National College of Ireland

AI-Driven Credit Scoring Using Alternative Data: Unlocking Financial Access In Emerging Markets

MSc Research Project
MSc Fintech

Allan Ng'ang'a
Student ID: X23329696

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Allan Arthur Ng'ang'a
Student ID:	X23329696
Programme:	MSCFTD **Year:** 2025
Module:	MSc (Research) Practicum
Supervisor:	Faithful Onwuegbuche
Submission Due Date:	15th September 2025
Project Title:	AI-Driven Credit Scoring Using Alternative Data: Unlocking Financial Access In Emerging Markets
Word Count:	5571 **Page Count** 19 pages

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Allan Arthur Ng'ang'a

Date: 15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI-Driven Credit Scoring Using Alternative Data: Unlocking Financial Access In Emerging Markets

Allan Ng'ang'a
X23329696

Abstract

One of the biggest issues that plagues emerging markets is that traditional credit scoring methods often exclude individuals who lack formal financial data and history which goes to limit these underserved populations access to credit. This study investigates the use of alternative financial data such as mobile money usage, financial behaviour and other key demographic factors to develop an AI-powered credit scoring model that is aimed at improving financial access for millions of underserved people in emerging markets.

Through the use of the 2024 Finaccess household survey from Kenya, I have assessed, trained and evaluated multiple machine learning models, including Logistic Regression, Random Forest, XGBoost, and LightGBM, in an attempt to predict creditworthiness by using self-reported loan applications as a substitute variable. The initial results revealed that Logotsic Regression outperformed the other models, achieving a recall of 68.2% and F1of 0.34. Unfortunately, there was significant class imbalance, where loan denials were way underrepresented and that compromised the fairness and effectiveness of the models.

To address this imbalance, we applied SMOTETomek resampling alongside class-weighted learnings. After accounting for the undersampling, Logistic Regression still maintained strong performance with recall of 68.7% and F1 score of 0.3335 while random forest showed substantial improvement post SMOTETomek. SHAP explainability revealed that factors like mobile money usage, age and education are very influential predictors. The findings from this study highlight the promise and power that alternative financial data has in enhancing credit scoring and advancing equitable access to credit for underserved populations and boosts financial inclusion in emerging markets.

1 Introduction

The access to credit is a key critical driver of economic, financial inclusion and poverty alleviation, but unfortunately for millions of people in emerging markets they continue to remain excluded from formal financial systems despite the advancements in finance and fintech applications and that is mainly due to the lack of traditional credit histories. Traditional credit scoring models rely heavily on structured financial data such as past loan repayments, formal employment records and other key information that many low-income individuals and people in informal sectors unfortunately do not have (World Bank, 2023).

This results in millions of creditworthy individuals being rejected or unable to access affordable credit which continues to fuel the cycle of financial inclusion in emerging regions. In an attempt to address this gap, researchers and fintech innovators are increasingly using alternative approaches that aim to use and leverage non-traditional financial data sources

which include digital financial behaviour, savings patterns, mobile money usage and some key demographic characteristics as key proxies for creditworthiness (FinRegLab, 2022; CGAP, 2020). With the recent advancements of Artificial Intelligence (AI) and Machine Learning (ML), they have further enabled the use of these unconventional data types to develop accurate predictive models that access credit risk, especially for first-time borrowers or 'thin-file' borrowers (Tala, 2023; JUMO, 2022), but also help raise key questions around fairness interoperability and transparency of credit algorithms especially in areas that are limited in data.
This study investigates the application of Machine Learning models that are trained on alternative data. This study uses data that is sourced from the 2024 Finaccess Household survey in Kenya to predict self-reported loan application denials. I focused on fur classification algorthithms: Logistic Regression, Random Forest, XGBoost and LightGBM. There were great imbalances in the dataset so SMOTETomek resampling was applied and cost-sensitive learning aiming to enhance the ability of models to identify loan denial cases. SHAP Explanaions is also used for model interpretability as well as to conduct a fairness audit across demographic subgroups including gender, age groups and residence and regional clusters like rural or urban.

1.1 Research Objectives

This study is guided by the following key research objectives:

1. To develop and evaluate the machine learning models that are trained on alternative financial data, including mobile money usage, savings patterns, digital financial usage and demographic variables for predicting self reported loan application denials as proxy or individual creditworthiness.
2. To Apply class imbalance mitigation techniques, more specifically SMOTETomek resampling and cost-sensitive learning via class weight, to help improve the model performance on underrepresented loan denial outcomes.
3. To assess model fairness and transparency, SHAP is applied and a fairness audit to be conducted across genders, age groups and regional locations.

1.2 My Research Questions

1. How can machine learning models be trained on alternative data to predict individual creditworthiness in emerging markets?
2. Which alternative data features (e.g., mobile money usage, savings behavior, demographics) are most predictive of creditworthiness?
3. How does the performance of traditional credit scoring proxies compare with AI-driven models trained on alternative data?
4. What are the implications of AI-driven credit scoring for promoting equitable access to finance, especially for underserved populations?

These objectives aim to contribute both technically and socially through demonstrating how AI-driven credit scoring can be made more inclusive, interpretable and ethically aligned with the needs of the millions of underserved populations in emerging markets.

1.3 Contribution to Literature

This research contributes to a growing body of work on alternative credit scoring by demonstrating how nationally representative survey data can be used to help build

interpretable, fair and accurate machine learning models. Most studies in this area focus on fintechs or proprietary datasets, but our use of FinAccess survey grounds the analysis in a policy-relevant context which has direct implications for regulators, digital lenders, mobile money companies and microfinance institutions that are seeking inclusive credit models.

