot encoding for all models as a standard method [33] [34].

With regard to this project: We built a LightGBM model to predict loan status based on both traditional and combined feature sets. LightGBM was our choice for quality, speed, and performance, especially on large datasets. We also added it to our analysis to check if it can deliver results better than or at least comparable to Random Forest and XGBoost. Having the LightGBM's performance metrics in different feature sets was helpful in understanding how good it is at loan prediction.

3.4 Evaluation

We used the standard classification metrics to assess the performance of the model and to see the complete scenario.

True positive (TP): Events when the model correctly forecasts a failure. Example: The model marked the client as a risky loan defaulter, and the client did really default.

True Negative (TN): Events when the model correctly forecasts no failure. Example: The model forecast a client would pay, and that is what happened.

Flase Positive (FP): Events when the model forecasts a failure, but the failure does not happen. So, the model tagged a client as high risk, but the client ended up paying the loan on time.

Flase Negative (FN): Events when the model doesn't forecast a failure, but the failure happens. e.g., The model forecast a client would pay, but the client ended up defaulting. This mistake is most of the time a very costly one in the loan prediction scenario.

Confusion matrix serves as the foundation for the computation of the following three metrics.

- **Accuracy:** This is a measure of how right the majority of the model's predictions were. For example, if your model made a correct prediction for 80 out of 100 loan applications for the categories 'Paid' and 'Defaulted', your accuracy would be 80%.

Mathematically:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4a}$$

- **Precision**: This focuses on the accuracy of positive predictions. In loan prediction, if 'Defaulted' is our positive class, Precision tells us, "Of all the loans the model predicted would default, how many actually did?" A high precision means fewer loans are incorrectly flagged as defaulted. [35].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.4b}$$

- **Recall**: This measures the model's ability to find all actual positive cases. In loan prediction, if 'Defaulted' is our positive class, Recall tells us, "Of all the loans that actually defaulted, how many did the model correctly identify?" A high recall means fewer actual defaulted loans are missed by the model. [35].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4c}$$

- **F1-score** This is the harmonic mean of Precision and Recall. It provides a single score that balances both metrics, which is especially useful when dealing with imbalanced datasets (like loan defaults, where 'Paid' loans are usually far more common than 'Defaulted' ones). A high F1-score indicates a good balance between catching defaulted loans and not incorrectly flagging too many healthy loans.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4d}$$

These metrics were computed on the held-out 20% test set and averaged across cross-validation folds to ensure a robust and reliable evaluation of model performance. In summary, this chapter has presented the proposed system model, which outlines the steps required to prepare for the experimental analysis in the subsequent chapter. The dataset was collected and preprocessed through data cleaning, normalization, and feature engineering. The feature preparation phase included feature selection, transformation,

and the construction of traditional, alternative, and combined feature sets. Three machine learning classifiers Random Forest, XGBoost, and LightGBM were trained and evaluated using both stratified train–test splitting and cross-validation techniques. At the end of this process, the prediction results were obtained using standard performance metrics. Therefore, Chapter 4 discusses the implementation of these steps and analyzes the experimental results.

4 EXPERIMENTS

The main focus of this chapter is on the experiment. It starts by discussing the experimental setup. Then it continues with the implementation procedure, explaining in detail how the dataset was collected, analyzed, and prepared the relevant features as well as how the model was built for experimentation.

4.0.1 Tool

Experimental realization has been done on a platform consisting of a 12th Gen Intel® Core™ i7-1260P processor (2.10 GHz) with 8.00 GB RAM and a 64-bit Windows 10 OS. Python (3.13) was the programming language used in the Jupyter Notebook environment for visualizing, cleaning, and training data through various models.

4.1 Dataset Description

The dataset of this research consists of 50, 000 records of Tele birr Super App users of Ethio Telecom. It combines conventional financial measures with alternative digital behavior data, thus presenting the full picture of individuals' credit profiles in the context of an emerging market. The conventional data part contains such variables as loan ID, age, gender, marital status, educational level, employment type, income, credit history, previous loan amount, previous loan status, and requested loan amount. On the other hand, the alternative data part comprises total top-up amount, telecom bill payment pattern, equb membership, and equb contribution amount. These are the variables that demonstrate user behavior through mobile service usage as well as community-based local savings groups (Equb), which in certain cases can be very revealing of a person's financial trustworthiness. Combining both types of data, this dataset offers a tool for a multidimensional analysis of creditworthiness and financial inclusion in a world of digital finance. The data file mixes different types of data: categorical and numerical, which is the normal situation for loan prediction projects. There are 9 categorical columns out of the total 15 (which is the object type of data) and this group in a way includes the features or

attributes gender, marital status, employment type, region, and equb membership. These are qualitative characteristics describing the applicants in terms of their demography and behavior. The other 6 columns are numeric (int64), e.g., age, income, previous loan amount, and equb contributions, reflecting the applicant's level and changes of financial resources.

Table 2: Dataset Column Description

No	Column Name	Non-Null Count	Data Type
0	Loan ID	50,000	String
1	Age	50,000	int64
2	Gender	50,000	category
3	Marital Status	50,000	object
4	Education Level	37,514	object
5	Employment Type	50,000	object
6	Income	50,000	int64
7	Previous Loan Amount	50,000	int64
8	Loan Status	39,991	object
9	Loan_Term_Months	50,000	int64
10	Total Top-up Amount	50,000	int64
11	Telecom Bill Payment Pattern	50,000	object
12	Equb Membership (Yes/No)	50,000	object
13	Equb Contribution Amount	50,000	int64
14	Region	50,000	object

The initial dataset for this research contained 50, 000 loan applicant records that had the demographic, financial, and behavioral attributes that can be used to predict loan status. When the data was being cleaned, the missing values in the Education Level column were filled in so as to keep all the records associated with them, whereas the rows with missing Loan Status values were removed. This left the final dataset with 39, 991 rows. Since no duplicates were found, it was confirmed that the dataset was both complete and unique for the next analysis. Data wrangling methods were used to preprocess the data ready for machine learning. With one-hot encoding, categorical variables were changed into numerical features, and then these were normalized so that all features would have the same level of contribution. The dataset was not changed in terms of the number of observations, which still stood at 39, 991 rows. After that, three

different feature sets were created: `df_traditional` (the usual demographic and financial features), `df_alternative` (non-traditional or behavioral features), and `df_combined` (both traditional and alternative features together), and all three contained the same number of rows. Stratified sampling was used to split the data sets as it ensured the class distribution remained the same as the original, which resulted in about 31, 992 rows for the training set and about 7, 999 rows for the test set for each feature set. The machine learning algorithms: Random Forest, XGBoost, and LightGBM were used to train models under two different conditions: one where only traditional features were used and the other where the combined feature set was used. For all the models, the addition of alternative features was a consistent factor in the improvement of the prediction performance, with the raising of the metrics such as accuracy, recall, and generalization. These findings show that alternative features are a good source of additional information and as a result, the prediction of loan status is more robust and dependable.

