DESIGN AND IMPLE MENTATION OF A CUSTOMER CHURN PREDICTION MODEL FOR ZIMBABWEAN BANKS USING MACHINE LEARNING TECHNIQUES.

AFRICA UNIVERSITY
(A United Methodist-Related Institution)

2024

DESIGN AND IMPLEMENTATION OF A CUSTOMER CHURN PREDICTION MODEL FOR ZIMBABWEAN BANKS USING MACHINE LEARNING TECHNIQUES.

BY

ANESU SHARLEEN MAKONI.

A DISSERTATION
SUBMITTED TO COLLEGE OF ENGINEERING AND APPLIED SCIENCES
AFRICA UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE BSC HONORS IN COMPUTER INFORMATION SYSTEMS 2024.

Abstract

The loss of customers in the banking industry due to account closures or the termination of banking services is known as customer churn. This could have serious repercussions for banks, such as decreased revenue, increased expenses to draw in new customers, and harm to their reputation. This study explores the problem of customer churn in Zimbabwean banks and looks into the efficacy of traditional machine learning algorithms in order to forecast and analyze customer churn.

The study uses a dataset of customer data that includes demographic and banking-related features to train and assess a number of models, including logistic regression, decision trees, random forests, and XGBoost. The results show that every model was able to accurately forecast customer attrition.

The study illustrates the value of anticipating customer attrition in the banking industry and shows how well machine learning algorithms perform in this regard. The study does point out certain limitations, though, such as the dataset's limited feature set and small sample size, which could limit how broadly the results can be applied.

The study's conclusion addresses how the results might be applied to banks in Zimbabwe and offers possible avenues for further investigation.

Keywords: Customer Churn, Banking Industry, Machine Learning, Predictive Modeling, Financial Services.

Declaration

I affirm that this dissertation proposal is entirely my own creation, with the exception of duly cited and acknowledged sources. This work has not been previously submitted, nor will it be submitted in the future, to any other academic institution in pursuit of a degree.

………*A.S.Makoni*……… …………30/07/2024…………….

ANESU MAKONI Date (Student)

………… …………30/07/2024………

MR B. MUKHALELA. Date (Supervisor)

Copyright

Acknowledgement

I want to start by giving thanks to God for leading me on this journey. This dissertation was completed successfully, and for that I am sincerely grateful to everyone who helped.

I want to express my profound gratitude to Africa University for providing the funding and learning opportunities that have enabled me to pursue my academic goals. My knowledge and proficiency in data analysis and machine learning have been greatly influenced by the university's resources and stimulating environment.

My supervisor, Mr. Mukhalela, has my sincere gratitude for his invaluable advice, assistance, and knowledge, all of which were crucial to my research process. Your helpful criticism and insights have significantly improved the caliber of this work, especially in the development and application of the customer churn prediction model.

I would especially like to thank the stakeholders and banking professionals who took part in my study. My knowledge has been enhanced and this research has been made possible by your willingness to share your experiences and insights about customer churn in Zimbabwean banks.

Additionally, I would like to thank the internet sources that made the dataset available for this study. We sincerely appreciate your help in creating a machine learning-based customer churn prediction model for Zimbabwean banks.

Finally, I would like to thank my family for their unwavering love and support. My greatest strength has been your faith in my abilities and your willingness to make sacrifices. I appreciate you being my pillar of support and being there for me at every turn.

All of you have given me encouragement and support, which is reflected in this dissertation. I appreciate your vital role in this journey.

Dedication

This dissertation is dedicated to my mother, whose endlessly supportive presence and encouraging words have been a constant source of inspiration. She has always believed in me, reminding me that the sky is the limit and that with hard work, I can achieve anything I set my mind to. Her unwavering care, wisdom and life lessons have been instrumental in shaping me into the person I am today. I am more than grateful for her guiding influence and for the privilege of having her in my life. This work is a testament to her profound impact.

List of Acronyms and Abbreviations

ROI – Return on Investment.

CRM – Customer Relationship Management.

B2B – Business-to-Business.

B2C – Business-to-Consumer

CLV – Customer Lifetime Value.

TPB – Theory of Planned Behavior.

GBDT – Gradient Boosted Decision Trees.

DSRM – Design Science Research Methodology.

DSR – Design Science Research

XGBoost – Extreme Gradient Boost.

ROC-AUC – The Receiver Operating Characteristic – Area Under the Curve.

KNN – K-Nearest Neighbors

RF- Random Forest

SVM- Support Vector Machine.

CrCard – Credit Card.

Contents

CHAPTER 1 INTRODUCTION

1.1 Overview

Businesses around the world have been compelled to adjust as a result of the development and digitalization of the world. One result of the rapid digitalization that has swept the globe is banking services, which present both opportunities and problems that call for contemporary solutions. In addition to altering how business is performed, digitalization has also increased the quantity of banking services available to customers due to the abundance of information available. This could be seen as a difficulty for businesses because it might get harder to keep clients. The market has become very dynamic and highly competitive nowadays because of the availability of a large number of service providers in the banking sector in Zimbabwe. According to (Manas Rahman and Kumar, 2020), The challenges of service providers are finding the changing customer behaviour and their rising expectations.

Consumers of the current age in Zimbabwe have higher aspirations than those of earlier generations, and they have a wider range of connectivity and innovative, tailored approach needs. They are better informed of emerging approaches to the banking services. Such sophisticated knowledge has altered their shopping habits, leading to a tendency of "analysis-paralysis" where they excessively consider the selling and purchasing scenario, ultimately improving their purchasing selections. As a result, the new generation of banking service providers face a significant difficulty in coming up with creative solutions to satisfy and provide value for their clients. The banking institutions need to recognize their consumers. (Liu and Shih, 2021)supported this argument by implying that increasing competitive pressures on organizations to develop innovative marketing approaches, to meet consumer expectations and enhance loyalty and retention. The digital transformation has been used to help businesses and banks to retain a competitive edge in this 4IR (Forth Industrial Revolution) era. The machine learning and data mining techniques are a commonly used information technology for the extraction of marketing expertise and further guidance for business decisions.

Customers can change organizations (Banks) fairly easily in search of higher service levels or lower costs. Companies believe that acquiring new clients is much more expensive and difficult than maintaining current ones but delivering reliable service on time and in budget to customers while maintaining a good working partnership with them is another significant challenge for them. To address these issues, they must take into account consumers' needs. Among these, one of their primary focus will be on client churn. According to (M Rahman, 2020), Customer churn takes place when clients or subscribers cease to engage incorporation with a company or service. In Zimbabwe there is an abundance of different banks to choose from, which has increased competition and made it more difficult to retain customers for Steward Bank Limited in this modern-day service market. It is now the duty for the bank to adjust by concentrating more on customer relationship management (CRM), particularly customer churn management, due to the availability of big data and substitutes. (Yasin Osman, 2021) bolstered this claim by suggesting that maintaining a low churn rate—defined as the number of clients quitting their service provider within a specific time frame—is essential to success in a subscription-based business, such as banks. A local bank is more likely to sell its goods and services to an existing client than to a brand-new one.

The expense of acquiring new customers can serve as a good example of this. Attracting new clients can be five to six times more expensive than keeping existing ones, according to Verbeke et al. (2022), underscoring the significance of reducing customer attrition at neighborhood banks.

Aside from the expense of attracting and keeping consumers, retention-related initiatives typically yield larger net returns on investments (ROI) than acquisition-related strategies, which could boost revenue from existing customers. Taking into account of one of the local banks, Steward Bank limited, in the past three years it lost a very huge number of its clients to other banks like BancABC, CABS and NMB due to different factors surrounding the poor service delivery by the bank. There has been a decrease in the number of low KYC accounts(ISAVE) customer transactions as they switched to open accounts with other banks due to a rise in complaints on Steward Bank Limited poor services. The Stewardbank financial 2021 AUDITED ABRIDGED FINANCIAL STATEMENTS show that the inflation-adjusted net operating income was down 34% from the previous year, with a loss of ZWL903 million compared to a loss of ZWL773 million. In order to try keep its client, the bank embarked on the upgrading of the Core Banking System, and its digital products to improve customer experience but most of its client already opened accounts with other banks.

Currently the bank is implementing marketing strategies to attract new customers which comes with extra costs than if it has just predicted the customers who were going to leave the bank and find ways to keep getting services from the bank. Moreso, offering retention incentives is the main tactic utilized in customer retention tactics to decrease churn. When creating retention strategies, businesses shouldn't concentrate on the entire customer base because these tactics are not free and not all customers have the motivation to leave.

