2022

Relevance of alternative data and machine learning in predicting default in a non-deposit taking SACCO in Kenya.

Juma, Silas Okeyo
Strathmore Business School
Strathmore University

RELEVANCE OF ALTERNATIVE DATA AND MACHINE LEARNING IN PREDICTING DEFAULT IN A NON- DEPOSIT TAKING SACCO IN KENYA

SILAS OKEYO JUMA
ADMISSION NO. 072286

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMMERCE IN FORENSIC ACCOUNTING AT STRATHMORE UNIVERSITY

STRATHMORE BUSINESS SCHOOL
STRATHMORE UNIVERSITY
NAIROBI, KENYA

SEPTEMBER, 2022

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name....Juma Silas Okeyo….

Signature: 

Date……September 5th, 2022

Approval

The thesis of Juma Silas Okeyo was reviewed and approved for examination by the following:

Name of Supervisor: …Dr David Mathuva……

Faculty Affiliation: …. Strathmore Business School …

Institution: …… Strathmore University …….

Signed……  ……Date…September 5th, 2022…

ABSTRACT

Credit risk is the most important and difficult risk to manage in any financial institution. In Savings and Credit Co-operatives (SACCOs) particularly, credit risk is critical to the financial performance as it directly affects whether loans advanced will contribute to profits or losses. Traditional methods of credit scoring widely used like linear regression, discriminant analysis and judgement-based models have been proven to give mixed and unreliable results. This is majorly because they consider a small number of linear variables and experience of the credit officer which may also be subjective. The purpose of this research was to examine the relevance of non-traditional (alternative) data and Machine Learning (ML) algorithms in predicting default in a selected large SACCO in Kenya. Using micro-level secondary data of 783 loans extracted from the SACCO systems for a period of one-year (July 2018-June 2019), Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost) algorithms were implemented through experimental research design. The results, after hyperparameter tuning of algorithms, reveal that when traditional and alternative data on borrower behavior are used, both LR and XGBoost showed greater improvement in default prediction than when traditional data was solely used. For Logistic Regression, the Area Under Curve, Accuracy, Precision and Recall improved by 5%, 12.1%, 12.9% and 1.43% respectively while in XGBoost, improvements of 15%, 2.41% and 2.41% for Area Under Curve, Accuracy and Recall were noted. Precision scores remained unaffected in this model. Overall, XGBoost showed superior performance than LR. Further, the predictors of default are spread across traditional as well as alternative features, with alternative features seemingly improving predictive power of the ML models. The novelty of this approach lies in the combination of data previously considered irrelevant and ML algorithms that aim to reduce dimensionality in the data and increase accuracy in predicting future behavior of borrowers. Unlike prior studies, this study employed a pragmatic approach to simulate practical appraisal procedures for SACCOs in Kenya using scarce micro-level default data. However, financial data availability, legal and regulatory limitations on private data usage were the major challenges. Future studies may also consider other forms of alternative data like analysis of social media activities, unemployment data, average household incomes, mortgage uptake data, inflation rate, consumer price index among others.

Key Words: Machine Learning, alternative data, Algorithms, Logistic regression, Extreme Gradient Boosting, loan default

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

AI – Artificial Intelligence

AUC – Area Under Curve

CRB – Credit Reference Bureau

DAU – Daily Active Users

DL - Deep Learning

DNN – Deep Neural Network

DQN - Deep Q Network

ERP – Enterprise Resource Planning

GDP – Gross Domestic Product

ICA – International Co-operatives Association

IFRS – International Financial Reporting Standards

KNN - K - Nearest Neighbors

LDA - Linear Discriminant Analysis

LR – Logistic Regression

ML – Machine Learning

NACOSTI - National Council of Science and Technology Innovation

NB - Nave Bayes

NPLs – Nonperforming loans

PCA - Principal Component Analysis

PD – Probability of Default

ROC - Receiver Operating Characteristic

SACCO - Savings and Credit Co-Operatives

SARSA - State-Action-Reward-State-Action

SASRA - SACCO Societies Regulatory Authority

SNS – Social Networks Sites

SVM - Support Vector Machines

XGBoost - Extreme Gradient Boost

DEFINITION OF TERMS

1. **Algorithm -** A machine learning algorithm is the method by which the Artificial Intelligence system conducts its task, generally predicting and improving output values from given input data (Xia et al., 2021) .
2. **Alternative data -** Alternative data is information gathered from nonstandard or non-traditional data sources, for example, web traffic, social media posts, geolocation data, among others (Jagtiani & Lemieux, 2019). In this study, Alternative data include 53 behavioral data of members on the online member portal. This form of data was previously considered irrelevant for credit scoring.
3. **Artificial Intelligence** – any system that perceives its environment and takes actions that maximize its chance of achieving its goals (Barbaglia et al., 2020).
4. **Default -** Default in this study essentially is a delinquency stage of 90 days or more, PAR90. This is similar to definition by Basel Committee on Banking Supervision (Barbaglia et al., 2020).
5. **Logistic Regression -** one of the most popular supervised ML models for estimating the Probability of Default (PD), because it is easy to develop, validate, calibrate, and interpret (Bracke et al., 2019) .
6. **Machine Learning** – the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data (Bracke et al., 2019).
7. **Non-Deposit Taking SACCO -** SACCOs that take deposits from members only in the form of shares (Savings). These amounts are refundable to members only when they leave the SACCO. They are expected to comply by applying to the Authority within six months (SASRA, 2020).
8. **SACCO -** people-centred enterprises owned, controlled and run by and for their members to realize their common economic, social, and cultural needs and aspirations. Members share equal voting rights regardless of the amount of capital they put into the enterprise (Aderitus, 2020).
9. **Traditional data -** traditional data sources including loan amount applied, credit duration, age of the applicant, gender, loan product type among others (Jagtiani & Lemieux, 2019).
10. **XGBoost -** XGBoost is an improvement of the gradient boosting algorithm and a decision tree based on the gradient boosting algorithm (Xia et al., 2021).

ACKNOWLEDGMENTS

This research proposal could not have been complete without the invaluable guidance, patience, support and encouragement of my supervisor, Dr David Mathuva, from whom I have learnt a lot. I am humbled, honoured and privileged to have worked under his supervision.

Secondly, for the support and encouragement throughout this journey I am immensely grateful for my family and friends, may God almighty bless you abundantly.

DEDICATION

I dedicate this thesis research paper to my family for their continued support and love. To the memory of mum, I dedicate this work. Although she was my inspiration to pursue education, she was unable to see my graduation. This is for her.

CHAPTER ONE: INTRODUCTION

1.1 Introduction

This chapter introduces the background of the study which makes the context clear, followed by the statement of the problem, general and specific objectives of the study and research questions. The parts which conclude this chapter are the scope, research significance and organization of the study.

1.2 Background to the study

Credit risk management is the most important and difficult risk to manage in any financial institution due to loan default which has significant implications on its balance sheet and its stakeholders (Mashange et al., 2022). Banks and other microfinance institutions are however highly regulated with rigid lending rules and restrictions. This leaves SACCOs as the only intermediary for financial and social-economic development to many (Aderitus,2020). In SACCOs, loan default not only weakens the financial muscles in terms of liquidity and loss of rebates to its members, but may also lead to its total collapse as found out by (Maina et al., 2016; Salaton et al.,2020; Mashange et al., 2022). Therefore, SACCOs must be able to gauge the likelihood of a borrower defaulting before advancing the loan as this will result in either profits or losses. According to SASRA (2020) annual report on performance of SACCOs in Kenya, a marginal increase from 6.14% to 6.15% in Non-Performing Loans (NPLs) represented an increase of Ksh. 4.79 billion nationally between 2017 and 2019 exposing members to risks of irreversible financial losses. Therefore, the urgent need to find superior ways to improve credit risk assessment cannot be overemphasized.

Despite the importance of credit risk assessment in SACCOs, the process has been undertaken using largely traditional methods like linear regression, discriminant analysis and judgement-based models. The bulk of these methods being manual or devoid of important parameters. For instance, judgement-based models in SACCOs is based on parameters such as the quality of existing loans (where applicable), reliability of income source (whether salaried, self-employed or out of employment), internal credit history within the same SACCO (i.e. history of default internally), value of deposits held in the SACCO, value of guarantor deposits attributable to the current member loan, and in some cases, a credit score report from a credit reference bureau (for some SACCOs) among others. A major drawback of some of these parameters has been the mixed and often unreliable results when making the ultimate decision whether to lend or not (Turiel &

Aste, 2020). This is because the methods utilized consider a small number of linear variables and experience of the credit officer, who may deploy highly subjective assessments of the borrower's ability to repay.

Finding a superior technique to manage credit risk can have a huge impact on the SACCO's balance sheet and financial performance as well as its survival (Nyamasyo, 2018). Many studies have shown that modern techniques like use of Artificial Intelligence (AI), Machine Learning (ML) algorithms and alternative borrower data provides better credit score rating compared to traditional methods. For example, Jagtiani and Lemieux (2019) found that the use of non-traditional data like type of loan, payment history, length of credit history, the current level of indebtedness has allowed borrowers with fewer or inaccurate credit scores to access credit that is appropriately risk priced. The use of modern credit assessment techniques could also reduce the cost of making credit decisions and credit monitoring through automation hence lowering the overall operating costs of the organization with the potential to pass the benefits to customers (Jagtiani & Lemieux, 2019). Additionally, the accuracy of a ML model; neural network, compared to a traditional scoring method was found to be 82.1% compared to 37.9% in predicting loan default. The traditional models have also proved that only 1.82% of default loans will be default while the neural network estimates that 52.6% of default loans will be default (Ereiz, 2019a). Similar findings are demonstrated by Turiel & Aste (2020) who found that loan default can be predicted in an automated way with results above 85% rejection call for acceptance and above 75% default recall for loan default using a two-phase ML model. In this study, the impact of employing non-traditional data and two Machine Learning (ML) algorithms in predicting default in a SACCO setting is evaluated.

Although artificial intelligence (AI) and machine learning (ML) are not new concepts, SACCOs are hesitant to adopt these methods as part of their credit risk assessment process. This is majorly due to their complexity, hyper technology associated with big data, alternative data availability and computing power (Addo et al., 2018). Their predictions are also often difficult to explain and validate (Bracke et al., 2019). The Logistic Regression (LR) or logit model is one of the most popular ML algorithm for estimating the Probability of Default (PD) because it is easy to develop, validate, calibrate, and interpret (Bracke et al., 2019). This is in contrast to the more advanced ML algorithms like Random Forest (RF), Extreme Gradient Boosting (XGBoost) and self-training algorithms like Deep Neural Networks (DNN) which provide better prediction

performance but their results may be difficult to explain and validate (Xia et al., 2021). SACCOs therefore still use traditional credit scoring methods based on a linear calculation of a small number of indicators, this despite the wide range of benefits and opportunities that can be derived from the adoption of AI & ML (Ereiz, 2019). For example, direct financial benefits in lowered losses, better pricing strategies, lower loan loss provisions and better profit margins among others. Operationally, there can be reduced documentation burden to borrowers, better turnaround time, faster loan processing and enhanced customer experience rates (Bacham & Zhao, 2017).

It is also expected that AI and ML could also be used to detect early warning signs of financial distress from borrowers by continuously analysing their cash flow patterns, transactions categorization and loan loss provisioning according to guidelines of IFRS 9 hence increasing compliance as well (SAS, 2020). Another aspect that AI & ML can be employed by SACCOs is in the prediction of prepayment risk. Although not as bad as default risk, borrowers who repay loans before maturity can cause an important loss of profits to the SACCO. It is therefore important to have a model that can predict whether a customer is likely to pay back the loan before maturity or not (Zahi & Achchab, 2020). However, the reliability of AI & ML algorithms depends on many factors including quality of data, model stability and policy changes among others. Generally, custom-made models give better output results than generic models as posited by (Jagtiani & Lemieux, 2019; Ereiz, 2019a). This is majorly due to configuration of various optimization parameters under which ML algorithms operate. This process is known as hyperparameter tuning.

1.2.1 Adoption of technology by SACCOs in Kenya

The history of SACCOs in Kenya can be traced back to1908 when the first cooperative was formed by European farmers in Lumbwa near present Kericho for their crop production and marketing (Hesbon, 2011). Africans owned SACCOs were however not present until the 1930s and were characterized by slow growth due to lack of trust amongst locals, non-regulatory framework and lack of encouragement by the colonials (Hesbon, 2011). Co-operative societies have continued to transform their mandate over time. Initially, SACCOs focused on resource mobilization especially farm inputs, agro-processing and marketing.

Today SACCOs are categorized as either deposit-taking or non-deposit taking. Deposit-taking SACCOs require members to have a savings account and deposit money that can easily be withdrawn similar to what commercial banks do. Non-deposit taking SACCOs on the other hand require a member to buy shares into the SACCO and save money but cannot withdraw it unless they are leaving the SACCO. This has expanded the scope of SACCOs and enabled their evolution to include areas of real estate, financial services and manufacturing. As a result, SACCO subsector is positioned as an important player for financial intermediary, wealth creation and accumulation as well as job creation (Alex, 2019). SACCOs have also continued to compete aggressively in areas that were previously a preserve of commercial banks. For example, processing of salaries through check off systems, issuing and use of ATM cards, clearing of cheques, issuing of banker's cheques among others (Patroba et al., 2016).

To stay in competition, SACCOs are continuing to adopt technology to address the rapid and ever dynamic members' needs and habits to increase their satisfaction through mobile-banking, electronic-banking, VISA/SACCO link cards linking them to their SACCO accounts at any time and from anywhere. Indeed, several studies have been carried out in this regard. For example, Peter (2019) in his study sought to establish the effects of financial innovations on the financial performance of SACCOs in Kenya. The study recommended the adoption of technology to improve financial performance especially on the use of mobile banking and improved online presence. In another study, David (2020) studied the effects of mobile banking services on the financial performance of deposit-taking SACCOs in Kenya. Five independent variables were investigated of which three; mobile deposit services, mobile bill payment services and mobile statement services were found to have a significant influence on the financial performance of SACCOs. On the other hand, mobile transfer services and mobile balance enquiry services were found to have no statistically significant effect on the financial performance. Similar studies have also been carried out by (Lucy,2019; Timothy, 2015 and Nabavi, 2019). Although many studies have supported the use of technology in improving the performance of SACCOs, high costs, technological expertise and complex implementation matrix and information asymmetry have led to the slow adoption by SACCOs compared to commercial banks (Davis, 2015). This study, therefore, aims to contribute to the body of knowledge towards improving the adoption of technology by SACCOs in credit risk management using Machine Learning (ML). ML is a modern

method of data analysis and a part of artificial intelligence on the basis that an algorithm is able to learn from previous data, identify patterns or distributions of datasets, and make decisions without explicit human intervention (Wang et al., 2020).

1.3 Statement of the Problem

Prediction of loan default using alternative data and machine learning has been studied robustly in the context of commercial banks especially abroad. For example, Moradi and Rafiei (2019) studied credit risk assessment with ML in Iranian banks, Turiel and Aste (2020) studied peer to peer lending with ML, Petropoulos et al. (2018), Addo et al. (2018), Barbaglia et al. (2020), Zahi and Achchab (2020) have all studied credit risk prediction using AI & ML in banking institutions in foreign jurisdictions especially in the United States, Europe and Asia. More prominently, Jagtiani and Lemieux (2019) studied the role of alternative data and ML in predicting loan default in Fintech lending in the United States. SACCO operations are however very different compared to commercial banks and this exposes them to unique challenges in regards to loan defaults.

With regards to SACCOs, most studies have focused on other aspects not related to the prediction of loan default. For example, Mitei (2016) and Theophillus (2018) researched the determinants of loan default in SACCOs in Kenya, Karagu & Okibo (2014) and Mwangi & Ombui (2018) as well as Faith (2016) who focused on factors affecting the financial performance of SACCOs in Kenya. Similarly, researchers Mapunda (2019), Kengia (2015) and Oynaka (2020) focused on factors influencing financial performance in SACCOs in Tanzania and Ethiopia. There is a knowledge gap when it comes to the adoption of modern techniques like ML in predicting loan defaults in SACCOs in Kenya. Despite the numerous support for this approach, Machine learning algorithms have parameters associated with them that are not parameters of the models, hyperparameters. For example, the number of levels in the decision trees of a Random Forest (RF) algorithm. The values of hyperparameters in machine learning models are most often determined by trial-and-error methods or automated brute force search methods like grid search (Yang & Shami, 2020). The absence of an exact and efficient way of finding the optimum values of the hyperparameters makes designing machine learning a challenging task as a wrong choice of values of the hyperparameters can lead to imprecise models. It is on this basis that this study examined the role of alternative data and ML in predicting loan default in a sample private sector SACCO in Kenya. More specifically, the study examined the features contributing to ML application in predicting default.

1.4 Research Objectives

1.4.1 General objective

The study sought to determine the relevance of alternative borrower data & Machine Learning in predicting loan default in SACCOs.

1.4.2 Specific objectives

To address the general objective, the study sought to:

1. Determine the features contributing to Machine Learning application in predicting loan default in SACCOs.
2. Determine the optimal hyperparameters for different machine learning algorithms in predicting loan default in SACCOs.
3. Determine which Machine Learning algorithm best predicts loan default in SACCOs.

1.5 Research questions

1. Which important features contributed to Machine Learning application in predicting loan default in SACCOs?
2. Which hyperparameters optimized different Machine Learning algorithms in predicting loan default in SACCOs?
3. Which Machine Learning algorithm best predicted loan default in SACCOs?

1.6 Scope of the study

This case study is drawn from a large Non Deposit Taking SACCO in the private sector, which has exhibited the highest default rate as per (SASRA, 2020). The study focused on a sample SACCO so as to perform in-depth analyses on micro-level data drawn at member-level, so as to yield results at an individual member level, compared to a generalized analysis using macro-level borrower data. Secondary data from July 2018 to June 2019 was used to train and test using Logistic Regression and XGBoost Machine Learning algorithms in Python 3.9. The findings and recommendations of this study are expected to be generalizable although limited only to other SACCOs with similar characteristics in the industry. Python, like IBM's SPSS Modeler, R-software, Microsoft's SQL Server and other data analytics software packages implement decision tree algorithms. Decision tree algorithms can handle both numerical and categorical data making it simple to understand and interpret. However, tree models can be non-robust

implying that a small change in the training data can result in a large change in the tree and consequently the final predictions, this may affect the results of the study.

1.7 Significance of the Study

This study will benefit the following:

1.7.1 Policymakers

Policymakers will gain an understanding and appreciate the importance of AI & ML in managing credit risk in SACCOs. This may lead to advocacy for changes in the regulatory framework to a more technological oriented SACCO movement in the country. Improved credit-risk management leads to the improved financial performance of the SACCOs hence curbing the rampant collapse, losses and social-economic hazards that have been experienced in the recent past.

1.7.2 Other researchers and future scholars

The research will provide researchers and scholars with a reference point and valuable information for future studies in related areas of innovation, AI & ML, credit risk management and the development of customized credit scoring models. The research will also facilitate the identification of additional gaps that can be studied.

1.7.3 Practitioners in default risk management

SACCO management have continued to improve on loan appraisal procedures, timely reminders on loan repayments and timely escalating defaulters for follow-up to its debt recovery and legal teams. However, loan default has rapidly increased from 1.4% in 2018 to 9.3% in 2020 despite measures taken to mitigate the problem. This has necessitated an innovative approach in the management of default risk. The study will have a positive contribution in this regard as well as cost-saving and efficiency in the credit scoring process. It will also help different stakeholders including management staff and directors to embrace technology as an efficient way of reducing overall credit risk and improved customer satisfaction. Additionally, the study will help fill the gap in the use of IA & ML in credit risk management and compliance with the guidelines of loan-loss provisioning.

