IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

XG Boost-driven feature selection for Paygo loan default optimized using hybrid meta-heuristic algorithms

Machariah Denis [1, *], Cheruiyot Dennis [2] and Mundia S [2]

[1] *School of Business, Economics and Management - Dedan Kimathi University of Technology (Kenya).*
[2] *School of Actuarial Science - Dedan Kimathi University of Technology (Kenya).*

International Journal of Science and Research Archive, 2025, 14(01), 034-042

Publication history: Received on 18 October 2024; revised on 28 December 2024; accepted on 31 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2025.14.1.2317

Abstract

Customer default is a persistent challenge impacting the loan repayment sector, particularly in the Pay-As-You-Go within Africa's renewable energy sector. In off-grid communities, renewable energy companies offer Solar Home Systems, where payments are made incrementally over time, using mobile money daily. Identifying potential defaulters early is essential for these companies' sustainability and profitability. Therefore, there is a pressing need for advanced prediction techniques to address these challenges. The primary goal of this research was to develop a hybrid meta-heuristic model that offers higher predictive accuracy in forecasting loan defaulters compared to traditional classifiers. The use of the Xgboost algorithm for feature selection where parameters of Xgboost algorithm are tuned to achieve optimal parameter points rather than default parameter points, while meta-heuristic optimization used was random forest optimized using Particle Swarm Optimization Algorithm. This meta-heuristic approach aims to achieve higher predictive accuracy by leveraging the strengths of individual classifiers.

The research results show that tuning the Xgboost algorithm parameter points to their optimal points in feature selection achieved a significant improvement in feature selection, with prediction accuracy reaching 81.289% to 93.456%.

These findings provide valuable insights into developing accurate and reliable Paygo prediction models. The hybrid model's improved accuracy suggests that it is better equipped to handle the complexities of loan default prediction in off-grid communities. Ultimately, this approach can facilitate the wider adoption of solar companies in Africa by mitigating the financial risks associated with loan defaults.

Keywords: Pay-As-You-Go; Xgboost; Meta-heuristic; Solar Home systems; PSO

1. Introduction

PAYGo business model in solar power systems has become increasingly popular in Africa as a way to provide access to electricity to people who are not connected to the grid commonly known as the Last Mile Distribution. These systems typically include a solar panel, battery, and a small device that controls the flow of electricity and manages payments. Customers pay for the electricity they use on a daily, weekly, or monthly basis, using mobile money or other digital payment methods. Upon completion of full payments, the ownership of the product is transferred to the customer [1] allowing them to affordably start with a smaller system and gradually add more components as their needs and incomes grow. Additionally, because customers pay for electricity regularly, it creates a steady cash flow for the solar company, which can be used to finance the expansion of the business. It can be difficult to reduce defaults in pay-as-you-go (PAYGO) systems since most target customers have mid-income to low-income levels and may find it difficult to make consistent payments. If implemented and managed properly, PAYGo systems can be profitable by generating a

[*] Corresponding author: Machariah Denis

consistent cash flow, maintaining low operating costs, and being able to monetize the data they collect from their PAYGo systems, such as by selling it to other businesses or using it to improve their operations.

1.1. Problem Statement

Pay-as-you-go (PAYGo) solar enterprises in Africa face significant challenges in predicting customer default and lifetime value due to dynamic customer behavior, macroeconomic factors, and limited historical data. This lack of precise customer insights hinders the growth of clean energy access and the financial sustainability of these businesses. To address this issue, this thesis proposes a novel approach utilizing advanced machine learning techniques to improve customer default prediction and CLTV estimation.

Objective

To develop a hybrid meta-heuristic optimization model that will maximize the accuracy of default prediction in the pay-as-you-go model by tuning parameters of the XgBoost algorithm for feature selection.

2. Literature Review

A research by [2] addressed the driving behavior features and risk indicator features extracted from vehicle trajectory data for modeling. Vehicle trajectory data from US Route 101 were collected using the Xgboost algorithm for feature selection showed that the approach was effective and reliable in identifying important features for driving assessment, and achieving an accurate prediction of risk levels.

