International Journal of Data Science and Analysis
2024; 10(3): 41-48
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20241003.11
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

A Hybrid Extreme Gradient Boosting Model for Credit Risk Modelling in the Presence of Inflation

Kenneth Kiprotich Langat[1, 2, *], Anthony Gichuhi Waititu[3], Philip Odhiambo Ngare[4]

[1]Department of Mathematics, Pan African Institute of Basic Science Technology and Innovation, Nairobi, Kenya
[2]Department of Mathematics, Egerton University, Nakuru, Kenya
[3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[4]Department of Mathematics, University of Nairobi, Nairobi, Kenya

Email address:
langatken26@gmail.com (Kenneth Kiprotich Langat)
[*]Corresponding author

To cite this article:
Kenneth Kiprotich Langat, Anthony Gichuhi Waititu, Philip Odhiambo Ngare. (2024). A Hybrid Extreme Gradient Boosting Model for Credit Risk Modelling in the Presence of Inflation. *International Journal of Data Science and Analysis, 10*(3), 41-48.
https://doi.org/10.11648/j.ijdsa.20241003.11

Received: 12 July 2024; **Accepted:** 31 July 2024; **Published:** 22 August 2024

Abstract: The recent developments in the credit and banking industry brought by technology has led to increased competition and the rise of risks and challenges. Credit scoring is one of the core items that keeps this industry competitive and profitable. The creation of credit score models to assess the ability of the loan applicant to repay his or her loan remains an active field of research. Practically, the existing models ignore the factor of inflation in determining the credit score of a loan applicant. Inflation affect the performance of the financing institution negatively because it makes some of the borrowers struggle to repay the loan and so leading to some bad debts that might end up being written off. By integrating the inflation factor to the Extreme gradient boosting algorithm led to improved accuracy of the model. In this paper, a new model that uses the inflation rate of a specific region or country in the regularization term of the extreme gradient boosting model has been developed. The evaluation of the model is by comparison with the other common models using ROC, Accuracy, precision and recall. The developed model emerge the second best in terms of performance but better than the standard extreme gradient boosting model.

Keywords: XGBoost, Inflation, Decision Tree, Credit Analysis

1. Introduction

Credit analysis is the process that credit manager of the lending company carries out in order to measure the ability of the borrower to repay his or her loan. Credit scoring is the tool that can rate the creditworthiness of the client using the personal information. Depending on the policies of the lender, the customer will be classified as either defaulter or non-defaulter using his or her score points. Credit scoring model is build using historical data of the current and previous customers of the lender. Thereafter, the lender will categorize client as bad if his or her character matches the ones of the clients classified as bad using the historical data, otherwise good. Loan application for the bad customers are not approved while those for good customers will be approved but the amount given and the interest rate charged will be pegged on his or her score. A more accurate credit scoring model translate to small percentage of misclassification and thus reducing the rate of losses made by the financing institution.

The main interest of any credit scoring model is to minimize the error of classifying bad clients as good and vice versa. XGBoost (Extreme Gradient Boosting) is a machine learning model with very high accuracy but does not consider inflation rate when used for credit scoring. This study has modified XGBoost so as to incorporate inflation rate.

Money lending is one of the ancient business in financial sector that has always played a key role in every economy of a country or region. It was majorly dominated by banks for many years before the emergence of technology. Its history can be traced back to before 2000 BC when the rich in the society used to keep their coins in the temple and all transactions

of money were done in the temple before the monk(s) or by the monk(s) [1]. Lending is the process of extending asset (money) to the borrower by the lender at an agreed interest rate for a certain agreed period of time [2].

The borrower is expected to return the asset to the lender before or at the end of the agreed time without failing. In money lending, some of the clients fail to repay their loans on time due to many reasons best known by them. This is the main risk that is experience by any loaning institutions and is normally tamed by minimizing the risk through various means. Banks and other financial Institutions mainly depend on the banking history of a client in order for them to decide the creditworthiness of the customer.

They achieve this by developing a scorecard using common features in all clients. The scorecard will allocate each feature some points which when summed up, it gives the credit score for an individual customer. Credit score is then used to separate the good clients from the bad and hence allowing loan officers to decide whether to give loan to customer or not and what amount does s/he qualify for. This has really minimize the chances of many people especially in developing countries to access loans from banks because they don't have banking history [3]. The availability of technology have totally changed this narrative through mobile loan lenders who are now making money by extending loans to the unbanked group of people. They are able to build a scorecard using the mobile phone data which enable them make judgement on the ability of the client to repay the loan without defaulting.