1.8 Organization of the thesis

This research thesis is made of five chapters. Chapter two presents the literature review where various concepts and theories supporting this study were elaborated and the empirical literature review analysed. The research gap and conceptual framework were also presented in this chapter. Chapter three describes the research methodology and

includes research philosophy, research design, data collection methods and tools, training and testing the models, hyperparameters tuning, data analysis, research quality in terms of reliability and validity of model results. This chapter was concluded by review of ethical issues and considerations of the study. In chapter four, data analysis and presentation of various statistical and ML model results were provided. Chapter five includes a summary of findings, conclusion as well as recommendations for future research and limitations of this study.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Literature review entails an overview of a wide range of information relating to the topic of study. This chapter focuses on various studies with theories relevant to the study and reviews empirical evidence relevant to the use of alternative borrower data and machine learning in predicting loan default in SACCOs.

2.2 Theoretical Review of Literature

The theoretical framework explains the path of research and grounds it firmly in a theoretical construct. The overall aim is to make research findings more meaningful, acceptable to the theoretical constructs in the research field and ensure generalizability (Adom et al., 2018). Theoretical framework consists of theoretical principles, constructs, concepts and tenants of a theory (Adom et al., 2018). Several theories and models relevant to the study exist, for example, the Technology Acceptance Model, Unified Theory of Acceptance and use of technology among others. In this study, however, DeLone & McLean theory and the Diffusion of Innovation Theory provide the most appropriate concepts and tenants to guide the research.

2.2.1 Information Systems Success Model - DeLone and McLean Theory

Developed by William H. DeLone & Ephraim R. McLean in 1992. This Information Systems Success Model is also known as Delone and McLean IS Success Model or simply (D&M IS Success Model). Since its first publication in 1992 and based on evaluation of the many contributions to it, the model was updated by DeLone & McLean (2003) and further to the present model Urbach & Müller (2010). The model seeks to provide a comprehensive understanding of the measure and evaluation of the Information Systems success by proposing a six-dimensional approach upon which IS success can be evaluated i.e. system quality, information quality, service quality, use, user satisfaction, and net benefits (Petter et al., 2008).

Other researchers are also in concurrence that the measure of an information system's success or failure is crucial in justifying its adoption or continued use in an organization (Muchangi et al., 2019). The success of a model is determined by the information system's technical quality and the output quality. While system quality focuses on the success of the information system at a technical level, output quality focuses on the success of the information in relaying the intended meaning (Muchangi

et al., 2019). It is on this basis that this theory is important to this study. The dimensional components earlier listed are discussed further below.

System quality consists of the desirable characteristics of the system itself in terms of usability aspects and performance characteristics, for example, the perceived ease of use (Urbach & Müller, 2010). Other measures of system quality include access Gable et al. (2008), convenience Iivari (2005), data accuracy Gable et al. (2008), ease of learning and customization Sedera et al. (2004), efficiency and flexibility Gable et al. (2008), integration, reliability Hamilton & Chervany (1981), system accuracy and turnaround time (Gable et al., 2008; Sedera et al., 2004).

Information quality constitutes desirable characteristics of Information system output thus focusing on the quality of information and its usefulness for the user. For example accuracy, adequacy, availability, completeness, conciseness, consistency, format, precision, relevance, scope, timeliness, understandability, puniness, understandability, usability (Gable et al., 2008; Hamilton & Chervany, 1981; Iivari, 2005; McKinney et al., 2002; Sedera et al., 2004).

Service quality refers to the quality of support the users get from the author in terms of training, assurance, interpersonal quality, responsiveness, empathy, flexibility (Chang & King, 2005). However, this dimension was not part of the original model but an improvement of it (Urbach & Müller, 2010).

Use. This attribute represents the manner and degree to which the information system is utilized by users. Objective measures will be instituted for example in tracking functions utilized and frequency of usage. Other measures useful for analysis include the intention to reuse, navigation patterns, number of transactions, and frequency (DeLone & McLean, 2003). However, perceived ease of use and perceived usefulness contributes to attitude to use, intention to use, and actual use as established by Davis (1989).

The user satisfaction attribute constitutes the level of utility the user derives from utilizing the system especially when the usage of the information system becomes mandatory. Some of the widely used user satisfaction indicators are the ones discussed by (Ives et al.,1983; Doll et al., 2004) and includes adequacy, effectiveness, efficiency, enjoyment, information satisfaction, overall satisfaction, system satisfaction.

Net benefit constitutes the extent to which an information system contributes to the success of different stakeholders. Initially separated as individual impact and organizational impact in the first model, the choice of what was to be measured in this study include decision effectiveness, awareness, individual productivity, job effectiveness, job performance, simplification, learning, task innovation, and usefulness as highlighted in previous studies by Davis (1989), Iivari (2005) and Gable et al. (2008).

2.2.2 Diffusion of Innovation theory

Diffusion of Innovation (DOI) theory was developed by E.M. Rodgers in 1962. It is now regarded as one of the most popular theories for studying the adoption of information technologies and understanding how these innovations gain momentum and spread through a specific population or social system over time (Robert, 2019). According to this theory, innovation is an idea, process, or technology that is perceived as new or unfamiliar to individuals within a particular area or social system (Zhang et al., 2015). Diffusion is the process by which the information about the innovation flows from one person to another overtime throughout the social system. The application of this theory hence helps to better understand the adoption of technology through AI & ML in detecting loan default in SACCOs. It also helps in designing better research methodology with attributes of innovation including the five user-perceived qualities; relative advantage, compatibility, complexity, trialability and observability in mind. Additionally, the application of DOI will help in the identification of obstacles that may impede the diffusion of usage of AI & ML in predicting loan default in SACCOs.

The four major determinants of the success of an Information Technology innovation are communication channels, the attributes of the innovation, the characteristics of the adopters, and the social system. Further, individuals of the social system have also been categorized into five groups depending on their attitude towards innovation, from innovators, early adopters, earlier majority, later majority and finally laggards who are the strongest resisters to the adoption of an innovation. As mentioned earlier prediction of loan default using alternative data & ML has been studied extensively in the context of commercial banks in more developed economies (Petropoulos et al., 2018 ; Jagtiani & Lemieux, 2019; Moradi & Rafiei, 2019; Moradi & Rafiei, 2019; Zahi & Achchab,

2020) among others. However, credit risk assessment in banks is very different from that of SACCOs owing to the unique lending model by SACCOs.

2.3 Empirical Review of Extant Literature

2.3.1 Role of Alternative data in predicting default

Alternative data is information gathered from nonstandard or non-traditional data sources, for example, web traffic, social media posts, geolocation data, point-of-sale transactions among others (Jagtiani & Lemieux, 2019). Social media users for example search, post and generate huge amounts of data leaving behind digital footprints which can be analysed to give insights on people's behavior, decisions and intentions and thus monitor key economic, social changes and trends (Blazquez & Domenech, 2018). Research studies in regards to social media have suggested that the number of posts and their frequency may lead to a better understanding of the expenditure patterns, lifestyle, and willingness of borrowers to repay debts (Blazquez & Domenech, 2018). Behavioral data such as frequency of web visits, previously used search term, pages viewed, amount of time spent on a page/website, individual actions performed, days since last visit etc. can be analysed. Useful insights about a customer's loyalty to an organization, participation in cross-selling strategies or intentions to borrow can then be drawn where the borrower has no credit history with the institution (Blazquez & Domenech, 2018).

Similarly, Jagtiani & Lemieux (2019) in their study sought to establish the role of alternative data and ML in Fintech lending. The study focused on determining whether the use of alternative data to build an internal credit scoring system can improve customers' ability to access credit and secondly whether the use of this data allowed Fintech lenders to better risk price their products than elsewhere where traditional credit scoring was implemented. Although alternative data used by different Fintech lenders varied from one another, some data analysed included bank account transactions e.g. rent payment, utility payments, education level, borrower's occupation, credit card transactions, insurance claims, investment choices, online shopping habits, etc. The study found that the use of alternative data allowed borrowers who could not have accessed credit to do so at cheaper costs. The model also performed well in predicting loan default. The use of alternative data, ML, and other complex AI algorithms could also reduce the cost of making credit decisions, credit monitoring, and overall reduction of operating costs (Jagtiani & Lemieux, 2019).

Mashange et al. (2022) utilizes Moody's credit rating model and a unique data set to estimate Markov chains in evaluating farmer cooperatives' credit quality. Whereas the study yields important insights on credit rating behavior of farmer cooperatives, it fails to examine the individual farmer credit behavior, which is the focus of the present study. Porath (2006) employs the area under Receiver Operating Curve (ROC) and information value to assess the probability of default of savings banks and credit cooperatives in Germany. Interestingly, it appears that the probability of default is driven by general macroeconomic factors together with the banks return, credit risk, market risk and capitalization. Overall, existing studies on default in cooperatives focus on the individual cooperative organizations, with extremely sparse studies focusing on individual borrowers of these cooperatives, which forms the focus of this study.

2.3.2 Machine Learning in credit risk management

In terms of the models used, Turiel & Aste (2020) demonstrated that peer-to-peer loan acceptance and default can be predicted in an automated way using Logistic Regression (LR). Other models tested included Support Vector (SV) ML algorithms together with linear and nonlinear Deep Neural Networks (DNN). Turiel & Aste (2020) developed a two-phased loan acceptance/rejection and default prediction model for approved loans. The study concluded that LR was the best performer in the first phase while DNN was the best performer in default prediction. Overall results were above 85% rejection call for loan acceptance and above 75% default call for loan default. This shows that ML can be used to improve credit risk models by reducing the default risk by as much as 70% without the need for human credit intervention. Their study sought to predict loan default using three different ML models i.e. decision forest, logistic regression model, and neural network model using a test data set of 30,065 loans. Results showed that both decision forest and logistic regression models failed to identify 5.35% of default loans while neural networks failed to identify 2.63% of default loans. At the same time, recall for default loan was 52.55% (Ratio of true positives over the total number of positives) with the precision of 15.52% (ratio of true positive) and accuracy of 82.1% (ratio of the sum of true positives and true negatives over the total number of predictions). Xia et al. (2021) also found that the Gradient Boosting model was superior together with the optimal predictive models. This reiterates the significant benefits that can be derived in using ML in predicting loan

defaults. However, the high rate of accuracy in results prediction depends on the quality of data.

Ereiz (2019) in his study, Predicting default loans using machine learning established that credit scoring using traditional methods are means to mixed and unreliable results. ML on the other hand provides a broader perspective of viewing the borrowers not only as a means to manage credit risk but also traverses to other business risks as well, for example, prepayment risk, financial fraud, money laundering, and risk of not complying with regulations. The same has been echoed by Zahi & Achchab (2020), Bracke et al. (2019) and Barbaglia et al. (2020).

On the other hand however, Kiefer & Mayock (2020) investigated why models built to predict failure fail. By analysing millions of mortgage accounts from 2014 to 2016, the study established that linear regression models and other AI & ML technologies widely used by lending institutions especially Fintechs could be very inaccurate when predicting loan performance in out-of-time samples (samples from an entirely different period than what was used to develop the model). From the study, model failure could majorly be attributed to intertemporal heterogeneity in the relationship between variables that are frequently used to predict performance and hence model instability is a significant source of risk to lending institutions. Another key finding of the study indicated that predictive accuracy deteriorated rapidly when models are trained in one macroeconomic environment but used to predict default in another, this can be reduced by using data from a wide range of economic conditions to smoothen the over volatility of parameter estimates. This parameter stability bounds the forecast performance away from over-predictions and under-predictions.

2.3.3 Machine Learning models

Zahi & Achchab (2020) define machine learning as a type of Artificial Intelligence that allows computers to find patterns within data and construct classification and prediction models without explicitly being programmed to do so. Generally, ML models can broadly be classified as Supervised and Unsupervised Machine Learning algorithms based on whether they are built to model labelled or unlabelled datasets (Yang & Shami, 2020). However, other scholars and researchers including Walusala et al. (2017), Aderitus (2020) and Sabato (2010) have suggested a third model combining concepts of both supervised and unsupervised models known as Hybrid or

Reinforced Machine Learning model in recent past. Details of each are as discussed in the following section.

2.3.3.1 Supervised Machine Learning

The main characteristic of supervised machine learning algorithm is that it consists of a target or outcome variable making it suitable for predictions (Walusala et al., 2017). The algorithm is developed using dataset that contain labelled dependent variable and independent variables, also known as features. The algorithm is then used to predicts future outcomes of dependent feature Sabato (2010), for example, a dataset of loans may be labelled as to identify defaults and accounts not in default. The algorithm will learn a general rule of classification that it will use to predict outcomes for other observations in the dataset. This makes supervised ML the most suitable model to use in this study. Examples of supervised ML include linear models, K - Nearest Neighbors (KNN), Support Vector Machines (SVM), Nave Bayes (NB), Decision-Tree-Based Models, and Deep Learning (DL) algorithms.

2.3.3.2 Unsupervised Machine Learning

In Unsupervised machine learning, there is no specific target or outcome variable to help predict or develop an estimate Walusala et al. (2017). This makes the algorithm suitable for use in clustering or segmenting features into various categories for specific intervention (Yang & Shami, 2020). Clustering methods mainly include K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), hierarchical clustering, and expectation maximization (EM); while two common segmentation algorithms are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

2.3.3.3 Reinforced Machine Learning

Reinforcement or Hybrid Machine Learning adopts concepts of both supervised and unsupervised machine learning. It allows the machine to learn from interaction with the environment. This allows changes in machine behavior and actions it takes depending on feedback from the environment. The goal is to train an algorithm that considers the ideal behavior that maximizes reward which can be learned once or through experience over time. For example in robotics, intelligent traffic lights, game theory and Internet of Things (IoT) devices (Yang & Shami, 2020; Aderitus, 2020).

Example of reinforced machine learning algorithms include Deep Q Network (DQN), Q-Learning and State-Action-Reward-State-Action (SARSA).

2.3.4 Concerns in the use of AI & ML in predicting loan defaults

Notwithstanding the benefits of using modern methods and big financial datasets for credit scoring, concerns about consumer data privacy and data protection violations, cyber-security, discrimination against minority groups, unfairness, exploitation by aggressive marketing, interpretability of the models and likelihood of unintended consequences are concerns because the models developed on historical data may learn and perpetuate historical bias (Jagtiani & Lemieux, 2019). Privacy and security have been cited as the major sources of dissatisfaction and non-adoption of these models (Robert, 2019). That said, there are also risks to borrowers and businesses from a lack of innovation in credit scoring as it may hinder improvements in financial inclusion and risk assessments (Ereiz, 2019a).

2.4 Summary of the Literature and Research Gap

Empirical literature shows Machine Learning provides better prediction in credit management than traditional methods. However, results conflict regarding the choice of different ML models in credit risk management with the majority suggesting customized tree-based models being more stable and accurate in prediction. In particular, there exist a gap in the use of alternative data and ML in predicting loan defaulters in SACCOs.

Empirical evidence has also shown that loan default poses a significant credit risk to any lending institution. In the SACCO subsector, default has continued to grow rapidly posing a threat to their sustainability and growth. Although credit management practices have rapidly moved from traditional methods to AI & ML in the banking sector in recent years, this trend has not received much research attention locally especially within the SACCO subsector. Internationally, AI & ML research has been conducted to predict not only loan default by different lending institutions but also help predict other business risks like prepayment risks, financial frauds, and bankruptcy. The researches show superior results were achieved in the use of AI & ML compared to traditional methods. These results are concurrent with observations made by both supporters and critics of AI & ML. However, the accuracy of results depended on the quality of data, size, and stability of the model being tested. This

study evaluated the suggestions of previous researchers on methods that produced the best prediction results for optimized observation. The gaps identified have been summarized in Table 2.1

Table 2. 1 Literature Matrix

	Articles Title	Author(s), Jurisdiction, Year	Findings	Research Gaps
1	Predicting Default Loans Using Machine Learning (OptiML)	Zoran Ereiz - Bosnia, 2019	Generic models not as good as customized models in default prediction. Microcredit institutions shy from using AI & ML despite being superior	Jurisdiction. Microcredit Organization, Traditional data used
2	Modelling car loan prepayment using supervised machine learning	Sara Zahi, Boujemâa Achchab - Morocco, 2020	The model has classified 83% of the data correctly (precision), that the classifier has an 85% accuracy rate and the positive rate of the model is 84%. Prepayment risk is less studied but can cause substantial losses	Jurisdiction. Focused on Prepayment prediction in Banking Industry. Traditional data used
3	The roles of alternative data and machine learning in Fintechs lending: Evidence from the Lending Club consumer platform	Julapa Jagtiani, Catharine Lemieux - United States, 2019	Alternative data leads to better credit scoring, default prediction. A shift of bad borrowers in traditional data to good ones	Jurisdiction. Fintechs & Banking industry
4	Credit Risk Analysis using Machine and Deep learning models	Peter Martey Addo, Dominique Guegan, Bertrand Hassani - Italy, 2018	Tree-based models are more stable than models based on multilayer artificial neural networks. Choice of features & choice of algorithms is critical	Jurisdiction. Banking Industry, alternative data used
5	Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modelling	Bacham Dinesh Zhao, Dr Janet Yinqing - the United States, 2017	Machine Learning offer better opportunities, challenges can be managed and regulated e.g. in data privacy	General
6	A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks	Somayeh Moradi and Farimah Mokhatab Rafiei - Iran, 2019	Dynamic models that assess borrower behavior monthly outperformed other ML models. Customers follow predictable patterns in times of economic crisis	Jurisdiction. Banking Industry
7	Peer-to-peer loan acceptance and default prediction with artificial intelligence	J. D. Turiel and T. Aste - the United Kingdom, 2020	Two phased models give better prediction results with LR in phase 1 and DNN in phase 2. AI can reduce default risk by up to 70%. Using DNN may lead to overfitting/overlook major flaws	Jurisdiction. P2P Online matching of lenders to borrowers. Loans used in the study were from small businesses only.