According to [3] personal credit scoring is a challenging issue addressed using machine learning techniques like Extreme Gradient Boosting. The choice of the Xgboost algorithm is based on the advantages of feature combination and feature selection power by using decision trees on data with a high dimension and a complex correlation. The adaptive particle swarm optimization (APSO)-Xgboost model for credit scoring. APSO is more suitable for the parameter optimization of Xgboost, and it improves the model prediction accuracy, resulting in the best accuracy of 76.85%.

Meta-heuristic optimization algorithms are a type of optimization algorithm that uses generic strategies to find approximate solutions by imitating natural phenomena like evolution, swarm intelligence, and natural selection. Ant Colony Optimization (ACO), developed by Gambardella Dorigo [4] is inspired by ants seeking their way between the colony and their food. These algorithms are best suited for solving problems that traditional optimization algorithms find difficult to solve.

Meta-heuristics are generic strategies to find approximate solutions. In practice, many optimization problems (searching, machine learning, etc.) are NP-hard, requiring much computing effort to solve them. Others include; Particle swarm optimization which is an evolutionary algorithm first described in 1995, it inhibits behavior observed in the school of fish and swarm of birds [5], Bat Algorithm, and Genetics algorithm among others. In meta-heuristics, the process is illiterate until an approximate solution is chosen; this allows it to search enormously huge spaces of candidate solutions while making little or no assumptions about the problem being optimized [6].

However, there are critical gaps in the current body of knowledge. Firstly, a consistent and optimal method for feature selection, particularly within the context of pay-as-you-go customer prediction, remains elusive. Existing research presents various approaches. Secondly, there is a complete lack of research focused on the application of Xgboost with hyperparameter tuning for feature selection in the pay-as-you-go (PAYGo) business model. This innovative model, originating in Kenya and gaining traction globally, presents a unique opportunity for analytical research. In predicting customer churn in business, various techniques have been used to acquire such knowledge. Data mining and machine learning models have worked to solve the churning prediction menace. However, with time they have not been providing effective model output in prediction. These algorithms can provide accurate and reliable results by allowing institutions to process massive amounts of data, which is critical for avoiding potential losses. When these algorithms are combined with human expertise, they can lead to well-informed and sound financial decisions [7]. Finally, the inherent differences between PAYG customers and traditional loan (commercial loan) customers pose a challenge. Data specific to PAYGo customers, their backgrounds, risk profiles, and historical data availability are often restricted within individual companies, making comprehensive analysis difficult.

From the research summary few research gaps emerge and the solution is provided in this paper. (i) There are good predictors but these system needs an algorithm to support the hypertuning of the estimators that quickly adapt to the dynamic business for self-learning as an autonomous system and (ii) The accuracy of a model depends mainly on preprocessing and feature selection processes thus a need of a mechanism to decide the data cleaning techniques that are to be applied simultaneously.

The review of the literature showed that hybrid meta-heuristic approaches have high prediction accuracy, yet identify a gap in terms of the quality and number of features used in these models. Specifically, the hybrid and meta-heuristic models were found to outperform single models due to their adaptability and flexibility on complex data that has little past on the customer. Therefore, the focus of this paper is to develop a standardized set of features that can improve the accuracy of the model. This involved filling the gap by exploring and understanding how these features can be used to enhance the performance of the model within the Pay-As-You-Go (PAYGo) model, specifically for solar companies in Africa. By bridging this gap and providing insights into the utility of specific features, this paper aimed to enhance the predictive capabilities of hybrid meta-heuristic optimization using the Xgboost algorithm in the PAYGo model

3. Material and methods

3.1. Xgboost Algorithm for Feature Selection

The Xgboost algorithm has several parameters, there are main parameters in Xgboost and their effects on model performance. Some of these parameters include;