Table 3: Dataset Preparation Steps and Number of Rows

Step	Preparation Step	Number of Rows
1	Initial Data Loading	50,000
2	After Removing Missing "Loan Status" Values	39,991
3	After Data Wrangling (Encoding/Normalization)	39,991
4	Feature Engineering (Traditional / Alternative / Combined)	39,991
5	Data Splitting – Training Set	31,992
6	Data Splitting – Testing Set	7,999

4.2 Data visualization

Data visualization is a very helpful means in data understanding and supporting decision-making processes. Our dataset entails 50000 rows and 15 columns containing loan information. A row denotes an individual loan application, while each column provides different aspects of the applicant's status and loan particulars. Major features consist of: the applicant's personal information, their income, employment details, loan amount, and the loan condition. Our first step is to look at the target, the Loan Status, which is a categorical feature. We will look at the frequency of each category, their percentage share,

and use a bar chart to visualize the distribution. In total, 50, 000 loan application data entries are in the dataset. Each application is differentiated by a unique Loan ID. More female applicants (25, 084) than males (24, 916) but the gender divide is almost equal. Four types of marital status are almost equal with Single (12, 630), Married (12, 584), Divorced (12, 437), and Widowed (12, 349). Education level feature contains Tertiary (12, 538), Primary (12, 506), and Secondary (12, 470) levels, however, about 25% of the data in this column is missing. Employment type is also nearly equal across Student (12, 654), Employed (12, 532), Self-employed (12, 468), and Unemployed (12, 346). Target variable Loan Status reveals that 30, 009 applicants were ("Paid") while 9, 982 were ("Defaulting"). This situation indicates a class imbalance which should be taken into account when the model is being designed. The behavioral characteristics include telecom bill payment pattern, where there are almost equal numbers of late (16, 994), no payment (16, 556), and on-time (16, 450) behaviors. Equb membership, a traditional Ethiopian savings group, is an alternative financial feature, where majority of the applicants (29, 955) are members of an Equb, while 20, 045 are not. As for the region, the data shows diversity with the highest proportion being from Addis Ababa (12, 563), followed by Oromia (9, 999), Amhara (9, 958), SNNPR (7, 442), Other (5, 076), and Tigray (4, 962). These categorical distributions imply that the dataset contains good loan outcomes features from both traditional and alternative perspectives for modeling. The below highlights provide insight into the main distribution patterns of the numerical features. Distribution of Age: This histogram presents the number of cases in each age category. The distribution seems to be quite broad but with a concentration of younger and middle-aged individuals. Thirty bars are used here to fine-tune the details.

Box Plot of Monthly Income (in Birr): This box plot illustrates the range and the median of the income among the loan applicants. The graph effectively communicates the fact that there is a drastic positive skew in the distribution of the incomes, with many outlying figures being far greater than the standard values.

Box Plot of Loan Amount (in Birr): The graph shows that the shape of the loan amounts distribution mirrors that of the incomes one: strongly right-skewed with quite

several outliers. It leads to the idea that the concentration of customers with small loans is higher whereas few customers have large loans.

Distribution of Loan Term (in Months): The histogram of the loan term duration suggests a little variety in the loan terms possible, not continuous ones but in discrete intervals (e.g., 6, 12, 24, 36 months).

Box Plot of Total Top-up Amount (in Birr): The visualization indicates that the behavior of total top-up amount is very much like that of income and loan amount in terms of distribution and presence of outliers.

Box Plot of Monthly Telecom Bill (in Birr): Similar to income, it represents distribution of monthly telecom bills and points out the highly right skewedness of the data.

Box Plot of Equb Contribution (in Birr): This is another right-skewed feature with many outliers, consistent with other income-, expense-, and loan-related characteristics. The scatter plot in Figure 4.2 shows Income on the x-axis and Previous Loan Amount on the y-axis for the 50,000-loan-application dataset. The highest concentration of points appears in the low-to-mid ranges of both income and loan amount. This pattern supports the right-skewed nature of both variables, where only a small number of applicants have very high incomes or previous loan amounts, while most observations cluster at lower values. Although the points are densely distributed, the plot suggests only a weak direct relationship between income and previous loan amount in this dataset. In addition, because the figure does not separate observations by loan status (Paid or Defaulted), it does not clearly show how repayment behavior varies with these two variables alone. This indicates that, beyond income and prior borrowing, loan repayment should be analyzed using additional behavioral and demographic features. Overall, this scatter plot highlights the complexity of loan risk prediction and the need for multi-feature modeling.