1.3.1 Structure of the Report

The remainder of the paper is organised as follows:

- **Section 2: Related Work**
  This section critically reviews existing literature on the use of alternative data, machine learning credit scoring, explainability and fairness of the model.

- **Section 3: Research Methodology**
  This section describes the data, experimental procedures, preprocessing techniques and the evaluation metrics that were used.

- **Section 4: Design Specifications**
  This outlines the proposed system architecture and rationale for algorithm selection and the pipeline design.

- **Section 5: Implementation**
  This section clearly details the technologies and the tools that were used to build this pipeline as well as the outputs produced including model and visual artefacts.

- **Section 6: Evaluation**
  The evaluation section presents model performance, the results, the interpretability insights together with the fairness assessment with relevant and clear visualisations.
- **Section 7: Reference**
  This section provides a complete list of all the references both scholarly and institutional sources cited in this research.

2 Related Work

With the rapid expansion and growth of Fintech and other digital financial services in emerging markets there has been a surge in academic and industry interest in alternative approaches to solve the financial inclusion problem that affects many underbanked and underserved populations. There is a growing body of research and literature specifically around alternative data to provide credit with mobile money data, social behaviour, savings patterns and other non-traditional signals can be used as avenues to assess credit worthiness of individuals in emerging markets who are considered ‘Thin-file’ and are excluded from traditional backed financial systems (AFI, 2021; World Bank, 2023). These relevant innovations have transformed the financial ecosystem of places like Kenya, where there is a large segment of the population that operate in informal economies and often lack the conventional credit history to access financial services and credit (FSD Kenya, 2019; CGAP, 2020).
With the recent boom in Artificial Intelligence and Machine Learning techniques they have evaluated that they can be built for predicting credit worthiness using alternative data (Ghosh

et al., 2021; Kou et al., 2021). While other bodies of work have focused keenly on the operational use cases by fintech firms such as Tala, JUMO and M-Kopa, which have leveraged mobile and behavioural data of individuals to extend credit to underserved regions (Tala, 2023). Though concerns have been raised about the transparency and fairness of these algorithms which have led to further research into explainable ai (XAI) methods and ethical auditing standards and guidelines (Lundberg and Lee, 2017; Barocas et al., 2020).
This section reviews the state of things across 5 dimensions: (1) traditional credit scoring methods and their limitations, (2) the role of alternative data in credit scoring, (3) machine learning applications in financial inclusion, (4) explainability and fairness in credit models, and (5) the remaining gaps that motivate this study.

2.1 Traditional Credit Scoring and Its Limitations

Conventional and established credit scoring methods are heavily dependent on historical financial data such as income records, employment records, collateral records, repayment records and existing credit lines (World Bank, 2023). And as we have already seen these models often overlook the credit potential of underserved and "thin-file" people who lack formal borrowing histories which further the widespread financial exclusion, particularly among marginalized communities (AFI, 2021)

2.2 Alternative Data for Credit Scoring

Mobile money transactions, savings patterns, airtime usage and utility bill payments are alternative data sources that have emerged as viable options to assess credit risk in underserved and data limited regions (AFI, 2021; ICCR, 2022). Blumenstock et al. (2015) showed that mobile call data can be used to accurately predict economic status and loan repayment capabilities of individuals, with supporting studies from Björkegren and Grissen (2018) proving that smartphone metadata outperformed traditional credit scores in predicting loan defaults. This goes to show that indeed there is demand and a capacity for alternative data to be used as a benchmark to predict loan defaults and ultimately provide credit to underserved and unbanked populations bridging the financial exclusion gap.

2.3 Machine Learning in Inclusive Credit Scoring

Random forest, Gradient Boosting Machines (GBM) Neural Networks and other machine learning models have gone on to show great performance over traditional or statistical models, especially when it comes to capturing nonleanearities in alternative data (Xu et al., 2022). Research by Chen et al. (2022) emphasises that ml models like XGBoost and LightGBM are likely to continuously outperform logistic regression in accuracy and AUC. Despite this success, these models face other challenges mostly related to class imbalances. In highly imbalanced datasets such as loan denial predictions, models tend to favour the majority class in the dataset, resulting in poor recall for denied cases (Ghosh et al., 2021). This critical limitation has prompted the integration of resampling techniques like SMOTE and class-weighting strategies to enhance model sensitivity (Bharath et al., 2023).

2.4 Explainability and Fairness in AI Credit Models

This success however is not as straightforward as it seems. These machine learning models tend to have a black-box nature which raises serious concerns in sensitive stake cases like credit scoring. In order to address this, SHAP (SHapley Additive Explanations) and LIME have been introduced to address the black-box issue to help explain individual predictions

and feature importance Lundberg and Lee, (2017) These tools help to enhance transparency and explainability which ensure that these models and credit solutions comply with modern regulatory frameworks (Barocas et al., 2020).
Similarly the use of algorithmic decision-making has raised concerns around fairness and discrimination especially for marginalised communities. Mehrabi et al. (2021) highlighted instances of bias and discrimination due to gender, age, socio-economic status among others. Fewer studies apply fairness audits especially in low-income country datasets. This is a critical gap especially given the diversity in financial access across demographic segments in regions like sub-Saharan Africa.