Table 1 Xgboost Algorithm Booster Parameters

Parameter	Description	Typical Values
eta	Analogous to the learning rate in GBM.	0.01-0.2
min_child_weight	Defines the minimum sum of weights of observations required in a child.	Tuned using CV
max_depth	The maximum depth of a tree. Used to control over-fitting.	0.3 - 10
max_leaf_nodes	The maximum number of terminal nodes or leaves in a tree.	
gamma	Specifies the minimum loss reduction required to make a split.	Tuned depending on the loss function
max_delta_step	Allows each treeâ€™s weight estimation to be constrained.	Usually not needed, explore if necessary
subsample	Denotes the fraction of observations to be random samples for each tree.	0.5-1
colsample_bytree	Denotes the fraction of columns to be random samples for each tree.	0.5-1
colsample_bylevel	Denotes the subsample ratio of columns for each split in each level.	Usually not used
lambda	L2 regularization term on weights (analogous to Ridge regression).	Explore reducing overfitting
alpha	L1 regularization term on weight (analogous to Lasso regression).	Used for high dimensionality
scale_pos_weight	Used in case of high-class imbalance for faster convergence.	Greater Than 0

It is important to note that both parameters and hyper-parameters are variables of a model. They are differentiated by how they learn in the model during data training, in the case of parameters. They can take different values by learning from data while hyper-parameters are by setting up the values manually.

- **Importance level**: The Gain is the most relevant attribute to interpret the relative importance of each feature. Gain is the improvement in accuracy brought by a feature to the branches it is on. Weight - This is the percentage representing the relative number of times a particular feature occurs in the trees of the model. The frequency for features is calculated as its percentage weight over the weights of all features.

- **learning_rate and n_estimators -** This is step size shrinkage used in the update to prevent overfitting. After each boosting step, we can directly get the weights of new features, and eta shrinks the feature weights to make the boosting process more conservative. Conventional ranges are: [0,1]
- **Tune max_depth and min_child_weight -** These have the highest impact on the model outcome.
- **Tune gamma -**The range of that parameter is [0, ∞], thus a "good" gamma is very dependent on both the data set and the other parameters. To ensure the optimal gamma, re-calibrating 5 boosting rounds were used.
- **subsample, colsample_bytree, and regularization -** Subsample: This hyperparameter controls the fraction of the training data that is randomly sampled to grow each tree. It is a value between 0 and 1. A value of 1.0 means using all training data for each tree, while a value less than 1.0 means subsampling. Colsample_bytree: This hyperparameter controls the fraction of features (columns) to be randomly sampled for each tree. It is also a value between 0 and 1. A value of 1.0 means using all features for each tree, while a value less than 1.0 means subsampling the features.