The number of mobile loan borrowers has been increasing exponentially. In Kenya as at December, 2018, 13.42 million adults were using mobile loans [4]. The research carried out by [5]in Kenya and Tanzania shows that the default rate of the mobile loans is still high in the two countries. The default rate in Kenya was 12 percent while in Tanzania was 31 percent and the loans that were paid late by clients was 50 percent and 56 percent respectively. According to Eric Njagi, director of M-Shwari, the overdue loans in the first month after Covid 19 was reported in Kenya shift from 18 percent in the month of March to 23 percent in the month of April. This indicates the need of considering unexpected events when coming up with a scorecard tool.

According to [6], the mobile phone data that qualify to be used are the ones for the clients that has been registered in a telecommunication company for at least six months. This is because they normally deregister any contact that has not been active for the last six months. Other than that, six months is a reasonable time for that contact to accumulate data. The number of phone users has kept on increasing worldwide [7] which is evidence enough that in future everyone would be having a phone and this implies that market for digital loans keeps on expanding and thus the mobile loan business is here to stay.

Considering this, then there is need to keep on improving the credit tools used to manage the risk of defaulters. Use of few features when developing scorecard would yield best probability estimates and at the same time protect the privacy of the phone owner [8]. In order for you to have a more accurate credit score model, lender need to use as many features/factors as possible. This is what we intend to do as a way building a more accurate scorecard to be used by short-term loan lenders.

XGBoost is a popular and powerful ensemble machine learning algorithm that has been used in various fields such as computer vision, natural language processing, and most importantly, credit scoring. The literature review of XGBoost scorecard models for mobile phone loans considers several studies, including: Chen and Guastrin [9] stated that XGBoost is a scalable and efficient tree boosting system that can handle large-scale data sets. They highlight the advantages of XGBoost over other tree-based algorithms and describes the key features that make it a powerful tool for predictive modeling.

Friedman [10] provided an overview of gradient boosting machines and their application to credit scoring. He discussed the greedy function approximation technique used in XGBoost and showed how it can improve the performance of credit scoring models. [11] discussed the use of random forests in credit scoring and how it can be used in combination with XGBoost to improve the performance of credit scoring models. The hybrid approach that combine genetic algorithms with dual scoring models to improve the performance of credit scoring models was carried out by [12].

They demonstrated the effectiveness of this approach by use of extreme gradient boost model as their base model. In the area of fraud detection, [13] used classification models to detect the frauds in credit cards. They compared these several models in terms of performance and their results proved that XGBoost can achieve high accuracy and good generalization performance. The research work of [14] provides an overview of the application of data mining techniques in credit risk evaluation that include extreme gradient boosting.They discuss the advantages of using XGBoost in credit risk evaluation and showed how it can be used to improve the performance of credit scoring models. The use of XGBoost to evaluate personal credit analysis was explore by [15] to show how this model can be used to improve the performance of credit scoring models.

They discussed key features that makes the model powerful tool for predictive modeling and illustrate by an example on how it can be used in practice. Overall, the literature suggests that XGBoost is a powerful and efficient algorithm that can be used to improve the performance of credit scoring models. It has been applied to various fields and has shown good results in terms of accuracy, generalization, and scalability. Additionally, it has been combined with other techniques such as Random Forest, Genetic Algorithm to achieve better results.

2. Method

2.1. Extreme Gradient Boosting

XGBoost is a model that combine weak learning models (decision trees), one at a time in order to end up with a strong

learning model. strength or weakness of a model is in terms of performance.

The model entails;

n - Total number of samples (loan clients)

m - Number of variables/features

x_i - variable information of the i^{th} sample, $x_i \in \mathbb{R}^m$

y_i - The value of the i^{th} sample

$\hat{y_i}$ - The predicted value of the i^{th} sample

$\hat{y_i}^t$ - The predicted value up to the t^{th} tree

$l(y_i, \hat{y_i})$ - The loss function of the i^{th} sample

$L(y, \hat{y})$ - The loss function of total sample

$\Omega(f_k)$ - Regularization term of objective function to prevent overfitting, f_k represent the k^{th} decision tree

$D = \{(x_i, y_i | x_i \in \mathbb{R}^m, y_i \in \mathbb{R}, x_i = \{x_{i,1}, x_{i,2}, ..., x_{i,m} | i = 1, 2, ..., n\})\}$

The model need to build t-trees and their prediction are as follows;

$$\hat{y}_i^0 = 0 \tag{1}$$

Equation (1) is the prediction of the 0^{th} tree.

$$\begin{aligned} \hat{y}_i^1 &= f_1(x_i) \\ &= \hat{y}_i^0 + f_1(x_i) \end{aligned} \tag{2}$$

Equation (2) is the prediction of the 1^{st} tree.

$$\begin{aligned} \hat{y}_i^2 &= f_1(x_i) + f_2(x_i) \\ &= \hat{y}_i^1 + f_2(x_i) \end{aligned} \tag{3}$$

Equation (3) is the prediction of the 2^{nd} tree.

$$\vdots$$

$$\begin{aligned} \hat{y}_i^t &= f_1(x_i) + f_2(x_i) + ... + f_t(x_i) \\ &= \sum_{k=1}^{t} f_k(x_i) \\ &= \sum_{k=1}^{t-1} f_k(x_i) + f_t(x_i) \\ &= \hat{y}_i^{t-1} + f_t(x_i) \end{aligned} \tag{4}$$

Equation (4) is the prediction of the t^{th} tree.