Avoid wasting money and resources by instead identifying the consumers who are most likely to leave and offering them incentives to stay. Preventing churn is more important in business-to-business (B2B) settings, such as NMB/CABS to OK supermarkets, where each customer makes more frequent, higher-value transactions. Retention in this situation may yield significant financial rewards because of the high transactional value that B2B clients represent. A bank's reputation can be greatly impacted by churn prediction and prevention, and a better reputation boosts revenue (Amin et al., 2013). Since customers are the source of profits, a reduction in the churn rate has a big effect on the company's earnings. Typically, churn identification is thought of as a prediction problem. In recent years, machine learning has been applied in many academic fields and has proven useful for a variety of difficult problems.

1.2 Statement of the problem

Customer churn is a significant problem faced by Zimbabwean banks, as it results in a loss of revenue and customer loyalty. Despite their efforts to retain customers, the banks continue to experience a high churn rate especially the Steward Bank Limited, highlighting the need for a more effective approach to customer retention. The problem addressed in this study is the inability for banks to accurately predict customer churn, leading to a lack of targeted retention strategies. The banks rely on traditional methods of customer retention such as mass marketing and promotional offers, which are not effective in retaining high-risk customers. There is a lack of specific research on customer churn prediction in the context of local banks in Zimbabwe. Most studies have focused on developed economies with limited research conducted in the context of Zimbabwean banks. The need for a study that explores the use of machine learning techniques in customer churn prediction for Zimbabwean banks is highlighted by this research gap.

Aim of the study

Design and Implementation of a Customer Churn Prediction Model for Zimbabwean Banks using machine learning techniques.

1.3 Research objectives

a) To explore the key factors contributing to customer churn from Zimbabwean Banks.

b) To use algorithms like logistic regression, decision trees, and random forests to create and implement a machine learning-based customer churn prediction model for banks in Zimbabwe.

c) To assess the performance of the developed model in predicting customer churn.

d) To investigate the effectiveness of the developed model in identifying high-risk customers and reducing customer churn at within Zimbabwean Banks.

1.4 Research Questions

a) What are the key factors contributing to customer churn from Zimbabwean Banks?

b) Which machine learning techniques can be applied to help Zimbabwean banks predict customer attrition?

c) How accurate is the developed model in predicting customer churn?

d) What is the effectiveness of the developed model in identifying high-risk customers and reducing customer churn from Zimbabwean Banks?

1.4 Assumptions/Hypothesis:

1.1.1 Assumptions:

a) Customer churn is a significant problem for Zimbabwean banks and the banks have access to relevant data on customer behavior and attributes.

b) Machine learning techniques can be effectively applied to customer churn prediction and the developed model will be able to identify high-risk customers with a high degree of accuracy.

c) The data used for model development is representative of the bank's customer base and the results can be generalized to the entire customer population.

d) The bank has the necessary resources and infrastructure to implement the developed model and integrate it into their existing customer retention strategies.

1.6.2 Hypothesis:

H_0**:** Machine Learning techniques have high precision on predicting customer churn from Zimbabwean Banks.

$H_{1:}$ Machine Learning techniques have low precision on predicting customer churn from Zimbabwean Banks.

1.5 Significance of the Study

1.1.2 Significance of the study to the researcher

a) Contribution to expertise: this study demonstrates the researcher's expertise in machine learning techniques and customer churn prediction, enhancing their reputation as a researcher in this field.

b) Development of research skills: conducting this study has honed the researcher's research skills, including data collection analysis and interpretation as well as critically evaluating existing literature.

c) Opportunities for future research: this study provides opportunities for collaboration with industry professionals and academics, expanding the researcher's professional network and potential for future research partnerships.

d) Collaboration and networking: this study provides opportunities for collaboration with industry professionals and academics, expanding the researcher's professional network and potential for future research partnerships.

e) Publication and dissemination: publishing this study in a reputable journal and presenting it at conferences will disseminate the researcher's findings to a wider audience increasing their visibility as a researcher and potential for citations.

f) Professional growth: completing this study demonstrates the researcher's ability to design, implement and evaluate a complex research project, enhancing their confidence and self-efficacy as a researcher.

g) Potential for awards and recognition: this study may be eligible for awards or recognition in the field of machine learning and customer churn prediction, further enhancing the researcher's reputation and career prospects.

1.1.3 Significance of the study to other financial institutions

a) Improved customer experience: predicting customer churn allows financial institutions to proactively address potential issues and provide better customer experiences. By identifying patterns and factors leading to customer churn, banks can make data-driven improvements in their products, services and customer interactions, ultimately enhancing overall customer satisfaction.

b) The implementation of a strong customer churn prediction model can give financial institutions a competitive edge. Banks can maintain a strong customer base, grow their market share, and stand out in a highly competitive market by successfully keeping their customers.

c) Strategic decision making: the insights gained from customer churn prediction models can inform strategic decision-making processes within financial institutions. By understanding customer behavior and preferences, banks can tailor their marketing campaigns, product offerings and customer engagement strategies to align with customer needs, resulting in better business outcomes.

d) Contribute to research: this study contributes to the existing body of knowledge in customer churn prediction within the banking sector. The methodologies and techniques employed in this research can serve as a reference for future studies and encourage further exploration in this domain.

e) Competitive advantage: Financial institutions can gain a competitive edge by putting in place a strong customer churn prediction model. Banks can stand out in a very competitive market, grow their market share, and keep a solid customer base by successfully keeping customers.

f) Contribute to the body of knowledge already available in the banking industry regarding the prediction of customer attrition, as this study does. The methods and strategies used in this study can spur more research in this area and be used as a guide for subsequent investigations.

1.1.4 Significance of study to marketing professionals

a) Improved customer retention strategies; the findings of this study can provide marketing professionals in other industries with insights on designing effective customer retention strategies. By accurately predicting customer churn, marketing professionals can proactively engage with customers at risk of churning, offering personalized incentives, loyalty programs or targeted marketing campaigns to increase customer loyalty and reduce churn rates.

b) Enhanced customer segmentation: customer churn prediction models can help marketing professionals identify different segments of customers based on their likelihood to churn. This segmentation enables marketers to tailor their messaging, promotions and engagement strategies for each segment, effectively addressing their specific needs and preferences. By delivering relevant and targeted marketing efforts, marketing professionals can improve customer satisfaction and retention rates.

c) Resource allocation optimization: by predicting customer churn, marketing professionals can allocate their resources more efficiently. Instead of investing heavily in acquiring new customers. This shift in resource allocation can lead to cost savings and improved return on investment (ROI) for marketing campaigns.

d) Customer lifetime value optimization: customer churn has a direct impact on the lifetime value of customers. By accurately predicting churn, marketing professionals can identify high-value customers who are at risk of leaving. These customers can be targeted with retention strategies to maximize their lifetime value, resulting in increased profitability for the company.

e) Competitive advantage: implementing a robust churn prediction model gives marketing professionals a competitive advantage. By proactively identifying customers likely to churn, marketing professionals can implement retention strategies ahead of competitors, reducing customer attrition and maintaining a larger share of the market.

f) Data-driven decision making: customer churn prediction models provide valuable insights into customer behavior and preferences. Marketing professionals can use these insights to make data-driven decisions on product development, pricing, messaging and customer experience improvements. This approach ensures that marketing efforts are aligned with customer needs, resulting in more effective strategies and improved business outcomes.

g) Contribution to marketing research: this study contributes to the existing body of knowledge in customer churn prediction within the marketing field. The methodologies and techniques employed in this research for marketing professionals in other industries encouraging further research and exploration in the domain of customer retention.