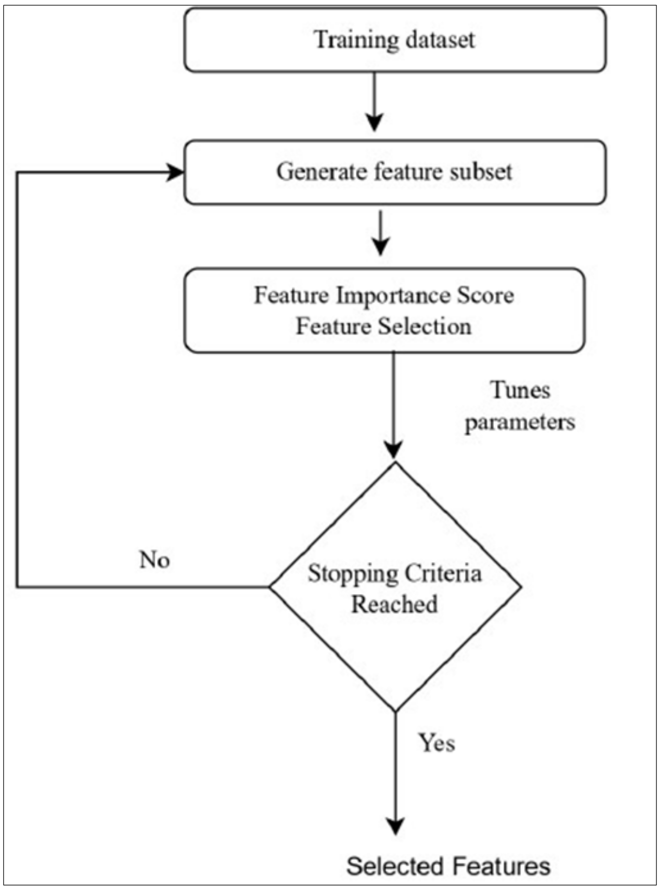

Figure 1 Flow chart of Feature selection using Xgboost algorithm

3.2. Meta-heuristic PSO-RF model

The position of the particle population is randomly selected within the boundaries the value of the search area and the value of the objective function are calculated for each particle position. PSO can simulate the behaviors of swarms to optimize a numeric problem iteratively. Its usage gets up-voted as the expected final result of the particle swarm converges to the best solution. Cost (fitness) value function minimization or maximization is the goal of PSO.

Random forests represent a powerful ensemble learning technique within the realm of supervised learning. They leverage the combined predictive power of multiple classifiers to achieve enhanced accuracy and robustness. This approach aggregates predictions from various individual models, ultimately leading to a more reliable outcome.

Even while straightforward predictive models and data processing techniques can significantly boost the capabilities of intelligent models, the structure and hyperparameters of the model can still be adjusted to further increase forecasting performance. The goal of the meta-heuristic optimization methods is to improve upon the initial data optimization. The entire process of obtaining input data is a standard feature selection, wherein input data is produced using meta-heuristic algorithms and utilized as training data to provide forecasting outcomes.

The combination of the three algorithms each playing a different role ensured that the model had the highest accuracy to fit the objectives of the paper.

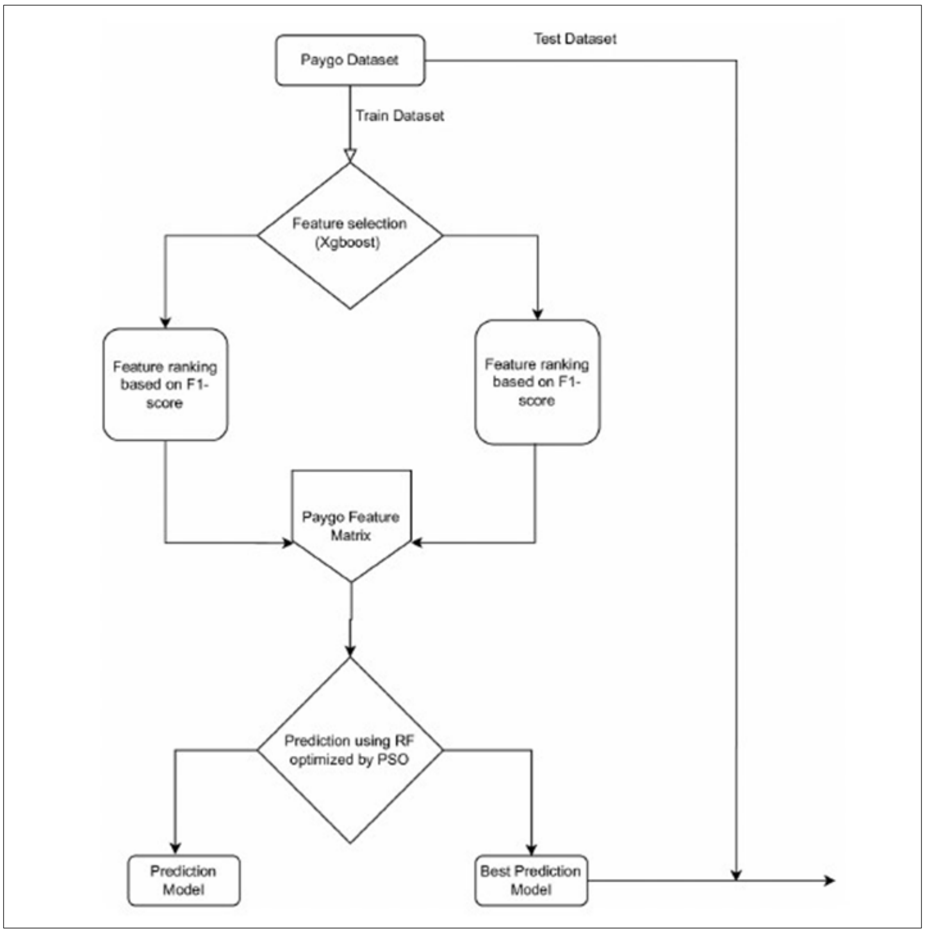
Paygo Dataset
Test Dataset
Train Dataset
Feature selection (Xgboost)
Feature ranking based on F1-score
Feature ranking based on F1-score
Paygo Feature Matrix
Prediction using RF optimized by PSO
Prediction Model
Best Prediction Model