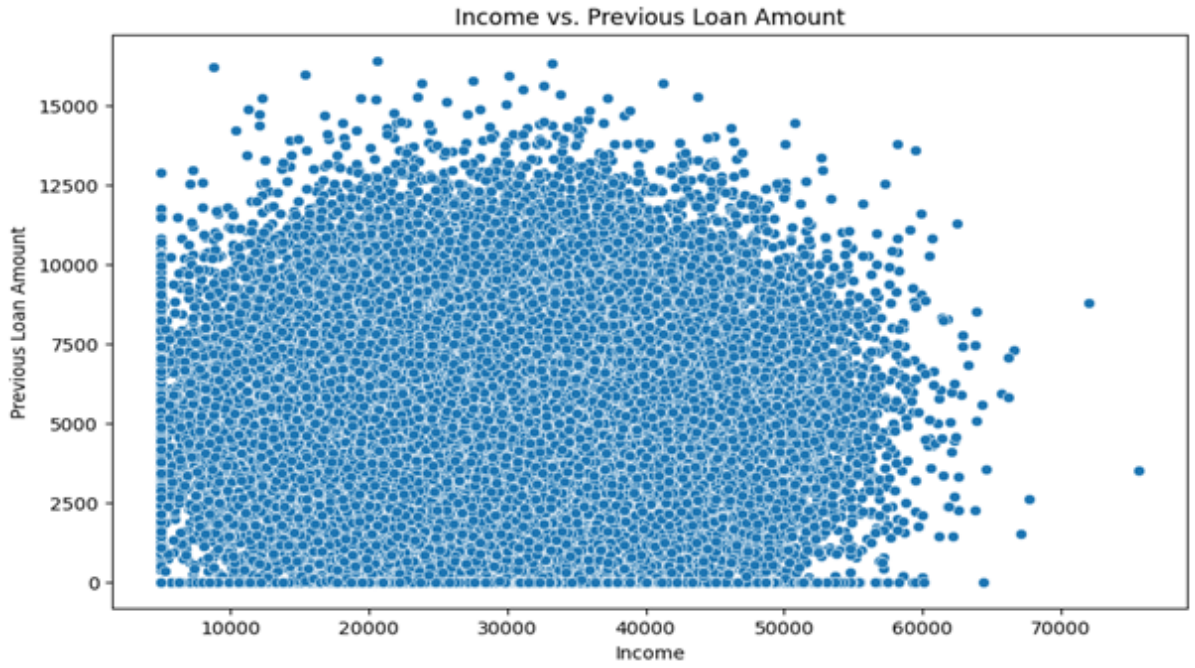

Figure 4.2: Scatter Plot of Income vs. Previous Loan Amount

Figure 4.3 displays bar charts depicting how different categorical variables relate to loan status (Paid or Defaulted). The series of visualizations came about after a data issue was resolved, specifically a leading space in the `Loan Status` column. The revised charts offer a firm ground for understanding which categorical factors lead to a higher rate of default or successful loan repayment.

Gender vs. Loan Status: Figure 4.3 shows a comparison of loan performance between men and women on the bar chart, indicating almost no significant difference in loan outcome between genders.

Marital Status vs. Loan Status: The results from the four marital statuses Single, Married, Divorced, and Widowed, point towards a general trend where more loans are paid than defaulted in all groups. However, small differences catching the eye indicate that Single and Widowed might be more prone to defaulting on loans thus they are worth a more detailed examination.

Education Level vs. Loan Status: Higher education levels correspond to better loan repayment percentages than defaults for Primary, Secondary, and Tertiary categories.

Default proportions look quite similar visually for all three types, which means that the education factor here may not be the strongest one when it comes to differentiating between loan cases.

Employment Type vs. Loan Status: While the default rate is marginally raised for the Self-employed and Unemployed groups, these two categories might be indicative of more income instability or less access to stable financial resources as a possible reason.

Telecom Bill Payment Pattern vs. Loan Status: The behavioral characteristic in question here is going to lead us to a quite drastic differentiation. On the one hand, those applicants who pay their telecom bills on time continuously are coming up with fewer default cases, and on the other, those who are either late or no payments majority of the times are presenting a greater frequency of loan defaults. What we see here is a linkage that is quite strong between a person's telecom payment behavior and their reliability in loan repayment.

Equb Membership vs. Loan Status: Being a member of an *Equb* (a traditional Ethiopian rotating savings group) is an activity that applicants doing it have shown a substantially lower rate of loan default compared to those who are not members. This points to the advantage of including alternative financial data in credit risk modeling and also casts Equb member status as a hallmark of financial discipline and social trust.

Region vs. Loan Status: The feature of regional distribution of loan results has revealed changes in the way various geographical areas of the country achieve moving towards more loan repayments or loan defaults. Applicants from Addis Ababa have the highest numbers of loans paid while some other regions, such as Tigray and *Other*, have relatively higher default rates when taking into account their total number of applications, respectively. This can be attributed to regional economic differences, disparities in financial literacy or inaccessibility of financial resources.

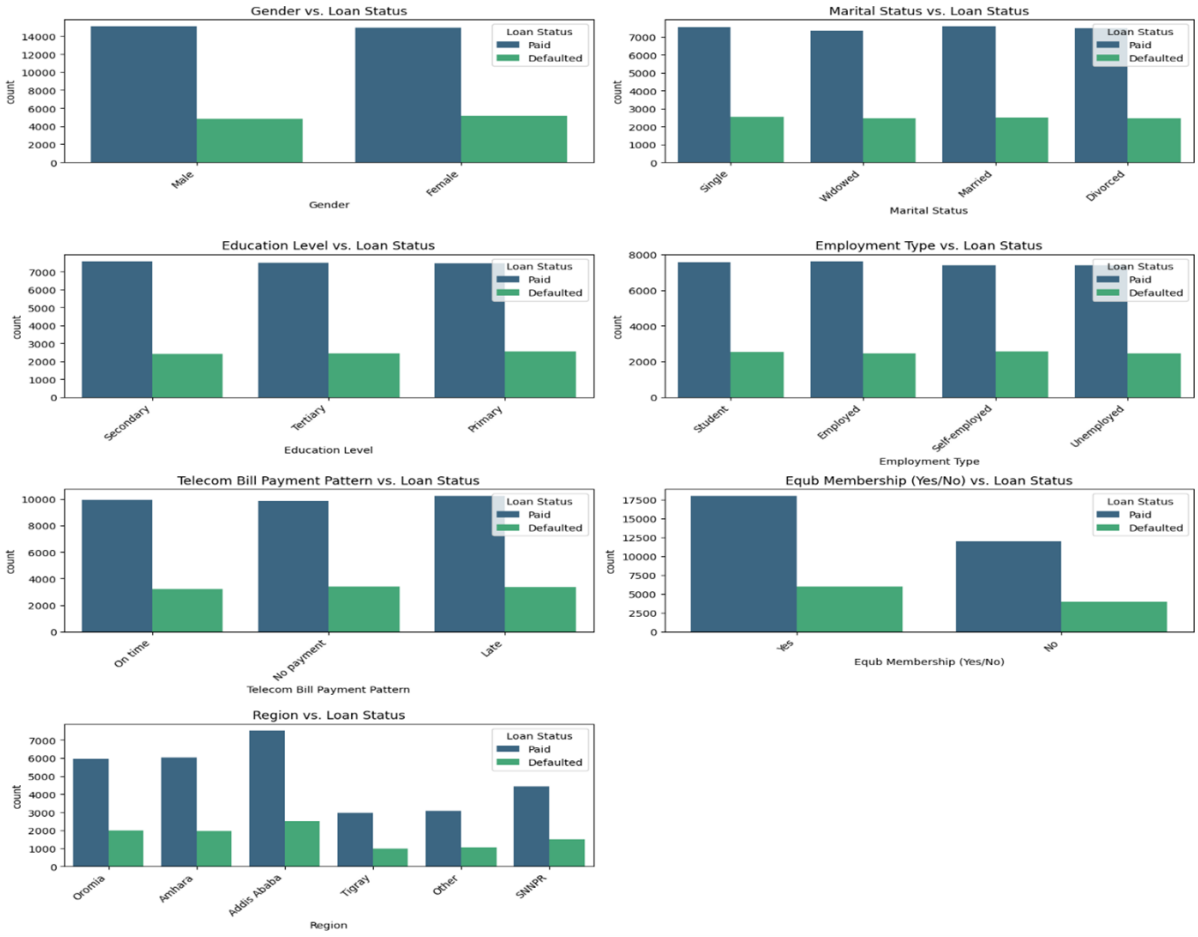

Figure 4.3: Count plots of categorical features vs. loan status.