	Articles Title	Author(s), Jurisdiction, Year	Findings	Research Gaps
8	Forecasting Loan Default in Europe with Machine Learning	Luca Barbagliaa, Sebastiano Manzana and Elisa Tosetti - Italy, 2020	Geographical heterogeneity in the variable is important, need for regionally tailored risk assessment and policies in Europe. Boosting models perform significantly better in providing predictions than LR	Jurisdiction. Banking Industry, Traditional data used
9	Machine Learning Explainability in Finance: An Application to Default Risk Analysis	Philippe Bracke, Anupam Datta, Carsten Jung and Shayak Sen - the United Kingdom, 2019	Notable model uncertainties do remain which stakeholders ought to be aware of. GTB classifier over those other classifiers based on its superior predictive performance	Jurisdiction. Banking Industry
10	Big Data sources and methods for social and economic analyses	Desamparados Blazquez, Josep Domenech - Spain, 2018	Decision trees outperform other models of ML	General
11	Why Do Models That Predict Failure Fail?	Hua Kiefery, Tom Mayockz - the United States, 2020	Model instability is a significant source of risk for lenders, such as Fintechs. Understanding model risks is paramount to ensure safe & sound lending practices. predictive accuracy deteriorates rapidly when models that are trained in one type of macroeconomic environment are used to predict loan performance in out-of-time samples characterized by many deferent economic conditions	Jurisdiction. Banking Industry
12	SACCOs credit rating prediction in Tanzania by using machine learning approach: A case of KKKT Arusha Road SACCOs Ltd.	Ngimbwa Aderitus - Tanzania, 2020	Some factors have high weights in influencing SACCOS members' credit rating than others. By using RF the metrics score increased as the number of features increase with the best at 11 variables	Jurisdiction. Traditional data set used
13	A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression	Sylvester Walusala, Dr Richard Rimirub, Dr Calvin Otieno - Kenya, 2017	Hybrid models give better credit score predictions than simple LR models in SACCOs	Jurisdiction. The research used traditional data.
14	The Relationship Between Loan Default And The Financial Performance Of SACCOs In Kenya	Nancy Jemoek Keitany - Kenya, 2013	There is a strong negative relationship between the loan default and the profitability of these SACCOS	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML

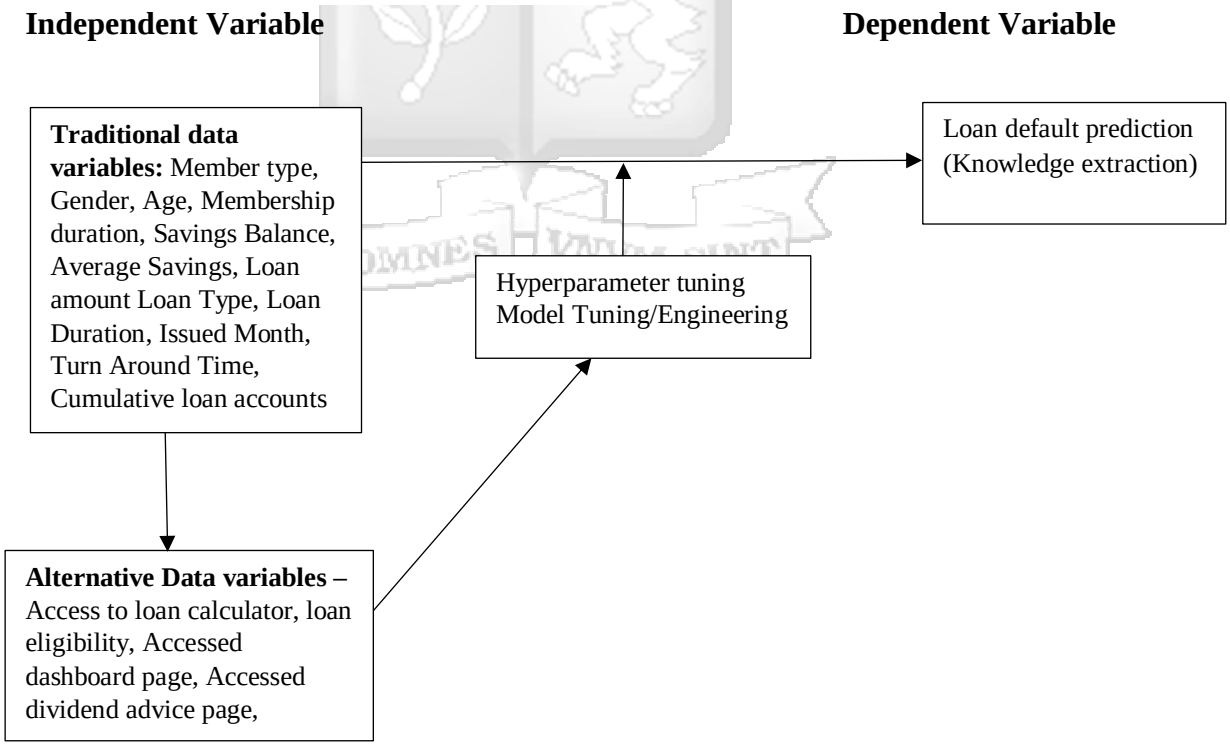
	Articles Title	Author(s), Jurisdiction, Year	Findings	Research Gaps
15	Adoption And Integration Of Information And Communication Technology, And Performance Of Deposit Taking SACCO'S In Nairobi City County	Wachira Davis Thanu - Kenya, 2015	There was a strong positive relationship between adoption and performance, while there was a positive but weak relationship between integration and performance,	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML
16	Determinants of Loan Defaults in SACCOs in Kenya: a Case of Metropolitan National SACCO Ltd	Nyamasyo, Kimatu Theophillus - Kenya, 2018	Loan default is significant of the rise due to factors e.g. increased transaction costs and low recovery rates.	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML
17	Influence Of Information Technology In Enhancement Of Sustainable Competitive Advantage Of SACCOs In Kisii County	Momanyi Mochere Patroba1 Kepha Osoro Dr. Michael Nyagol,Dr. Fredrick Odoyo - Kenya, 2016	Information technology has a positive impact on the image, goodwill and growth of SACCOs in Kisii County. IT has helped reduce other risks in SACCOs, growth in membership	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML
18	Effect Of Financial Innovations On The Financial Performance Of SACCOs In Kenya	Sang Peter Kiplimo - Kenya, 2019	Adoption of financial innovation enhanced the financial performance of the SACCO	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML
19	Determinants of Loan Default by Savings and Credit Co-Operative Societies' Members in Baringo County, Kenya	Allan Mitei, Mary Bosire, Robert Kibet Kirui - Kenya, 2016	Economic and terms of the loan factors significantly affect loan default however social factors does not significantly affect the loan default. social factors have a weak relationship with loan default. economic factor has a moderate positive significant association with loan default	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML
20	Performance Analysis of Debit Card Services on Deposit-Taking SACCOs' Financial Performance: A Case of Kenya	David Muchangi Mugo, Dr Stephen Muathe, Dr Stephen Titus Waithaka - Kenya, 2019	debit card services (cash withdrawal services, deposit services, account statements, bill payments services, and balance enquiry services) had a statistically significant positive effect on the financial performance	Although it is in the Kenyan context, Research not focused on predicting loan default with or without AI & ML

2.5 Conceptual framework

According to Adom et al. (2018), a conceptual framework is a structure that guides a researcher in the whole process of investigating the problem being studied. It shows how different ideas in the study relate to each other. It also clarifies the problem to be investigated, making it easier to analyse concepts to be included in the study.

Figure 2.1 below represents the conceptual framework used to guide this study. It shows the independent variables categorized into two broad streams; firstly, traditional data sources including loan amount applied, credit duration, age of the applicant, gender, loan product type among others. Secondly, Alternative data including the behavioral data of members on the online member portal. After determination of appropriate predictor variables for predicting loan default, data was collected from the ERP System. Data pre-processing then followed to have appropriate datasets for ML algorithm in analysing appropriate factors for SACCOs loan default. Finally, knowledge extraction from the algorithm output after classification of loans as either default or non-default.

Figure 2.51 Conceptual Framework



Source: Researcher, (2021)

Table 2. 2 Operationalization of Variables

	Traditional Data						
	Variable	**Description**	**Short code**	**Measurement**	**Interpretation**	**Literature**	**Theories**
1	Member type	Principal	mbrtyp_1	mbrtyp_1=1	Professionals of the field registered by the professional board		DeLone & McLean theory
		Introduced	mbrtyp_2	mbrtyp_2=0	Spouses, family, employees and anyone introduced to the SACCO by the principal member		
2	Gender	Male	sex_1	sex_1=1		(Aderitus, 2020), (Ereiz, 2019b), (Mitei, 2016)	DeLone & McLean theory
		Female	sex_2	sex_2=0			
3	Age	Age of Applicant at time of application (18-80)	Numerical	Number of years	Difference between when a member was born and time of application	(Aderitus, 2020), (Ereiz, 2019b), (Zahi & Achchab, 2020), (Mitei, 2016)	DeLone & McLean theory, asymmetric information theory
4	Membership duration	How long a member has been with the SACCO	Numerical	Number of months	Difference between when a member joined the SACCO and time of applying for the loan	(Aderitus, 2020)	DeLone & McLean theory, asymmetric information theory
5	Savings Balance	Amount of Savings at application date	Numerical	Kenya Shillings	Amount of deposits/savings at the time of loan application	(Ereiz, 2019b)	DeLone & McLean theory, asymmetric information theory
6	Average Savings	Average monthly savings since joining the SACCO	Numerical	Kenya Shillings	Quotient of Savings balance and membership duration i.e. Variable 5 divided by variable 4		
7	Loan amount	Amount of loan requested	Numerical	Kenya Shillings	Amount of loan approved	(Aderitus, 2020), (Ereiz, 2019b), (Zahi & Achchab, 2020),(Mitei, 2016)	DeLone & McLean theory, asymmetric information theory
8	Loan Type	Type of loan applied				(Aderitus, 2020), (Ereiz, 2019b), (Zahi & Achchab, 2020)	DeLone & McLean theory
		Development Loan	lnty_1	lnty_1=1,else 0			
		Emergency Loan	lnty_2	lnty_2=1,else 0			
		Intern Loan	lnty_5	lnty_5=1,else 0			

	Variable	Description	Short code	Measurement	Interpretation	Literature	Theories
		Medifinance Loan	lnty_7	lnty_7=1,else 0			Asymmetric information theory
		Insurance Loan	lnty_8	lnty_8=1,else 0			
		Equity Release Loan	lnty_9	lnty_9=1,else 0			
		Asset Finance Loan	lnty_10	lnty_10=1,else 0			
		School Fees Loan	lnty_11	lnty_11=1,else 0			
		Midterm Swift Loan	lnty_12	lnty_12=1,else 0			
		Flexi Development Loan	lnty_13	lnty_13=1,else 0			
		Staff Mortgage Loan	lnty_14	lnty_14=1,else 0			
9	Loan Duration	Maximum period for loan repayment	Numerical	Number of months	Loan tenure	(Aderitus, 2020), (Ereiz, 2019b), (Zahi & Achchab, 2020), (Mitei, 2016)	DeLone & McLean theory
10	Issued Month	Month a loan is disbursed			Months of the year	(Aderitus, 2020)	DeLone & McLean theory, asymmetric information theory
		January	ismth_1	ismth_1=1,else 0			
		February	ismth_2	ismth_2=1,else 0			
		March	ismth_3	ismth_3=1,else 0			
		July	ismth_7	ismth_7=1,else 0			
		August	ismth_8	ismth_8=1,else 0			
		September	ismth_9	ismth_9=1,else 0			
		October	ismth_10	ismth_10=1,else 0			
		November	ismth_11	ismth_11=1,else 0			
		December	ismth_12	ismth_12=1,else 0			
11	Turn Around Time	Time in Days between loan application and disbursement	Numerical	Number of days	Difference in days between when loan application is received and date the loan is disbursed	(Aderitus, 2020), (Mitei, 2016)	DeLone & McLean theory, asymmetric information theory

	Variable	Description	Short code	Measurement	Interpretation	Literature	Theories
12	Cumulative loan accounts	The total number of all loans ever issued	Numerical	Count	Total number of previous loans a borrower have taken with the SACCO since joining	(Aderitus, 2020)	DeLone & McLean theory
13	Active loan accounts	The number of active loan accounts at time of application	Numerical	Count	Number of active loans a borrower is servicing at the time of new application		DeLone & McLean theory
14	Loan repayment method	Repayment Standing Order, Direct debit, check off	Categorical		Method of loan recovery/ How the loan is being repaid	(Ereiz, 2019b)	DeLone & McLean theory
		Checkoff system	recmth_1	recmth_1=1,else 0			
		External Standing Order	recmth_2	recmth_2=1,else 0			
		Dividends & Interest	recmth_3	recmth_3=1,else 0			
		Direct Debit	recmth_4	recmth_4=1,else 0			
		Post-dated cheques	recmth_5	recmth_5=1,else 0			
15	Loan Interest Rate	Annual Interest Rate of Loan taken			Interest charged p.a	(Aderitus, 2020), (Zahi & Achchab, 2020), (Mitei, 2016)	DeLone & McLean theory
		12%	int_1	Interest rate %			
		13.5%	int_2	Interest rate %			
		14%	int_3	Interest rate %			
16	Repayment amount	Amount of monthly repayment (Principal + Interest)	Numerical		Amount a borrower pays back monthly as principal plus interest	(Aderitus, 2020)	DeLone & McLean theory

	Alternative Data: 53 footprints of whether a member;					
	Variable	**Short code**	**Measurement**	**Interpretation**	**Literature**	**Theories**
	Accessed dashboard page, Accessed dividend advice page, Accessed AGM documents, Accessed FAQS page, Accessed latest news page, Accessed Loan Application forms page, Accessed membership forms page, Accessed Nominations page, Accessed notices page, Accessed other downloads page, Accessed policies page, Accessed Strategic plan page, Accessed feedback page, Sent feedback, Accessed loans calculator page, Calculated for loan, Accessed guaranteed loans page, Accessed Eligibility page, Accessed loans page, Accessed loans status trail page, Accessed guarantorship requests page, Accessed online loans list page, Accessed online loans page, Accessed outstanding loans page, Accessed Account balances page, Accessed deposit adjustment page, Login Failed, Accessed login page, Login Successful, Account Locked, Accessed password change page, Changed password page, Accessed password recovery page, Accessed password Reset page, Accessed profile information page, Accessed profile update page, Accessed account statement page, Accessed Dividends Statement page, Accessed loans guaranteed page, Accessed loans guarantors page, Accessed loans statement page, Downloaded loans statement, Accessed old statement page, Accessed statement page	Numerical	Count - number of times (frequency)	Borrower's online footprints were analysed from the SACCO's ERP on their interaction with the online portal and categorized as variables. A total of 53 possible movements can be traced distinctively to monitor behavior	(Jagtiani & Lemieux, 2019) (Ereiz, 2019b), (Addo et al., 2018), (Blazquez & Domenech, 2018)	Information Systems Success theory Diffusion of Innovation theory Evolutionary theory

Source: Researcher, (2021)

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research design, target population, sampling techniques, data collection methods, and data analysis procedures. It also encompasses reliability, validity, and ethical considerations for the study.

3.2 Research Philosophy

Pragmatic paradigm is an ideal philosophy for this study. Scientific processes were followed in formulation of various experimental conditions, feature selection as well as tuning the ML algorithms. Pragmatism is a research philosophy based on the epistemology that embraces plurality of methods and accepts that there can be single or multiple realities that are open to empirical inquiry (Kaushik & Walsh, 2019). Knowledge of the multiple realities is therefore gained through an integration of multiple research methods. Through this integration, researchers gain better understanding of the problem under study from the views of people, scientific modelling or from testing of facts and figures. This mixed method approach enhanced a more detailed understanding of research questions and results leading to a balanced conclusion on the challenges and opportunities about the use of alternative data and ML in predicting loan default.

3.3 Research Design

According to Akhtar (2016), a research design is the glue that holds all of the elements in a research project together. This study adopted an experimental research design using quantitative approach. It constitutes an outline for the collection, measurement, and analysis of data. This design was selected to allow comparison, description and inferences of the findings. The goal was to assess the relevance of alternative borrower data and Machine Learning in predicting loan default in SACCOs in contrast to using traditional data solely. Secondary data was sourced from the organization's Enterprise Resource Planning (ERP) system and the members' online portal and analysed using Python 3.9 tools. Non-numeric variables for example month a loan is issued, gender, loan type, recovery mode among others were transformed to numeric data and encoded using an automated process of One-Hot Encoding technique in the Python 3.9 library. The appropriateness of this study design was therefore guided by the need to establish the usefulness of alternative data and ML in credit risk management in SACCOs.

3.4 Population and sampling

A target population is a group of individuals, objects, or items from which samples can be taken for measurement (Mapunda, 2019). Data in this study was drawn from a sample SACCO in the private sector, primarily targeting professionals in the medical field, with a rather closed field of membership. The sampled SACCO has an asset base of Ksh. 4.6 billion as of 2020, and a loan book of Ksh 3.0 billion. The SACCO has a membership base of 5,600. As of 2020, the default rate in the SACCO had increased from 1.4 percent to 9.3 percent. The micro-level member data was extracted from the SACCOs enterprise resource planning system for the mainstream data and the members' online portal for the alternative data for the period between July 2018 and June 2019. Consent and ethical approvals were sought from the SACCO's officials before extraction and analysis of the data. The dataset targeted 1,020 loans for the period between July 2018 and June 2019. The final analysis is based on a 783 loan-sample which represents 76.8 percent of the loans disbursed in the sample period. 236 loans were still active as of the date of analyses. Data was analysed using Python 3.9 tools.

3.5 Methods and Instruments of Data Collection

A research instrument is a tool used to collect data. The instruments employed in this study enabled the collection of both descriptive and numerical data. Collection of secondary data was through extraction from the society's ERP to Microsoft excel before loading onto Python 3.9 library for further analysis. All these methods were applied for collecting detailed and accurate data that assured valid and reliable inferences about the population under study.

3.6 Data Analysis

3.6.1 Definition of default loans

The Basel Committee on Banking Supervision defines default essentially as a delinquency stage of 90 days or more, PAR90 (Vidal & Barbon, 2019). This definition is similar to that adopted by SASRA new regulations 2020 and this study. It was important to decide on the best definition for the study because it directly affects who the model classified as a defaulter or not. All loans disbursed accounts in the dataset were classified in mutually exclusive categories relating to the performance of the loan. Bad loans are the loans that the organization would not have disbursed while good loans are those that it would be happy to repeat.

3.6.2 Data pre-processing

During loan application, borrowers fill in loan forms which provide varied datasets about them. Some of the data is important for credit management while some not. With the assistance of an employee from the credit department, a table was created containing records deemed important for credit risk management with attributes characterizing the loan itself and the borrower, for example, type of loan, interest rate, loan amount, average savings at time of application among other attributes. This information was available either before, during or after a loan account is closed in one way or another. Alternative data from the SACCO online member portal was analysed separately and its attributes linked to the individual member's accounts in the master datasheet. All identifiable member details e.g. names were then removed from the dataset to protect the identities of the borrowers. To comply with the provisions of Data Protection Act 2019, borrower's unique identification numbers were serialized by a factor to remain identifiable to the researcher but not any other third parties.

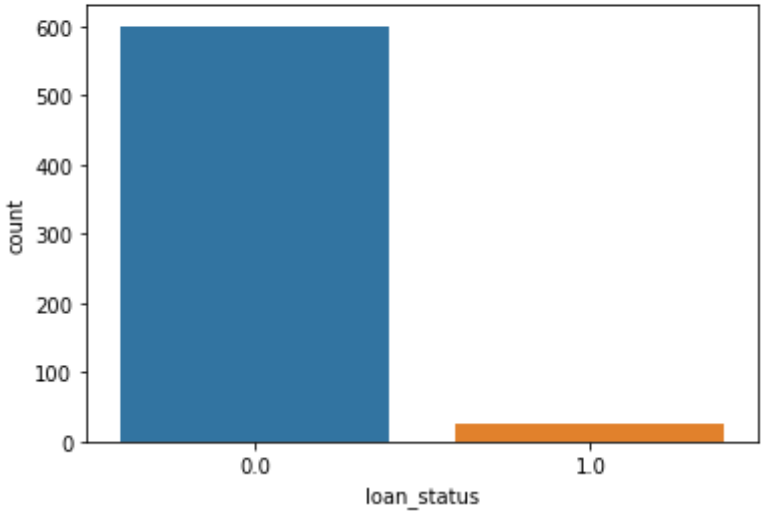
3.6.3 Data Cleaning

The dependent variable in the dataset is a binary indicator with the value of 1 flagging a defaulted loan. Loan status variable explains whether the loan has been fully paid, is current or has been defaulted. This study aimed to predict whether borrowers paid their loans without defaulting or being written off in one way or another hence data under 'Current' category was ignored as these are ongoing loans. Consideration was only on closed cases. Default statuses were then coded as 1, while Fully Paid coded as 0.

From preliminary analysis of the dependent variable, Figure 3.1, shows there is a significant imbalance between the number of defaults and the number of clients who paid on time. This is useful in determining the model evaluation criteria, in this case, use of accuracy score measure can be avoided since the result will not be credible. Instead, the ROC-AUC measure is used as it is better at handling data imbalances (Turiel & Aste, 2020). This problem of class imbalance was mitigated through regularization as well as by balancing the weights at the time of training of the model itself.