Figure 2 Overall perfection flow using meta-heuristic PSO-RF

For classification and regression tasks, the widely utilized ensemble learning method Random Forest is used. To create predictions, it integrates several decision trees, which offer accuracy and robustness.

Particle Swarm Optimization (PSO), on the other hand, is a meta-heuristic optimization algorithm that draws inspiration from the behavior of swarms or flocks. Although the Random Forest algorithm itself does not require optimization, you can utilize Particle Swarm Optimization to fine-tune its hyper-parameters. Hyper-parameters are settings made before the training process that are not determined by the data. It is possible to enhance the performance of the Random Forest model by optimizing certain hyper-parameters.

4. Result

4.1. Dataset

This research on Hamara Africa Energy, a Kenyan company deploying the PAYGo model for solar and LPG in rural counties, relied on a sector-specific dataset obtained from the company itself. This approach was necessary because general solar industry datasets were not sufficient for the research goals. While such datasets might contain information on panel efficiency or battery capacity, they would lack the critical details specific to the PAYGo model. The utilized dataset, containing 27 features and 9456 data points, likely captured PAYGo-specific metrics like payment installments, mobile money transactions, and user behavior within that payment structure. This targeted information, unavailable in broader solar industry data, proved essential for understanding the specific challenges and opportunities faced by companies operating in the PAYGo solar sector.

Table 2 Features on the dataset

Feature	Description
Loan_ID	Unique ldentifier
Application_type	application mode: individual, group, partnership or sponsorship
Education	Level of education for the applicant: Primary, Secondary, College or university
Self-employed	Self employed: Yes or No
Yearly Income	Yearly income of Customer
Dependants	Number of dependants of the customer
Previously Defaulted	Has the customer previously defaulted
Debt to Income	Ratio of debt to income of the customer
Product Amount	Cost of the product
Interest Charged	Interest charged on the loan
Present Balance	Current balance in open account
Location_code	Location of the customer
Inquiries	Calls made to customers to remind/demand payment
Coapplicant_insured	Sum assured by the guarantor incase of default
Loan_Term	Duration of the financing
Monthly Installments	Installments paid on monthly basis
Unpaid Amount	Cumulative debt the customer owes the company
Product	They type of solar product financed to the customer
Ave_Monthly_moneyIn	Average money in from the M-pesa Statement
Ave_Monthly_MoneyOut	Average Money out from the M-pesa Statement
Customer_Phone	Custome phone number
Customer_CRB_Listed	Is the customer Listed in the credit bureau
ID-Number	ID-number of the custmer
Guarantor_Name	Guarantor's Name
Guarantor_M-pesa Statement	Has guarantor provided M-pesa statement
Customer_Mpesa Statement	Has customer provided M-pesa statement
Guarantor_ CRB Listing	Is guarantor listed on credit bureau

4.2. Optimal Parameter Values

Parameters under the importance type are gain, weight, and a combination of gain and weight. This comparison showed the most important parameter to be used.

Table 3 Accuracy under optimal values of gain and weight parameters

Importance_type	Gain	Weight	Gain+weight
Best accuracy score	0.81164	0.81177	0.81289
Optimization time	229.47secs	249.23secs	254.01secs
Mean squared error	0.1904	0.1904	0.11235

Stopping search of the model: Swarm best objective change less than 0.001

Learning_rate controls the shrinkage of each tree's contribution and typical values range from 0.01 to 0.1, while n_estimators determine the number of trees in the ensemble.

Table 4 Learning_rate and n_estimate

Parameters	Values
learning_rate	0.1
n_estimators	253

The higher the learning rate, the more computational power is required.