At every iteration, a weak model $f_k(x_i)$ (decision tree) is generated. On the t^{th} iteration, $\hat{y}_i^t$ (the t^{th} prediction value) is the sum of $\hat{y}_i^{t-1}$, prediction of the previous iteration and the decision tree results of the t^{th} round, $f_t(x_i)$.

The model is a combination of loss function and regularization term to get the objective function, L^t.

$$minL^t(y, \hat{y}^t) = min(\sum_{i=1}^{n} l(y_i, {\hat{y_i}}^t) + \Omega(f_t)) \tag{5}$$

The loss function $l(y_i, {\hat{y_i}}^t)$ is estimated using Taylor approximation up to second order.

$$\begin{aligned} \sum_{i=1}^{n} l(y_i, {\hat{y_i}}^t) &= \sum_{i=1}^{n} l(y_i, {\hat{y_i}}^{t-1} + f_t(x_i)) \\ &= \sum_{i=1}^{n} \Big(l(y_i, {\hat{y_i}}^{t-1}) + \frac{\partial l(y_i, {\hat{y_i}}^t)}{\partial \hat{y}_i^t} f_t(x_i) + \frac{\partial^2 l(y_i, {\hat{y_i}}^t)}{2\partial (\hat{y}_i^t)^2} f_t^2(x_i) \Big) \\ &= \sum_{i=1}^{n} \Big(l(y_i, {\hat{y_i}}^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \Big) \\ &= l(y_1, {\hat{y_1}}^{t-1}) + g_1 f_t(x_1) + \frac{1}{2} h_1 f_t^2(x_1) + l(y_2, {\hat{y_2}}^{t-1}) + g_2 f_t(x_2) + \frac{1}{2} h_2 f_t^2(x_2) \\ &+ ... + l(y_n, {\hat{y_n}}^{t-1}) + g_n f_t(x_n) + \frac{1}{2} h_n f_t^2(x_n) \end{aligned} \tag{6}$$

$l(y_1, {\hat{y_1}}^{t-1})$, $l(y_2, {\hat{y_2}}^{t-1})$, ..., $l(y_n, {\hat{y_n}}^{t-1})$ are dropped out since the are not affect by the output of the tree(s) and this simplify equation (7) to;

$$\begin{aligned} \sum_{i=1}^{n} l(y_i, {\hat{y_i}}^t) &= g_1 f_t(x_1) + \frac{1}{2} h_1 f_t^2(x_1) + g_2 f_t(x_2) + \frac{1}{2} h_2 f_t^2(x_2) + ... + g_n f_t(x_n) + \frac{1}{2} h_n f_t^2(x_n) \\ &= \sum_{i=1}^{n} \Big[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \Big] \end{aligned} \tag{7}$$

Going back, equation (5) become;

$$L^t = \sum_{i=1}^{n} \Big[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \Big] + \Omega(f_t) \tag{8}$$

Among the n samples, some will share same leaf node and thus making the summation to run from 1 to the number of leaf nodes in a tree.

The set of samples sharing same leaf is represent by, $I_j = \{i | q(x_i) = j\}$, which is in the j^{th} leaf node in the tree. $q(x_i)$ is

a function that maps the samples x_i to the leaf j. Letting w to be the score of the leaf, then $f(x) = wq(x)$, $w \in \mathbb{R}^T, q : \mathbb{R}^d \longrightarrow \{1, 2, ..., T\}$

Considering that;

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_{t,j}^2$$

and

$$f_t(x_i) = w_t q(x_i)$$

Where,

T_t - Number of leaf nodes in the t^{th} tree

γ - Contraction (pruning) coefficient of the number of leaf nodes

$w_{t,j}$ - the score of the j^{th} leaf node in the t^{th} tree

λ - penalty coefficient of the score of leaf nodes.