1.1.5 Significance of study to customer service teams

a) Proactive customer retention: the findings of this study can significantly benefit customer service teams in other industries by enabling them to adopt a proactive approach to customer retention. By accurately predicting customer churn, customer service teams can identify at-risk customers and take proactive measures to address their concerns, provide personalized support and offer tailored solutions. This approach helps to mitigate churn and provide customer satisfaction.

b) Customer experience enhancement: customer churn prediction models provide valuable insights into customer behavior and preferences. By leveraging these insights, customer service teams can enhance the overall customer experience. The identified patterns and factors leading to customer churn can be utilized to identify pain points in the customer journey, improve service quality and optimize customer interactions, thereby increasing customer satisfaction and loyalty.

c) Resource allocation optimization: predicting customer churn allows customer service teams to allocate their resources efficiently. By identifying customers who are likely to churn, teams can prioritize their efforts and allocate resources accordingly. This targeted resource allocation helps allocate resources accordingly. This targeted resource allocation helps in optimizing manpower, time and budget ensuring that the team focuses on retaining valuable customers rather than spreading resources thinly across the entire customer base.

d) Tailored customer engagement: customer churn prediction models enable customer service teams to segment customers based on their churn likelihood. This segmentation allows for tailored engagement strategies for different customer segments. By understanding the specific needs and preferences of each segment, customer service teams can provide personalized support, targeted offers and relevant communication, resulting in a more effective and impactful customer engagement.

e) Retention strategy evaluation: the implementation of a customer churn prediction model enables customer service teams to evaluate the effectiveness of their retention strategies. By comparing the predicted the predicted churn outcomes with the actual churn events, teams can assess the performance of different strategies and identify areas of improvement. This data driven evaluation helps teams refine their retention approaches and optimize their efforts in reducing churn.

f) Competitive advantage: implementing a robust churn prediction model gives customer service teams a competitive advantage. By actively working to retain customers, teams can differentiate their organization from competitors. This advantage leads to increased customer satisfaction, loyalty and ultimately market share.

g) Contribution to customer service research: this study contributes to the existing body of knowledge in customer churn prediction within the customer service discipline. The methodologies and techniques employed in this research can serve as a reference for customer service teams in other industries, encouraging further research and exploration in the domain of customer retention and churn prediction.

1.1.6 Significance of study to data analysts and scientists

a) Methodological insights: this study provides valuable methodological insights for data analysts and scientists in the field of customer churn prediction. It offers a detailed description of the data collection process, preprocessing techniques, feature selection and engineering methods, model selection, evaluation and implementation. By referencing this study, data analysts and scientists can gain a deeper understanding of the practical steps involved in designing and implementing a customer churn prediction model using machine learning techniques.

b) Predictive modeling techniques: the study showcases the application of various machine learning algorithms for customer churn prediction. Data analysts and scientists can benefit from the analysis and comparison of different models, as well as the performance evaluation metrics used in this study. This information can guide them in selecting appropriate algorithms and evaluation methods for their own predictive modeling projects.

c) Feature importance analysis: the study conducts a feature importance analysis to identify the key factors influencing customer churn. This analysis can provide data analysts and scientists with insights into the drivers of churn in different contexts. By understanding these factors, analysts can focus their efforts on collecting and analyzing relevant data, improving the accuracy and interpretability of their churn prediction models.

d) Practical implementation considerations: this study explores the practical implementation of the churn prediction model in a real-world business setting. It discusses the challenges, limitations and considerations that arise during the implementation process. Data analysts and scientists can learn from these insights and better prepare for similar challenges when deploying predictive models in their own organizations.

e) Business impact assessment: the study examines the business impact of the churn prediction model for Steward Bank Limited. It evaluates the effectiveness of the model in terms of customer retention and provides insights into the potential financial benefits of implementing such a model. Data analysts and scientists can use this information to assess the potential return on investment (ROI) and business value of implementing churn prediction models in their respective industries.

f) Contribution to data science research: this study contribute to the existing body of knowledge in customer churn prediction within the field of data science. The methodologies, techniques and empirical findings presented in this research can serve as a reference for other researchers and practitioners in the field. It encourages further research and exploration into the development and application of predictive models for customer churn prediction.

1.6 Delimitations of the study

a) Focus on customer churn: this study specifically focuses on predicting customer churn within the context of Zimbabwean Banks. It does not address other aspects of customer behavior or operational challenges that the banks might face. The study delimits its scope to understanding and predicting customer churn only.

b) Data availability and quality: the study's findings and predictions are limited by the availability and quality of data provided by local bank customers. The accuracy and reliability of the churn prediction model depend on the completeness, accuracy and relevance of the historical customer data used for analysis. Any limitations or gaps in the data could impact the accuracy of the model and subsequent predictions.

c) Machine learning techniques: this study employs machine learning techniques for customer churn, the choice of models and techniques used in this study might have limitations. Other machine learning algorithms or advanced techniques beyond the scope of this study could potentially yield different results.

d) Time constraints: this study is conducted within a specified timeframe and with a specific amount of historical data available. The findings and predictions are based on the data and conditions up until the study's cutoff date. Changes in customer behaviorg3, market conditions or bank strategies beyond this timeframe might not be accounted for in the analysis.

e) External factors: the study does not take into account external factors such as macroeconomic conditions, regulatory changes or competitive landscape shifts that could influence customer churn. While these factors can significantly impact customer behavior, the study focuses solely on internal data and does not consider external variables that might affect churn prediction accuracy.

f) Ethical considerations: the study adheres to ethical guidelines and data privacy regulations. Any personal or sensitive information used for analysis is anonymized and treated with strict confidentiality. The study does not explore or analyze customer data beyond the scope necessary for churn prediction, ensuring the privacy and security of customer information.

1.7 Limitations of the study

a) Sample size: the effectiveness and generalizability of the customer churn prediction model developed in this study could be influenced by the size of the dataset available from local banks customers. If the dataset is limited in terms of the number of customers or the duration of historical data, the predictive power of the model may be affected. A larger and more diverse datasets could provide more accurate and robust predictions.

b) Data quality; the accuracy and reliability of the churn prediction model heavily rely on the quality of the data provided by local banks. Incomplete, inconsistent or erroneous data can introduce biases and affect the performance of the model. While efforts are made to preprocess and clean the data, it is essential to acknowledge that data quality limitations could impact the validity of the findings.

c) Variable selection: the selections of variables or features used in the churn prediction model is crucial. The effectiveness of the model is contingent on identifying the most relevant and significant factors influencing churn. While the study employs feature selection techniques, there is a possibility that some important variables might be overlooked or excluded, potentially affecting the model's predictive accuracy.

d) Assumptions of stationarity: the churn prediction model assumes that the underlying patterns and relationships between customer attributes and churn remain constant over time. However, customer behavior, market dynamics and internal business factors may change over time, leading to a violation of this assumption. The model's accuracy may decrease if significant shifts occur in the underlying patterns during the model's deployment period.

e) Model Overfitting or Under-fitting: when developing a machine learning model, there is a risk of either over-fitting or under-fitting the data. Over-fitting occurs when the model performs well on the training data but fails to generalize to unseen data. Under-fitting on the other hand, happens when the model is too simplistic and fails to capture the complex relationships within the data. The study attempts to mitigate these risks through appropriate model evaluation and validation techniques, but it is important to acknowledge but it is important to acknowledge that model performance can still be influenced by these issues.

f) Lack of causality: the customer churn prediction model developed in this study focuses on identifying patterns and making predictions based on historical data. However, the model does not establish causal relationships between customer attributes and churn. While the model can provide insights into the factors associated with churn, it cannot definitively determine the cause-and-effect relationships between those factors and churn events.

g) External validity: the findings and conclusions of this study are specific to Steward Bank Limited and may not de directly applicable to other banks or financial institutions. The specific characteristics of bank's customer base, market conditions and business strategies may limit the generalizability of the results to different contexts. Replication of the study in other organizations is necessary to validate the findings and evaluate their applicability in diverse settings.

CHAPTER 2 REVIEW OF RELATED LITERATURE

2.1 Introduction

In the context of Zimbabwean banks, this chapter seeks to provide a thorough theoretical and empirical basis for the development and application of a machine learning-based customer churn prediction model. This chapter's literature review will cover and look into the following important areas:

a) Defining customer churn and its significance in the banking sector: This review will examine the idea of customer churn, how it hurts banks' competitiveness and profitability, and how important it is to have efficient customer retention plans.

b) Theoretical frameworks for understanding customer churn: the chapter will examine the relevant theoretical perspectives, such as the customer lifetime value (CLV) theory and the theory of planned behavior (TPB), which provide a foundation for understanding the factors influencing customer churn.

c) Application of machine learning in customer churn prediction: the review will synthesize the existing research on the use of machine learning techniques, such as classification algorithms, decision trees and neural networks in developing accurate and reliable customer churn prediction models.

d) Relevance of the theoretical and empirical literature to the proposed study. The review will critically analyze the applicability of the existing knowledge to the specific context of Zimbabwean Banks., identifying gaps and opportunities for furthering the research in this domain.

By providing a comprehensive review of the relevant literature, this chapter will establish the theoretical and empirical foundations necessary for the design and implementation of a customer churn prediction model tailored to the needs and challenges faced by local banks. The insights gained from this review will guide the research methodology, the selection of appropriate machine learning techniques and the interpretation of the study's findings.