Tune max_depth and min_child_weight} **-** Heavy Gridsearch is performed. Takes 6054.17 seconds to complete tuning to achieve optimal "depth" and "child_weight

Table 5 Optimal parameter point for depth, weight, and best score

Parameter	min	Optimal	max
max_depth	3	5	7
min_child_weight	0.5	1	2
Best_score	0.89	0.933	0.933

Deeper trees can capture more complex interactions but may overfit.

Tune gamma

Table 6 Optimal gamma parameters

Parameter	min	Optimal	max
max_depth	0.0	0.2	0.5

Subsample, colsample_bytree and regularization **-** Subsample: This hyperparameter controls the fraction of the training data that is randomly sampled to grow each tree. It is a value between 0 and 1. A value of 1.0 means using all training data for each tree, while a value less than 1.0 means subsampling. Colsample_bytree: This hyperparameter controls the fraction of features (columns) to be randomly sampled for each tree. It is also a value between 0 and 1. A value of 1.0 means using all features for each tree, while a value less than 1.0 means subsampling the features.

Table 7 Values for Subsample, colsample_bytree and regularization

Parameter	Optimal
Subsample	0.8
Colsample_bytree	0.8
Reg_alpha	0.1

4.3. Selected Features

The effectiveness of Xgboost for feature selection and model optimization in customer risk assessment is demonstrated by this experiment. Using Collaboration Matrix to compare the results of the above feature importance score results

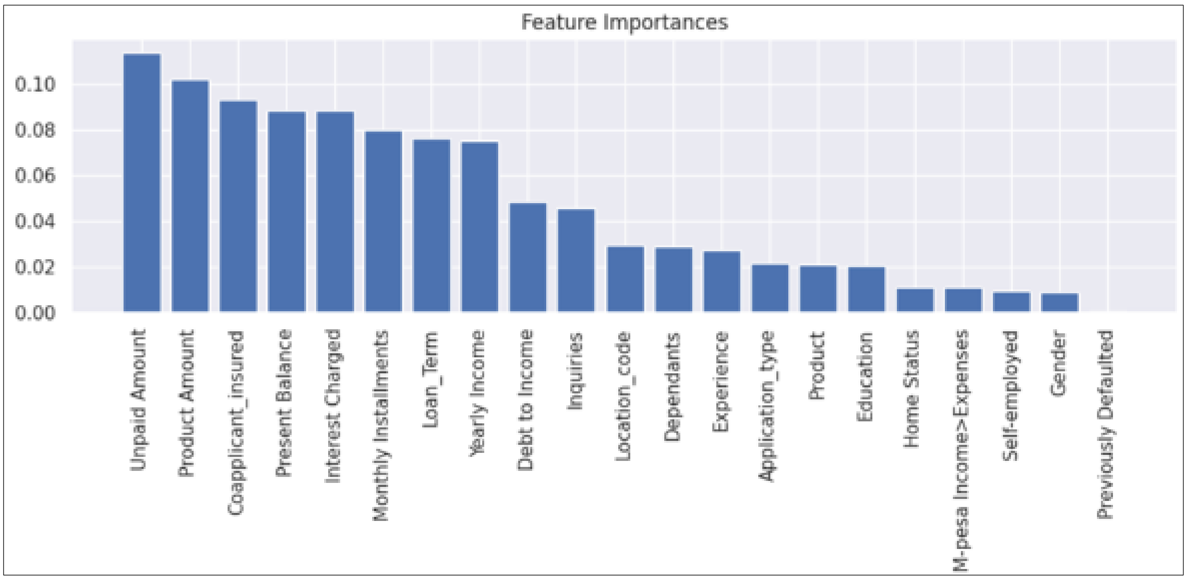

Figure 3 Selected feature using Xgboost algorithm optimal parameters

4.4. Prediction Accuracy

For benchmark purposes and a comparison of the performance of the model, individual model prediction on training data was done.