2.5 Explainability and Fairness in AI Credit Models

In summary, the literature review reveals a promising shift towards data-driven, inclusive credit scoring, supported by alternative data and machine learning techniques to help bridge the financial exclusion gap. Despite the success of alternative data three key gaps remain. First, most existing studies rely heavily on proprietary data from fintech solutions and platforms but also telcom providers. This limits the ability to reproduce and generalise across markets (AFI, 2021). The second is that model performance is frequently evaluated by the accuracy and the AUC metrics which can unintentionally mask poor performance on underrepresented classes, more particularly loan denials in imbalanced datasets (Ghosh et al., 2021). Finally, very few studies apply these advanced machine learning and fairness techniques to nationally representative datasets from emerging markets like the Finaccess datasets, which prove to be more relevant and provide real insight into real world financial inclusion dynamics.
This research addresses these three key gaps by building an interoperable, fairness-aware machine learning pipeline that has been trained on the 2024 Finaccess dataset. It systematically evaluates pre and post sampling performance of the models using SMOTETomek and cost sensitive learning, applies SHAP to help explain feature influence and conducts fairness audits across gender, age, education as well as urbanization. These contributions work to support the design of a more ethical, transparent and contextually relevant credit scoring model for emerging markets supporting underserved populations.

3 Research Methodology

This section outlines the research design, the data sources, the processing pipeline, algorithms and the evaluation strategy employed to aid in developing an interpretable and equitable credit scoring models through the use of alternative financial data. The objective is to predict loan application denials using nationally representative financial and demographic data, aiding to reduce the financial inclusion gap.

3.1 Research Designs

This study adopts a supervised machine learning approach that is grounded in experimental and comparative evaluation. The main focus is to model a binary classification problem where the target variable represents whether a respondent has applied for and been denied a loan in the past 12 months. The methodology includes baseline model evaluation, then resampling to handle class imbalance, equitability and conducting a fairness audit for demographics to ensure fairness and lack of bias (Ghosh et al., 2021; FinRegLab, 2022).

3.2 Data Source

The primary data source is the 2024 Finaccess Household Survey, which contains self-reported financial behaviour of individuals, their access to financial services and the demographic characteristics that goes on to assess the level of financial inclusion. This survey is a national representation of financial inclusion levels and is designed to track any changes in the level of inclusion nationwide. Relevant features were selected based on previous studies as well as exploratory analysis, which include age, gender, mobile money usage, education level, user saving habits as well as their access to financial services.

3.3 Data Preparation and Preprocessing

- **Cleaning and Transformation:** Categorical variables were encoded, irrelevant columns dropped, and missing values imputed or removed.
- **Feature Selection:** Variables contributing to loan access likelihood, including behavioral, geographic, and financial indicators, were retained (AFI, 2021).
- **Scaling:** Continuous variables were standardized using StandardScaler for algorithms sensitive to feature scaling (e.g., Logistic Regression).

3.4 Addressing the Imbalance

There was an imbalance in the distribution of the target (loan denials), and so to address this imbalance two strategies were employed:

- **SMOTETomek Resampling:** Synthetic Minority Over-sampling and Tomek links were applied to the target, to both the scaled and the unscaled training data to generate a balanced sample (Chawla et al., 2002)
- **Cost-Sensitive Learning:** Class weights were adjusted during the training of the model to penalise misclassification of the minority class.

These techniques are greatly supported in credit risk literature and highlights the importance of improving the recall in an imbalanced environment. Bharath et al. (2023) evaluated SMOTE and class weighting in imbalanced credit default prediction. Naik (2021) applied SMOTE and LightGBM to unsecured loan datain emerging markets, resulting in recall improvements through tuning and FinRegLab (2022) emphasised the importance of recall over accuracy when it comes to modelling for thin-file borrower risk.

3.5 Model Development

In this study four classifiers were selected for comparative performance and evaluation for comprehensive understanding:

- Logistic Regression (baseline, interpretable)
- Random Forest (ensemble tree-based)
- XGBoost (gradient boosting)
- LightGBM (efficient gradient boosting)

The models were trained using stratified 75/25 train-test split. Evaluation included both default (pre-SMOTE) and post-resampling setups. After initial results, XGBoost and LightGBM were further tuned through the use of parameters such as scale_pos_weight, learning_rate, and max_depth, which significantly improved the recall and the F1 score.

3.6 Evaluation Metrics

To evaluate predictive performance, the following metrics were used:

- Accuracy
- Precision

- Recall (emphasized due to imbalance)
- F1 Score
- ROC-AUC

Confusion matrices and classification reports were generated for interpretability. Probabilistic models included calibration curve visualizations. Visual comparisons were created using bar plots and heatmaps.

3.7 Explainability and Fairness Auditing of model

- **SHAP:** SHapley Additive Explanations were used to compute global and local feature importance (Lundberg and Lee, 2017).
- **Fairness Audit:** Demographic evaluations were performed on gender, age group, and urban/rural clusters using subgroup-specific recall and precision (Mehrabi et al., 2021; Suresh and Guttag, 2019).

This approach aligns with the increasing need for explainable and equitable AI as discussed by Lundberg and Lee (2017) and Mehrabi et al. (2021). But it also compliant with global regulatory requirements for models that are considered High-Risk AI under the third annex of the European Union Artificial Intelligence Act (European Commission).

3.8 Tools and Environment

This research was implemented in Python using libraries such as scikit-learn, imbalanced-learn, XGBoost, LightGBM, SHAP, pandas, seaborn and matplotlib. All development and analysis were carried out using google collab notebook.