4.3 Data Cleaning

In this experiment, normalization, punctuation removal, and stop word removal were applied to clean the dataset.

4.3.1 Normalization

We decided to take the route of feature normalization for our study with selected numerical features only in order to make them comparable across variables and to boost the performance of machine learning models. More precisely, *Income* and *Equb Contribution Amount* were the two features targeted for normalization as they had broad value ranges and were measured in different units/scales. Ahead of normalization being applied, a descriptive statistical analysis was executed to determine the distribution and magnitude of the variables. The summary statistics showed that both features varied substantially: Income ranged from 5, 000 to 75, 621 Birr whereas Equb Contribution Amount ranged from 0 to 95, 904 Birr. Such discrepancies in scale can adversely influence distance-based and tree-based learning algorithms if not taken care of. In order to resolve this problem, **Min–Max Normalization** was used. This method adjusts each feature into the same scale of 0 to 1, the one given by Equation 1: [36]

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

MinMaxScaler was selected because it preserves the original distribution shape while ensuring that all values fall within a uniform range. After normalization, both variables

Table 4: Descriptive Statistics of Features Before Normalization

Feature	**Mean**	**Std. Dev.**	**Minimum**	**Maximum**
Income	30,060.53	9,929.84	5,000	75,621
Equb Contribution Amount	9,594.20	14,927.92	0	95,904

were successfully scaled to the $[0, 1]$ range. The minimum values became 0, and the maximum values became 1, confirming the correct application of the MinMaxScaler. The transformed features exhibited reduced variance while maintaining their relative

distribution patterns. This normalization ensured that Income and Equb Contribution Amount contributed proportionally during model training and prevented features with larger numeric ranges from dominating the learning process.

4.4 Feature Engineering

To build predictive models, the raw data was initially subjected to a series of feature engineering and transformations to adapt it better for the machine learning algorithms. The preprocessing consisted mainly of filling missing values for categorical variables, removing features that don't affect the prediction, and turning the semi-features into meaningful ones.

4.4.1 Features Removed or Transformed

During the preprocessing stage, several features of the initial dataset were dropped or changed as follows:

- **Categorical Columns (8 features):** Original categorical variables such as "Gender", "Marital Status", "Education Level", etc., were transformed into one-hot encoded vectors. In effect, this method split each categorical feature into multiple binary features and completely removed the original categorical features from the dataset.

- **Loan ID (1 feature):** Since every Loan ID was a unique value and had no predictive capability, the Loan ID column was removed.

- **Loan Status:** The Loan Status feature was first one-hot encoded into two separate columns, `Loan Status_paid` and `Loan Status_defaulted`. When the model was trained, `Loan Status_paid` was considered the dependent variable, thus excluding `Loan Status_defaulted` from the feature set.

4.4.2 Features Added through One-Hot Encoding

One-hot encoding of eight categorical features made them extend to twenty-six new binary features. Thus there was a net increase of eighteen features versus the original data set. Below is the list of the newly added features:

- **Gender:** `Gender_female`, `Gender_male`
- **Marital Status:** `Marital Status_divorced`, `Marital Status_married`, `Marital Status_single`, `Marital Status_widowed`
- **Education Level:** `Education Level_primary`, `Education Level_secondary`, `Education Level_tertiary`
- **Employment Type:** `Employment Type_employed`, `Employment Type_self-employed`, `Employment Type_student`, `Employment Type_unemployed`
- **Loan Status:** `Loan Status_paid`, `Loan Status_defaulted` (`_paid` used as target variable)
- **Telecom Bill Payment Pattern:** `Telecom Bill Payment Pattern_late`, `Telecom Bill Payment Pattern_no payment`, `Telecom Bill Payment Pattern_on time`
- **Equb Membership (Yes/No):** `Equb Membership (Yes/No)_no`, `Equb Membership (Yes/No)_yes`
- **Region:** `Region_Addis Ababa`, `Region_Amhara`, `Region_Oromia`, `Region_Other`, `Region_SNNPR`, `Region_Tigray`

4.4.3 Feature Grouping

After being preprocessed for predictive modeling, the features were segregated into three distinctive groups.

- **Traditional Features (18 features):**

  Numerical: `Age`, `Income`, `Previous Loan Amount`, `Loan_Term_Months`, `Total Top-up`

Amount

One-Hot Encoded: Gender_female, Gender_male, Marital Status_divorced, Marital Status_married, Marital Status_single, Marital Status_widowed, Education Level_primary, Education Level_secondary, Education Level_tertiary, Employment Type_employed, Employment Type_self-employed, Employment Type_student, Employment Type_unemployed

- **Alternative Features (6 features):**

  Numerical: Equb Contribution Amount

  One-Hot Encoded: Telecom Bill Payment Pattern_late, Telecom Bill Payment Pattern_no payment, Telecom Bill Payment Pattern_on time, Equb Membership (Yes/No)_no, Equb Membership (Yes/No)_yes

- **Combined Features (24 features):** This feature set contains all the features of the Traditional as well as Alternative groups.

  In each predictive modeling session, the feature Loan Status_paid was made use of as the target variable.

4.5 Handling Missing Values

The dataset was preprocessed so that it can be fed to the machine learning model. First, the data was examined for missing values, inconsistencies and features that don't really help with predicting. Different techniques were employed to deal with missing data based on the feature type and how much the feature was needed for making a prediction.

Education Level The categorical feature *Education Level* had around 24.97% missing values. This was resolved by **mode imputation**, i.e., the missing entries were filled with the value of the most frequently occurring category in the column. Mode imputation is a really good option for categorical variables when the proportion of missing values is moderate and there is no indication of systematic bias. Besides, this method results in a

bigger dataset, preserves consistency, and doesn't cause any artificial category creation, or lean on complicated statistical assumptions. Through imputing with the most prevalent education level, the representativeness of the dataset is maintained, and the feature engineering efforts are supported.

Loan Status On the contrary, the *Loan Status* column, which is the target variable of the prediction, had about 20.02% missing values. Instead of trying to fill the missing values, all the rows with missing target values were deleted. Filling the target variable may lead to bias, incorrect labels and will lower the integrity and quality of the models in terms of prediction. Incomplete rows were removed to make sure that each data point was a source of trustworthy information for supervised learning. Dataset size dropped from 50, 000 to 39, 991 records after the rows removal, but it was still a great number for good model training.

Duplicate Check Besides missing value treatment, the dataset was looked for duplicate entries in order to avoid redundancy and data leakage. No duplicate records were found this way, ensuring that every observation contributes unique and valid information to the modeling process.