Table 8 Model accuracy perfomance

Parameter	Default parameters	Tuned hyperparameters
Best accuracy	81.289%	93.301%
Test accuracy	81.379%	93.456%

The above-tuned parameters run three times to generate a class reports. The hybrid model performs much better than the individual models prescribed in the benchmark section. This shows that hyper-tuning and optimization are paramount to achieving the overall performance of a model.

Table 9 Prediction Model class report

	Precision	Recall	F1-score	Support
0	0.81	1	0.89	8220
1	0.03	0.01	0.05	1236
Accuracy			0.93301	9456
Macro avg	0.9	0.51	0.47	9456
Weighted avg	0.5	0.81	0.73	9456

5. Discussion

A value of 0.933 suggests the model can very effectively distinguish between those who will repay their loans (good) and those who are more likely to default (bad). This allows Paygo solar companies to make more informed decisions about loan approvals, potentially reducing defaults and improving their financial sustainability.

The model's discriminatory strength and the trade-off between sensitivity and specificity are generally revealed via the ROC curve and AUC, which aid in understanding how well the model performs across a range of decision thresholds.

The decision threshold for classifying loans as good or bad might need adjustments based on the company's risk tolerance and the cost associated with different types of errors (e.g., approving a bad loan vs. rejecting a good one).

The effectiveness of XGBoost for feature selection and model optimization in customer risk assessment is demonstrated by this experiment. Highlights from this research include;

Baseline Performance - An accuracy of 81.2% was obtained by using XGBoost with the default hyperparameters. This creates a foundation for development

Impact of Hyperparameter Tuning: XGBoost hyperparameter adjustments greatly improved the model's capacity to extract pertinent features from the 24-dimensional dataset. This optimization method increased accuracy.

Accuracy and Feature Selection: A subset of 18 features were identified by the revised model, which resulted in an impressive 93.4% accuracy on the training data and a highly generalizable performance of 93.3% on the test data. This is a significant advancement over the baseline model.

The results illustrate the effectiveness and dependability of this strategy in finding essential characteristics for customer assessment and enabling accurate risk level predictions, despite the computational overhead.

6. Conclusion

Unveiling a distinctive methodology for predicting customer attrition by leveraging diverse attributes. The discoveries outlined in this research can underscore the importance of integrating the option for loan repayment of Solar Home Systems (SHS) in Africa, especially in the operations of a renewable energy enterprise. This approach enables the facilitation of access to sustainable energy solutions without the burden of initial expenses, thereby fostering a more environmentally friendly and sustainable ecosystem.

Compliance with ethical standards

Acknowledgments

We would like to acknowledge Hamara Africa Energy for the provision of data to undertake this study.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Gennaro Cuofano. What Is The Pay-As-You-Go Business Model? The Pay-As-You-Go Business Model In A Nutshell - FourWeekMBA [Internet]. FourWeekMBA. 2024 [cited 2024 Nov 25]. Available from: https://fourweekmba.com/pay-as-you-go-business-model

[2] Hossein Abbasimehr, Reza Paki, Aram Bahrini. A novel XGBoost-based featurization approach to forecast renewable energy consumption with deep learning models. Sustainable computing (Print). 2023 Apr 1;38: 100863–3.

[3] Qin C, Zhang Y, Bao F, Zhang C, Liu P, Li P. XGBoost Optimized by Adaptive Particle Swarm Optimization for Credit Scoring. 2021 Mar 23; 2021:1–18.

[4] Luca Maria Gambardella, Dorigo M. Ant-Q: A Reinforcement Learning approach to the traveling salesman problem. Elsevier eBooks. 1995 Jul 9;252–60.

[5] Aote SS, Raghuwanshi MM, Malik L. A brief review on particle swarm optimization: limitations & future directions. International Journal of Computer Science Engineering (IJCSE). 2013 Sep;14(1):196-200.

[6] Desale S, Rasool A, Andhale S, Rane P. Heuristic and meta-heuristic algorithms and their relevance to the real world: a survey. Int. J. Comput. Eng. Res. Trends. 2015 May;351(5):2349-7084.

[7] Gao L, Xiao J. Big data credit report in credit risk management of consumer finance. Wireless Communications and Mobile Computing. 2021;2021(1):4811086.