So,

$$L^t = \sum_{j=1}^{T_t} \left[G_j w_{t,j} + \frac{1}{2}(H_j + \lambda) w_{t,j}^2 \right] + \gamma T_t \tag{9}$$

Where,

$$G_j = \sum_{i \in I_j} g_i$$

and

$$H_j = \sum_{i \in I_j} h_i$$

To minimize L^t, $G_j w_{t,j} + \frac{1}{2}(H_j + \lambda) w_{t,j}^2$ is differentiated w.r.t. $w_{t,j}$ and equated to zero. The optimal score of the j^{th} leaf in the t^{th} tree is;

$$w_{t,j}^* = -\frac{G_j}{(H_j + \lambda)} \tag{10}$$

Thus,

$$min L^t = -\frac{1}{2} \sum_{j=1}^{T_t} \frac{G_j^2}{(H_j + \lambda)} + \gamma T_t \tag{11}$$

Alternative Solution

Considering equation (9),

$$L^t = \sum_{j=1}^{T_t} \left[G_j w_{t,j} + \frac{1}{2}(H_j + \lambda) w_{t,j}^2 \right] + \gamma T_t$$

It is clear that part of it, that is, $G_j w_{t,j} + \frac{1}{2}(H_j + \lambda) w_{t,j}^2$ is a quadratic equation that forms an upward parabola. The minimum solution of a parabola function is normally the value of x, which is $-\frac{b}{2a}$ at the vertex of that function. For this case, $b = G_j$ and $a = \frac{1}{2}(H_j + \lambda)$ and hence the solution $w_{t,j}^* = -\frac{G_j}{(H_j + \lambda)}$ for leaf$j$ of tree t.

2.2. Modifying the Model

During wars, pandemic, catastrophic et cetera times, economies around the world experienced significant disruptions that lead to economic shocks. One of the consequences of such shocks is inflation, where the prices of goods and services increase and hence reducing the purchasing power of money. In the context of credit scoring, this inflation can affect borrowers' ability to repay loans, leading to higher default risks.

To mitigate the risks associated with increased inflation during economic shocks without increasing interest rates (which might drive borrowers to other lenders), Weights of the XGBoost model used for credit scoring can be adjusted. This adjustment help in controlling and reducing the amount clients are allowed to borrow, ensuring they are not overburdened with debt they may struggle to repay.

2.2.1. Adjusting Weights with Inflation

To incorporate the inflation rate into the XGBoost model, a modification of the regularization term is required. The regularization term typically involves parameters lambda (λ) or alpha (α) that control the weight penalty. Since we are using Ridge regularization, we then consider lambda.

2.2.2. Standard Regularization Term

The normal ridge regularization is given by the equation;

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_{t,j}^2 \tag{12}$$

We introduce the inflation rate (π) into the regularization term to adjust the weights.

2.2.3. Modified Regularization Term

After incorporating the inflation, regularization becomes,

$$\begin{aligned} \Omega'(f_t) &= \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_{t,j}^2 + \pi \sum_{j=1}^{T_t} w_{t,j}^2 \\ &= \gamma T_t + \frac{1}{2}(\lambda + 2\pi) \sum_{j=1}^{T_t} w_{t,j}^2 \end{aligned} \tag{13}$$

Where; π - Inflation rate at the current time

Study assumptions

This study makes the following two assumptions;

i The inflation must increase during economic shock times

ii The small change caused by inflation on lambda won't cause any over-fitting/under-fitting on the model.

2.3. Proposed Model

Based on the objective function of the normal Extreme Gradient Boosting model given by;

$$\min L^t(y, \hat{y}^t) = \min \Big(\sum_{i=1}^{n} l(y_i, \hat{y_i}^t) + \Omega(f_t) \Big) \tag{14}$$

The study replaces the standard regularization function $\Omega(f_t)$ with the modified regularization function $\Omega'(f_t)$ and so equation (14) becomes;

$$\min L'^t(y, \hat{y}^t) = \min \Big(\sum_{i=1}^{n} l(y_i, \hat{y_i}^t) + \Omega'(f_t) \Big) \tag{15}$$

Note that;

$$\Omega'(f_t) = \gamma T_t + \frac{1}{2}(\lambda + 2\pi) \sum_{j=1}^{T_t} w_{t,j}^2$$

and

$$f_t(x_i) = w_t q(x_i)$$

Where:

T_t - Number of leaf nodes in the t^{th} tree
γ - Contraction (pruning) coefficient of the number of leaf nodes
$w_{t,j}$ - the score of the j^{th} leaf node in the t^{th} tree
λ - penalty coefficient of the score of leaf nodes.
The study use the modified Ridge regularization, equation (13). Considering this, equation (15) becomes;

$$\begin{aligned}\min L'^t(y,\hat{y}^t) &= \min\Big(\sum_{i=1}^{n} l(y_i,\hat{y_i}^t) + \Omega'(f_t)\Big)\\ &= \min\Big(\sum_{i=1}^{n} l(y_i,\hat{y_i}^t) + \frac{1}{2}(\lambda+2\pi)\sum_{j=1}^{T_t} w_{t,j}^2 + \gamma T_t\Big)\end{aligned} \tag{16}$$

2.4. Optimizing the Modified Model

Assuming that data is give as follows, $(x_1,y_1),(x_2,y_2),$ $(x_3,y_3),...,(x_n,y_n)$ where x_i represents the independent variable and y_i represents the dependent variable. The optimization steps are given as:

$$\begin{aligned}\hat{y_i}^t &= \sum_{k=1}^{t} f_k(x_i)\\ &= \hat{y_i}^{(t-1)} + f_t(x_i)\end{aligned} \tag{17}$$

where $\hat{y_i}^t$ is the predicted value of the model in the round t and the proposed XGBoost model algorithm is formed by continuous iteration, and each iteration is trained by adding a lesson of decision tree to the prediction value $\hat{y_i}^t$ of the previous round.