2.2 Theoretical Framework

The proposed research on customer churn prediction for Zimbabwean Banks is grounded in the theoretical framework of customer relationship management (CRM), customer lifetime value (CLV) and the application of machine learning techniques.

a) Customer Relationship Management (CRM) Framework:

   CRM is a strategic approach that focuses on building and maintaining long-term relationships with customers (Payne & Frow, 2005). It involves the systematic collection and analysis of customer data to better understand their needs, preferences and behaviors (Ngai, 2005). In the banking and financial services industry, CRM has become increasingly important, as it enables organizations to identify and retain their most valuable customers, thereby improving profitability and competitive advantage (Ngai et al., 2009).

b) Customer lifetime value (CLV) Framework:

   CLV is a fundamental concept in CRM that refers to the total net profit a company can expect from a customer over the entire duration of their relationship (Gupta et al., 2006). Accurately predicting and maximizing CLV is essential for banks, as it allows them to focus their resources on retaining and nurturing their most valuable customers (Kumar &Reinartz, 2016).

c) Machine learning in bank churn prediction:

The use of machine learning techniques for customer churn prediction is rooted in the ability of these algorithms to detect patterns and trends in large customer datasets, which can be used to accurately predict the likelihood of customer attrition (Coussement & Van den Poel, 2008). Various machine learning techniques such as logistic regression, decision trees, support vector machines and ensemble methods, have been successfully applied in the context of customer churn prediction (Verbeke et al., 2012). The integration of CRM, CLV and machine learning provides a powerful framework for a local bank to develop a comprehensive customer churn prediction model, which can be used to identify at-risk customers and implement targeted retention strategies to maximize the bank's long-term profitability.

Customer churn is more than just a catchphrase in the banking industry; it's an impending catastrophe. It is urgently necessary to solve this issue, as financial institutions lose a startling 25–30% of their customer base annually, including bank churners. (Anon, 2022) However, how can one forecast something as erratic as human nature? To forecast client attrition, machine learning is used in this situation. This program identifies clients who are most likely to shop elsewhere by sorting through massive amounts of data, including transaction histories and credit scores. Furthermore, machine learning algorithms recommend focused actions in addition to projections in order to keep these at-risk clients.

Impact of Churn on Banking Profitability

A bank's bottom line is directly impacted by customer attrition. The bank incurs additional fees to acquire new clients in addition to losing the money earned by that customer's departure. Predictive bank churn can assist in identifying clients who are at danger, enabling the bank to take precautionary action. Profitability depends on the cost of churn prediction because it is less expensive to keep an existing client than to get a new one.

Significance of Churn Prediction in Banks and Financial Institutions

For banks and other financial institutions, churn prediction is essential as it facilitates proactive customer retention. Banks are able to recognize the warning signals early on by utilizing machine learning algorithms for churn prediction. This keeps their revenue stream consistent by allowing them to move quickly to keep customers from leaving. In today's competitive market, predicting bank client attrition is not just a necessary strategy, but also a must.

.Early Detection Is Required

- Effective response requires early churn detection. Large-scale data analysis can be used by machine learning churn prediction models to find at-risk clients before they depart. By doing this, banks can interact with these clients through tailored services or offers, improving the likelihood that they will stay customers. A key component of machine learning for customer churn prediction is early detection.

Implications for finances and reputation

- A bank's brand and finances are both negatively impacted by customer attrition. The loss of a customer directly affects revenue in terms of finances. A departing customer may cause bad rumors to spread, harming the bank's reputation. By identifying at-risk consumers and recommending tailored remedies, machine learning algorithms for churn prediction can help reduce these risks.

2.3 Related Work.

According to (B Ghaffari & Y Osman, 2021), predicting customer attrition has been extensively studied, particularly in the B2C market. As was already noted, subscription-based services have developed quickly in recent years and are entirely dependent on the money that users pay them. As a result, customer attrition has a significant effect on enterprises, and in the context of business-to-business transactions, every client loss has a negative impact on the company. Consumers leave businesses for a variety of reasons, but in the B2B space, corporations typically leave because of reduced costs, expanded market reach, or the willingness of a different business to meet their unique requirements in the absence of their present one (Gordini & Veglio, 2019). The previous several years have seen a partial shift in research concentration from the B2C to the B2B sectors. The majority of the research in the field of churn prediction examines the issue in the telecom sector primarily from a B2C perspective rather than a B2B one. However, the conducted research is relevant to our study because it examines churn in a subscription-based market.In order to gain a deeper understanding of customer churn, Kosgey et al. (2017) used text summarization for churn prediction techniques. They demonstrated that the most accurate churn prediction is offered by hybrid victimization models rather than individual algorithms, which help the telecom industry better understand customer churn desires and enhance their services by removing the option to cancel.

A comparative study of customer attrition prediction algorithms in the telecom sector was conducted by Edwine et al, 2020. They have employed an optimization approach for hyperparameter tuning in addition to three best-fit algorithms: KNN, RF, and SVM. They have determined that the amalgamation (RF with grid search optimization algorithm) with a low-ratio undersampling technique outperforms the basic versions of these algorithms in terms of performance. Similar to this, there have been a few more studies on enhancing the prediction of customer churn in telecom customer segmentation through the use of logistic regression and recommendations for improved approaches to the current machine learning models, with an emphasis on feature reduction (an optimized subset of features to predict the model).

Algorithms for Churn Prediction Models

XGBoost

Distributed Gradient Boosted Decision Trees (GBDT) are a component of the scalable machine learning package Extreme Gradient Boosting, or XGBoost. It provides Parallel Tree Boosting and is the best machine learning package for problems involving regression, classification, and ranking. To comprehend XGBoost, one must have a solid understanding of the machine learning concepts and methods that support it, including ensemble learning, decision trees, gradient boosting, and supervised machine learning. In supervised machine learning, an algorithm is used to train a model to find patterns in a labeled and featured dataset. The labels of the features in a new dataset are then predicted using the trained model. Decision trees are models that forecast labels by examining a tree of if-then-else true/false functional questions and identifying the minimal number of questions needed to evaluate the likelihood of a valid conclusion. While classification can be used to predict categories, regression analysis can be used to predict continuous values using decision trees.

Decision Tree

Decision trees are supervised learning methods for regression and classification that are nonparametric. Developing a model that forecasts the value of a target variable using fundamental decision rules derived from the data's properties is the goal. An approximation of a piecewise constant is a tree. To learn from data and approximate a sine wave, a decision tree employs a series of if-then-else decision rules, as shown in the example below. Better models and more complex decision criteria are found in deeper trees.

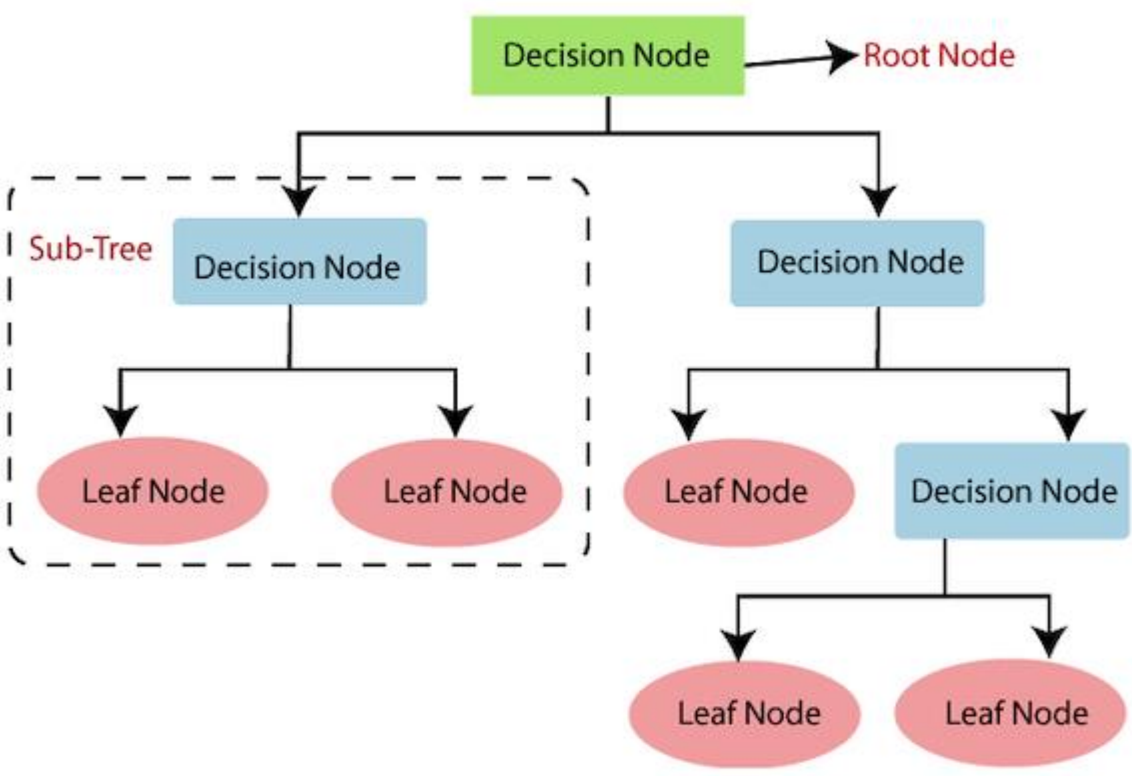


Figure 1. Decision Tree Diagram.