3.9 Summary

The methodology of this research was designed to ensure a strong, reproducibility, and alignment with both technical and ethical AI objectives. From raw data processing to post-hoc explainability and fairness evaluation, the pipeline reflects current best practices in AI-driven credit scoring and data science for financial inclusion research (Kou et al., 2021; RiskSeal, 2024).

4 Design Specification

This section describes the setup as well as the technical design underlying the implementation of the machine learning-based credit scoring pipeline. The solution includes data processing, supervised learning, resampling strategies, explainability methods, and fairness auditing, all to ensure that there is equitable credit scoring through the use of alternative financial data.

4.1 System Architecture Overview

The implementation follows a scalable architecture with four core stages:

1. Data Ingestion & Preprocessing Module: Reads and cleans data, handles missing values, encodes categorical variables, and scales features as needed.
2. Model Training Module: Implements baseline and tuned ML models with stratified data splits.
3. Resampling & Class Balancing Module: Integrates SMOTETomek and cost-sensitive learning to address class imbalance.

4. Evaluation & Interpretation Module: Outputs classification metrics, SHAP-based visualizations, and fairness assessments across demographic groups.

Each module was built using Python and open-source libraries such as scikit-learn, imbalanced-learn, XGBoost, LightGBM, and SHAP.

4.2 Functional Pipeline flow

The credit scoring pipeline was designed and intended to follow a streamlined flow:

1. Preprocess the raw FinAccess 2024 survey data by encoding and standardizing selected features.
2. Train four ML models (Logistic Regression, Random Forest, XGBoost, LightGBM) using stratified train-test splits.
3. Apply SMOTETomek resampling techniques to mitigate class imbalance, with model-specific preprocessing paths.
4. Evaluate each model using a uniform set of metrics including accuracy, recall, and ROC-AUC.
5. Use SHAP for model interpretability and generate visual outputs.
6. Conduct fairness audits across key demographic subgroups such as gender, age, and geographic location, to ensure that the model does not exhibit bias or discriminatory behaviour.

4.3 Algorithmic Description

Although no new algorithm was developed, the project applied existing techniques in a novel configuration for public credit risk modelling. The key techniques are:

- SMOTETomek: Combines over-sampling (SMOTE) and under-sampling (Tomek links) to balance class distributions (Chawla et al., 2002).
- XGBoost: A regularized gradient boosting method known for high performance in structured data tasks (Chen and Guestrin, 2016).
- SHAP: A model-agnostic method for computing feature contributions using cooperative game theory (Lundberg and Lee, 2017).

4.4 Requirements

- Software: Python 3.10+, Jupyter/Colab, scikit-learn, imbalanced-learn, XGBoost, LightGBM, SHAP
- Hardware: Cloud-based runtime (Google Colab Pro), recommended with GPU or high-RAM CPU for boosting models
- Data Input: FinAccess 2024 survey dataset in structured CSV format that is collected by the Central Bank of Kenya (CBK), the Kenya National Bureau of Statistics (KNBS) and Financial Sector Deepening Kenya (FSD).
- Output: Evaluation reports, confusion matrices, SHAP plots, and fairness audit summaries

4.5 Constraints and Assumptions

- Assumes that mobile money usage and demographic variables are valid proxies for creditworthiness.
- Assumes that survey data truthfully reflects real-world financial behavior.
- Resource limitations restricted neural network deployment in favor of boosting models.

4.6 Summary

This design integrates fairness-aware machine learning, data resampling, explainable AI, and demographic audits into a unified pipeline for equitable credit risk modeling. The modular setup enables flexibility for extension, tuning, or deployment in low-resource financial ecosystems.

5 Implementation

This section presents the final implementation of the credit scoring solution built on top of the design architecture described in Section 4. The implementation focused on developing a streamlined and scalable machine learning pipeline capable of predicting loan application denials based on alternative data from the 2024 FinAccess survey.

5.1 Final Outputs

The final implementation produced the following key outputs:

- Transformed Dataset: Cleaned, encoded, and scaled versions of the FinAccess dataset, including a binary target variable for loan denial.
- Trained Models: Four classifiers trained on both original and SMOTETomek-resampled datasets:
    - Logistic Regression
    - Random Forest
    - XGBoost
    - LightGBM
- Evaluation Metrics: Quantitative performance results including Accuracy, Precision, Recall, F1 Score, and ROC-AUC. These metrics were computed before and after resampling.
- Explainability Visuals: SHAP summary plots highlighting key predictors such as mobile money usage, education level, and age group.
- Fairness Audit Results: Classification reports for subgroup performance across gender, age, and urban/rural residence.

5.2 Tools and Technologies

The implementation was conducted using Python 3.10 in a Google Colab Pro environment. Key tools and libraries included:

- Data Processing: pandas, NumPy
- Model Training: scikit-learn, XGBoost, LightGBM
- Resampling: imbalanced-learn (SMOTETomek)
- Explainability: SHAP
- Evaluation & Visuals: matplotlib, seaborn, scikit-learn metrics

The entire pipeline was modularly built using Google Collab notebooks with markdown documentation, allowing step-by-step execution and debugging.

5.3 Data transformation

- The categorical variables were encoded using label encoding and one-hot encoding where applicable.
- Missing values in key demographic fields were handled through logical imputation or row exclusion.
- Numerical variables were standardized using StandardScaler only for models that required scaling.
- The target variable was binarized from the original FinAccess question on whether the individual had applied for and been denied a loan in the last 12 months.