To sum up, these preprocessing steps such as missing value handling, removal of non-predictive identifiers and data integrity checks led to a dataset that was clean and consistent. This setting is a very good basis for feature engineering, model development and reliable prediction of loan repayment outcomes.

4.6 Cross-Validation

During stage one, this paper mainly employed 10-fold stratified cross-validation for hyperparameter tuning of Random Forest and LightGBM models with the combined feature set. This made sure the evaluation of the models was reliable and fair. Initially, the `StratifiedKFold` technique was chosen for splits of 10, together with `shuffle` set to `True` and `random_state` set to `42`. Via this, the dataset is partitioned into 10 equal parts (folds) while keeping the same ratio of loan statuses, *Defaulted* and *Paid*, in each fold as

in the overall training dataset. It is important to keep this distribution in loan prediction scenarios especially when there exists a class imbalance.

Finding the best hyperparameters for the models was done through `RandomizedSearchCV`. This method enables an economical searching among various hyperparameter combinations without the need for exhaustive evaluations. Meanwhile, the 10-fold cross-validation setting was utilized. Every fold served as validation set once. The F1-score was chosen as the main metric of measurement since it combines precision and recall and is thus very effective for imbalanced datasets like loan default prediction.

According to cross-validation, the final tuned Random Forest model has the highest cross-validated F1-score of 0.8574. In like manner, the best cross-validated F1-score of LightGBM model after tuning was 0.8574 as well. In conclusion, using 10-fold stratified cross-validation not only guaranteed that the hyperparameters selected for both models were robust and not a result of a single random split of the training data but also came along with increased faith in the capability of the final models.

4.7 Imbalanced Data and Dataset Splitting

Data imbalance occurs when the distribution of the target variable is unequal across its classes (Figure 4.4). In this loan prediction dataset, exploratory data analysis revealed that the *Loan Status* variable is imbalanced, with a substantially higher number of *Paid* instances compared to *Defaulted* instances. In this context, *Defaulted* refers to borrowers who failed to repay their loans according to the agreed terms. To mitigate the negative impact of data imbalance during model training and evaluation, **stratified sampling** was employed when splitting the dataset into training and testing sets. Stratified sampling ensures that the proportion of instances in each class of the target variable is preserved in both subsets, reflecting the original class distribution. By using `stratify = y` in the `train_test_split` function, the minority class (*Defaulted*) was adequately represented in both the training and testing sets for each of the three feature groups (*Traditional*, *Alternative*, and *Combined*). This approach enables the models to learn from a representative distribution of both classes and provides a more reliable evaluation of performance

on unseen data, particularly for metrics such as *precision* and *recall*, which are more informative than accuracy in the presence of class imbalance.

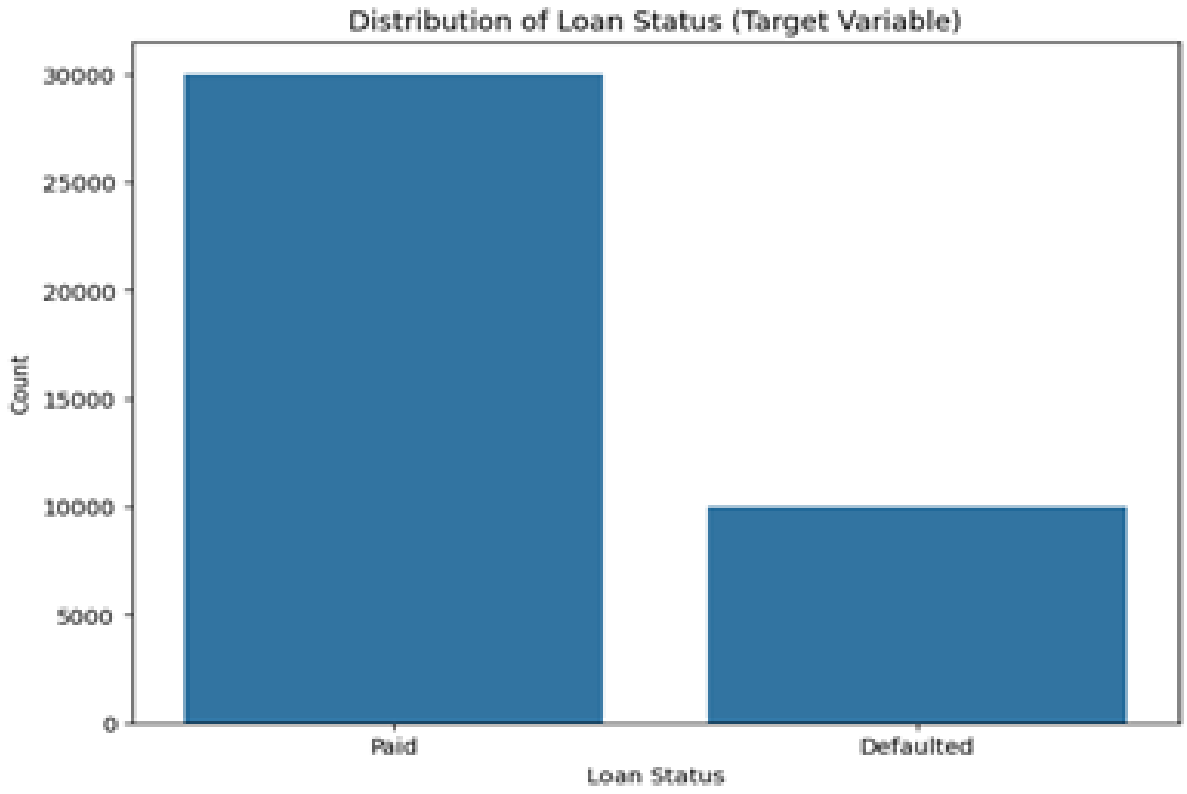

Figure 4.4: Distribution of loan status

5 Experiment Setup

In the experiment, different feature sets were tested to see which ones can best help machine learning models predict loan repayments. After data loading, the loan prediction dataset was put in a pandas DataFrame, and a try–except block was implemented for the case when there may be an error in a file being read. Then, **exploratory data analysis (EDA)** was performed to check the dataset structure, find missing values, and get the distributions of numerical and categorical features through summary statistics and visualizations.

Data cleaning steps were taken to find missing, inconsistencies and `Loan Status` column was fixed and outliers in numerical attributes were taken care of using the **Interquartile Range (IQR)** method besides making sure the data types were correct. During the **data wrangling** step, one-hot encoding was used to transform categorical variables, and features such as *Income* and *Equb Contribution Amount* were normalized by **MinMaxScaler** method so that all features would be of comparable scale.