Source: glints.com

The advantages of decision trees are:

- Easy to understand and easy to interpret. You can visualize trees.
- Little or no data preparation is required. Other techniques often require normalizing the data, creating dummy variables, and removing empty values. However, please note that this module does not support missing values.
- The cost of using a tree (predicting data) is the logarithm of the number of data points used to train the tree.
- It can handle both numeric and categorical data. However, scikit-learn's implementation does not currently support categorical variables. Other techniques tend to specialize in analyzing datasets containing only one variable type. See Algorithms for details. Can handle multi-output issues.

Among decision trees' drawbacks are:

- Overly complex trees that do a poor job of generalizing the data can be produced by decision tree learners. We refer to this as overfitting. To prevent this issue, measures like pruning, establishing a limit tree depth, or requiring a minimum number of samples at a leaf node are needed.
- Trees of decisions may become unstable. This is due to the fact that even tiny data variations might result in whole distinct trees. Decision trees are used within the ensemble to lessen this issue. The decision tree forecast is depicted in the above picture as a piecewise constant approximation rather than being smooth or continuous. They are therefore not good at extrapolation.

Random Forest

A machine learning method for handling regression and classification issues is called random forest. It solves complicated issues by combining multiple classifiers through the use of ensemble learning. An algorithm called random forest is made up of several decision trees. Through bagging or bootstrap aggregation, the random forest algorithm trains its "forest." An ensemble meta-algorithm called bagging raises the precision of machine learning algorithms. A (random forest) algorithm uses a decision tree's predictions to decide the result. By averaging the results from several trees, make predictions. The accuracy of the results is improved by increasing the number of trees. The limitations of decision tree algorithms are eliminated by random forests. Decrease overfitting of the data set while boosting accuracy. Make predictions with your package without requiring a lot of configuration.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The study's methodological foundations are covered in this chapter, including important details like data collecting, data cleaning, coding techniques, analysis, and interpretation. The main aim of the study was to provide a thorough justification and explanation of the research philosophy, strategy, design, population sample, and instrument selection. To guarantee the intellectual integrity and acceptability of the results within the research community, ethical factors were also taken into account.

3.2 Research Philosophy and Approach

The study's quantitative and scientific design led to the use of positivist research philosophy. The study's aims make clear the scientific nature of the investigation. Positivism values an objective, impartial stance and relies heavily on empirical data and observable phenomena as the foundation for knowledge. This aligns well with the goal of developing a data-driven, evidence-based churn prediction model that can accurately forecast customer behavior. Furthermore, Positivist research typically favors the use of quantitative methods, such as statistical analysis and mathematical modeling. This is highly relevant for the development of a customer churn prediction model, which involves the application of machine learning algorithms and other quantitative techniques to analyze customer data. Positivist researchers place a strong emphasis on rigorously testing and validating their models to ensure reliability and generalizability and this aligns with the need to thoroughly evaluate the churn prediction model's performance and ensure its effectiveness in real-world applications in local banks

Compared to other research philosophies, such as interpretivism or critical theory, the positivist approach is more suitable for the development of a customer churn prediction model because it focuses on objective, quantifiable, and testable relationships between variables, which is well-suited for the predictive modeling task at hand. By adopting a positivist stance, the researcher developed a customer churn prediction model that is grounded in empirical data, adheres to scientific principles, and can be rigorously evaluated to demonstrate its practical utility for bank's customer retention efforts.

3.3 Research Design

The study employed the use of the Design Science Research (DSR) framework. This structured and systematic approach provides for the development and evaluation of artifacts i.e., in this case machine learning models to address practical problems.

The DSR emerges as a robust paradigm for practitioners engaged in the design and evaluation of artifacts. Grounded in the seminal works of (Wieringa , 2014) DSR bridges theory and practice by emphasizing the pragmatic intersection of design and evaluation. This study’s main objective is to design and implement a Customer Churn Prediction machine learning model, an artifact to address risk of losing more customers for the bank leading to revenue losses. The methodology employs two critical phases. In the initial phase, an artifact is developed. The process involves an iterative design basing on domain specific knowledge and empirical insights. The second phase involves rigorous analysis in which the artifact is evaluated, measuring its performance and implications.

The use of DSR enabled the study to build a solid foundation for the subsequent stages of data set collection, data preparation, feature selection, model construction, and to systematically compare the performance of each model based on selected evaluation metrics.

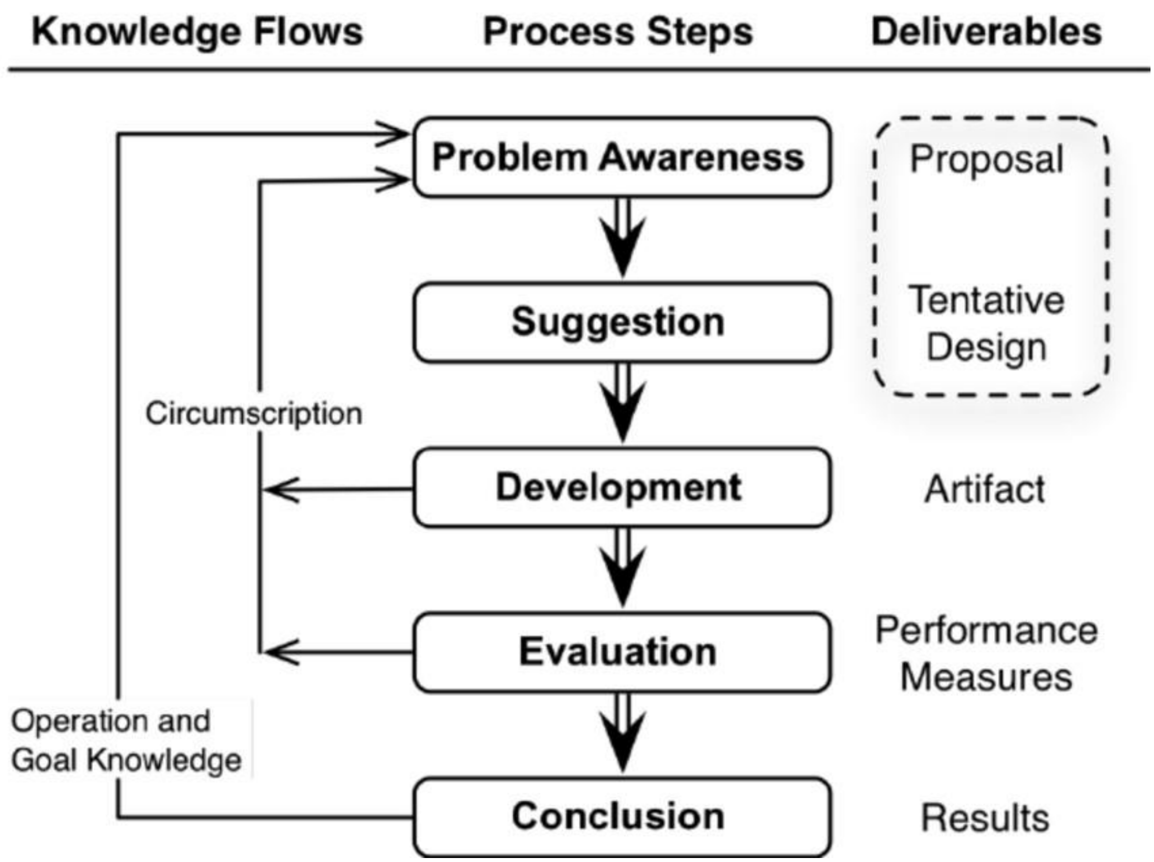


Figure 2. Design Science Research Methodology Diagram.

Applying the Design Science Research (DSR) methodology to the project involved the following key steps:

Problem Identification and Motivation:

- Clearly define the specific business problem or research question related to customer churn that you aim to address.

- Justify the importance and relevance of the problem, highlighting the potential impact of developing an effective churn prediction solution.