Generally, the formula for the objective function is:

$$obj(\phi) = L(\phi) + \Omega(\phi) \tag{18}$$

where ϕ is the parameter to be estimated, $L(\phi)$ is the loss function and $\Omega(\phi)$ is the regularization term. Thus, we minimize $obj(\phi)$ which gives the criterion for selecting $f(x)$

$$\begin{aligned}L'^t &= \sum_{i=1}^{n} l\Big(y_i,\hat{y}_i^{(t)}\Big) + \Omega'(f_t)\\ &= \sum_{i=1}^{n} l\Big(y_i,\hat{y}_i^{(t-1)} + f_t(x_i)\Big) + \Omega'(f_t) + constant\end{aligned} \tag{19}$$

Taylor expansion is used to expand the approximate objective function and remove the constant term. Using equation (7), then equation (19) become;

$$\begin{aligned}L'^t &= \sum_{i=1}^{n}\Big[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\Big] + \Omega'(f_t)\\ &= \sum_{i=1}^{n}\Big[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\Big] + \gamma T_t + \frac{1}{2}(\lambda+2\pi)\sum_{j=1}^{T_t} w_{t,j}^2\end{aligned} \tag{20}$$

where:

$$g_i = \partial_{\hat{y}(t-1)} l(y_i,\hat{y}^{t-1})$$

$$h_i = \partial^2_{\hat{y}(t-1)} l(y_i,\hat{y}^{t-1})$$

Among the n samples, some will share same leaf node and thus making the summation to run from 1 to the number of leaf nodes in a tree.

The set of samples sharing same leaf is represent by, $I_j = \{i|q(x_i)=j\}$, which is in the j^{th} leaf node in the tree. $q(x_i)$ is a function that maps the samples x_i to the leaf j. Letting w to be the score of the leaf, then $f(x) = wq(x)$, $w \in \mathbb{R}^T, q : \mathbb{R}^d \longrightarrow \{1,2,...,T\}$

By letting,

$$f_t(x_i) = w_{t,j}$$

and

$$G_j = \sum_{i\in I_j} g_i$$

and

$$H_j = \sum_{i\in I_j} h_i$$

Then the final objective function is:

$$\begin{aligned}L'^t &= \sum_{j=1}^{T_t} G_j w_{t,j} + \Big[\frac{1}{2}\sum_{j=1}^{T_t}\Big(H_j+\lambda+2\pi\Big)w_{t,j}^2\Big] + \gamma T_t\\ &= \sum_{j=1}^{T_t}\Big[G_j w_{t,j} + \frac{1}{2}\Big(H_j+\lambda+2\pi\Big)w_{t,j}^2\Big] + \gamma T_t\end{aligned} \tag{21}$$

The optimal value of $w_{t,j}$ is obtained by differentiating equation (21) with respect to $w_{t,j}$ and equating to zero. Thus the optimal weight for each leaf of a tree is;

$$w'_{t,j} = -\frac{G_j}{H_j+\lambda+2\pi} \tag{22}$$

The final objective function becomes:

$$\min L'^{t} = -\frac{1}{2}\sum_{j=1}^{T_t}\frac{G_j^2}{H_j+\lambda+2\pi}+\gamma T_t \tag{23}$$

3. Results

This paper uses one data set to demonstrate the applicability of the new model in credit scoring. The new model is compared to traditional models used in credit scoring.

3.1. Data

The data fundamentally consist of customer information, such as demographic information, amount of money borrowed, frequency of loan repayment (weekly or monthly), outstanding loan balance, number of repayments and number of days in arrears. Table 1 shows the description of the features used.