Figure 3. 1 Dependent Variable

```
0.0    0.958466
1.0    0.041534
Name: loan_status, dtype: float64
```



Source: Researcher, (2021)

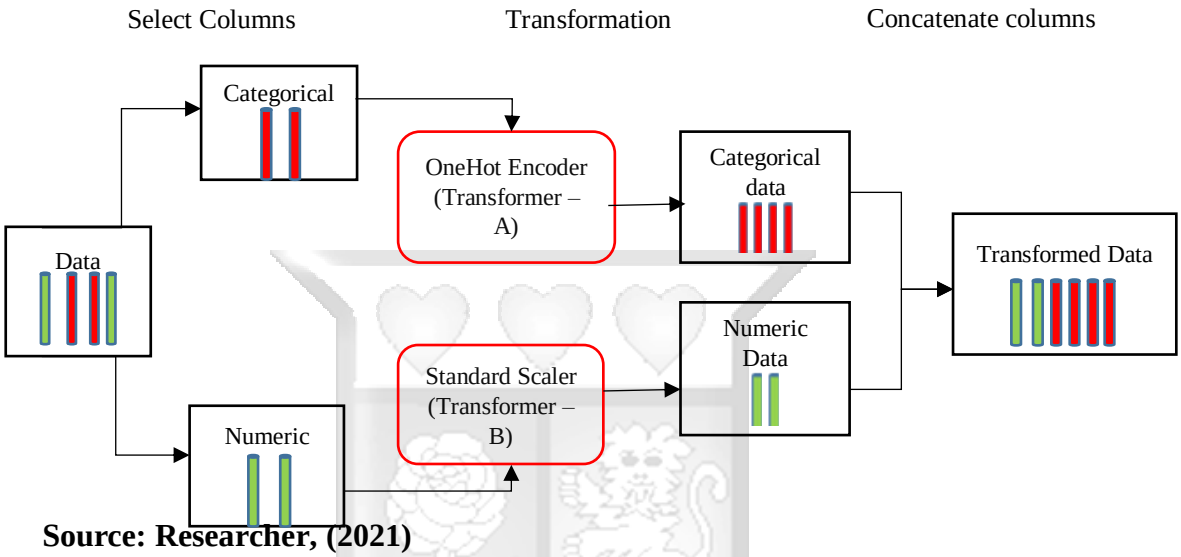
Information which could only be known after approval of the loans were deemed irrelevant or unavailable in predicting default and hence removed from the dataset. Inclusion of these features could lead to an overfitted model resulting in low performance outcome. Based on this argument, 'outstanding interest', 'outstanding balance', 'approved amount', 'top-up amount' are deemed irrelevant and dropped from the dataset. Other irrelevant features dropped include date of birth as 'Age' data is available, similarly, date of registration is dropped as membership duration data could easily be established.

3.6.4 Processing pipeline

The dataset was composed of both numeric and categorical variables. However, for ML algorithms to work properly, categorical data must be transformed to numeric form. The next step therefore involved separation and encoding of categorical data which included member type, gender, loan type, issued month and recovery modes. For faster results with minimal errors, an automated process within Python library, One- Hot Encoding technique was applied. The results of which were then combined with standardized numeric data. Secondly, numeric variables in this study were in different scales, for example, age ranges between 18 years and 79 years while savings is between 0 and Kshs 14,887,043. It was therefore necessary to standardize the variables as the parameter coefficients and the model weights will be very different

even when the two variables contribute equally to the model as suggested by Vidal & Barbon (2019). To transform numeric variables so that they are on the same scale, Standard Scaler toolkit in Python was used. This process ensures normal distribution of data having a mean value 0 and standard deviation of 1. It also ensures data consistency and better machine learning outcomes. This process is summarized in Figure 3.2

Figure 3. 2 Processing Pipeline



Source: Researcher, (2021)

3.6.5 Splitting training and testing data for the model

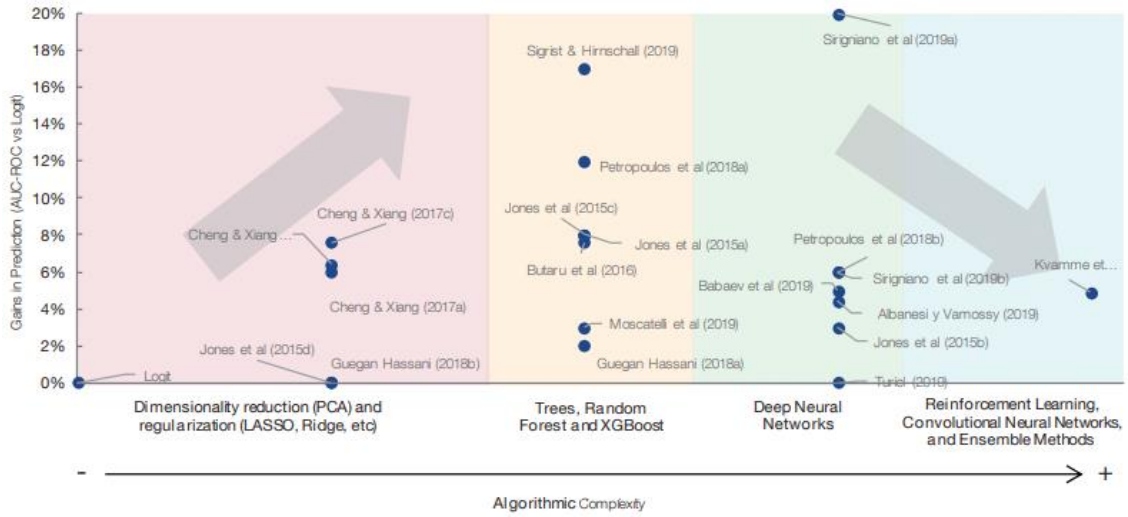
In this study, data was split into two sets where training data was 80% representing 627 loans and the remaining 20% representing 156 loans used as test data for both LR and XGBoost models. As suggested by Bracke et al. (2019), early period data was used as training set while later period data used to evaluate the models. This approach mimics a real-world situation in which a model is trained on past data and then used to predict the performance of subsequent cohorts of loans.

3.6.6 Machine Learning deployment

Quantitative techniques were used for the analysis of data collected. Data was summarized and categorized according to the broad categories of traditional data and alternative data. Analysis of data was initially through Microsoft Excel before being uploaded onto Python 3.9 Machine Learning software. This gave accurate and consistent responses to the research questions. Tests were also applied to establish the existence or otherwise of important gaps between the predicted values and the observed ones. In regards to ML, empirical literature Yang & Shami (2020), Addo et al. (2018), Turiel & Aste (2020), Bracke et al. (2019), Blazquez & Domenech (2018),

Kiefer & Mayock (2020) and Walusala et al. (2017) have suggested that more advanced, complex models lead to better prediction results than traditional ones as shown in Figure 3.3 It is for this reason that this study adopted two supervised ML models, Logistic Regression (LR) and Extreme Gradient Boost (XGBoost)

Figure 3. 3 Dilemma between prediction power and ML complexity



3.6.7 Logistic Regression

The Logistic Regression (LR) or logit model is one of the most popular supervised ML models for estimating the probability of Default (PD) because it is easy to develop, validate, calibrate, and interpret. It also has flexibility on preposition of data and ability to handle qualitative indicators unlike its predecessor linear regression (Bracke et al., 2019). In this study, the aim of using LR as the first model was to serve as a benchmark and help build intuition around some of the explainability metrics affecting default. There are two types of LR, the binary LR whose dependent variable has only two outcomes and multinomial LR which has multiple outcomes (Zahi & Achchab, 2020). Binary LR was applied in this study to allow determination of how a set of predictor variables is related to a dichotomous target variable. The equation;

$$Logit(P(Y=1)) = \log it(p) + \beta o + \beta_1 X_1 + \beta_2 X_2 + ...\beta_n X_n \qquad (1)$$

Where;

$X_1, X_2, ...X_n$ denote the set of n explanatory variables

$\beta_o, \beta_1 .....\beta_n$ denotes the set of n+1 parameters

Y denotes the dependent variable

If from equation 1, P(Y=1) =p, the Logit function then is as follows;

$$\log it(p) = \log \frac{p}{1-p} = \log \frac{P(Y-1/X)}{P(Y=0/X)} = \log(Odds) \quad (2)$$

The model measures the estimated probability of the predicted output, which varies between 0 and 1, and is based on a sigmoid function which has the following form

$$f(t) = \frac{1}{1+e^{-t}} \quad (3)$$

3.6.8 Extreme Gradient boosting model (XGBoost)

XGBoost is an improvement of the gradient boosting algorithm and a decision tree based on the gradient boosting algorithm (Li et al., 2021). Through a large number of iterations, each iteration produces a weak classifier, and each weak classifier is trained on the bias of the result of the previous classifier. This can be summarized as;

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{k} fx(x_i) \quad (1)$$

Where fx is the regression tree, k is the number of regression trees, $fx(x_i)$ is the score of the i^{th} observation given by the k^{th} tree. To make the prediction results more accurate, a penalty term is added to the prediction function, to reduce the occurrence of overfitting and increase the generalizability of the model function. The model objective function will hence be as follows:

$$obj(\phi) = \sum_i l(\hat{y}, y_i) + \sum_k \Omega(f_k) = \Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \quad (2)$$

Where y_i is the true value of the training sample, $\hat{y}_i$ is the predicted value of the training sample. l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. Penalty term to prevent the model from being too complicated is represented by Ω . γ is a parameter that controls the number of T nodes, λ is the parameter that controls the weight of the leaf node. However, the equation is difficult to optimize the Euclidean Spaces through traditional methods. Therefore, XGBoost adopts the greedy idea and adds f_t into the objective function to improve the performance of the model. The objective function becomes

$$obj^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3)$$

$\hat{y}_i^t$ is the predicted value of the i^{th} instance in the t^{th} interaction. In general, the gradient of the objective function is difficult to obtain but the second-order Taylor expansion may be used to simplify the equation and stabilize the tree structure. A greedy algorithm is used to continuously divide leaf nodes and iteratively add subtrees. The following equation is used to evaluate the split node of the final model.

$$Obj_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_{i+\lambda}} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_{i+\lambda}} - \frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I} h_{i+\lambda}}\right] - \gamma$$

Where I_R and I_L are the Right and Left nodes of the instance set after leaf node I is split, respectively.

3.6.9 Hyper parameters

Hyper parameters are parameters that are used to either configure an ML model e.g. penalty parameter in SVM and learning rate in NN or used to specify the algorithm to be used to achieve research objectives (Yang & Shami, 2020). While model parameters can be initialized and updated throughout data learning process, hyper parameters can only be configured before training an ML model because they define the architecture of the ML model. Guided by the research literature, the hyper parameters for this study are summarized in table 3.1

Table 3. 1 Optimal Hyper parameters

ML Algorithm	Main Hyper Parameters	Optimal Hyper Parameters	Alternative Hyper Parameter Optimization - Automated
Logistic Regression	Penalty, C, Solver		BO-TPE, SMAC
XGBoost	n_estimators min_child_weight max_depth gamma Learning_rate Subsample colsample_bytree reg_alpha reg_lambda	200 6 7 0.5 0.1 0.6 0.9 2 0.05	GA, PSO, BO-TPE, SMAC, BOHB

Source: Yang & Shami, (2020), Li et al., (2021)

3.7 Research Quality

3.7.1 Performance evaluation of the ML models

A number of metrics can be used to evaluate performance of LR models for example Hosmer–Lemeshow test, confusion matrix, misclassification costs, Area Under the curve and Receiver Operating Characteristic (AUC - ROC) among others. In this study however, confusion matrix and AUC-ROC tests are most suitable as suggested by previous literature by Zahi & Achchab (2020), Xia et al. (2021) and Shen et al. (2020). The confusion matrix is summarized in Table 3.2

3.7.2 Confusion Matrix

Table 3. 2 Confusion Matrix

		Predicted Class (Default)	
		+	-
Observed Class (Default)	+	True Positive – (TP)	False Positive – (FP)
	-	False Negative – (FN)	True Negative – (TN)

Source: (Zahi & Achchab, 2020)

Each row of the confusion matrix represents a real or observed class whereas each column represents the predicted class. True Positive, (TP) will mean that the model correctly predicts the value defined as Positive while False Positive will mean that the model will incorrectly predict the positive value. The same interpretation will apply to False Negative and True Negative. Various conclusions can thereafter be deduced from the observations including model accuracy, precision, Recall/sensitivity and specificity as summarized in Table 3.3

Table 3. 3 Evaluation matrix

	Metric	**Formula**	**Definition**
1	Accuracy	$= \frac{TP + TN}{TP + TN + FP + FN}$	Proportion of the Total number of predictions that turn out to be correct
2	Precision	$= \frac{TP}{TP + FP}$	Number of the correctly predicted cases that actually turn out to be positive
3	Sensitivity/Recall	$= \frac{TP}{TP + FN}$	Number of the actual positive cases that are predicted correctly with the model

Source: (Zahi & Achchab, 2020)

3.7.3 Area Under Curve (AUC)

The AUC test was applied as the second evaluation tool to evaluate the quality of models prediction; this method gives reliable results irrespective of what decision threshold is determined Xia et al., (2021). AUC measures the entire two-dimensional area under the ROC curve and calculated as follows for a binary classification

$$AUC = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1}$$

Where n_0 and n_1 denote the number of non-default and default loans in test set, respectively. $S_0 = \sum \text{rank}_j$ is the rank of probability predications of j^{th} default loans. The larger the AUC the better the prediction model performance.

3.7.4 Validity and Reliability

Content validation was applied to establish whether instruments include items capable of addressing all variable indicators and their measures. Validity is the appropriateness and usefulness of inferences made based on the collected data. Expert empirical data and piloting was used to prove the face, content, and construct validity of the machine learning model to be adopted. Piloting of the instruments was carried with different models and datasets to provide various results for comparison in the study.

Reliability is the consistency of scores obtained from an instrument. Instruments are said to be reliable if they give consistent results with repeated measurements of the object.

3.8 Ethical Issues in Research

3.8.1 Data protection and privacy

All identifiable member details e.g. names and membership numbers were omitted from the dataset to protect the identities of the borrowers. The borrower's unique identification numbers were however randomized by a factor to remain identifiable to the researcher but not any other third parties. Passwords protection was implemented on all workbooks and databases as extra caution on data protection.

3.8.2 Ethical Considerations of the Study

The mandatory research authority was secured from Strathmore University Institutional Ethical Review Committee after review of the research proposal and measures to protect personal data privacy. License from the National Council of Science and Technology Innovation (NACOSTI) was also acquired prior to

commencement of the research. Further, only one research assistant was engaged to ensure minimal risk in compromising data privacy. This was after induction on expectations of the study including avoidance of deception, privacy and integrity requirements. The Data Protection Act, 2019 additionally provides guidelines on acquisition, processing and disposal of personal data. This was fully complied with by consent of the Board of Directors on behalf of the members. Every secondary source of information that was used in this study have been duly acknowledged.

CHAPTER FOUR: PRESENTATION OF RESERCH FINDINGS

4.1 Introduction

This chapter puts together the findings of the study in response to the research questions. The findings include the factors (features) which influence SACCO members' credit behavior. The results from Logistic Regression and XGBoost Machine Learning algorithms using traditional data in contrast to a composite of traditional and alternative data is analysed in this chapter. Additionally, performance of the ML algorithms is compared based on their default settings (parameters) versus performance after hyper-parameters tuning.

4.2 Sample and descriptive statistics

Although the researcher was able to collect a heterogeneous sample in terms of duration of membership, age, sex, savings amount and various indicators of online member behavior for a period of one year, the researcher makes no claim that the sample is representative of the entire population nor the SACCO industry at large. The final sample consisted of 783 loans extracted from the SACCO ERP and online platform for the period July 2018 to June 2019. From the statistical study five percent of the sampled loans had been defaulted, while the remaining 95 percent were performing. The default rate compares with prior similar studies such as Barbaglia et al. (2020) who established an average default rate of 4.73 percent for 7 countries in the Europe. The default rate in this study is however lower than the 20 percent in Rwandese SACCOs as established by Papias & Ganesan (2009). Given the SACCO setting, a significantly lower default rate than in mainstream banks was anticipated, and this proposition is consistent with Nitani & Legendre (2021). Although the sample chosen is homogeneous i.e. from a single SACCO, there was observable heterogeneity in the features extracted from its systems and this permitted further analyses on the most important ones in explaining default by members. Table 4:1, panels a and b present summary descriptive statistics on the extracted borrower features.

Table 4. 1 Borrower features

Panel a: CATEGORICAL VARIABLES			Number		Percentage		
Attribute	**Full Description**	**Abbreviation**	**Defaulted**	**Not Defaulted**	**Defaulted**	**Not Defaulted**	**Total**
Member Type	Principal Member	Member	31	615	5%	95%	646
	Introduced member	Introduced	7	130	5%	95%	137
		Total	**38**	**745**	**5%**	**95%**	**783**
Gender	Male	Male	31	457	6%	94%	488
	Female	Female	7	288	2%	98%	295
		Total	**38**	**745**	**5%**	**95%**	**783**
Loan Type	Asset Finance Loan	Ast	0	5	0%	100%	5
	Development Loan	Dev	12	279	4%	96%	291
	Dividends Advance Loan	Dva	0	66	0%	100%	66
	Emergency Loan	Emg	12	229	5%	95%	241
	Equity Release Loan	Eqt	3	6	33%	67%	9
	Flexi Development Loan	Flx	1	3	25%	75%	4
	Insurance Loan	Inc	0	16	0%	100%	16
	Intern Loan	Int	4	30	12%	88%	34
	Medifinance Loan	Mfs	2	16	11%	89%	18
	Mid Term Development Loan	Mid	2	60	3%	97%	62
	Mid Term Swift Loan	Mis	0	5	0%	100%	5
	School Fees Loan	Scl	0	5	0%	100%	5
	Staff Mortgage Loan	Sml	0	1	0%	100%	1
	Swift Development Loan	Swf	2	24	8%	92%	26
		Total	**38**	**745**	**5%**	**95%**	**783**

Panel a: CATEGORICAL VARIABLES			Number		Percentage		
Attribute	**Full Description**	**Abbreviation**	**Defaulted**	**Not Defaulted**	**Defaulted**	**Not Defaulted**	**Total**
Month of Issue	January	Jan	3	102	3%	97%	105
	February	Feb	2	62	3%	97%	64
	March	Mar	1	63	2%	98%	64
	April	Apr	2	57	3%	97%	59
	May	May	1	59	2%	98%	60
	June	Jun	3	66	4%	96%	69
	July	Jul	10	60	14%	86%	70
	August	Aug	6	55	10%	90%	61
	September	Sep	1	60	2%	98%	61
	October	Oct	5	54	8%	92%	59
	November	Nov	2	45	4%	96%	47
	December	Dec	2	62	3%	97%	64
		Total	**38**	**745**	**5%**	**95%**	**783**
Recovery mode	Checkoff from payroll	Checkoff	15	430	3%	97%	445
	Direct Debit from the bank	Direct Debit	1	22	4%	96%	23
	Deduction from Dividends	Dividend	0	66	0%	100%	66
	External Standing Order	External STO	22	224	9%	91%	246
	Issuance of Postdated cheques	Postdated cheques	0	3	0%	100%	3
		Total	**38**	**745**	**5%**	**95%**	**783**

Panel b: NUMERICAL VARIABLES			Number		Percentage		
Attribute		**Description**	**Defaulted**	**Not Defaulted**	**Defaulted**	**Not Defaulted**	**Total**
Members' Age (Years)	**Age**	18 - 29	7	133	5%	95%	140
		30 - 39	17	408	4%	96%	425
		40 - 49	11	127	8%	92%	138
		50 - 59	1	54	2%	98%	55
		60 - 69	1	18	5%	95%	19
		70 - above	1	5	17%	83%	6
		Total	**38**	**745**	**5%**	**95%**	**783**
Membership Duration (Months)	**Membership Duration**	0 - 49	9	224	4%	96%	233
		50 - 99	16	348	4%	96%	364
		100 - 149	11	100	10%	90%	111
		150 - 199	2	57	3%	97%	59
		200 - 249	0	16	0%	100%	16
		Total	**38**	**745**	**5%**	**95%**	**783**
Savings balance at time of application (Ksh)	**Savings**	0 - 999,999	28	523	5%	95%	551
		1,000,000 - 3,800,000	9	192	4%	96%	201
		3,800,001 - 6,600,000	1	20	5%	95%	21
		6,600,001 - 9,400,000	0	4	0%	100%	4
		9,400,001 - 12,200,000	0	4	0%	100%	4
		12,200,001 - 15,000,000	0	2	0%	100%	2
		Total	**38**	**745**	**5%**	**95%**	**783**
Loan Amount approved (Ksh)	**Loan Amount**	0 - 999,999	27	610	4%	96%	637
		1,000,000 - 3,800,000	7	113	6%	94%	120
		3,800,001 - 6,600,000	4	16	20%	80%	20
		6,600,001 - 9,400,000	0	4	0%	100%	4
		9,400,001 - 12,200,000	0	0	0%	0%	0
		12,200,001 - 15,000,000	0	2	0%	100%	2
		Total	**38**	**745**	**5%**	**95%**	**783**

Panel b: NUMERICAL VARIABLES			Number		Percentage		
Attribute		**Description**	**Defaulted**	**Not Defaulted**	**Defaulted**	**Not Defaulted**	**Total**
Turnaround Time (Days)	**TAT**	0 - 69	38	742	5%	95%	780
		70 - 133	0	1	0%	100%	1
		134 - 197	0	0	0%	0%	0
		198 - 261	0	0	0%	0%	0
		262 - 325	0	1	0%	100%	1
		326 - 389	0	1	0%	100%	1
		Total	**38**	**745**	**5%**	**95%**	**783**
Cumulative Loan accounts (Count)	**Cumulative**	0 - 5	28	569	5%	95%	597
		6 - 10	7	147	5%	95%	154
		11 - 15	1	24	4%	96%	25
		16 - 20	2	0	100%	0%	2
		21 - 25	0	2	0%	100%	2
		26 - 30	0	3	0%	100%	3
		Total	**38**	**745**	**5%**	**95%**	**783**
Other active Loan Accounts at time of new application	**Active Loan Accounts**	1	35	685	5%	95%	720
		2	3	57	5%	95%	60
		3	0	1	0%	100%	1
		4	0	2	0%	100%	2
		Total	**38**	**745**	**5%**	**95%**	**783**
Loan Repayment amount (Ksh)	**Repayment**	1,303 - 99,999	35	710	5%	95%	745
		100,000 - 203,000	3	26	10%	90%	29
		203,001 - 306,001	0	6	0%	100%	6
		306,002 - 409,002	0	1	0%	100%	1
		409,003 - 512,003	0	0	0%	0%	0
		512,004 - 615,175	0	2	0%	100%	2
		Total	**38**	**745**	**5%**	**95%**	**783**

Table 4.1, panel a, presents the descriptive statistics based on the categorical features of the data extracted from the SACCO's systems. According to the results, it appears that the default rates between principal and introduced members compare at 5 percent of the total loans disbursed. However, on average, male borrowers are most likely to default at 6 percent compared to their female counterparts at 2 percent. It is however interesting to note that the male borrowers are associated with the highest value loans. In terms of the loan type, the results show that the development loan is the most popular at 37 percent of the total loans taken followed by the emergency loan at 31 percent. Emergency loans however experiences higher default rate at 5 percent than the development loans at 4 percent. The results show that the equity release and flexi development loans are associated with highest default rates, which is an important indicator for the SACCO to reconsider the structure and design of the two loan types. A closer examination of the borrowing pattern shows that most loans were issued in January (13 percent) while most defaults occurred in July (14 percent), August (10 percent) and October (8 percent) in that order. The common repayment method was through salary checkoffs, which is at 57 percent. Any deviation from this recovery method resulted into increased default. For instance, it's observed that loans repaid using external standing orders were most defaulted at 9 percent compared to those repaid through checkoffs (3 percent) or direct debits (4 percent).