During **feature engineering**, the dataset was divided into *Traditional features*, *Alternative features*, and *Combined features*, each organized into separate DataFrames including the target variable for independent model evaluation. Here an 80/20 train–test split was carried out by stratified sampling using a to maintain the proportion of Paid and Defaulted loans the two subsets.

Model training consisted of training the three Random Forest, XGBoost and LightGBM classifiers independently on the Traditional and Combined feature sets with the default hyperparameter settings. First model evaluation was performed by applying the models to the respective test sets and then standard classification metrics such as Accuracy, Precision, Recall, and F1-score were reported for comparison of the two feature sets. As a measure of their reliability, the models together with the feature sets were subjected to stratified 5-fold cross-validation with the performance metrics being the mean of the folds which represent the best estimates of the expected performance on new, unseen data.

newpage

Experiment 1: Impact of Alternative Data Features on Model Performance (RQ1) This experiment was conducted to assess the value of incorporating non-traditional data sources into the loan prediction process. *Traditional features* typically include demographic information, income, credit history, and previous loan details. *Alternative features*, in this context, refer to data points such as telecom bill payment patterns, *Equb* (informal savings group) membership, and Equb contribution amounts. The hypothesis is that these alternative data points provide additional insights into an individual's financial behavior and stability, potentially leading to more accurate loan default predictions, particularly for individuals with limited traditional credit history. By comparing the performance of models trained solely on traditional features against those trained on a combination of traditional and alternative features, this experiment evaluates whether these non-traditional data sources offer a significant improvement in predictive power.

Experiment 2: Model Comparison for Loan Prediction (RQ2) This experiment was conducted to identify the most effective algorithm among the selected models for this loan prediction task and dataset. The evaluated models were **Random Forest**, **XGBoost**, and **LightGBM**, all of which are ensemble learning methods known for strong predictive performance. Random Forest is a bagging-based ensemble method, while XGBoost and LightGBM are gradient boosting algorithms. Each model exhibits different strengths and weaknesses, and their performance varies depending on dataset characteristics. By training and evaluating each model on the same data (both Traditional and Combined feature sets) and comparing their performance across multiple metrics (*accuracy*, *precision*, *recall*, and *F1-score*), this experiment determines which model is best suited for predicting loan outcomes in this scenario.

6 RESULT AND DISCUSSION

This chapter shows the results from the model experimentation, in which the machine learning models' performances were evaluated through the four most important metrics: *accuracy*, *precision*, *recall*, and *F1-score*. Accuracy gives us the overall correctness of the model's predictions by counting correct predictions (true positives plus true negatives) and comparing them to total instances. Precision tells how trustworthy the positive predictions are by giving the ratio of the true positives among all positive predictions, and it is main concern in reducing false positives. Recall (or sensitivity) measures the ability of the model to find all the really positive cases by looking at the proportion of true positives to the total number of positives, highly considering the reduction of false negatives. At last, the F1-score which is the harmonic mean of precision and recall gives a good measure of the model performance balance. It is mainly used to balance the precision and recall trade-off in the case of such situations when the accuracy is misleading due to the presence of the unbalanced datasets. In short, these metrics together help us understand thoroughly the strong points and weaknesses of the models that were tested in this work.

6.1 Experimental Results

Experiment 1: Impact of Alternative Features on Model Performance (RQ1) This research explores the potential benefit of alternative data features in loan prediction models over using only traditional features. This research explores the potential benefit of alternative data features in loan prediction models over using only traditional features.

Judging by the performance figures shown in Table 5, the use of alternative features has brought out different reactions in the models. To take an example, Random Forest showed some improvements after being endowed with the joint feature set for training not only in *Accuracy* but *Recall*, and *F1-score* as well, which demonstrates that alternative features supply some additional information when it comes to loan prediction. The XGBoost, similarly, experienced a slight increase in all metrics, though the difference was

minor. Conversely, the LightGBM outcomes are either very similar or if changed slightly (Accuracy and F1-score) have been achieved through combined features, which means that alternative features barely enhanced its prediction power and potentially caused a little bit of noise.

Briefly put, alternative features aligned somewhat with the Random Forest and XGBoost, yet produced very little effect and rather a negative one, even on LightGBM. To put it differently, the usefulness of alternative data varies with the model, and that one should not take for granted that it is going to bring about performance improvement all the time.

Table 5: Single Train–Test Split Results (Evaluation on 20% Test Set)

Feature Set	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Traditional	Random Forest	0.7442	0.7499	0.9890	0.8530
Traditional	XGBoost	0.7425	0.7511	0.9822	0.8513
Traditional	LightGBM	0.7500	0.7504	0.9992	0.8571
Combined	Random Forest	0.7483	0.7500	0.9970	0.8560
Combined	XGBoost	0.7368	0.7498	0.9745	0.8475
Combined	LightGBM	0.7493	0.7502	0.9985	0.8567

Analyzing the machine learning models gave us quite a few valuable clues about their effectiveness at prediction and the nature of the dataset as well.Random Forest and LightGBM were the best performers among the tested algorithms. These two models achieved a mean F1-score of 0.8574 based on the 10-fold cross-validation, which is a very good sign of their ability to predict loan outcomes. XGBoost, on the other hand, came up with a mean F1-score of 0.8495 when run with default settings. The gap between the scores emphasizes the role of hyperparameter tuning since the right set of parameters helps a model to better 'fit' the data and thereby gain in prediction power.

Apart from high recall and strong F1-scores which the models exhibited, a point to note is that they were, in fact, picking up the majority class (Paid loans) quite well, which is no mean feat. One thing that also stands out in the cross-validation results is the difference between folds which is almost negligible, showing that the models are giving quite a steady and reliable predictions and their performance is not heavily influenced by a particular data split.

On the other hand, when we dive deeper into the prediction results, we see a big number of errors in the classification. For example, the optimized LightGBM model made 1, 996 false positive calls or instances where defaulted loans were categorized as paid. Unlike false positives, there was only one true negative instance identified. Such an imbalanced situation almost automatically put a brake on the extent to which a detailed disparity between correct and false classifications of defaulted loans could be explored.

We resorted to SMOTE and threshold optimization methods to fix the issue of imbalanced classes. On one hand, we saw that these methods made the model more confident in recognizing the minority class (Defaulted loans). Unfortunately, they also made the model quite 'default-prone.' Due to this, the minority class precision dipped drastically while the model's ability to correctly identify Paid loans got compromised leading to a decrease in the overall accuracy.

Generally speaking, the findings show that models have high stability and are capable of pinpointing paid loans quite well. Due to the skewed distribution of the dataset, it is almost impossible for them to precisely detect the defaulted loans. The results reveal the importance of using suitable class balancing methods and thoroughly evaluating models in the scenario of financial datasets with imbalanced class distributions.