Table 1. *Description of variables.*

Variable	Definition
Deposit	Amount deposited on MPESA
Received	Amount Received in MPESA
Transfer	Amount on transfers from Bank to MPESA
Credits	MPESA credits
Airtime	Number of times airtime is less than Ksh 2
Amount	Total amount on okoa jahazi
Status	Binary, a defaulter or non-defaulter

3.2. Data Preparing

The data-set used had 10,329 observations/clients and 9 features. The loan status column was generated by classifying all the ”loan age days” greater than 90 days as defaulters while those less than or equal to 90 days as non-defaulters. Conventionally, default = 1 and non-default = 0. The distribution is as shown in figure 1 below;

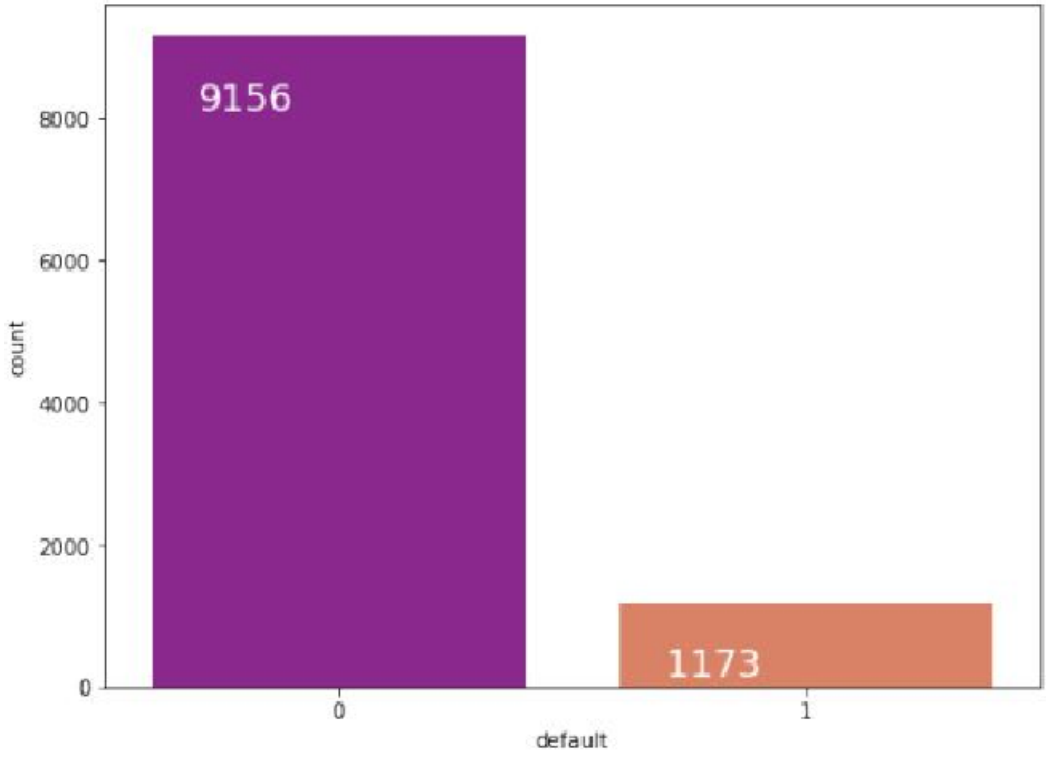

Figure 1. *Data Distribution of Defaulters vs Non-defaulters.*

This shows that there are 11.36% defaulters and 88.64% of non-defaulters and hence clear indication that the data set is highly imbalanced. Imbalanced data lead to getting pretty high accuracy by predicting the majority class but failing to capture the minority class, which is the most often the point of creating a model.

Random forest is one of the machine learning models that has the ability of performing well on imbalanced data-set. It is considered as a highly accurate and robust method because of the number of decision trees participating in the process. In addition, they don’t suffer from the over-fitting problem since it takes the average of all the predictions, which cancels out the biases. On training the data-set on the random forest model, the performance in form of confusion matrix and its report was as in figure 2.

	precision	recall	f1-score	support
0	0.90	0.99	0.94	2747
1	0.71	0.15	0.25	352
accuracy			0.90	3099
macro avg	0.81	0.57	0.60	3099
weighted avg	0.88	0.90	0.87	3099

Figure 2. *Random Forest Confusion Matrix for Imbalanced Data.*

The ROC-AUC score for this model is 64% which is above average but not as good as it should be. F1 score is also low and the cause of it is the imbalance data. The model is able to identify members of majority class well while being careful on the minority class. This is evidence in the classification report.

A recall score of 0.16 shows that the model has failed to predict 84% of the minority class and a precision score of 0.8 shows that 80% of the predictions on the minority class are correct. By inspecting confusion matrix, most of the predicted results are false negative and true negative. This results has a strong bias in predicting the results as the majority class since there are more training data samples than the minority class. Considering this explanation, it is clear that there is need of carrying out data balancing.

3.3. Data Balancing

After classifying the customers as defaulters and non-defaulters, we encountered an imbalance data where 88.6% of the entire data set belonged to the non-defaulters class. In such a scenario the danger is that any model fitted to the data might end up predicting the majority risk class all the time even though the model diagnostics show that the fitted model is good. To overcome this challenge we adopted the machine learning synthetic minority over-sampling technique for nominal and continuous to over-sample the minority classes to achieve fair representation for all classes in the data set. SMOTE works by generating instances that are close in feature space using interpolation between positive cases that are close to each other. It randomly selects a minority class instance and finds its nearest neighbour. Then it creates synthetic models by randomly choosing one of the neighbours and form a line segment in the feature space. After this the majority class constituted only 55.6%.