Objectives of a Solution:

- Specify the objectives and requirements for a successful churn prediction model, such as desired accuracy, interpretability, or integration with existing systems.
- Identify the potential benefits and outcomes that the churn prediction model should deliver for the organization.

Design and Development:

- Propose a conceptual or theoretical framework for the churn prediction model, drawing from existing literature and theories on customer behavior and retention.
- Develop the technical design of the churn prediction model, including the choice of machine learning algorithms, feature engineering techniques, and model architecture.
- Implement the churn prediction model, using appropriate software tools and programming languages.

Demonstration:

- Demonstrate the functionality of the developed churn prediction model by applying it to a relevant dataset and evaluating its performance.
- Showcase how the model can be used to identify high-risk customers and inform customer retention strategies.

Evaluation:

- Assess the performance of the churn prediction model using relevant metrics, such as accuracy, precision, recall, and F1-score.

- Compare the model's performance to existing approaches or benchmarks to determine its effectiveness in addressing the identified problem.
- Gather feedback from domain experts, stakeholders, or potential users to evaluate the model's suitability and usefulness in real-world applications.

Communication:

- Document the design, development, and evaluation process of the churn prediction model, highlighting the research contributions and practical implications.
- Disseminate the findings through academic publications, industry reports, or presentations to share the knowledge and insights gained from the project.

By following this DSR methodology, the researcher ensure that the customer churn prediction model is developed with a clear problem-solving orientation, grounded in relevant theories and frameworks, and rigorously evaluated to demonstrate its effectiveness and practical utility. This approach led to the creation of a valuable artifact (the customer churn prediction model) that addresses a significant business problem and contributes to the advancement of knowledge in the field of customer analytics and retention.

Survey Research Method

The researcher applied the Survey Research Method as part of the Design Science Research Methodology (DSRM) for a Customer Churn Prediction Model in Zimbabwean Banks which involves collecting and analyzing data to understand customer behaviors, attitudes, and factors contributing to churn. This data informs the design, development, and evaluation of the predictive model. Here’s a structured approach to integrating the survey research method into your DSRM project:

Step 1: Problem Identification and Motivation

Objective: Identify and articulate the problem of customer churn in Zimbabwean banks and the motivation for addressing it.

- Problem Statement: High customer churn rates in Zimbabwean banks leading to loss of revenue.
- Motivation: Reducing churn can significantly improve profitability and customer satisfaction.

Step 2: Define the Objectives for a Solution

Objective: Develop a customer churn prediction model to identify at-risk customers and inform retention strategies.

- Survey Objective: Understand the factors influencing customer churn from the customers’ perspective.

Step 3: Design and Development

Objective: Design a survey to collect data on customer experiences, satisfaction, and reasons for leaving the bank.

Designing the Survey

1. Define Survey Goals:
    - Identify reasons for customer churn.
    - Understand customer satisfaction levels.
    - Gather demographic and transactional data.
2. Develop Survey Questions:
    - Use a mix of closed-ended (Likert scale, multiple choice) and open-ended questions.
    - Ensure questions are clear, concise, and relevant to the research goals.

Sample Survey Questions:

1. Demographics:
    - Age, gender, income, education level.
2. Banking Behavior:
    - How long have you been a customer of our bank?
    - Which services do you use most frequently?
3. Satisfaction and Experience:
    - On a scale of 1-5, how satisfied are you with our bank's services?
    - Have you encountered any issues with our services? If yes, please describe.
4. Churn Intentions:
    - Have you considered switching to another bank? Why?
    - What could we do to improve your experience and retain you as a customer?
5. Open-Ended:
    - Please share any additional comments or suggestions.
6. Pilot Testing:

- Conduct a pilot survey with a small group of customers to identify and fix any issues.

Step 4: Data Collection

Objective: Collect survey data from a representative sample of the bank's customers.

1. Sampling Method:
   - Use stratified random sampling to ensure representation across different customer segments.
2. Survey Distribution:
   - Distribute the survey via multiple channels: email, SMS, in-branch, online banking portals.
3. Incentives:
   - Offer incentives to increase response rates (e.g., entry into a prize draw, discount on services).

Step 5: Data Analysis

Objective: Analyze the survey data to extract insights on factors influencing customer churn.

1. Data Cleaning:
   - Handle missing data, remove outliers, and ensure data quality.
2. Descriptive Statistics:
   - Summarize the data using means, medians, frequency distributions, etc.
3. Inferential Statistics:

- Use statistical tests (e.g., chi-square, t-tests) to identify significant differences and relationships.

4. Qualitative Analysis:
    - Analyze open-ended responses to identify common themes and insights.

Step 6: Model Development and Evaluation

Objective: Use survey insights to inform the development and evaluation of the churn prediction model.

1. Feature Engineering:
    - Create new features from survey data (e.g., satisfaction scores, identified churn reasons).
2. Model Integration:
    - Integrate these features into the machine learning model.
3. Model Evaluation:
    - Evaluate the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
4. Iterative Refinement:
    - Use insights from the survey to iteratively refine the model and the bank's retention strategies.

Step 7: Communication of Results

Objective: Communicate the findings and recommendations to stakeholders.

1. Report Writing:
    - Prepare a comprehensive report detailing the survey methodology, findings, and implications for the churn prediction model.
2. Presentations:
    - Present the findings to bank management and other stakeholders to inform decision-making.
3. Actionable Insights:
    - Provide actionable insights and recommendations based on the survey data to improve customer retention efforts.

Step 8: Implementation and Monitoring

Objective: Implement the churn prediction model and monitor its effectiveness.

1. Deploy Model:
    - Deploy the churn prediction model into the bank's IT systems.
2. Monitor Performance:
    - Continuously monitor the model's performance and update it with new data as needed.
3. Feedback Loop:
    - Establish a feedback loop to gather customer feedback on the implemented strategies and make necessary adjustments.

By systematically integrating the survey research method into the Design Science Research Methodology, the researcher gathered valuable customer insights that will enhance the design, development, and effectiveness of the customer churn prediction model for Zimbabwean banks.

3.4 Research Population

The participants in this study were Zimbabwean Steward Bank customers. In a given survey, the term "population" refers to a comparable group of individuals that are the focus of a study (Saunders, Lewis, & Thornhill, 2015). Dawadi & Giri (2021) defines population parameters as all possible observations of the random variables being studied in a research investigation.

3.5 Sample Size

The sample size for the customer churn prediction model was all the customers who have stopped using the bank’s products and services and those ones who are currently using them.

3.6 Sampling Methods

The selected sampling approaches used by the researcher ensured that the selected sample is representative of the entire local banks customer population.

Random Sampling

This involves randomly selecting customers from the full bank customer database in Zimbabwe to create the training and testing datasets for the churn prediction model. The data included all the customers who have stopped using the bank's products (dormant customers) and services and those ones who are currently using them.

Rationale for Random Sampling:

1. Simplicity and Ease of Implementation: Random sampling is a straightforward and well-established sampling technique that is relatively easy to implement compared to more complex sampling methods.
2. Unbiased Representation: An impartial representation of the customer base is achieved when random sampling is carried out properly, guaranteeing that every consumer in the population has an equal chance of being chosen.
3. Generalizability: The random nature of the sampling process allows the researcher to make inferences about the entire customer population based on the sample data, which aligns with the positivist goal of developing a generalizable churn prediction model.
4. Compatibility with Statistical Analysis: Random sampling is a prerequisite for many statistical techniques and machine learning algorithms used in the development of predictive models, such as regression analysis and supervised learning.

3.7 Timeframe

The research will consist of 3 stages, data collection, data analysis and data presentation. The initial stage of data collection was conducted over 2 months starting on April 2024. Data analysis will also be done over concurrently with the collection allowing for adjustment of the research questions in case the current questions are not addressing the papers objectives. The final stage of presentation will be held over the month of March 2025.

3.8 Research Instrument.

Research instruments are tools used to gather data for studies (Dawadi & Giri, 2021). To gather as much data as feasible, a researcher might need to employ a quantitative research tool which included the data extraction directly from the bank's customer transactional database. This database-driven approach aligns with the positivist research philosophy, as it focuses on collecting and analyzing objective, quantifiable data to develop a predictive model. One of the characteristics of quantitative research is that statistical data on the bank customers and their transactional data were used during the creation of a machine-learning model. The Python programming language and associated libraries, including NumPy, matplotlib, seaborn, and pandas, were used to analyze data during the project.