Table 6: Model Performance Comparison

Model	Mean				Std			
	Accuracy	F1-score	Precision	Recall	Accuracy	F1-score	Precision	Recall
Tuned LightGBM	0.7504	0.8574	0.7504	1.0000	0.0001	0.0001	0.0001	0.0000
Tuned Random Forest	0.7504	0.8574	0.7504	1.0000	0.0001	0.0001	0.0001	0.0000
XGBoost	0.7398	0.8495	0.7506	0.9784	0.0016	0.0011	0.0006	0.0025

LightGBM was dominant in delivering the best performance with both feature sets as it incurred the highest mean accuracy and F1-scores while having negligible standard deviations which indicates great stability and reliability. Thus, these findings point to LightGBM being the most success model for loan repayment prediction in this paper.

Experiment 2: Model Comparison (RQ2)

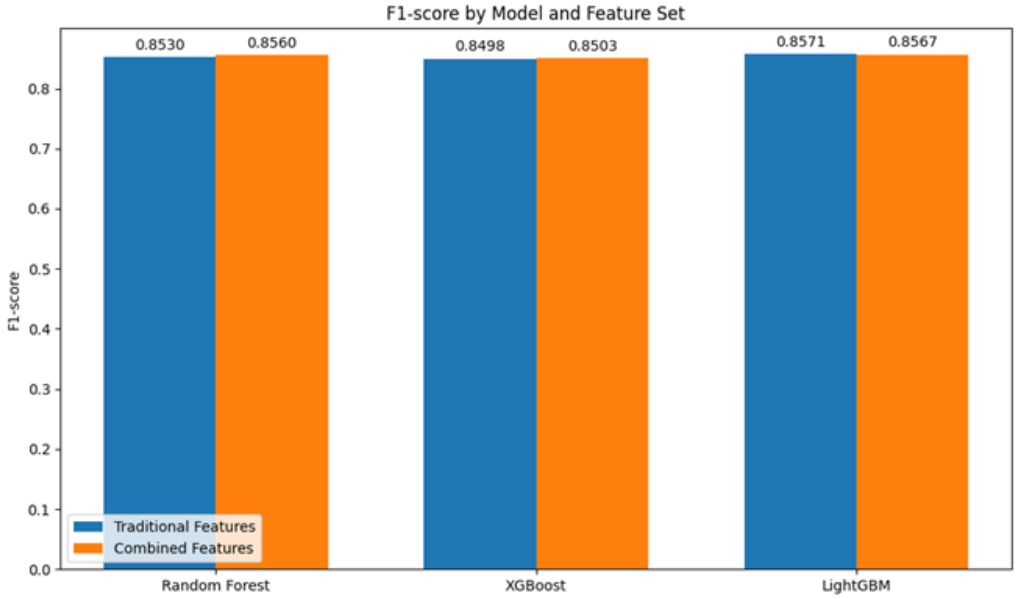

Figure 6.5: F1-score by Model and Feature Set

7 Conclusion and Recommendations

This paper aimed at figuring out the best way of using alternative data alongside traditional data and experimenting with a variety of state-of-the-art machine learning algorithms separately and in combination to forecast loan defaults. The findings add three main points to the ongoing debate, i.e., the subtle and complex function of alternative data, the surge of highly sophisticated algorithmic methods, and the indispensable role of thorough methodological experimental designs. On the feature impact aspect, the paper believes that alternative features' effects are not only quite limited, but also differ substantially from one case to another. Apart from supporting a few works of literature, which indicate that non-traditional data can be helpful to traditional credit history, the results here imply that the alternative features are less predictive as compared to other usual financial indicators in the present data set. Most importantly, their usefulness was not consistent but was strictly tied to certain models, which means that the benefit of alternative data cannot be assessed from a single point of view, but it has to be considered within the actual algorithmic context in which it is used. The paper provides very strong evidence that the gradient boosting methods, particularly LightGBM and XGBoost, out-strip all other competitors by a wide margin in this sort of prediction problem as per the testings. Due to their superior prediction even over Random Forest, every time, the authors argue that these models have the capability to learn very complex, non-linear data patterns that are usually present in financial data. This is consistent with their status as the most leading-edge tools for modeling credit risk. Furthermore, apart from being computationally efficient and less memory-consuming, LightGBM also has a novel boosting mechanism that leads to dramatic performance gains on classification tasks without a significant amount of overfitting. Methodological exploration isolated and guaranteed some fundamental ideas of machine learning. As is well known, performance measures can vary considerably at random due to an individual train-test partition of the data. The simple use of a single train-test split will most likely result in grossly inaccurate estimates of generalization error, whereas cross-validation, if properly performed, provides a solid basis for model selection and makes the performance estimation less variable. In

summary, the research did not detect that the bringing in of alternative data used had a major effect on the predicting capability. Conversely, it proposes a very convenient, readily implementable way for a thorough evaluation of the scenario, which could indeed be a case that a methodological framework is adopted. It concludes that the current success in loan prediction is definitely not due to a single data source alone, but rather results from the fruitful integration of in-depth feature analysis, extremely powerful and well-designed algorithms such as gradient boosting, and stringent methodological norms.

7.1 Future Research Directions

Eventually, the next research may think of involving more and detailed alternative data sources e.g. credible cash flow statements, rental payment records, telephone usage behaviors, etc. in identifying the features that could be even more effectively combined with traditional financial indicators for the purpose of prediction. Besides, this ought to be accompanied by a very thorough audit of model interpretability using e.g. SHAP, to ensure fairness, accountability, and transparency. Also, cost-sensitive analyses should be done to ascertain the real financial impacts of false positives and false negatives in lending decisions. Finally, the objective of forthcoming investigations should be to come up with a more sophisticated or hybrid set-up of models and to explore the different stages of data-model relationships over economic cycles in order to establish robust, ethical, and sustainable lending systems.

References

[1] Xgboost: A scalable tree boosting system. Online; accessed 01-May-2025. [Online]. Available: https://github.com/dmlc/xgboost

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

[3] Random forest: A complete guide for machine learning. Online; accessed 01-May-2025. [Online]. Available: https://builtin.com/data-science/random-forest-algorithm

[4] Xgboost: Extreme gradient boosting explained | ultralytics. Online; accessed 01-May-2025. [Online]. Available: https://www.ultralytics.com/glossary/xgboost

[5] D. Björkegren and D. Grissen. Behavior revealed in mobile phone usage predicts loan repayment. Online; accessed 01-May-2025.

[6] Alternative data transforming credit scoring for the unbanked. Online; accessed 01-May-2025. [Online]. Available: https://www.google.com/search?q=CGAP.+%282021%29.+Alternative+data+transforming+credit+scoring+for+the+unbanked.