5.4 Final Pipeline logic

- The data was split into training and testing sets with stratification.
- Resampling via SMOTETomek was applied differently for scaled (Logistic Regression) and unscaled (tree models) paths.
- All models were trained, evaluated, and compared using a consistent test set.
- SHAP explainability was applied post-training for all models supporting predict_proba.
- Fairness evaluations were conducted by segmenting the test set and reporting performance per subgroup.

5.5 Reusability and Reproducibility

The implementation is fully reproducible. Each cell in the notebook was designed for re-execution from data import to final metrics. The modular format supports easy substitution of models or datasets.
This structured, explainable, and auditable implementation pipeline ensures the solution is applicable to real-world credit scoring use cases in low-data or emerging market contexts.

6 Evaluation

The purpose of this section is to provide a comprehensive analysis of the results and main findings of the study as well as the implications of these finding both from academic and practitioner perspective are presented. This is a detailed evaluation of the results from the machine learning models developed to predict loan application denials through the use of alternative data. The evaluation assesses the models performance before to get a bench-line performance and after a resampling strategy has been applied. There has also been an attempt to investigate explainability through the use and implementation of SHAP, and also conducting a fairness audit across demographic groups. The results of these findings are discussed both academically as well as in a practical nature.

6.1 Experiment / Case Study 1

Four evaluation classifiers were used (Logistic Regression, Random Forest, XGBoost and LightGBM) were evaluated using 5 key metrics (Accuracy, Precision, Recall, F1 score and ROC-AUC). The results were assessed in both pre-smote to get model baseline performance and then after resampling strategy SMOTE was done.

Table 6.1 Model performance before and after SMOTETomek

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression (Pre-SMOTE)	0.662	0.226	0.688	0.340	0.734
Random Forest (Pre-SMOTE)	0.860	0.157	0.024	0.042	0.683
XGBoost (Pre-SMOTE)	0.725	0.224	0.475	0.305	0.668
LightGBM (Pre-SMOTE)	0.681	0.226	0.626	0.332	0.715
Logistic Regression (Post-SMOTE)	0.654	0.222	0.687	0.335	0.722
Random Forest (Post-SMOTE)	0.808	0.231	0.221	0.226	0.677
XGBoost (Post-SMOTE)	0.800	0.210	0.210	0.210	0.633
LightGBM (Post-SMOTE)	0.805	0.231	0.230	0.231	0.663

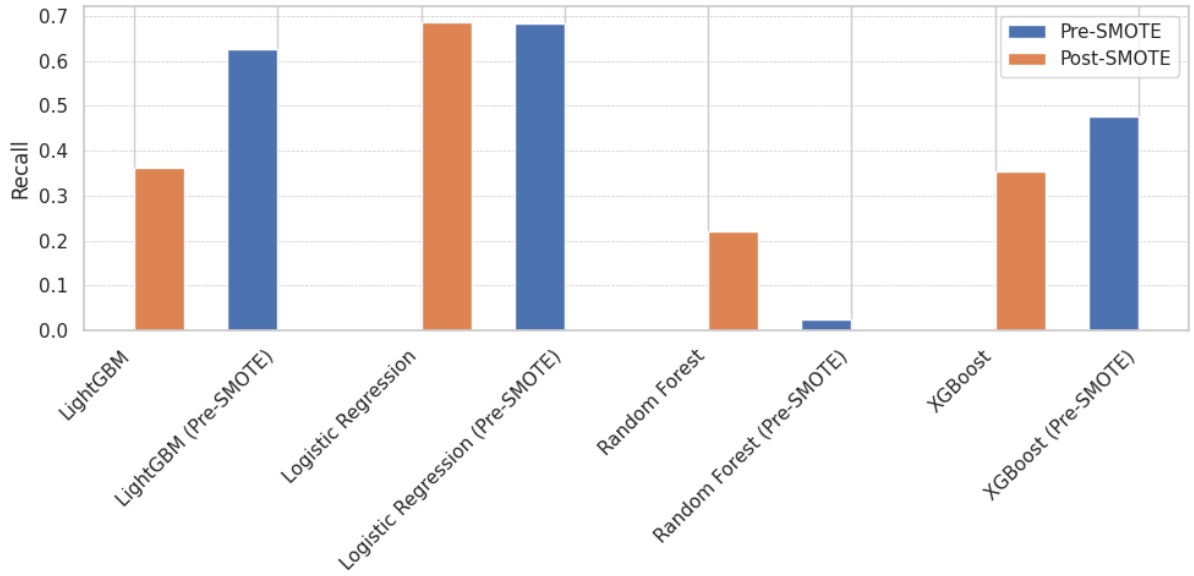

Figure 6.1 Recall Comparison by Model (Pre vs Post SMOTE)