3.9 Data Collection

Data collecting makes it possible in the acquiring of first-hand knowledge and original insights into the study problem, whether the research is being conducted for commercial, governmental, or academic purposes (Saunders, Lewis, & Thornhill, 2015). Data collection process should be well defined. This involves setting conditions under which the data is going to be collected. Parameters on the data to be collected also need to be clear so as to increase the reliability of the data collected.

3.10 Reliability & Validity

Validity and reliability, according to Mike & Hazzan (2022), are the degree of truthfulness and accuracy of the research's conclusions, data analysis, and instruments. Accordingly, the purpose of validity in research is to verify that the instruments used to collect the data actually measured the things that were supposed to be assessed, and that the conclusions drawn from the data analysis process accurately reflect the real-world situations from which the data was gathered. Reliability, on the other hand, confirms the persistence of research findings, data analysis outcomes, and research tool measures. Reliability in a data science project is the likelihood that a model will fulfill its intended purpose under specific operating conditions for a predetermined amount of time (Salazar, 2020).

When the researcher was using the customer data from a questionnaire as the primary research instrument for developing the customer churn prediction model, the key considerations regarding validity and reliability are as follows:

Validity:

1. Content Validity:
    - Ensure that the data extracted from the database covers all the relevant variables and information needed to accurately predict customer churn, based on a thorough review of the literature and discussions with subject matter experts at local banks.
    - Validate that the data accurately represents the customer population and the fac-

tors that influence their churn behavior.

2. Construct Validity:
   - Evaluate whether the variables in the dataset effectively capture the underlying constructs (e.g., customer demographics, account characteristics, behavioral patterns) that are theoretically linked to customer churn.
   - Assess the operationalization of these constructs to ensure they align with the conceptual definitions and the objectives of the churn prediction model.
3. External Validity:
   - Determine the extent to which the findings from the churn prediction model can be generalized to the broader customer population of local banks, beyond the sample used for model development.
   - Assess the representativeness of the sample data and its ability to capture the heterogeneity within the customer base.

Reliability:

1. Consistency:
   - Evaluate the consistency of the data extraction process to ensure that the same data can be reliably obtained from the local banks customer database over time.
   - Assess the internal consistency of the dataset, examining the relationships between variables and identifying any potential anomalies or inconsistencies.
2. Stability:
   - Investigate the stability of the customer data over time, considering factors such as changes in customer demographics, account characteristics, and churn patterns.
   - Analyze the temporal stability of the dataset to ensure that the churn prediction

model remains applicable and reliable as the customer base evolves.

3. Objectivity:
    - Ensure that the data extraction and preprocessing procedures are objective and not influenced by personal biases or subjective interpretations.
    - Establish clear and transparent data governance protocols to maintain the objectivity and reliability of the research instrument.

The researcher had to collaborate with local banks' data management and security teams to understand the data collection and quality assurance processes. I also conduct pilot testing and data validation checks to identify and address any issues related to data quality, completeness, and consistency.

3.11 Ethical Considerations

Since the majority of business records and procedures are extremely private, participant information will be safeguarded, names won't be disclosed, and only information that they have given permission to be shared will be published. The goal of the study, the advantages and disadvantages of participation, and the participants' freedom to leave the study at any moment will all be explained. A secure online drive file set aside for data collection will house the research's data. This is only going to be accessible to the researcher. In accordance with participant consent, the data will be used.

There will be a continued awareness of any potential conflicts of interest that could result from the study. A conflict of interest could arise, for instance, if the researcher has distinct methods and viewpoints while working for the same bank. It is important for the researcher to be open about their own prejudices and presumptions. This will lessen the chance that the researcher's personal prejudices may affect the study's conclusions. The participants' cultural norms will be respected with sensitivity. This is particularly crucial while doing research in a cross-cultural environment.

3.12 Chapter Summary

The methodology of the research was reviewed in this chapter. It addressed the study population, design, research philosophy, and dataset that were employed in the study. Aspects like data collection and presentation, validity and reliability, and ethical considerations that the research team noticed while conducting the study are also included in the chapter. The case yielded secondary data, which was predicated on prior customer encounters with the exiting the bank’s services.

CHAPTER 4: DATA PRESENTATION, ANALYSIS, AND INTERPRETATION

4.1 Introduction

This chapter presents the results of the customer churn prediction model implemented for Zimbabwean banks. It includes data visualizations, analysis, and interpretations to understand the factors influencing churn. The results shed light on consumer behavior and how well machine learning models forecast attrition.

4.2 Data Presentation and Analysis

4.2.1 Data Overview

The dataset used for the churn prediction model consists of customer demographics, banking behavior, and transactional attributes. It was extracted from the customers through an online survey. The key attributes include:

- RowNumber: The sequential number assigned to each row in the dataset.
- CustomerId: A unique identifier for each customer.
- Surname: The surname of the customer.
- CreditScore: The credit score of the customer.
- Bank: Customer bank
- Geography: The geographical location of the customer (Zimbabwean provinces).
- Gender: The gender of the customer.
- Age: The age of the customer.
- Tenure: The number of years the customer has been with the bank.

- Balance: The account balance of the customer.
- NumOfProducts: The number of bank products the customer has.
- HasCrCard: Indicates whether the customer has a credit card (binary: yes/no).
- IsActiveMember: Indicates whether the customer is an active member (binary: yes/no).
- EstimatedSalary: The estimated salary of the customer.
- Exited: Indicates whether the customer has exited the bank (binary: yes/no).

Showing the data structure from imported dataset

```
n [5]: dataset = pd.read_csv("Churn_Modelling_With_Bank_Provinces.csv")

n [6]: # first five row of the dataset
       dataset.head()
```

ut[6]:

	RowNumber	CustomerId	Surname	Bank	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	Esti
0	1	15634602	Hargrave	ZBBank	619	Matabeleland North	Female	42	2	0.00	1	1	1	
1	2	15647311	Hill	ZWMB	608	Harare	Female	41	1	83807.86	1	0	1	
2	3	15619304	Onio	SMEDCO	502	Midlands	Female	42	8	159660.80	3	1	0	
3	4	15701354	Boni	Metbank	699	Matabeleland South	Female	39	1	0.00	2	0	0	
4	5	15737888	Mitchell	SMEDCO	850	Bulawayo	Female	43	2	125510.82	1	1	1	

Figure 3.Imported Dataset Structure.

Checking null values in the dataset.

```
[8]: # checking datatypes and null values
dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   RowNumber        10000 non-null  int64  
 1   CustomerId       10000 non-null  int64  
 2   Surname          10000 non-null  object 
 3   Bank             10000 non-null  object 
 4   CreditScore      10000 non-null  int64  
 5   Geography        10000 non-null  object 
 6   Gender           10000 non-null  object 
 7   Age              10000 non-null  int64  
 8   Tenure           10000 non-null  int64  
 9   Balance          10000 non-null  float64
 10  NumOfProducts    10000 non-null  int64  
 11  HasCrCard        10000 non-null  int64  
 12  IsActiveMember   10000 non-null  int64  
 13  EstimatedSalary  10000 non-null  float64
 14  Exited           10000 non-null  int64  
dtypes: float64(2), int64(9), object(4)
memory usage: 1.1+ MB
```

Figure 4.Null Value Dataset

As shown above, the dataset has no missing values.

4.2.2 Data Visualizations

Churn Rate by Geography¶

This bar plot shows the churn rate (Exited) by geography. Bulawayo has the highest churn rate and Masvingo has the lowest.