Table 4.1, panel b presents the descriptive statistics based on the numerical features extracted from the SACCO systems. In terms of age, most defaulters fell between 30 – 50 years with some peaks in late 30s and early 70s. The mean age of defaulters was 38 years with a standard deviation of 10 while the mean age of non-defaulters was 36 years with a standard deviation of 9. Further, it is observed that both the defaulters and non-defaulters were SACCO members who had been in the SACCO for an average of 6 years with a standard deviation of 4 and 3 for non-defaulters and defaulters respectively. The results show that members with higher savings displayed better loan repayment than members with lower savings in absolute terms. Higher value loans (mean = Ksh 1,152,148/-) were likely to be defaulted compared to lower value loans (mean = Ksh 678,226/-). The repayment amounts for both defaulters and non-defaulters compared favorably at an average of Ksh 32,092/-. The results show that the sampled members had been with the coop for between 5 and 7 years. Finally, it is observed that majority of the defaulters had been with the coop for between 7-9 years.

4.3 Diagnostic tests

4.3.1 Multicollinearity test

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model. This violates one assumption of LR, independent variables (Magali, 2013). While Multicollinearity may not affect the accuracy of the model, there is risk of losing reliability in determining the effects of individual features on dependent variable posing a problem of interpretability (Daoud, 2018). However, Multicollinearity is not such a problem for non-regression methods like decision trees, clustering, and nearest-neighbors (Daoud, 2018).

Multicollinearity can be diagnosed by analysing the Variance Inflation Factors (VIF) of variables. According Magali (2013), the value of VIF=1 indicates no correlation, $1<VIF\leq10$ indicates moderate correlation while VIF>10 indicates presence of serious problem of Multicollinearity with suggestion, in severe case, of eliminating one independent variable. It is worth noting however that VIF cannot be applied to categorical variables as dummy variables are introduced to represent a categorical variable with two or more categories (Daoud, 2018). If the proportion of cases in the reference category is small, the indicator variables will have high to infinite VIFs, even if the categorical variable is not associated with other variables in the regression model (Daoud, 2018). In this study, there is no evidence of Multicollinearity and no further action was needed. Table 4.2 show the results of this test using traditional data (Table 4.2a) and traditional data and alternative data (Table 4.2b).

Table 4. 2 Diagnostic tests - VIF

	Features	VIF
0	membership_duration	1.63138
1	savings	2.38816
2	loan_amount	2.525685
3	loan_duration	1.72977
4	tat	1.111155
5	cummulative_loan_accounts	1.507893
6	repayment	2.789264
7	mbrtyp_1	2.843879
8	sex_1	2.293772
9	ismth_1	1.258981
10	ismth_5	1.137963
11	ismth_7	1.239614
12	ismth_8	1.149332
13	ismth_10	1.15356
14	recmth_2	1.667815

VIF Using Traditional data

	variables	VIF
0	membership_duration	1.615763
1	savings	2.455026
2	loan_amount	1.563305
3	cummulative_loan_accounts	1.488639
4	repayment	2.532751
5	accd_dashboard	8.46614
6	login_failed	8.240888
7	accd_pass_recovery	5.173151
8	statements_view_download	5.894135
9	lnty_6	1.048127
10	lnty_7	1.036544
11	ismth_7	1.019946
12	ismth_10	1.014399
13	ismth_11	1.013705
14	ismth_12	1.009853

VIF using traditional and Alternative data

Key: membership duration (membership_duration), member savings (savings), loan amount (loan_amount), loan duration (loan_duration), sex 1 (sex_1), sex 2 (sex_2), loan type 1 (lnty_1), turnaround time (tat), cumulative loan accounts (cumulative_loan_accounts), repayment amount (repayment), issued month 1, January (ismth_1), issued month 5, May (ismth_5), issued month 7, July (ismth_7), issued month 8, August (ismth_8), issued month 10, October (ismth_10), recovery method 2 standing order (recmth_2), accessed dashboard (accd_dashboard), login failed (login_failed), accessed password recovery (accd_pass_recovery), statement viewed and downloaded (statements_view_download)

4.3.2 Normality test

The Kolmogorov-Smimov (KS) test measures the maximum vertical separation between two cumulative distributions i.e. good and bad in a credit scorecard. The higher the separation between the two lines, the higher the KS, which translates into a more accurate scorecard (Vidal & Barbon, 2019). The tool is effective when comparing different models during their development because they use the same sample but it can be misleading when comparing models being applied to different products or on different samples (Vidal & Barbon, 2019). Using KS test, normality test was conducted on the features with different models and the findings of the presented in Table 4.3 a & b. The findings revealed that the data set was not normally distributed. This is because the p-value of the Shapiro Wilk Test was less than 0.05 for using both Logistic Regression and XGBoost. Standard scaler tool in Python 3.9 was used to normalize the data as described earlier.

Table 4. 3 Diagnostic tests - Kolmogorov test

a) Logistic Regression

		Kolmogorov-Smirnov test		Shapiro-Wilk test	
		statistic	P-value	statistic	P-value
	Features				
1	membership_duration	0.098184	5.06E-07	0.95554	1.21E-14
2	savings	0.256472	6.68E-46	0.555657	1.74E-40
3	loan_amount	0.294244	1.44E-60	0.51683	1.10E-41
4	loan_duration	0.277326	9.88E-54	0.818906	8.85E-29
5	sex_1	0.5	1.56E-181	0.614091	1.65E-38
6	sex_2	0.5	1.56E-181	0.614103	1.65E-38
7	lnty_1	0.5	1.56E-181	0.612164	1.41E-38
8	lnty_3	0.5	1.56E-181	0.309656	0.00E+00
9	lnty_5	0.5	1.56E-181	0.204123	0.00E+00
10	ismth_5	0.5	1.56E-181	0.29248	0.00E+00
11	ismth_7	0.5	1.56E-181	0.320588	0.00E+00
12	ismth_8	0.5	1.56E-181	0.295412	0.00E+00
13	ismth_9	0.5	1.56E-181	0.295412	0.00E+00
14	recmth_2	0.5	1.56E-181	0.584289	1.52E-39
15	recmth_3	0.5	1.56E-181	0.309656	0.00E+00

b) Extreme Gradient Boost

		Kolmogorov-Smirnov test		Shapiro-Wilk test	
		statistic	P-value	statistic	P-value
	Features				
1	membership_duration	0.098184	5.06E-07	0.95554	1.21E-14
2	savings	0.256472	6.68E-46	0.555657	1.74E-40
3	loan_amount	0.294244	1.44E-60	0.51683	1.10E-41
4	loan_duration	0.277326	9.88E-54	0.818906	8.85E-29
5	tat	0.377767	6.63E-101	0.1908	0.00E+00
6	cummulative_loan_accounts	0.194761	1.71E-26	0.766886	7.32E-32
7	repayment	0.251029	5.77E-44	0.536905	4.48E-41
8	mbrtyp_1	0.666377	0.00E+00	0.460075	2.65E-43
9	sex_1	0.5	1.56E-181	0.614091	1.65E-38
10	ismth_1	0.5	1.56E-181	0.402125	8.41E-45
11	ismth_5	0.5	1.56E-181	0.29248	0.00E+00
12	ismth_7	0.5	1.56E-181	0.320588	0.00E+00
13	ismth_8	0.5	1.56E-181	0.295412	0.00E+00
14	ismth_10	0.5	1.56E-181	0.28952	0.00E+00
15	recmth_2	0.5	1.56E-181	0.584289	1.52E-39

Key: membership duration (membership_duration), member savings (savings), loan amount (loan_amount), loan duration (loan_duration), sex 1 (sex_1), sex 2 (sex_2), loan type 1 (lnty_1), turnaround time (tat), cumulative loan accounts (cumulative_loan_accounts), repayment amount (repayment), issued month 1, January (ismth_1), issued month 5, May (ismth_5), issued month 7, July (ismth_7), issued month 8, August (ismth_8), issued month 10, October (ismth_10), recovery method 2 standing order (recmth_2), accessed dashboard (accd_dashboard), login failed (login_failed), accessed password recovery (accd_pass_recovery), statement viewed and downloaded (statements_view_download)

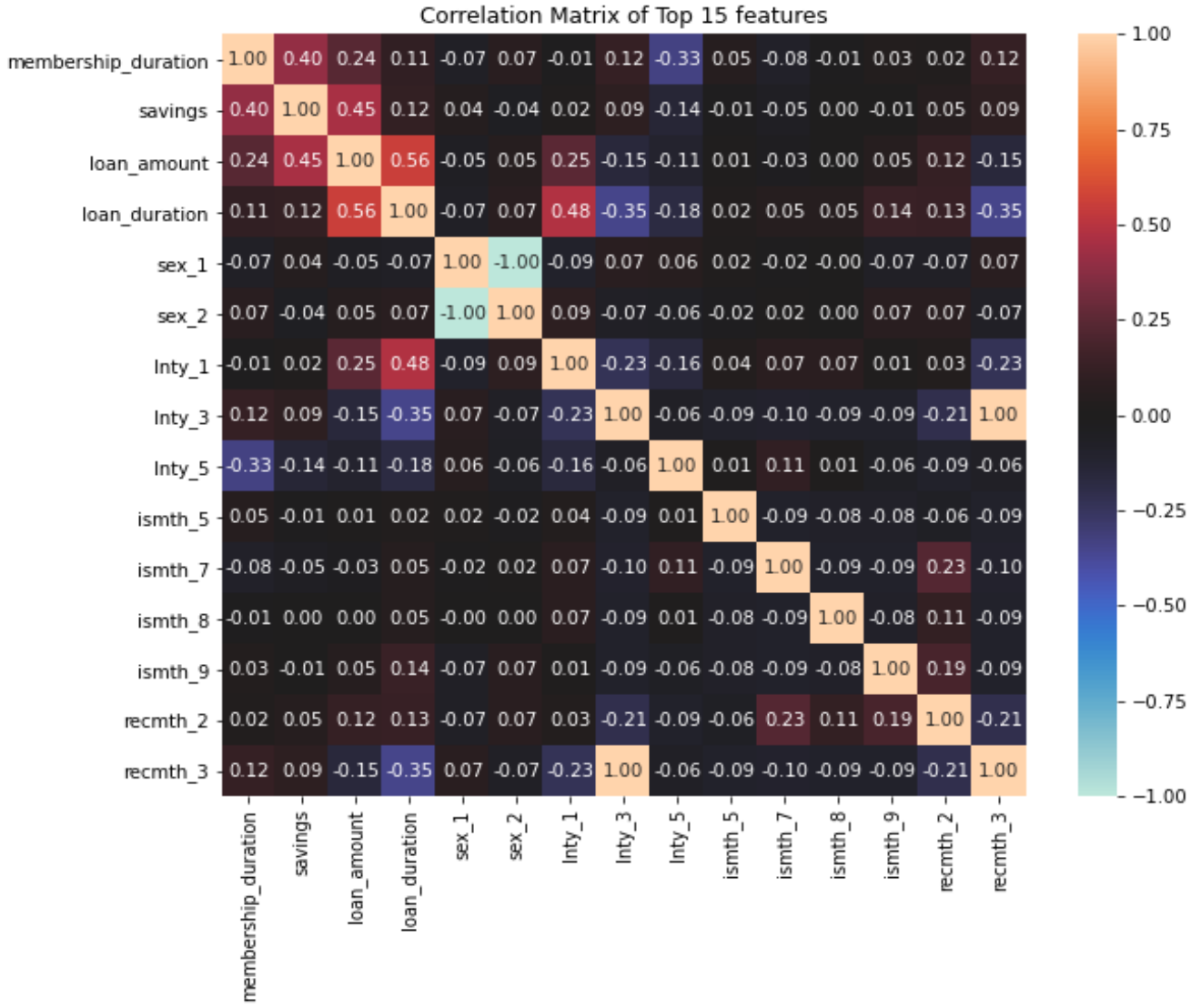
4.4 Feature Selection using Traditional data

In this study, an automated process, Recursive Feature Elimination (RFE) was implemented to select fifteen best features based on their importance to the ML algorithm. This is by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of all features and the importance of each feature is obtained either through a coefficient attribute or through a feature importance attribute. Then, the least important features are pruned from the set of features. This process is recursively repeated on the pruned set until the desired number of features is eventually reached. The top fifteen features are then used to train the two ML models and results evaluated. The different features and models are further discussed in the following sections.

4.4.1 Feature Selection using Traditional data by applying LR

The importance of features differs as shown in Figure 4.1. As shown, membership duration which represents how long a member has been with the SACCO and the amount of their savings have the highest importance followed by the loan amount, loan duration, cumulative number of active loan accounts the member has in that order. On the other hand, repayment method, and the month a loan is issued are the least important features in this model. These findings are nearly similar to the findings of Aderitus (2020) who used Logistic Regression and Random Forest algorithms to evaluate the probability of defaulting on their loan obligation.

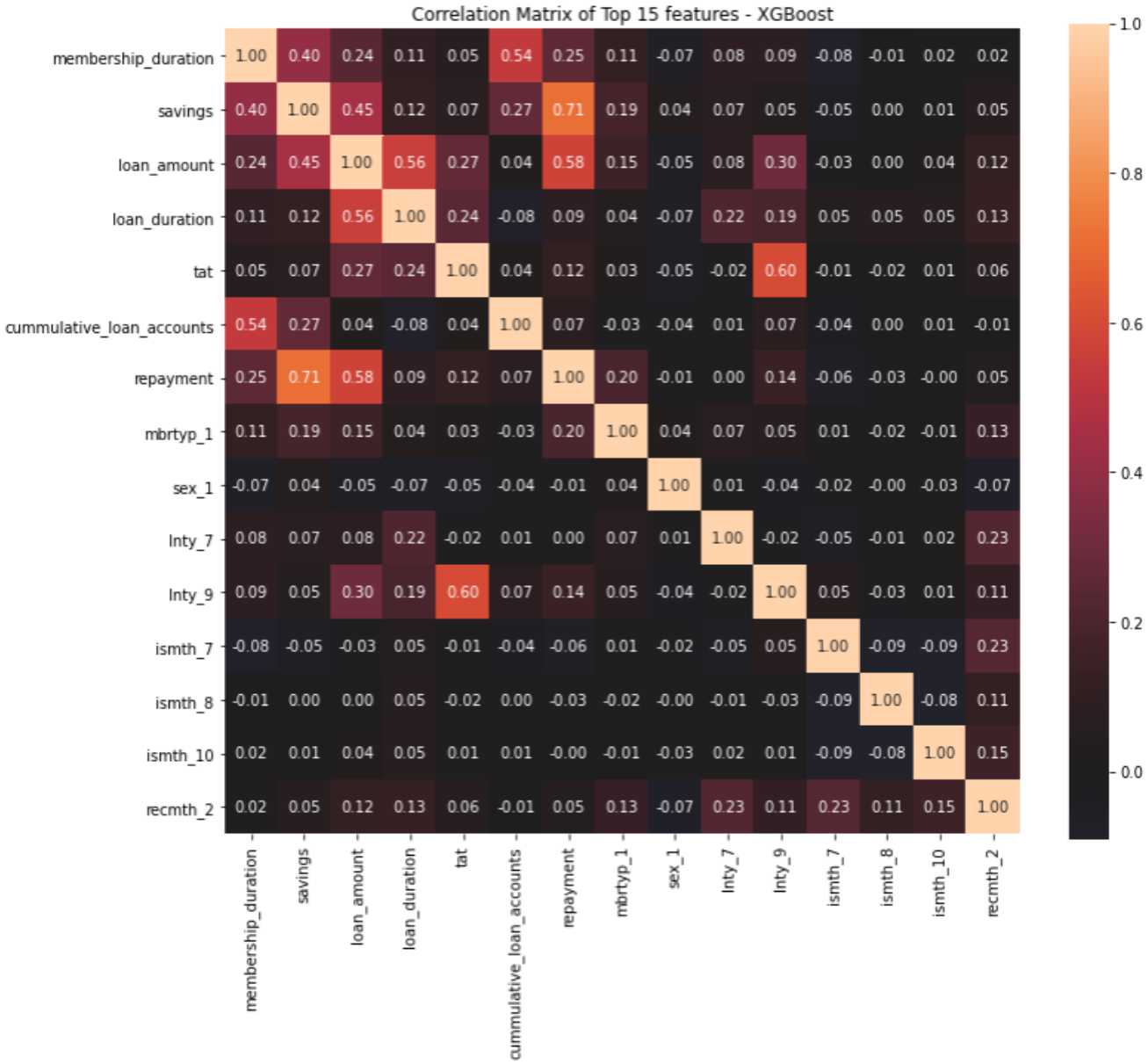
Figure 4. 1 Feature Selection using Traditional data by applying LR



4.4.2 Feature Selection using Traditional data by applying XGBoost

Using XGBoost ML algorithm to establish which factors influenced loan default, the results are almost as similar to those of LR with membership duration, savings, loan amount, loan duration, cumulative loan accounts, being the most important features while repayment method and month of issue being the least. However, loan repayment amount, membership type and interest rate were included as factors influencing default unlike in LR. This is exhibited in Figure 4.2