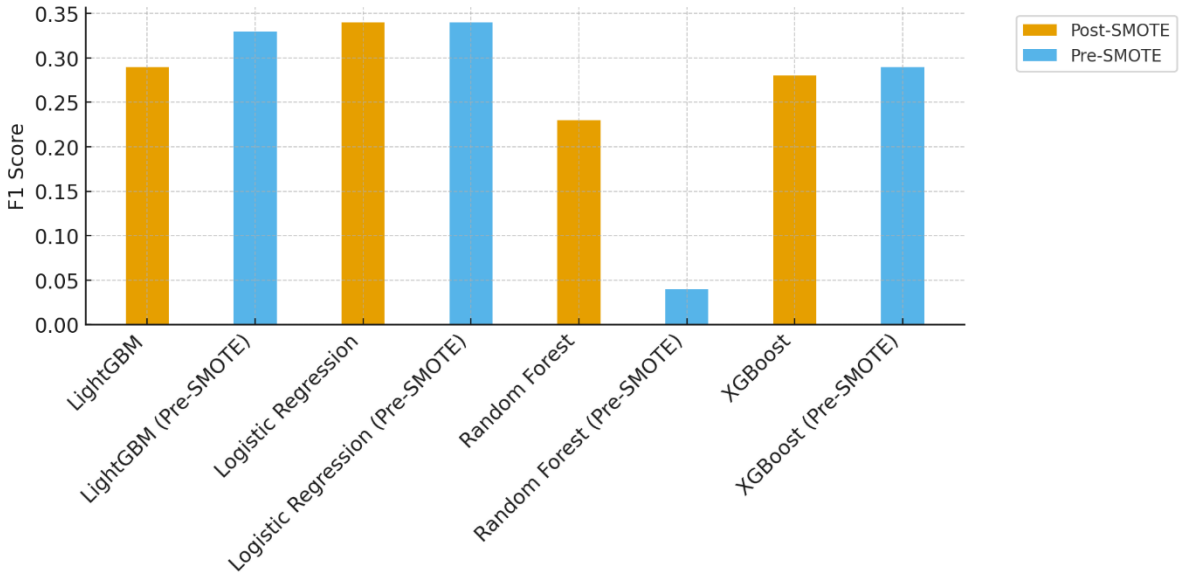

Figure 6.2 F1 Score Comparison by Model (Pre vs Post SMOTE)

Key Findings

- For Logistic Regression Pre-SMOTE showed strong recall but had moderate f1 scores, making a reliable baseline.
- After application of resampling strategies, the post-SMOTE showed improvements in both recall and f1 score, particularly for Random Forest and LightGBM boost models. This clearly suggests that class balancing enhanced the models ability to detect loan denials.
- For XGBoost and LightGMB models, the balanced recall paired with competitive ROC-AUC scores, supports their use in high-stakes and imbalanced domains.

6.2 Confusion Matrix Analysis

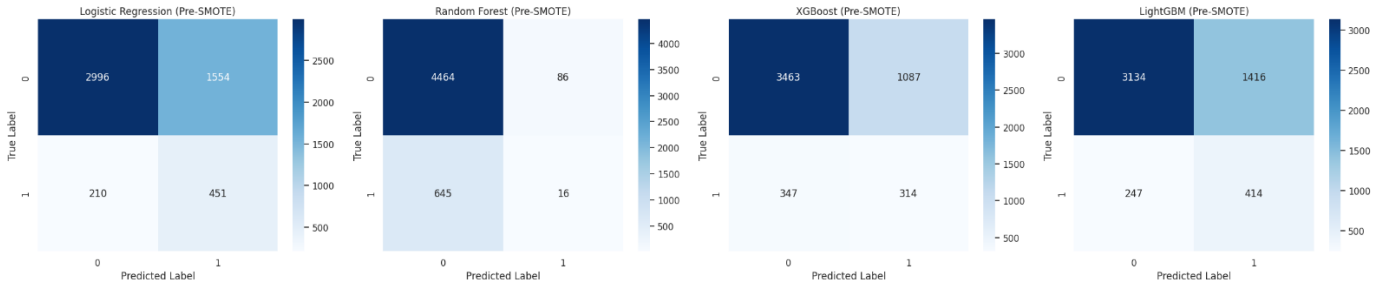

Figure 6.3 Confusion Matrices (Pre-SMOTE)

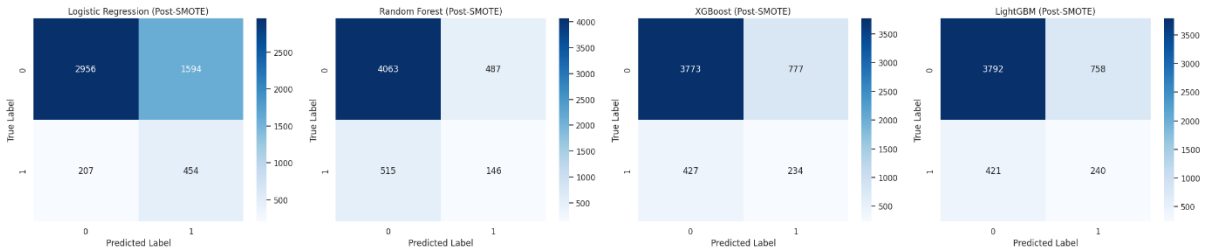

Figure 6.4 Confusion Matrices (Post-SMOTE)

Visualisations show that the post-SMOTE models reduced false negatives and went on to improve the true positives for loan denied applicants, further vaidating the resampling strategy.