3.4. Training-Test Split

For all the models fitted in this study, we split the balanced data into 80% for the training set and 20% for the testing set (validation). Unless otherwise stated, all analyses presented in this paper were done using the Python programming language.

3.5. Summary Statistics of Features

Table 2 presents the summary statistics of the numerical features used in this paper.

Table 2. *Summary Statistics of numerical features.*

Variable	Mean	SD	Q1	Q2	Q3
Deposit	18602.61	133156.35	2500	27900	96700
Received	30606.07	189238.75	2200	38540	175640
Transfer	5412.64	41061.80	0	0	0
Credits	54621.33	254327.81	4684	12774	35030
Airtime	84.18	54.90	35	84.179	95
Amount	309.80	1139.25	0	0	1500

3.6. Feature Selection

The dataset had more variables than used in this paper. Some of the variables were of no use such as Customer ID which is simply a unique customer identifier for all customers in the data. Other variables such as date were not used. The rest of the variables given in Table 2 were used in the models.

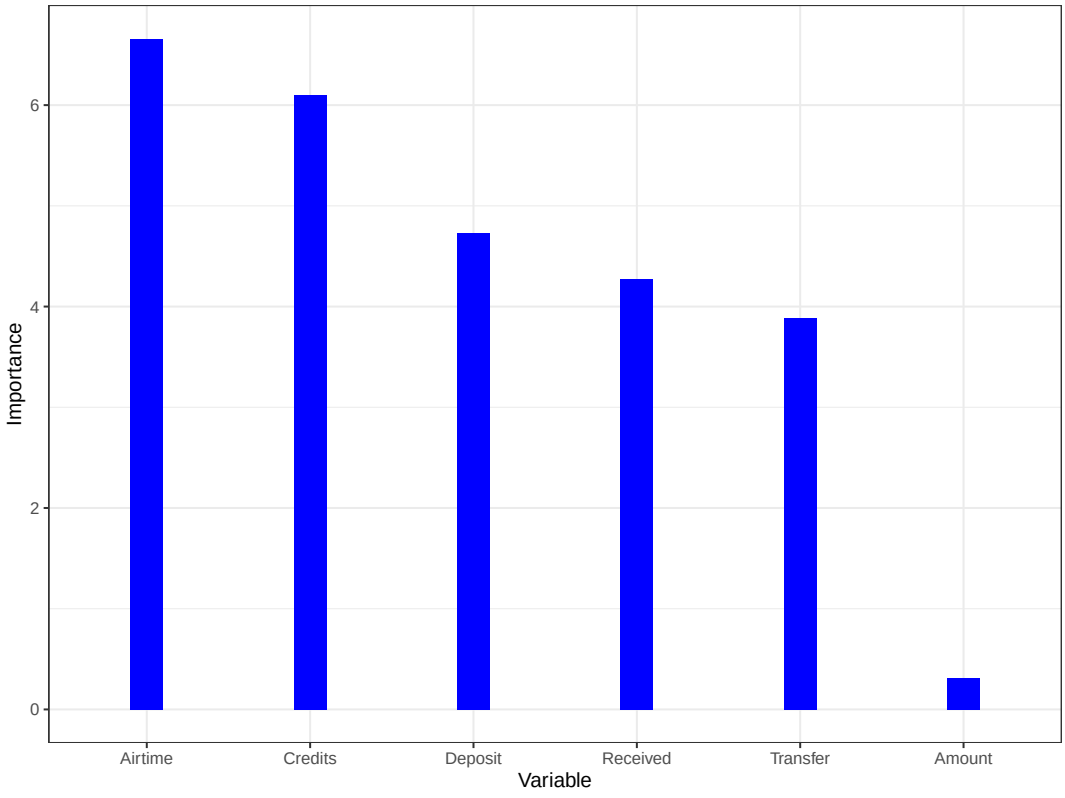

Figure 3. *Feature Importance.*

4. Model Comparison

This section gives the metrics used for model comparison. The idea was to determine which model performs best with our data, and as a first step, we considered each model's overall out-of-sample prediction accuracy, f1-score and recall for the model comparison. The results are shown in table 3. The modified XGB classifier has a relatively high accuracy implying that it can be effectively used to do credit scoring.

Table 3. *Test set comparison metrics.*

Model	Accuracy %	Precision	Recall
Random Forest Classifier	81.39	81.43	80.75
Modified XGB Classifier	80.76	80.75	80.18
XGB Classifier	77.79	77.61	77.25
KNN Classifier	63.02	62.43	62.14
SVM Classifier	57.14	62.92	52.16
Logistic Regression	54.40	51.99	51.42
Decision Tree Classifier	53.60	50.32	50.12

The ROC curves for the competing models are given in Figure 4.