[7] Improving access to credit through digital innovation in sub-saharan africa. Online; accessed 01-May-2025. [Online]. Available: https://www.google.com/search?q=World+Bank.+%282022%29.+Improving+access+to+credit+through+digital+innovation+in+Sub-Saharan+Africa.

[8] (2021) National financial inclusion strategy-ii (2021-2025) - national bank of ethiopia. Online; accessed 29-Apr-2025. [Online]. Available: https://nbe.gov.et/files/national-financial-inclusion-strategy-ii-2021-2025/

[9] Microfinance credit risk management tool guides credit scoring: What is a credit score? Online; accessed 2025.

[10] Unknown, "An intelligent credit card fraud detection model using data mining and ensemble learning," *Web article*, 2025, online; accessed 25-Dec-2025. [Online]. Available: https://ideas.repec.org/a/ajp/edwast/v9y2025i2p1391-1405id4651.html

[11] Using alternative data and artificial intelligence to expand financial inclusion: Evidence-based insights. Online; accessed 30-Apr-2025. [Online]. Available: https://www.povertyactionlab.org/blog/3-21-24/using-alternative-data-and-artificial-intelligence-expand-financial-inclusion-evidence

[12] The roles of alternative data and machine learning in fintech lending: Evidence from the lendingclub consumer platform. Online; accessed 30-Apr-2025. [Online]. Available: https://www.philadelphiafed.org/consumer-finance/the-roles-of-alternative-data-and-machine-learning-in-fintech-lending

[13] R. Hlongwane, K. K. K. M. Ramaboa, and W. Mongwe, "Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data," *PLoS One*, vol. 19, no. 5, May 2024.

[14] A. G. Sirishma, P. H. Kumar Reddy, Y. Tulasi, K. H. V. Reddy, P. R. Ys, and G. Ramasamy, "Machine learning for risk assessment: A comparative study of models predicting loan approval," *SSRN Electronic Journal*, Nov 2024.

[15] Predicting lending risk with machine learning models | online ethics. Online; accessed 30-Apr-2025. [Online]. Available: https://onlineethics.org/cases/george-mason-tech-ethics/predicting-lending-risk-machine-learning-models

[16] The role of machine learning and alternative data in expanding access to credit: Fintechs' regulatory advantage is to the detriment of consumers. Online; accessed 30-Apr-2025. [Online]. Available: https://bpi.com/the-role-of-machine-learning-and-alternative-data-in-expanding-access-to-credit-fintechs-regulatory-advantage-is-to-the-detriment-of-consumers/

[17] Alternative data in digital lending: Game-changing strategy for modern financial institutions. Online; accessed 30-Apr-2025. [Online]. Available: https://finezza.in/blog/alternative-data-in-digital-lending/

[18] Sciencedirect. Online; accessed 30-Apr-2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405918822000095

[19] E. Von Der Burg and S. Strömberg, "The power of credit scoring: Evaluating machine learning and traditional models in swedish retail banking," *Working paper*, 2025, working paper.

[20] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: A review of the literature," *Intelligent Systems in Accounting, Finance and Management*, vol. 18, no. 2-3, pp. 59–88, Apr 2011.

[21] N. Capon, "Credit scoring systems: A critical analysis," *J. Mark.*, vol. 46, no. 2, p. 82, Spring 1982.

[22] Alternative data is now mainstream; ai could be next special report from lowenstein sandler's investment management group. Online; accessed 29-Apr-2025. [Online]. Available: https://www.google.com/search?q=Alternative+Data+is+Now+%0D%0AMainstream%3B+AI+Could+Be+Next%0D%0ASpecial+Report+from+Lowenstein+Sandler%E2%80%99s+Investment+Management+Group

[23] The expanding horizon of alternative data in financial services - inrix. Online; accessed 29-Apr-2025. [Online]. Available: https://inrix.com/blog/the-expanding-horizon-of-alternative-data-in-financial-services/

[24] Y. L. Toh. Addressing traditional credit scores as a barrier to accessing affordable credit. Online; accessed 29-Apr-2025. [Online]. Available: www.KansasCityFed.org

[25] Z. Temelkov and V. G. Svrtinov, "Ai impact on traditional credit scoring models," *Working paper*, 2025, working paper.

[26] G. Chen and S. Rasmussen. Financial innovation for the unbanked: Balancing inclusion and risk. Online; accessed 29-Apr-2025.

[27] Alternative data for credit scoring. Online; accessed 29-Apr-2025. [Online]. Available: https://www.smefinanceforum.org/

[28] Ai in financial markets: Ethical and regulatory considerations. Online; accessed 29-Apr-2025. [Online]. Available: https://www.google.com/search?q=OECD.+%282021%29.+AI+in+financial+markets%3A+Ethical+and+regulatory+considerations.

[29] Y. Wen, "Analysis of credit default prediction based on logistic regression, random forest and knn model," *Advances in Economics, Management and Political Sciences*, vol. 85, p. 2024, 2024.

[30] W. Wang, X. Zuo, and D. Han, "Predict credit risk with xgboost," *Applied and Computational Engineering*, vol. 74, pp. 164–177, 2024, (Published online 11 July 2024).

[31] F. Zhao, "Evaluating machine learning models for personal credit risk: A comparison of logistic regression, random forest, and xgboost," *Applied and Computational Engineering*, vol. 158, pp. 232–241, 2025.

[32] D. Demirtürk, O. Mintemur, and A. Arslan, "Optimizing lightgbm and xgboost algorithms for estimating compressive strength in high-performance concrete," *Arabian Journal for Science and Engineering*, 2025.

[33] L. Yang, W. Zhou, C. Zhang, and F. Tang, "A comparative machine learning study identifies light gradient boosting machine (lightgbm) as the optimal model for unveiling the environmental drivers of yellowfin tuna distribution using shap analysis," *Biology*, vol. 14, no. 11, 2025.

[34] I. Budak, "Prediction of water quality's ph value using random forest and lightgbm algorithms," *MEMBA Journal of Water Sciences*, vol. 11, no. 1, pp. 42–49, 2025.

[35] K. Ileri, "Comparative analysis of catboost, lightgbm, xgboost, rf, and dt methods optimised with pso to estimate the number of k-barriers for intrusion detection in wireless sensor networks," *International Journal of Machine Learning and Cybernetics*, vol. 16, pp. 6937–6956, 2025.

[36] X. Cui, S. Yang, and J. Wang, "Performance evaluation and interpretation of hybrid models based on light gradient boosting machine to predict coal mining subsidence," *International Journal of Coal Science & Technology*, vol. 13, 2026.