Figure 4. 2 Feature Selection using Traditional data by applying XGBoost



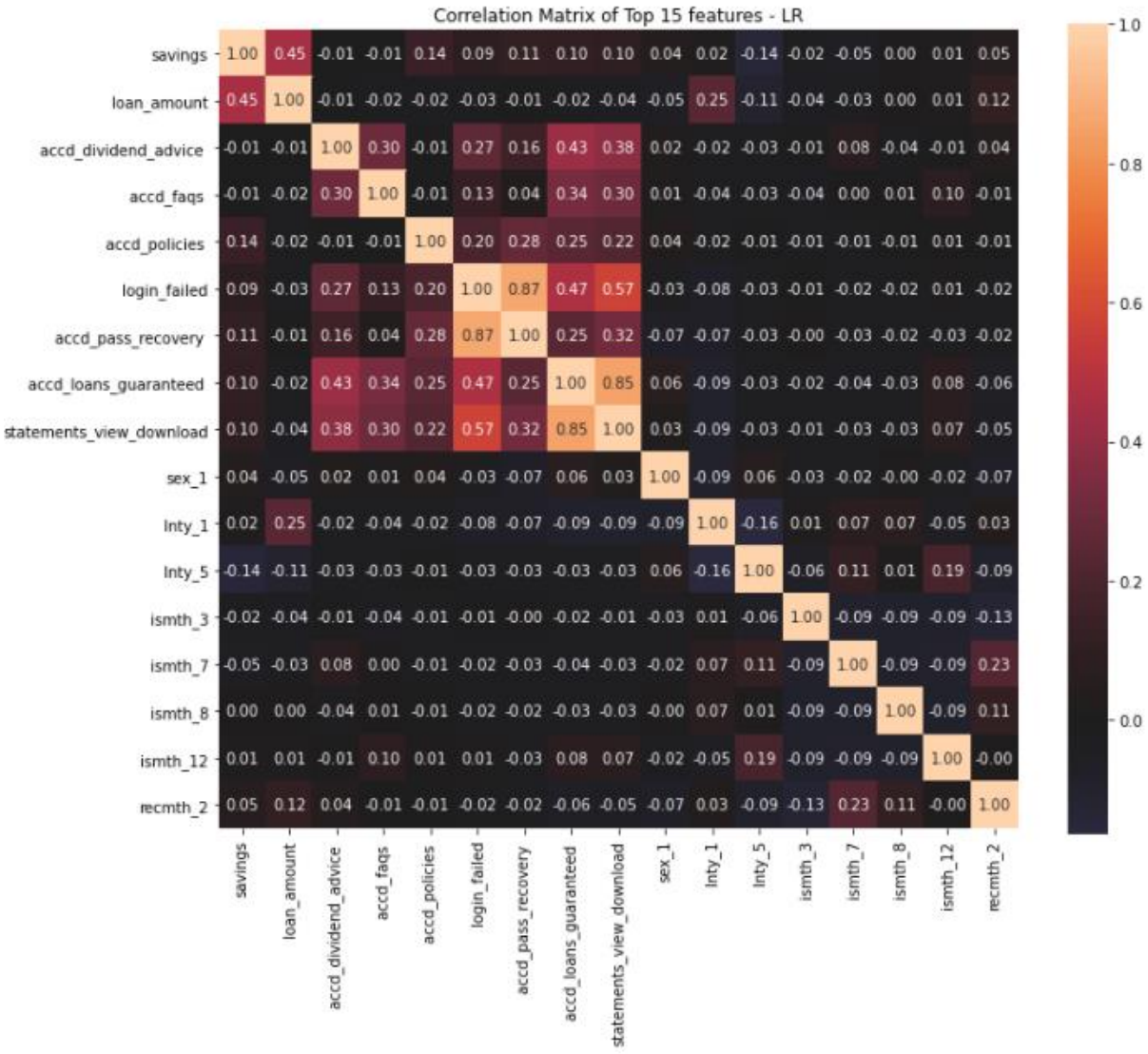
4.5 Feature Selection using alternative data

In this section, members' behavior was analysed based on data collected on the interactive online portal at granular level known as alternative data for this study. A total of fifty-three possible online movements was tracked and analysed as at the date of loan application.

4.5.1 Feature Selection using alternative data by applying LR

The introduction of alternative data on LR algorithm caused significant change on selection of top fifteen features influencing default. Loan duration, members accessing dashboard, dividend advice, read SACCO policies, accessed calculator function, their eligibility to guarantee or be guaranteed were ranked as most important features while issued month and loan repayment method continued to rank low as shown in Figure 4.3.

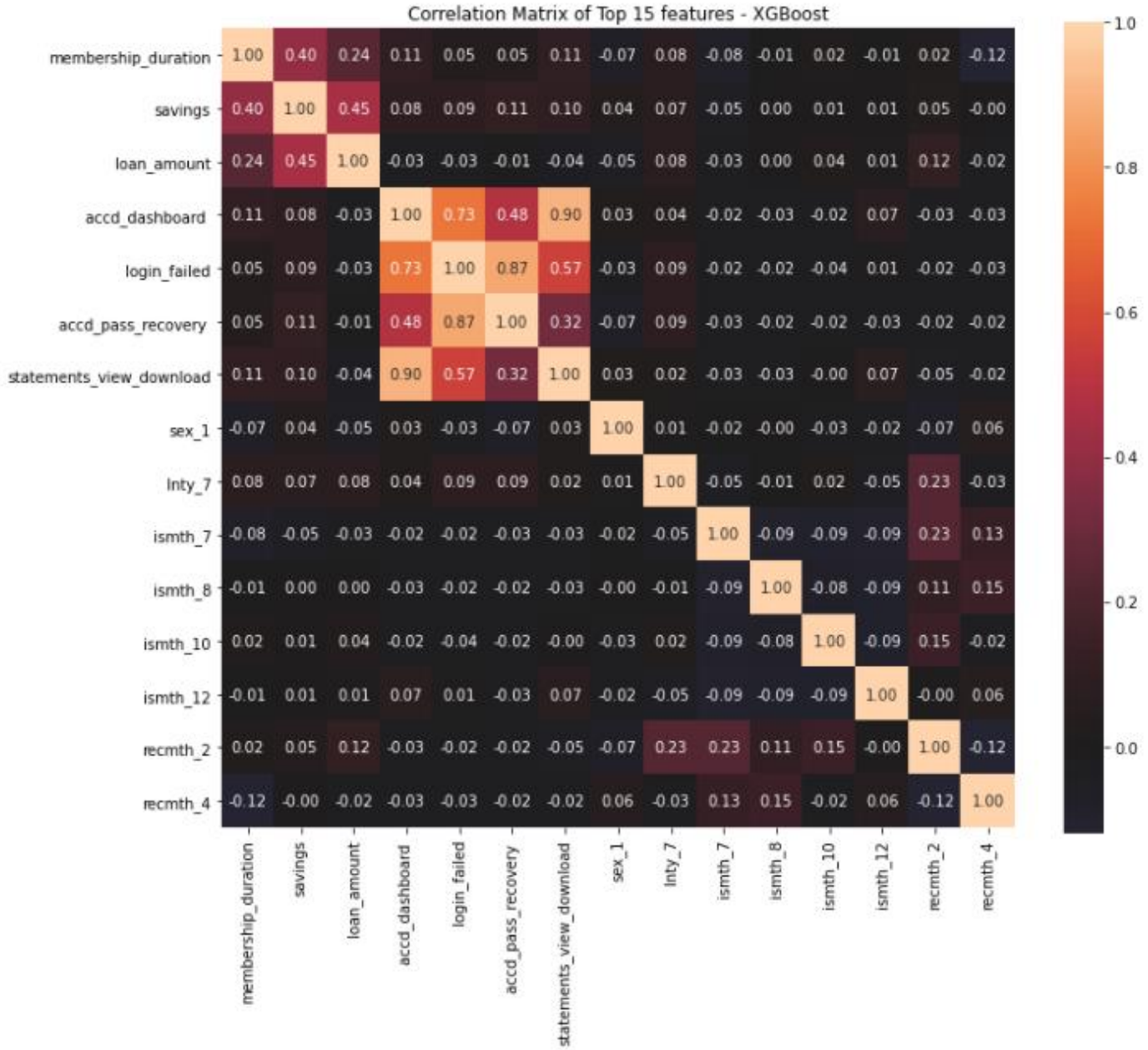
Figure 4. 3 Top 15 features with alternative data using LR



4.5.2 Feature Selection using alternative data by applying XGBoost

The results of XGBoost to rank membership duration, savings, loan amounts and number of active loan accounts as the top four features as was the case using traditional data only. Members' access to dashboard, and view of statements were other features of importance while repayment method and month of issue remained low ranking features as shown in Figure 4.4

Figure 4. 4 Top 15 features with alternative data using XGBoost



4.6 Hyper parameter tuning

Hyperparameter tuning is choosing a set of optimal hyperparameter for a machine learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins. In this study, performance of the two ML models was evaluated both on default settings and after hyperparameter tuning.

4.6.1 Effects of hyperparameter tuning on Logistic Regression

The best hyperparameters from the automated process, GridSearchCV, suggest that setting of C=0.05, intercept_scaling=2, multi_class='ovr', n_jobs=-1, random_state=0, solver='liblinear', verbose=True were the most optimal for the current dataset returning a model score of 68% Area Under Curve compared to default settings which had a score of 64%. This shows an improvement of 4% in performance. This study therefore suggests hyper parameter tuning for LR models of optimized performance. Automated hyperparameter tuning resources GridSearchCV provide better avenues in error minimization and reduced model training time.

The default settings of LR returned a model accuracy score of 64% Area Under Curve using traditional dataset solely. This is lower than the 73% score obtained using a combination of traditional data and alternative data on default settings. However, upon hyperparameter tuning using GridSearchCV, the model score using traditional data improved by 4% to 68%. Upon implementation together with best fifteen features, Area Under Curve measures, Accuracy, Precision and Recall were observed as 68%, 69.43%, 69.66% and 96.19% respectively for LR with traditional datasets while the same measures were 73%, 81.53%, 82.55% and 97.62% respectively for LR with traditional in combination with alternative data. These results suggest improvements in model AUC, Accuracy, Precision and Recall of 5%, 12.1%, 12.9% and 1.43% respectively in Logistic Regression using both traditional and alternative data compared to using traditional datasets alone. As found out by Turiel & Aste (2020). Logistic Regression algorithms provided better results for loan acceptance while DNN was better for default prediction. Of note however, is that both models showed substantial improvements on traditional credit screening methods with recall score above 70% and AUC-ROC of approximately 70% as well.

4.6.2 Effects of hyperparameter tuning on XGBoost

Previous similar studies by Yang & Shami (2020) and Li et al. (2021) suggested optimal hyperparameter tuning of n_estimators=200, min_child_weight=6, max_depth=7, gamma=0.5, Learning_rate=0.1, Subsample=0.6, colsample_bytree=0.9, reg_alpha=2, reg_lambda=0.05 for loan default prediction models. However, in this study, the suggested settings resulted in lower model score of 46.4% Area Under Curve compared to 50% using default machine settings. On the other hand, by using an automated machine tuning technique, LightGBM, the model's accuracy score improved to 58% on 53 iterations representing and improvement of 8% performance improvement. The resulting best parameters being; n_estimators=500, min_child_weight=1, max_depth=5, gamma=0.5, Learning_rate=0.1, Subsample=0.5, colsample_bytree=0.5, reg_alpha=0, reg_lambda=1. It is observed that using automated parameter tuning techniques provide faster results hence reducing model training time. The differences in model results from suggested hyperparameter settings and automated best parameters could be explained by the differences in datasets and environments of this study from empirical studies reviewed earlier.

This study also sought to evaluate the use of XGBoost ML algorithm in predicting loan default. The results of evaluation criteria using Area Under Curve measures, Accuracy, Precision and Recall scores of were observed as 58%, 92.36%, 98.62% and 92.36% respectively using traditional dataset only. On the other hand, using both traditional and alternative datasets, the scores were observed as 73%, 94.77%, 98.62% and 94.77% respectively. These results also show improvements of scores of 15%, 2.41% and 2.41% for Area Under Curve measures, Accuracy and Recall. Precision scores remained affected by introduction of alternative data in this model. These results have been summarized in Figure 4.5. The results of this study further supports the findings of Xia et al. (2021) whose study not only showed that boosting ML algorithms had superior predictive performance but also had better engineering optimization capabilities to outperform other models. Their study included three sets of data categorized as loan characteristics, borrower's creditworthiness, and borrower's solvency. Moreover, microeconomic variables were added to reflect the dynamics of business cycle on repayment with results suggesting Gradient boosted algorithms being more responsive to new data than Logistic Regression models. In yet another study, Moradi & Rafiei (2019) while investigating model efficiency in

predicting loan default in Iranian banks found that Traditional static models like Logistic Regression proved to work reasonably well in predicting credit risks during periods of stasis, but they fail to do so in the face of economic and political fluctuations, where dynamic models like Random Forest and Gradient Boosting models were proposed. However, Kiefer & Mayock (2020) in their study why do models that predict failure fail concluded that in both types of predictive models, accuracy deteriorates rapidly when models that are trained in one type of macroeconomic environment are used to predict loan performance in out-of-time samples characterized by much different economic conditions. Unlike in boosted algorithms, this model instability can be at least partially alleviated for traditional statistical models through the use of data from a wide mix of economic conditions.

4.7 Summary of Results

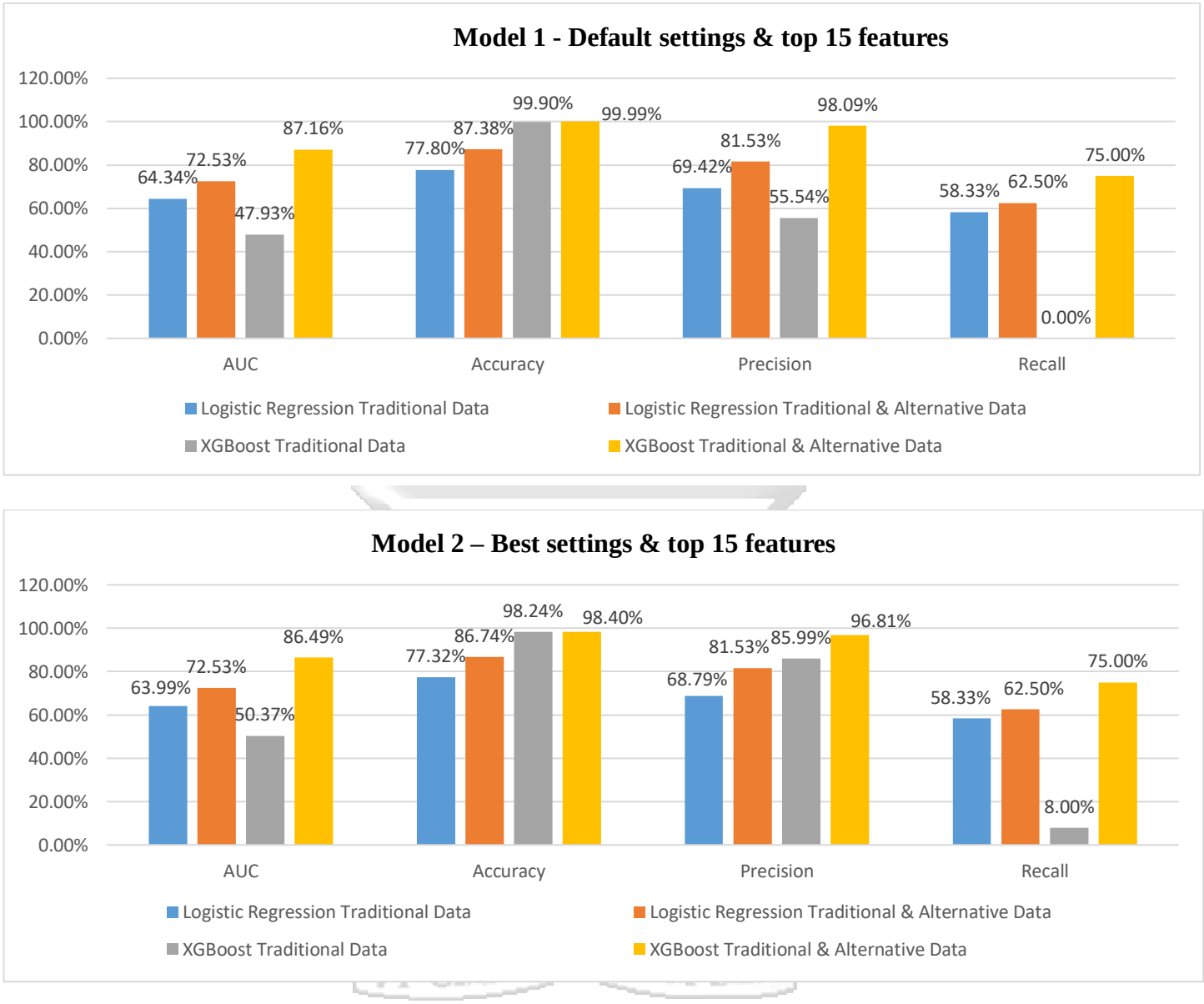
To determine the best performing ML model using the selected features, the two models i.e. the LR and XGBoost were fitted using the best hyperparameters as explained in previous sections and assess their accuracy, precision, recall and AUC in line with Li et al. (2021). This is prior to performing the multivariate regressions under each. Table 4.4 presents a summary of the results obtained under each ML model. In Table 4.4, panel A, the default settings of the LR returned a model accuracy score of 64.34 percent AUC which is lower than the 72.53 percent obtained using both traditional and alternative data. Similar results are observed under the accuracy, precision and recall attributes. This implies that the LR model using both traditional and alternative features yield better results compared to using traditional features only. This provides prima-facie support of alternative data in predicting default by members in the SACCO. The results show that using the XGBoost ML approach, the model's AUC is 47.93 percent under traditional data, and this improves to 87.16 percent when both traditional and alternative data are considered. This observation is consistent with Li et al. (2021), and is replicated under the accuracy, precision, and recall attributes. This provides further evidence that the use of both traditional and alternative features yields better results when predicting the default of borrowers. Similar findings in support of the use of both traditional and alternative data is revealed when predicting default using best settings and top borrower features selected using the recursive feature selection approach discussed earlier. These results mirror the argument presented by Turiel

& Aste (2020) as well as Liu et al. (2022) that ML algorithms provide better insights for loan acceptance and better default prediction.

Table 4. 4 Evaluation of ML models

Panel A - Model 1 - Default settings & top 15 features						
	Logistic Regression		**XGBoost**			
	Traditional Data	**Traditional & Alternative Data**	**Traditional Data**	**Traditional & Alternative Data**	**Effect of data change**	
AUC	64.34%	72.53%	47.93%	87.16%	8.19%	39.23%
Accuracy	77.80%	87.38%	99.90%	99.99%	9.58%	0.09%
Precision	69.42%	81.53%	55.54%	98.09%	12.11%	42.55%
Recall	58.33%	62.50%	0.00%	75.00%	4.17%	75.00%
Panel B - Model 2 – Best settings & top 15 features						
	Logistic Regression		**XGBoost**			
	Traditional Data	**Traditional & Alternative Data**	**Traditional Data**	**Traditional & Alternative Data**	**Effect of data change**	
AUC	63.99%	72.53%	50.37%	86.49%	8.54%	36.12%
Accuracy	77.32%	86.74%	98.24%	98.40%	9.42%	0.16%
Precision	68.79%	81.53%	85.99%	96.81%	12.74%	10.82%
Recall	58.33%	62.50%	8.00%	75.00%	4.17%	67.00%
Effect of Model change						
AUC	-0.35%	0.00%	2.44%	-0.67%		
Accuracy	-0.48%	-0.64%	-1.66%	-1.59%		
Precision	-0.63%	0.00%	30.45%	-1.28%		
Recall	0.00%	0.00%	8.00%	0.00%		

Figure 4. 5 Graphical representation of results



Upon identification of which models provide better default prediction capabilities, the specific borrower features that are significant in default prediction were examined. Table 4.5 reports the estimation results using both the LR and XGBoost ML models under both traditional and alternative data. According to the results, it appears that the only significant traditional features are savings, loan amount, gender of the borrower, and in some cases, the loan type. For instance, the LR model 3 using both traditional and alternative data shows that borrowers with higher savings are less likely to default compared to those with lower savings (coefficient -0.326, z-value = 2.13). The same model shows that borrowers with higher loan amounts are more likely to default, and this is highly significant at the 1 percent level (coefficient = 0.839, z-value = 5.63). The results in Model 2 using XGBoost

with traditional data shows that principal members (Mbrtyp_1) seem to default more compared to introduced members, and this is highly significant at the 1 percent level (coefficient = -2.498, z-value = -9.86) consistent with Papias & Ganesan (2009). Models 1, 2 and 3 seem to show that gender plays an important role in default. According to the results, it appears that male borrowers are more likely to default compared to their female counterparts, and this is highly significant at the 1 percent level. These results are in agreement with those of Mapunda (2019) who sought to study the factors affecting loan repayment efficiency in selected SACCOS in Tanzania. It was established that female borrowers were more careful and ensured repayments were done in time compared to male borrowers. Similar results were also found by Aslam et al. (2020) in their study Predicting likelihood for loan default among bank borrowers. They also found that males and young borrowers were more responsible for loan default.