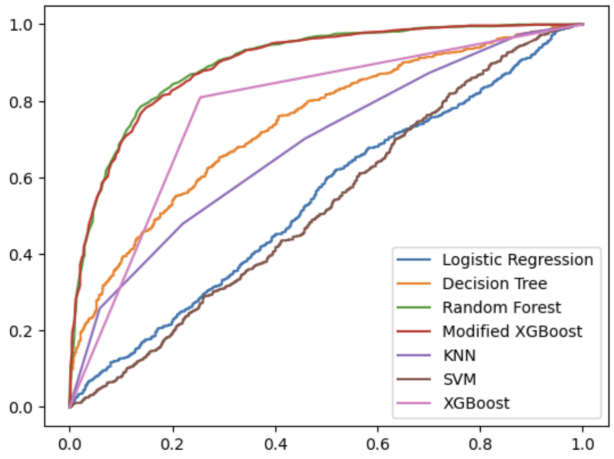

Figure 4. *ROC Curves for competing models.*

5. Conclusions

In this paper a new extreme gradient boosting model was developed that takes care of the inflation in a country. The effects of the inflation rates on the parameters is investigated and the inflation rates have a considerable effect on the hyperparameters of the model. By developing the insights, the modified XGBoost is able to solve real world scale problems using a minimal amount of resources. After this improvement it can handle the problem of credit scoring with imbalanced data in the presence of inflation rate, which helps reduce the losses of creditors and banks.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Kurylowicz, Lukasz. Short History of Banking and Prospects for Its Development in Poland at the Beginning of the 21st Century. *Available at SSRN 3051610*. 2004. http://dx.doi.org/10.2139/ssrn.3051610

[2] Greenlaw, D., Hatzius, J., Kashyap, A. K., & Shin, H. S.. *Leveraged losses: lessons from the mortgage market meltdown. In Proceedings of the US monetary policy forum* 2008; pp 7-59.

[3] Hodgson, G. M.. 1688 and all that: property rights, the Glorious Revolution and the rise of British capitalism. *Journal of Institutional Economics*. 2017, 13(1), 79-107. https://doi.org/10.1017/S1744137416000266

[4] F. S. D. Kenya, 2016 finaccess household survey,Financial Sector Deepening and Central Bank of Kenya. http://fsdkenya.org/publication/finaccess2016/Accessed *Leveraged losses: lessons from the mortgage market meltdown*. 2016; pp 2019.

[5] Izaguirre, J. C., Mazer, C. R., Graham, C. L., & Center, Digital credit market monitoring in Tanzania *Slide Deck*. 2018.

[6] Dushimimana, B., Wambui, Y., Lubega, T., & McSharry, Use of machine learning techniques to create a credit score model for airtime loans *Journal of Risk and Financial Management*. 2020, 13(8), 180. https://doi.org/10.3390/jrfm13080180.

[7] M. Whitney and H. Richter Lipford, Participatory sensing for community building, in CHI11 Extended *Abstracts on Human Factors in Computing Systems*. 2011, 1321-1326.

[8] Shema Effective credit scoring using limited mobile phone data. In *Proceedings of of the Tenth International Conference on Information and Communication Technologies and Development*, 2019; pp. 1-11.

[9] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I. & Zhou, T., Xgboost: extreme gradient boosting *R package version 0.4-2,*. 2015, 1(4), pp 1-4.

[10] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001, 1189-1232.

[11] L. Brieman, J. H. Friedman, R. A. Olshen, & C. J.Stone Classification and regression trees. wadsworth *Inc. Monterey, California*. 1984.

[12] T. Chen, H. Li, Q. Yang, and Y. Yu, General functional matrix factorization using gradient boosting. In *International Conference on Machine Learning*, 2013; pp. 436-444.

[13] D. Shen, G. Wu, & H.-I. Suk, Deep learning in medical image analysis, *Annual review of biomedical engineering*. 2017, 19, 221-248. https://doi.org/10.1146/annurev-bioeng-071516-044442

[14] Nguyen, Giang, Stefan Dlugolinsky, Martin Bobak, Viet Tran, Álvaro Lopez Garcia, Ignacio Heredia, Peter Malik, and Ladislav Hluchy, Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey, *Artificial Intelligence Review*. 2019, 52, 77-104.

[15] K.Wang, M. Li, J. Cheng, X. Zhou, and G. Li, Suk, Research on personal credit risk evaluation based on xgboost, *Procedia computer science*. 2022, 119, 1128-1135. https://doi.org/10.1016/j.procs.2022.01.143