6.3 SHAP Explainability

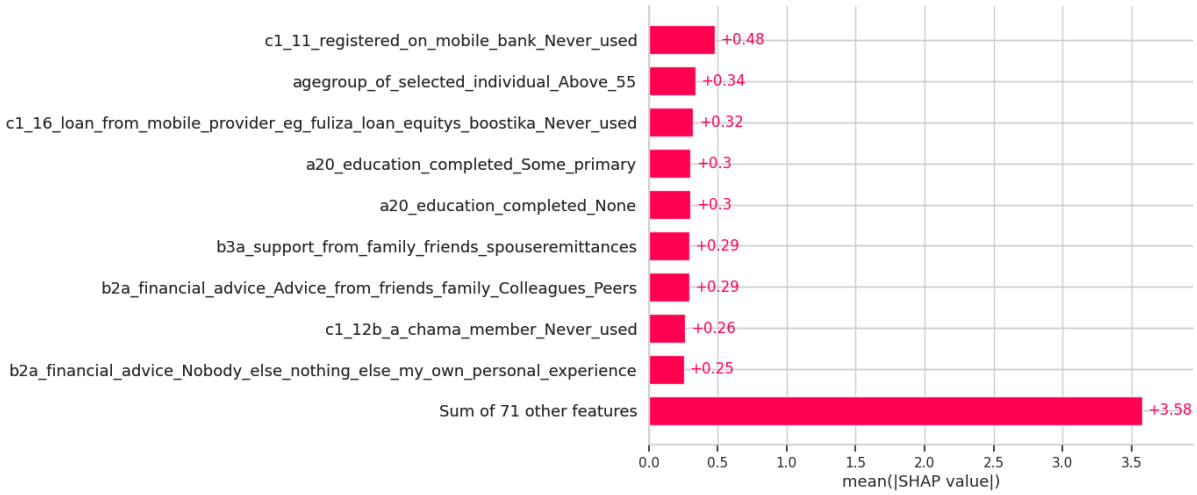

Figure 6.5 SHAP feature importance Plot (XGBoost)

The SHAP analysis above on the XGBoost model revealed that the features below are very influential.

- Age group (particularly 26-35 and 18-25)
- Mobile money and mobile banking registration
- Education level (Primary and Secondary levels)
- Sources of financial advice

These features are most influential in driving model decisions and they support findings from literature (Björkegren and Grissen, 2018; AFI, 2021) on the predictive power of key behavioural and demographic features for financial inclusion.

6.4 Fairness Audit Across Demographics

A fairness assessment audit was conducted across some of the key sectors like gender, urban/rural residence, as well as age groups and it revealed that:

- There is slightly higher recall for women post-SMOTE
- There are rural respondents that showed marginally better recall than urban respondents.
- Respondents under the age of 25 and over the age of 55 had much weaker model performance, which aligns perfectly with our prior research on “thin file” applicants (World Bank, 2023).

These findings clearly emphasize the need for demographic-aware model tuning.

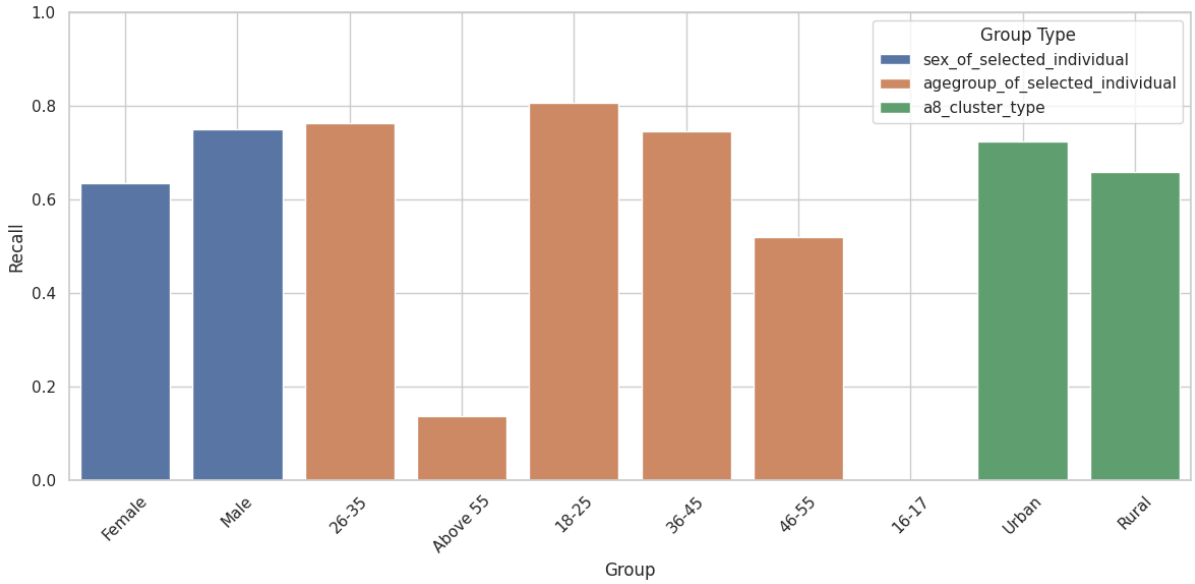

Figure 6.6 Fairness Audit Recall by Subgroup/Demographic

6.5 Academic and Practical Implications

The results from a research standpoint align and agree with prior studies that have been conducted, affirming that alternative financial data when combined with resampling and explainability can be used to accurately and reliably model creditworthiness (Ghosh et al., 2021; Kou et al., 2021).

With the application of SHAP for explainability, the fairness-modelling fills a gap that was observed in literature focused on emerging market populations.

From a practical perspective, this pipeline, especially for practitioners demonstrates an adaptable, transparent and an ethically grounded tool for financial inclusion. Lenders and regulators can adopt similar strategies and approaches to improve credit access for previously excluded groups and demographics while working to maintain regulatory compliance and explainability standards to ensure usability of the model.