The results show that the loan types susceptible to default are development loan (Lnty 1), intern loan (Lnty_5), swift development loan (Lnty_6) and medi-finance loan (Lnty_7) compared to other loans. In terms of the alternative data features, the results show that the number of times a member accessed the dashboard, logins failed, accessed loans guaranteed and statements view downloaded are significant predictors of default. For instance, it appears that members who frequently access their account dashboard are less likely to default, and this is highly significant at the 5 percent (model 3) and 1 percent (model 4). This might imply that SACCO members who frequently access their account dashboard are more active and concerned about their financial status in the SACCO and feel more obliged to repay their loans when they become due. It is also a show of loyalty and patronage to the SACCO and a commitment to its economic prosperity. The results show that members with more failed logins are less likely to default, and this is highly significant at the 5 percent (model 3) and 1 percent levels (model 4). This seems to corroborate earlier findings that more committed members to the SACCO are likely to be active, hence experience less login failures, and have lower default probabilities. The result might be taken to mean that members with higher login failures seem to have been in touch with the SACCO in a long time hence their commitment to the SACCO might be interpreted to be lower. These could be those members who rarely check their account dashboards and are more likely to default than not. Next, the results show that members who accessed their guaranteed loans are more likely to default than those who didn't (coefficient = -2.548, z-value=-3.89). This might point to potential moral hazard that comes with the uptake of SACCO loans which are guaranteed by fellow members (joint liability)

as argued by Ghatak & Guinnane (1999). It might imply that the guarantorship imposes some degree of acquaintanceship in a borrower, and this leads to default. It appears that borrowers compare their own costs and benefits of default, an argument held by Qinlan & Izumida (2013). In such a case, a borrower is shielded by the guarantors from experiencing the real consequence of default and might take advantage of this incentive to default. The results further show that borrowers who accessed their statements for download are more likely to default, and this is highly significant at the 5 percent and 1 percent in models 3 and 4 respectively. This might imply that a borrower who is in default is likely to frequently check their statement in view of assessing the extent of the liability owing to the SACCO. In line with the earlier findings on the strength and predictive power of the models, diagnostic features depicted in Table 4.5 seem to improve as both traditional and alternative data are utilized to predict default.

Table 4. 5 Estimation Results

	Data analysed	**Traditional Data**		**Traditional & Alternative Data**	
	Model Number & Type	**[1] - Logistic Regression**	**[2] - XGBoost**	**[3] - Logistic Regression**	**[4] - XGBoost**
	Dependent Variable	Loan Status			
	Predictor	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**
1	Membership duration	-0.030 (-0.21)	0.174 (1.07)		-0.118 (-1.08)
2	Savings	-0.256 (-1.33)	-0.161 (-0.72)	-0.326** (-2.13)	-0.031 (-0.22)
3	Loan amount	0.166 (1.11)	0.149 (0.36)	0.839*** (5.63)	-0.301** (2.52)
4	Loan duration	0.770*** (4.48)	0.131 (0.78)		
5	TAT		0.011 (0.07)		
6	Cumulative loan acc		0.020 (0.13)		0.002 (0.02)
7	Repayment		0.228 (1.08)		-0.096 (-0.61)
8	Mbrtyp 1		-2.948*** (-9.86)		-0.096 (-0.61)
9	Accd dashboard			3.403*** (5.14)	2.907*** (4.68)
10	Login failed			-1.056** (-2.10)	-0.969*** (-3.13)
11	Accd pass recovery			0.096 (0.21)	0.058 (0.27)
12	Accd loans guaranteed			-2.548*** (-3.89)	
13	Statement view download			-1.079** (-2.40)	-2.003*** (-4.29)
14	Sex 1	-1.752*** (-8.42)	-0.738*** (-3.06)		
15	Sex 2			-3.565*** (-6.51)	
16	Lnty 1	-2.948*** (-8.09)		-3.368*** (-7.57)	
17	Lnty 5	-0.236 (-0.38)		-1.658*** (-2.86)	
18	Lnty 6				-2.947*** (-3.05)

	Data analysed	Traditional Data		Traditional & Alternative Data	
	Model Number & Type	**[1] - Logistic Regression**	**[2] - XGBoost**	**[3] - Logistic Regression**	**[4] - XGBoost**
	Dependent Variable	Loan Status			
	Predictor	**Coefficient**	**Coefficient**	**Coefficient**	**Coefficient**
19	Lnty 7				
20	Ismth 1		-1.518** (-2.44)		-1.852** (-2.39)
21	Ismth 5	-2.528** (-2.40)	-1.971* (-1.89)		
22	Ismth 7	0.308 (0.70)	0.402 (0.36)	0.645 (1.27)	-1.982*** (-5.25)
23	Ismth 8	-0.307 (-0.61)	-0.247 (-0.48)	-0.198 (-0.36)	
24	Ismth 9	-3.496*** (-2.89)			
25	Ismth 10		-0.529 (-0.98)		-2.657*** (-4.85)
26	Ismth 11				-3.105*** (-4.17)
27	Ismth 12				-4.414*** (-4.24)
28	Recmth 2	-0.795*** (-2.75)	-0.056 (-0.18)	-1.888*** (-5.82)	
	Pseudo R^2 (-)	0.417	0.250	0.458	0.640
	AIC	454.868	410.183	469.255	832.676
	BIC	510.825	480.130	529.876	902.623
	No. of observations	783	783	783	783

Note: ***, ** and * denote significance at the 1,5, and 10 percent levels respectively. The z-values are in parentheses. XGBoost represents Extreme gradient boosting model

4.8 Summary of the Chapter

In this chapter, findings related to the research objectives were provided. Discussion concerning appropriate factors influencing SACCOS members' credit ratings were specified. Also, the selected Machine Learning algorithms that are Logistic Regression and Extreme Gradient Boosting were trained using traditional dataset as well as a combination of traditional and alternative dataset for credit rating. The results of which have also been discussed in detail in previous sections of this chapter.

CHAPTER FIVE: DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents a summary of the findings, conclusions and recommendations for practice and further research on the research problem. The purpose of the study was to assess the relevance of alternative borrower data and Machine Learning algorithms in predicting loan default in SACCOs as well as determining important features that contribute to prediction of loan default. Finally, optimal hyperparameter tuning for different ML algorithms were also evaluated.

5.2 Summary of the Findings

In this section, the results and findings of the study have been discussed in detail to answer the research questions that this paper sought to answer.

5.2.1 Findings on features affecting prediction of loan default in SACCOs

The findings of this study showed that there are factors that have more weight in influencing SACCO members' default behavior than others. Using Logistic Regression model and traditional data only, membership duration which represents how long members have been with the SACCO, member savings have the highest importance followed by the loan amount, loan duration, cumulative number of active loan accounts the member has. On the other hand, repayment method, and the month a loan is issued are the least important features. Introduction of alternative data on LR algorithm caused significant change in selection of top fifteen features influencing default. Loan duration, frequency of a member accessing dashboard, dividend advice, read SACCO policies, accesses calculator function, views their eligibility to guarantee or be guaranteed ranked as top features while issued month and loan repayment method continued to rank low.

Using XGBoost ML algorithm, the results are almost similar to those of LR with membership duration, savings, loan amount, loan duration, cumulative loan accounts, being the most important features while repayment method and month of issue being the least important. With introduction of alternative data, these results remained stable except for the inclusion of Members' access to dashboard, and view of statements being additional features of importance. Repayment method and month of issue remained low ranking.

5.2.2 Findings on effects of hyperparameters tuning in ML algorithms

It is observed that using automated parameter tuning techniques leads to better results in ML performance, this could be due to difference in datasets and environment of study from the settings suggested by empirical studies. In this study, GridSearchCV and LightGBM automated hyperparameter tuning were used for LR and XGBoot respectively. Improved accuracy score was noted after hyperparameter tuning was performed.

5.2.3 Findings on ML algorithms for prediction of loan default in SACCOs

Implementation of Logistic Regression trained with fifteen best features resulted an Area Under Curve score, Accuracy, Precision and Recall 68%, 69.43%, 69.66% and 96.19% respectively using traditional datasets while the same measures were 73%, 81.53%, 82.55% and 97.62% respectively for LR using a combination of both traditional and alternative data. These results suggest that adopting Logistic Regression using both traditional and alternative data leads to better results compared to using traditional datasets only. The second algorithm tested was XGBoost ML whose results for Area Under Curve scores, Accuracy, Precision and Recall were observed as 58%, 92.36%, 98.62%, and 92.36% respectively using traditional dataset only. On the other hand, scores of 73%, 94.77%, 98.62% and 94.77% respectively were observed when the model is trained using both traditional and alternative data. As observed from results of the two models, the use of alternative data has led to improved performance of both models. However, XGBoost has achieved better scores compared to LR suggesting responsiveness and stability in handling new data. These results reiterate the findings of Blazquez & Domenech (2018) on the importance of alternative data default prediction.

5.3 Conclusions

This study employed a pragmatic approach to simulate practical appraisal procedures for a SACCO in Kenya using scarce micro-level default data. The study calls for utilization of ML algorithms and borrower data beyond traditional features to improve the prediction of default in SACCO borrowers. By applying Machine Learning models, attributes of a borrower can be entered into the dynamic model, evaluated and accurate decisions about them made in a fast and non-subjective manner.

5.4 Contribution to Knowledge

The novelty and contribution of this study lies in the use of two emerging fields of alternative data and Machine Learning techniques. First, it demonstrates that SACCOs should embrace both traditional and alternative data when checking the probability of default for their members. Second, the study adds on the dearth of studies on credit risk assessment in SACCOs which have experienced increased defaults over time. Third, the study makes theoretical contribution by revealing the potential moral hazard that seems to manifest with loan guarantorship schemes employed by SACCOs. Fourth, the study makes some contribution in the ongoing discussions and implementation debates following IFRS 9 implementation and the role of Big Data in credit risk assessment. More specifically, the study shows that both traditional and alternative borrower features need to be considered when assessing the probability of default, and ML algorithms seem to provide better predictive ability in this regard. This would be very useful in determining the expected credit losses (ECLs) with improved precision compared to relying on purely traditional data and static modelling approaches which have cost financial institutions (Moradi & Mokhatab Rafiei, 2019).

5.5 Recommendations

The recommendations suggested includes the followings;

5.5.1 Recommendations for policy

Data availability is one of the areas where policymakers can help drive innovation in machine learning. The Government has a key role to play in encouraging open data initiatives by the creation of new open standards, for example shape and form for metadata. Government should explore ways to stimulate the safe and rapid delivery of these to support machine learning in Kenya. In areas where there are datasets unsuitable for general use, creation of policy frameworks or agreements which make data available to specific users under clear and binding legal constraints to safeguard its use, and set out acceptable uses.

The Government, industry and academic professionals should help ensure that relevant insights into ML are included in the education curriculum considering the educational needs of young people through the lens of implications of machine learning and associated technologies for the future. In addition to the relevant areas

in ML taught to future users, developers, and citizens, the ethical and social implications should be included within teaching activities.

Policies on Research funding should ensure that data handling, including the cost of preparing data and metadata, and other associated costs is supported as key part of research funding, and that researchers are actively encouraged to apply for funds to cover the same.

5.5.2 Recommendations for practice

Machine Learning approaches could be used to analyze SACCOS members' credit behaviors due to their efficiency and accuracy in prediction. Through ML techniques, credit risks can be reduced saving SACCOS of financial losses with the benefits being passed to members. As the study found out and discussed in the previous chapter, alternative data is useful component in predicting loan default and should be used in addition to traditional data in credit risk management to improve prediction accuracy and feature selection.

Secondly, ML usage leads to quicker processing compared to using human expertise which may also be subjective. This will save time which might be used to enhance other business developments. ML may also lead to long term competitive advantage due to quick data processing where ML is pioneering.

5.6 Areas of Further Research

This study was based on two supervised ML algorithms, LR and XGBoost algorithms. Future research may consider implementing other models. For example, Deep Neural Networks, Random Forest, Reinforced Learning, Ensemble Methods and Convolutional Neural Networks in analysing SACCOS loan default in Kenya. Furthermore, this study focused on predicting loan default in a Non-Deposit Taking SACCO. Collection of data from other SACCOS databases may be considered in order to acquire different features which could not be obtained in this research. Additionally, analysis software like MATLAB and R studio may be used for data analysis for comparison and validation.

While this study considered alternative data from client online portal in predicting loan default, other forms of alternative data, economic and other control factors can be incorporated in future studies. These factors include analysis of social media

activities, unemployment data, average household incomes, mortgage uptake data, inflation rate, consumer price index among others.

5.7 Limitations of the Research

5.7.1 Data availability

Access to clients' financial information is highly controlled and it is not clear to what extent the results of this study based on data from one SACCO generalize to the larger industry. Financial data is fragmented and mostly unreachable with private companies and other organizations holding much of it, and even publicly available data can be extremely difficult to access and combine for study. The problem of data availability is one of the areas where policymakers can help drive innovation in machine learning by encouraging open data initiatives. Such initiatives have been implemented elsewhere. For example, the European Data warehouse (ED) is a centralized securitization repository implemented by the European Central Bank (ECB) as part of the loan-level initiative that collects, validates, and distributes standardized loan-level data for several European countries. Through this vehicle, banks provide asset-backed securities as collateral in the ECB refinancing operations.

5.7.2 Legal and regulatory limitations

Implementation of Machine Learning algorithms and alternative data have raised concerns of confidentiality and ethical considerations in past studies. The Data Protection Act, 2019 have grey areas and does not prescribe the usage of other forms of alternative data especially social media data. This may lead to reputational damage and litigations. Another threat to privacy emanates from the ease to identify individuals in a de-identified dataset, for example by use of open source algorithms to back-calculate information about the individuals the algorithm was initially trained on. This is especially easy when the data obtained can be combined with other open source data like social media accounts by web scrapping or email addresses. Privacy therefore is a major challenge for machine learning policy, and protecting privacy is an area that regulations and legal frameworks have continued to lag behind and expose researchers to significant risks.

5.7.3 Scope of the study

This study was based on a SACCO in the private sector of a developing country. The researcher appreciates that there could be differences in borrower behaviors in other SACCOs. However, care was taken to ensure that most of the features examined in this study can be replicated elsewhere. The dataset was also based on a single period, and the results could vary due to economic shocks affecting SACCO members such as a financial crisis or a global pandemic like the Coiv-19 pandemic among others.

References

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. In *SSRN Electronic Journal* (Issue 08). https://doi.org/10.2139/ssrn.3155047

Aderitus, N. (2020). *Saccos Credit Rating Prediction in Tanzania By Using Machine Learning Approach : a Case of Kkkt Arusha Road Saccos Ltd . Masters of Science Information Technology the University of Dodoma*.

Adom, D., Hussein, E., & Adu-Agyem, J. (2018). Theoretical and Conceptual Framework: Mandatory Ingredients of a Quality Research. *International Journal of Scientific Research*, *7*(1), 438–441.

Akhtar, M. I. (2016). Research design Research design. *Research in Social Science: Interdisciplinary Perspectives*, *September*, 68–84.

Alex, I. (2019). *A Critical Analysis of Kenya's Legal Framework for Deposit Taking Saccos: Towards a More Efficient Regulatory Framework for Transparency and Accountability*.

Aslam, M., Kumar, S., & Sorooshian, S. (2020). Predicting likelihood for loan default among bank borrowers. *International Journal of Financial Research*, *11*(1), 318–328. https://doi.org/10.5430/ijfr.v11n1p318

Bacham, D., & Zhao, D. J. Y. (2017). *Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling*. Moody's Analytics Risk Perspectives.

Barbaglia, L., Manzan, S., & Tosetti, E. (2020). Forecasting Loan Default in Europe with Machine Learning. *SSRN Electronic Journal*, 1–33.

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*(September 2017), 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine Learning Explainability in Finance: An Application to Default Risk Analysis. *SSRN Electronic Journal*, *816*. https://doi.org/10.2139/ssrn.3435104

Chang, J. C. J., & King, W. R. (2005). Measuring the performance of information systems: A functional scorecard. In *Journal of Management Information Systems*. https://doi.org/10.1080/07421222.2003.11045833

Daoud, J. I. (2018). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, *949*(1). https://doi.org/10.1088/1742-6596/949/1/012009

David, M. (2020). Effect of Mobile Banking Services on Financial Performance of Deposit-Taking Saccos in Kenya. *International Journal of Social Science and Economic Research*, *05*(05), 1186–1222. https://doi.org/10.46609/ijsser.2020.v05i05.009

Davis. (2015). *Adoption And Integration Of Information And Communication Technology, And Performance Of Deposit Taking Sacco's In Nairobi City*

County. *December*.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/249008

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*. https://doi.org/10.1080/07421222.2003.11045748

Doll, W. J., Deng, X., Raghunathan, T. S., Torkzadeh, G., & Xia, W. (2004). The meaning and measurement of user satisfaction: A multigroup invariance analysis of the end-user computing satisfaction instrument. *Journal of Management Information Systems*. https://doi.org/10.1080/07421222.2004.11045789

Ereiz, Z. (2019a). Predicting Default Loans Using Machine Learning (OptiML). *27th Telecommunications Forum, TELFOR 2019, February*, 3–7. https://doi.org/10.1109/TELFOR48224.2019.8971110

Ereiz, Z. (2019b). Predicting Default Loans Using Machine Learning (OptiML). *27th Telecommunications Forum, TELFOR 2019, November 2019*, 3–7. https://doi.org/10.1109/TELFOR48224.2019.8971110

Faith, C. (2016). *Selected factors influencing financial performance of Savings & Credit cooperatives Societies in Kenya*. Kabarak University.

Gable, G. G., Sedera, D., & Chan, T. (2008). Re-conceptualizing information system success: The IS-impact measurement model. *Journal of the Association for Information Systems*. https://doi.org/10.17705/1jais.00164

Ghatak, M., & Guinnane, T. W. (1999). The economics of lending with joint liability: theory and practice. *Journal of Development Economics*, *60*(1), 195–228. https://doi.org/10.1016/S0304-3878(99)00041-3

Hamilton, S., & Chervany, N. L. (1981). Evaluating information system effectiveness - Part I: Comparing evaluation approaches. *MIS Quarterly: Management Information Systems*.

Hesbon, O. (2011). *Financial challenges facing savings and credit co-operative societies in kenya: The case of saccos in Nairobi* [University of Nairobi]. http://erepository.uonbi.ac.ke/handle/11295/12852

Iivari, J. (2005). An Empirical Test of the DeLone-McLean Model of Information System Success. *Data Base for Advances in Information Systems*. https://doi.org/10.1145/1066149.1066152

Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. *Communications of the ACM*. https://doi.org/10.1145/358413.358430

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform.

Financial Management, *48*(4), 1009–1029. https://doi.org/10.1111/fima.12295

Karagu, J. M., & Okibo, B. (2014). Financial Factors Influencing Performance of Savings and Credit Co- Operative Organization in Kenya. *International Journal of Academic Research in Accounting Finance and Management Sciences*, *4*(2), 295–306. https://doi.org/10.6007/IJARAFMS/v4-i2/892

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a researcKaushik, V. and Walsh, C. A. (2019) 'Pragmatism as a research paradigm and its implications for Social Work research', Social Sciences, 8(9). doi: 10.3390/socsci8090255.h paradigm and its implications for Social Work research. *Social Sciences*, *8*(9), 1–17.

Kengia, J. (2015). *Factors Influencing Loan Repayment Among (Saccos): a Case of Chamihado Saccos , Dodoma Factors Influencing Loan Repayment Among (Saccos): a Case of Chamihado Saccos , Dodoma*.

Kiefer, H., & Mayock, T. (2020). Why Do Models that Predict Failure Fail? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3616889

Li, Z., Li, S., Li, Z., Hu, Y., & Gao, H. (2021). Application of XGBoost in P2P Default Prediction. *Journal of Physics: Conference Series*, *1871*(1). https://doi.org/10.1088/1742-6596/1871/1/012115

Liu, Y., Yang, M., Wang, Y., Li, Y., & Xiong, T. (2022). Applying machine learning algorithms to predict default probability in the online credit market: Evidence from China. *International Review of Financial Analysis*, *79*, 101971. https://doi.org/10.1016/J.IRFA.2021.101971

Lucy, K. (2019). *Effects of adoption of technology on performance of DT Saccos*. *126*(1), 1–7.

Magali, J. J. (2013). Factors Affecting Credit Default Risks For Rural Savings and Credits Cooperative Societies (SACCOS) in Tanzania. *European Journal of Business and Management*, *5*(32), 60–73. http://www.iiste.org/Journals/index.php/EJBM/article/view/9559

Maina, J. N., Kinyariro, D. K., & Muturi, H. M. (2016). Influence Of Credit Risk Management Practices On Loan Delinquency In Savings And Credit Cooperative Societies In Meru County. *International Journal of Economics, Commerce & Management*, *IV*(2), 763–773.

Mapunda, L. (2019). *FACTORS AFFECTING LOAN REPAYMENTS EFFICIENCY AMONG SACCOs BORROWERS IN TANZANIA : A CASE OF SELECTED SACCOs IN KIBAHA TOWN COUNCIL*.

Mashange, G., Featherstone, A. M., & Briggeman, B. C. (2022). Evaluating changes in credit rating quality of U.S. farmer cooperatives. *Journal of Co-Operative Organization and Management*, *10*(1), 100153. https://doi.org/10.1016/J.JCOM.2021.100153

McKinney, V., Yoon, K., & Zahedi, F. (2002). The measurement of Web-customer satisfaction: An expectation and disconfirmation approach. *Information Systems Research*. https://doi.org/10.1287/isre.13.3.296.76

Mitei, A. (2016). *Determinants of Loan Default by Savings and Credit Co-Operative Societies ' Members in Baringo County , Kenya*. *8*(30), 158–164.

Moradi, S., & Mokhatab Rafiei, F. (2019). A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. *Financial Innovation*, *5*(1), 15. https://doi.org/10.1186/s40854-019-0121-9

Muchangi, D., Muathe, S., & Titus, S. (2019). Performance Analysis of Debit Card Services on Deposit-Taking SACCOs' Financial Performance: A Case of Kenya. *The African Journal of Information Systems*, *11*(2), 118. https://digitalcommons.kennesaw.edu/ajisAvailableat:https://digitalcommons.kennesaw.edu/ajis/vol11/iss2/3

Mwangi, D. K., & Ombui, K. (2018). Factors Affecting Financial Performance of Deposit Taking Saccos in Nairobi County, Kenya. *International Journal of Scientific and Research Publications (IJSRP)*, *8*(10), 153–159. https://doi.org/10.29322/ijsrp.8.10.2018.p8220

Nabavi, S. mohammed. (2019). *Effects of adoption of technology on performance of DT Saccos: K-Unity Sacco*. *2*. http://repositorio.unan.edu.ni/2986/1/5624.pdf

Nitani, M., & Legendre, N. (2021). Cooperative lenders and the performance of small business loans. *Journal of Banking & Finance*, *128*, 106125. https://doi.org/10.1016/J.JBANKFIN.2021.106125

Nyamasyo, K. T. (2018). *Determinants of Loan Defaults in Saccos in Kenya : a Case of Metropolitan National Sacco Ltd*.

Oynaka, N. N. (2020). Factors Influencing the Financial Performances of Saving and Credit Cooperative Societies in Case of Derash and Alle Woreda in SNNPRG, Ethiopia. *European Journal of Business and Management*, *12*(16), 32–46. https://doi.org/10.7176/ejbm/12-16-04

Papias, M., & Ganesan, P. (2009). Repayment behaviour in credit and savings cooperative societies. *International Journal of Social Economics*, *36*(5), 608–625. https://doi.org/10.1108/03068290910954059

Patroba, M. M., Kepha, O., Nyagol, M., & Odoyo, F. (2016). Influence Of Information Technology In Enhancement Of Sustainable Competitive Advantage Of Saccos In Kisii County. *IOSR Journal of Humanities and Social Science*, *21*(3), 103–117. https://doi.org/10.9790/0837-210301103117

Peter, K. (2019). *EFFECT OF FINANCIAL INNOVATIONS ON THE FINANCIAL PERFORMANCE OF SACCOS IN KENYA* (Issue December). Strathmore University.

Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2018). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *The Use of Big Data Analytics and Artificial Intelligence in Central Banking*, *50*(August), 30–31. https://www.bis.org/ifc/publ/ifcb49_49.pdf

Petter, S., DeLone, W., & McLean, E. (2008). Measuring information systems

success: Models, dimensions, measures, and interrelationships. *European Journal of Information Systems*, *17*(3), 236–263. https://doi.org/10.1057/ejis.2008.15

Porath, D. (2006). Estimating Probabilities of Default for German Savings Banks and Credit Cooperatives. *SSRN Electronic Journal*, *July*, 214–233. https://doi.org/10.2139/ssrn.2793958

Qinlan, Z., & Izumida, Y. (2013). Determinants of repayment performance of group lending in China. *China Agricultural Economic Review*, *5*(3), 328–341. https://doi.org/10.1108/CAER-08-2012-0083

Robert, N. (2019). *EFFECT OF BUSINESS OPERATIONAL FACTORS ONADOPTION OF E-BANKING BY SACCOS IN KISUMU COUNTY, KENYA* (Vol. 8, Issue 5). Maseno University.

Sabato, G. (2010). Credit Scoring. *Encyclopedia of Quantitative Finance*. https://doi.org/10.1002/9780470061602.eqf09019

Salaton, K. E., Gudda, P., & Rukaria, G. (2020). Effect of Loan Default Rate on Financial Performance of Savings and Credit Cooperative Societies Innarok, County Kenya. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, *10*(2), 65–75. https://doi.org/10.6007/ijarafms/v10-i2/7345

SAS. (2020). *Building Artificial Intelligence in Credit Risk : A Commercial Lending Perspective*. https://www.sas.com/en_us/html

SASRA. (2020). *The Sacco Supervision annual report 2019* (Vol. 254, Issue 20).

Sedera, D., Gable, G. G., & Chan, T. (2004). Measuring Enterprise Systems Success: The Importance of A Multiple Stakeholder Perspective. *ECIS*.

Shen, F., Wang, R., & Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, *26*(2), 405–429. https://doi.org/10.3846/tede.2019.11337

Timothy, G. (2015). *Adoption of mobile banking services by Nairobi county saccos*.

Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*, *7*(6), 191649. https://doi.org/10.1098/rsos.191649

Urbach & Müller. (2010). The Updated DeLone and McLean Model of Information Systems Success. *Springer*, *28*, 461. https://doi.org/10.1007/978-1-4419-6108-2

Vidal, M. F., & Barbon, F. (2019). *Credit Scoring in Financial Inclusion* (Issue July). www.cgap.org

Walusala, S., Rimiru, R., & Otieno, C. (2017). A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression. *International Journal of Computer (IJC) International Journal of Computer (IJC*, *27*(1), 84–

102. http://ijcjournal.org/

Wang, Y., Zhang, Y., Lu, Y., & Yu, X. (2020). A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data. *Procedia Computer Science*, *174*, 141–149. https://doi.org/10.1016/j.procs.2020.06.069

Xia, Y., He, L., Li, Y., Fu, Y., & Xu, Y. (2021). A dynamic credit scoring model based on survival gradient boosting decision tree approach. *Technological and Economic Development of Economy*, *27*(1), 96–119. https://doi.org/10.3846/tede.2020.13997

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zahi, S., & Achchab, B. (2020). Modeling car loan prepayment using supervised machine learning. *Procedia Computer Science*, *170*, 1128–1133. https://doi.org/10.1016/j.procs.2020.03.055

Zhang, X., Yu, P., Yan, J., & Ton A M Spil, I. (2015). Using diffusion of innovation theory to understand the factors impacting patient acceptance and use of consumer e-health innovations: A case study in a primary care clinic Healthcare needs and demand. *BMC Health Services Research*, *15*(1), 1–15. https://doi.org/10.1186/s12913-015-0726-2

APPENDIX 1: Authorization letter for research

Ole Sangale Rd, Madaraka Estate,
P.O Box 59857 00200, Nairobi, Kenya.
Cell: +254 703 414/6/7, Twitter: @SBSKenya
Email: info@sbs.ac.ke or visit www.sbs.strathmore.edu

Strathmore UNIVERSITY
BUSINESS SCHOOL

To Whom It May Concern

Dear Sir/Madam,

14th June 2021

Re: Facilitation Of Research – Silas Juma

This is to introduce Silas Juma who is a Master of Commerce (MCOM) Student at Strathmore University Business School, admission number MCOM/072286. As part of our MCOM Program, Silas is expected to do applied research and undertake a project. This is in partial fulfilment of the requirements of the MCOM course. To this effect, Silas would like to request for appropriate data from your organization.

Silas is undertaking a research paper on "**ROLE OF ALTERNATIVE DATA AND MACHINE LEARNING IN PREDICTING DEFAULT: CASE OF KENYA MEDICAL ASSOCIATION SACCO** ". The information obtained shall be treated confidentially and shall be used for academic purposes only.

Our MCOM seeks to establish links with industry, and one of these ways is by directing our research to areas that would be of direct use to industry. We would be glad to share our findings with you after the research, and we trust that you will find them of great interest and of practical value to your organization.

We appreciate your support and shall be willing to provide any further information if required.

Yours sincerely,

Njoki Kiagĩri
Associate Manager – Graduate Programs.
Strathmore University Business School.

Strathmore Business School is a Proud member of:

Association of African Business Schools AABS EFMD AACSB

APPENDIX 2: Ethical Review

Strathmore UNIVERSITY

13th July 2021

Mr Juma Silas,
juma.silas@strathmore.edu

Dear Mr Juma,

RE: Role of Alternative Data and Machine Learning in Predicting Default: Case of Kenya Medical Association Sacco

This is to inform you that SU-IERC has reviewed and **approved** your above **master's** research proposal. Your application reference number is **SU-IERC1083/21.** The approval period is **13th July 2021 to 12th July 2022.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 48 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 48 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and also obtain other clearances needed

Yours sincerely,

for: Dr Virginia Gichuru,
Secretary; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
13 Jul 2021
TEL: +254 (0)70303400
P.O BOX 59857-00200
NAIROBI-KENYA

Cc: Prof Fred Were,
Chairperson; SU-IERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

APPENDIX 3: NACOSTI License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **553793**

Date of Issue: **09/August/2021**

RESEARCH LICENSE

This is to Certify that Mr.. Silas Okeyo Juma of Strathmore University, has been licensed to conduct research in Nairobi on the topic: ROLE OF ALTERNATIVE DATA AND MACHINE LEARNING IN PREDICTING DEFAULT: CASE OF KENYA MEDICAL ASSOCIATION SACCO for the period ending : 09/August/2022.

License No: **NACOSTI/P/21/12287**

553793

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

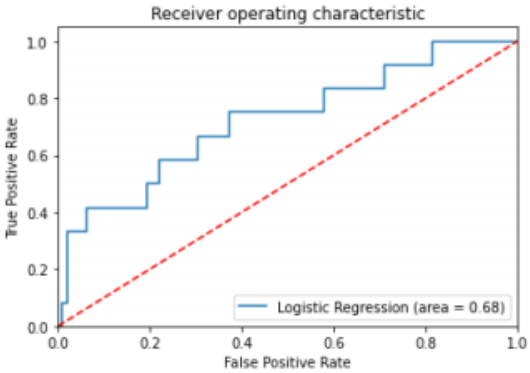
APPENDIX 4: Sample of Code

Model Evaluation

Using defined Matrices

Using Individual predictions with sample data

```
logit_roc_auc = roc_auc_score(y_test, model_logreg.predict(x_test))
fpr, tpr, thresholds = roc_curve(y_test, model_logreg.predict_proba(x_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()
```




```
# Import Confusion Matrix Package
from sklearn.metrics import confusion_matrix

# Compute confusion matrix
confusion_matrix = confusion_matrix(y_test, y_pred)

# Print Confusion Matrix
print(confusion_matrix)
```

```
[[101  44]
 [  4   8]]
```

B) Extreme Gradient Boosting (XGBoost)

Model 1 - Based on default Settings & Entire Dataset

```
X=Data3.drop(['loan_status'], axis=1)
Y= Data3['loan_status']

X.sample(5)
```

	age	membership_duration	savings	average_saving	loan_amount	loan_duration	tat	cummulative_loan_accounts	active_loar
147	0.013416	0.872790	-0.268848	-0.477608	0.617547	1.611196	0.926410	-0.585244	
677	-0.752573	-0.327219	-0.019919	0.162109	-0.225251	-0.630510	-0.208204	0.313865	
282	3.296226	-0.643011	-0.449267	-0.366749	-0.310304	0.490343	0.720117	-0.285541	
592	0.013416	0.114890	1.112420	1.149913	0.617547	-0.630510	-0.208204	0.014162	
114	-0.096011	-1.379858	-0.545547	0.188391	-0.464945	-0.630510	-0.053483	-0.884947	

```
#Split the data into training, test
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=0)
```

```
# manually handling imbalance of dependent variable Y
# Inversely assign the weights - since default is a small %tage, It gets more weight to balance
weight_ratio = float(len(Y_train[Y_train == 1]))/float(len(Y_train[Y_train == 0]))
w_array = np.array([1]*Y_train.shape[0])
w_array[Y_train==0] = weight_ratio
w_array[Y_train==1] = 1- weight_ratio

xgboost=xgb.XGBClassifier()
xgboost.fit(X_train, Y_train, sample_weight=w_array)
boost_score = xgboost.score(X_test,Y_test)
boost_score
```

```
[21:44:22] WARNING: ..\src\learner.cc:1061: Starting in XGBoost 1.3.0, the default evaluation metric used with
the objective 'binary:logistic' was changed from 'error' to 'logloss'. Explicitly set eval metric if you'd like
to restore the old behavior.
C:\ProgramData\Anaconda3\lib\site-packages\xgboost\sklearn.py:888: UserWarning: The use of label encoder in XGB
Classifier is deprecated and will be removed in a future release. To remove this warning, do the following: 1)
Pass option use_label_encoder=False when constructing XGBClassifier object; and 2) Encode your labels (y) as in
tegers starting with 0, i.e. 0, 1, 2, ..., [num_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)
0.9235668789808917
```

```
print("Weight assigned to Non default is,", weight_ratio)
print("Weight assigned to Default is,", 1-weight_ratio)
```

```
Weight assigned to Non default is, 0.043333333333333335
Weight assigned to Default is, 0.9566666666666667
```

```
Y_pred = xgboost.predict(X_train)
accuracy = xgboost.score(X_train, Y_train)
print("Accuracy: %.2f%%" % (accuracy * 100.0))

# make predictions for test data
Y_pred = xgboost.predict(X_test)
accuracy = xgboost.score(X_test, Y_pred)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

Feature Selection

Recursive feature elimination (RFE): RFE is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_ attribute or through a feature importance attribute. Then, the least important features are pruned from current set of features.This procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

```
#from sklearn import datasets
from sklearn.feature_selection import RFE

#Deploying Model
logreg = LogisticRegression(class_weight={0:0.04,1:0.96},random_state=0, solver = 'lbfgs')
rfe = RFE(logreg, 20)

#data_sel = x.columns.values[np.where(rfe.ranking_==1)].tolist()

# Fit in the algorithm
rfe = rfe.fit(x,y.values.ravel())

# We will consider for True Values
print(rfe.support_)
# Print Ranking - If ranking is 1 means selected, other values doesnt carry much weigh
SupportRFE=rfe.support_
RankingRFE=rfe.ranking_
buffer=[]
index=0
for i in RankingRFE:
    if(i!=1):
        buffer.append(x.columns[index])
    index+=1
data_sel=x.drop(buffer,axis=1)
data_sel.columns
#data_sel.head()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\utils\validation.py:70: FutureWarning: Pass n_features_to_se
lect=20 as keyword args. From version 1.0 (renaming of 0.25) passing these as positional arguments will result
in an error
  warnings.warn(f"Pass {args_msg} as keyword args. From version "
[False False  True False  True False False False False False False  True
  True False  True False False False False False False  True False False
 False False False False False False False False False False False False
  True  True False  True False  True False False False  True  True  True
 False  True False  True False False False False False False False False
 False False False  True False False False False  True  True False False
  True False  True False False False]
Index(['savings', 'loan_amount', 'accd_dashboard ', 'accd_dividend_advice ',
       'accd_faqs ', 'accd_policies ', 'login_failed', 'accd_pass_recovery ',
       'accd_loans_guaranteed ', 'statements_view_download', 'sex_1', 'sex_2',
       'lnty_1', 'lnty_3', 'lnty_5', 'ismth_3', 'ismth_7', 'ismth_8',
       'ismth_12', 'recmth_2'],
      dtype='object')
```

Hyperparameter Tuning

In [79]:

```
#Selection of parameters for this model
model_xgboost = xgb.XGBClassifier(learning_rate=0.1,
                                  max_depth=4,
                                  n_estimators=50,
                                  subsample=0.7,
                                  colsample_bytree=0.5,
                                  eval_metric='auc',
                                  gamma=0,
                                  verbosity=1)

eval_set = [(Xx_train, yy_train)]

model_xgboost.fit(Xx_train,
                  yy_train,
                  early_stopping_rounds=10,
                  eval_set=eval_set,
                  verbose=True)
```

```
[0]     validation_0-auc:0.76862
[1]     validation_0-auc:0.80447
[2]     validation_0-auc:0.85447
[3]     validation_0-auc:0.88641
[4]     validation_0-auc:0.88135
[5]     validation_0-auc:0.87416
[6]     validation_0-auc:0.88138
[7]     validation_0-auc:0.88423
[8]     validation_0-auc:0.88375
[9]     validation_0-auc:0.95512
[10]    validation_0-auc:0.96488
[11]    validation_0-auc:0.96415
[12]    validation_0-auc:0.96700
[13]    validation_0-auc:0.97461
[14]    validation_0-auc:0.97584
[15]    validation_0-auc:0.97992
[16]    validation_0-auc:0.98163
[17]    validation_0-auc:0.98104
[18]    validation_0-auc:0.98082
[19]    validation_0-auc:0.98238
[20]    validation_0-auc:0.98255
[21]    validation_0-auc:0.98065
[22]    validation_0-auc:0.98179
[23]    validation_0-auc:0.98252
[24]    validation_0-auc:0.98361
[25]    validation_0-auc:0.98384
[26]    validation_0-auc:0.98523
[27]    validation_0-auc:0.98518
[28]    validation_0-auc:0.98563
[29]    validation_0-auc:0.98468
[30]    validation_0-auc:0.98624
[31]    validation_0-auc:0.98674
[32]    validation_0-auc:0.98697
[33]    validation_0-auc:0.98641
[34]    validation_0-auc:0.98647
[35]    validation_0-auc:0.98630
[36]    validation_0-auc:0.98579
[37]    validation_0-auc:0.98691
[38]    validation_0-auc:0.98719
```

```
C:\ProgramData\Anaconda3\lib\site-packages\xgboost\sklearn.py:888: UserWarning
s deprecated and will be removed in a future release. To remove this warning,
coder=False when constructing XGBClassifier object; and 2) Encode your labels
```