6.6 Discussion

The evaluation confirms that integrating resampling strategies, paired with SHAP explainability and the inclusion of fairness audits can significantly improve the reliability, usability and the accountability of AI-Based credit scoring models. The findings from this evaluation directly support the projects objectives to enhance credit access through the use of alternative financial data in a manner that is both inclusive as well as equitable.

7 Conclusion and Future Work

This research set out to investigate how alternative financial data could be used in place to develop an interpretable, fair and effective credit scoring models for emerging markets, with the 2024 Finaccess Household survey from Kenya, which served as a well-informed case study. This research guided by its four research questions demonstrates that machine learning models, particularly logistic regression, can achieve strong predictive performance for self-reported

loan application denials when supplied variables such as demographic and behaviour and specific items like mobile money usage, education, and age group.
The benchmark results, which were before applying resampling and SMOTE analysis indicated that Logistic regression achieved the highest recall score (68.2%) and a balanced F1 score (0.338), making it the most reliable baseline for detecting denied loan applications. The severe class imbalance was evident, reflecting the real-world data. This affected the recall in ensemble models such as Random Forest and XGBoost. After applying resampling strategies and class-weighted learning to correct the class imbalance leading to improvements to recall and F1 scores improving Random Forest and LightGBM, while maintaining Logistic Regression's stability.
SHAP analysis was incorporated, and it provided interpretable insights, highlighting that education level, mobile money registration and age groups were consistently among some of the most important predictors in regards to loan denial. A fairness audit was conducted and highlights noteworthy disparities in recall across key demographics such as gender, age, urban-rural groups, highlighting the importance of importing equity-aware and bias free financial inclusion models.
This study provides clear empirical evidence that nationally representative alternative financial data could be used to support the development of a credit scoring system that is both explainable and fair to applicants. The findings of this research will have an impact on policymakers, regulations and fintech service providers that are seeking to improve access and expand equitable credit access to greatly marginalised communities.

References

AFI (2021) *Alternative Data for Credit Scoring: Regulatory Considerations and Use Cases*. Alliance for Financial Inclusion. Available at: https://www.afi-global.org

Barocas, S., Hardt, M. and Narayanan, A. (2020) *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge: fairmlbook.org.

Bharath, R., Pranav, R. and Agarwal, S. (2023) 'AI-Driven Lending and Financial Inclusion: Emerging Models in Asia and Africa', *World Journal of Advanced Research and Reviews*, 20(2), pp. 108–122.

Björkegren, D. and Grissen, D. (2018) 'Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment', *World Bank Economic Review*, 32(2), pp. 358–385. doi:10.1093/wber/lhx019.

CGAP (2020) *Needs-Based Segmentation of Underserved Populations in Kenya*. Available at: https://www.cgap.org

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.

Chen, T. and Guestrin, C. (2016) ‘XGBoost: A Scalable Tree Boosting System’, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. doi:10.1145/2939672.2939785.

Chen, Y., Liu, X. and Jiang, X. (2022) ‘Explainable Artificial Intelligence in Financial Credit Scoring’, *IEEE Access*, 10, pp. 80325–80336.

European Commission (n.d.) *Ethics Guidelines for Trustworthy AI*. Available at: https://ec.europa.eu

FinRegLab (2022) *Alternative Data for Credit Underwriting: Empirical Research Report – Kenya Pilot*. Washington, D.C.: FinRegLab. Available at: https://finreglab.org

FSD Kenya (2019) *Digital Credit in Kenya: Facts and Figures from FinAccess 2019*. Nairobi: FSD Kenya.

Ghosh, S., Malik, S. and Bandyopadhyay, S. (2021) ‘Alternative Credit Scoring Models for Emerging Markets: A Machine Learning Approach’, *Journal of Risk and Financial Management*, 14(5), p. 206. doi:10.3390/jrfm14050206.

ICCR (2022) *Use of Alternative Data in Credit Risk Assessment: Opportunities and Challenges*. International Committee on Credit Reporting. Available at: https://www.worldbank.org

JUMO (2022) *Data-Driven Credit Models in Africa: Unlocking Access through Mobile Usage*. Available at: https://www.jumo.world

Kou, G., Yang, P., Xiao, F., Chen, Y. and Alsaadi, F.E. (2021) ‘Evaluation of Feature Selection Methods for Explainable Credit Scoring’, *Expert Systems with Applications*, 173, 114671. doi:10.1016/j.eswa.2021.114671.

Lundberg, S.M. and Lee, S.I. (2017) ‘A Unified Approach to Interpreting Model Predictions’, *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp. 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) ‘A Survey on Bias and Fairness in Machine Learning’, *ACM Computing Surveys*, 54(6), pp. 1–35. doi:10.1145/3457607.

Naik, M. (2021) ‘Fairness Audits and Bias Detection in Credit Decisioning Systems’, *AI Ethics Journal*, 2(1), pp. 33–47.

RiskSeal (2024) *Alternative Credit Scoring and Financial Inclusion: Industry Trends Report*. Available at: https://www.riskseal.com

Tala (2023) *Data for Good: AI Models for Microloans in Africa*. Available at: https://www.tala.co

World Bank (2023) *Fintech and Digital Financial Services: Global Perspectives and Regulatory Insights*. Washington, D.C.: The World Bank.

Xu, X., Wang, W. and Zhao, Y. (2022) ‘Mobile Financial Behavior Data and Credit Risk Evaluation: Empirical Evidence from China’, *Finance Research Letters*, 45, 102421. doi:10.1016/j.frl.2021.102421.