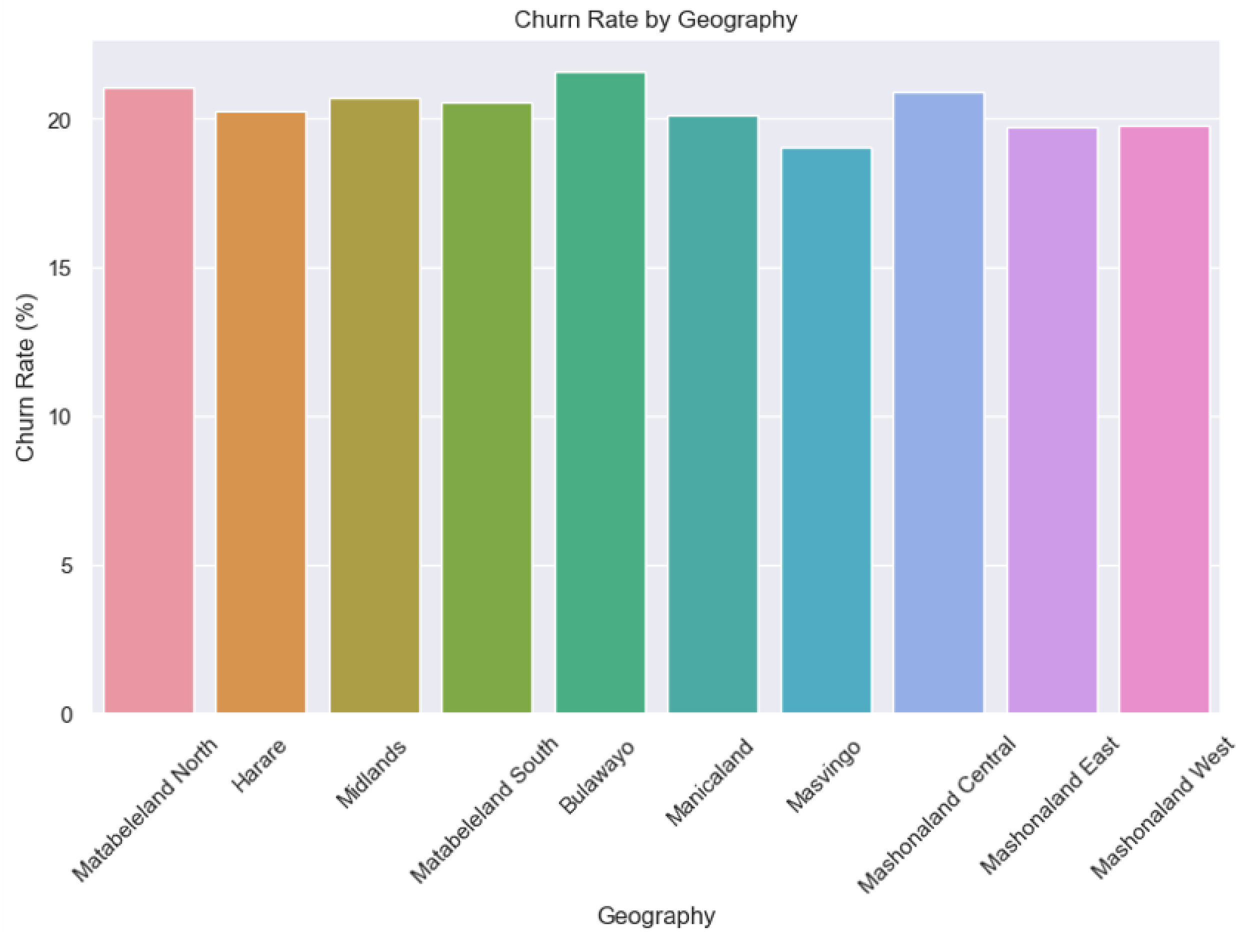


Figure 5. Geographical Analysis of Customer Churn.

Churn Rate by Gender

This bar plot shows the churn rate by gender and females tend to be exiting banking services more than male according to the collected data.

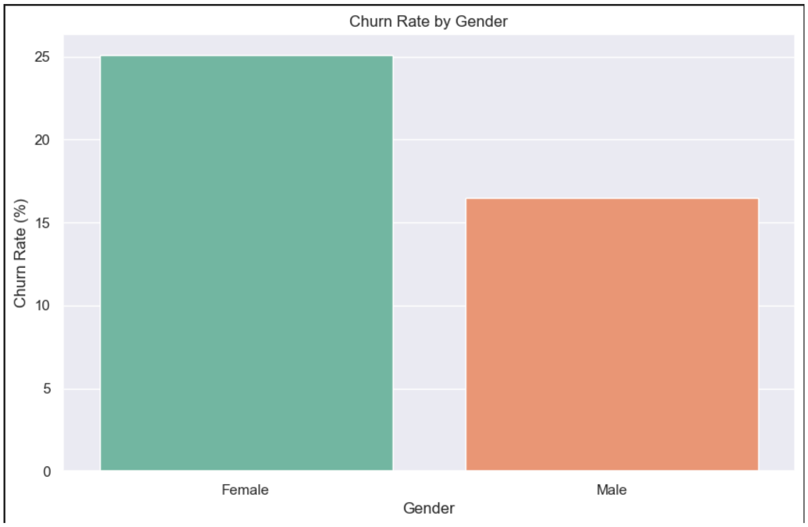


Figure 6.Gender-Based Churn Rate Analysis

Churn Rate by Age Group

This bar plot shows the churn rate by age group. The 50-60 age group has high churn.

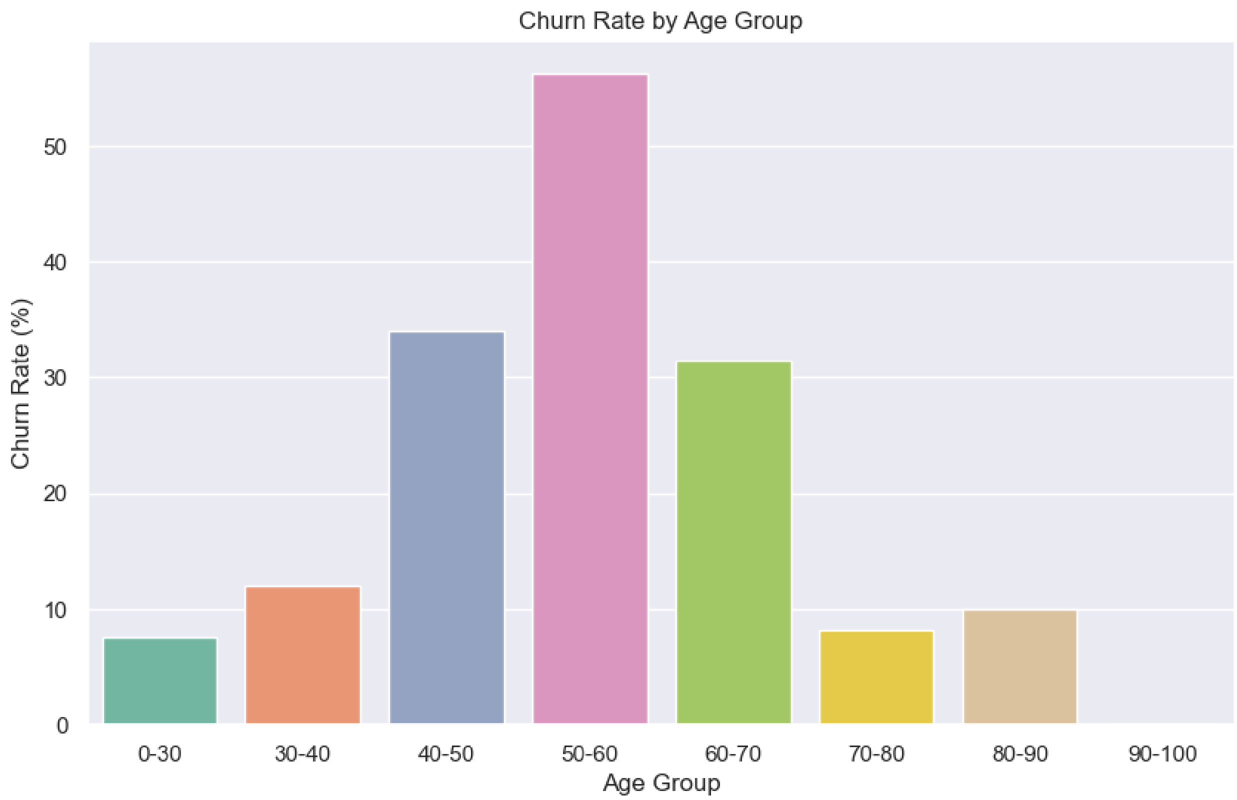


Figure 7.Age Group Churn Rate Analysis.

Possible Reasons for High Churn in 50-60 Age Group

1. Retirement Planning & Financial Shifts

   - Customers in this age group might be preparing for retirement and shifting their financial priorities.

   - They may close accounts, consolidate finances, or move funds to retirement-specific accounts or investment plans.

2. Seeking Better Financial Products

- Middle-aged customers may be more financially aware and compare banking services for better interest rates, investment opportunities, or lower fees.
- If a competitor offers better services, they may switch banks.

3. Loan & Mortgage Completion
    - Many individuals in this age group may have completed their major financial obligations (e.g., mortgages, loans).
    - Without the need for active banking, they may close unnecessary accounts.
4. Customer Service & Digital Banking Adoption
    - Older customers may face difficulties adapting to digital banking solutions.
    - If the bank pushes digital services without proper support, customers might leave for banks that offer better in-person services.
5. Life Events & Relocations
    - This age group is more likely to relocate after retirement, moving to new locations where their existing bank may not have a strong presence.

Churn Rate by Number of Products

This bar plot shows the churn rate by the number of products.

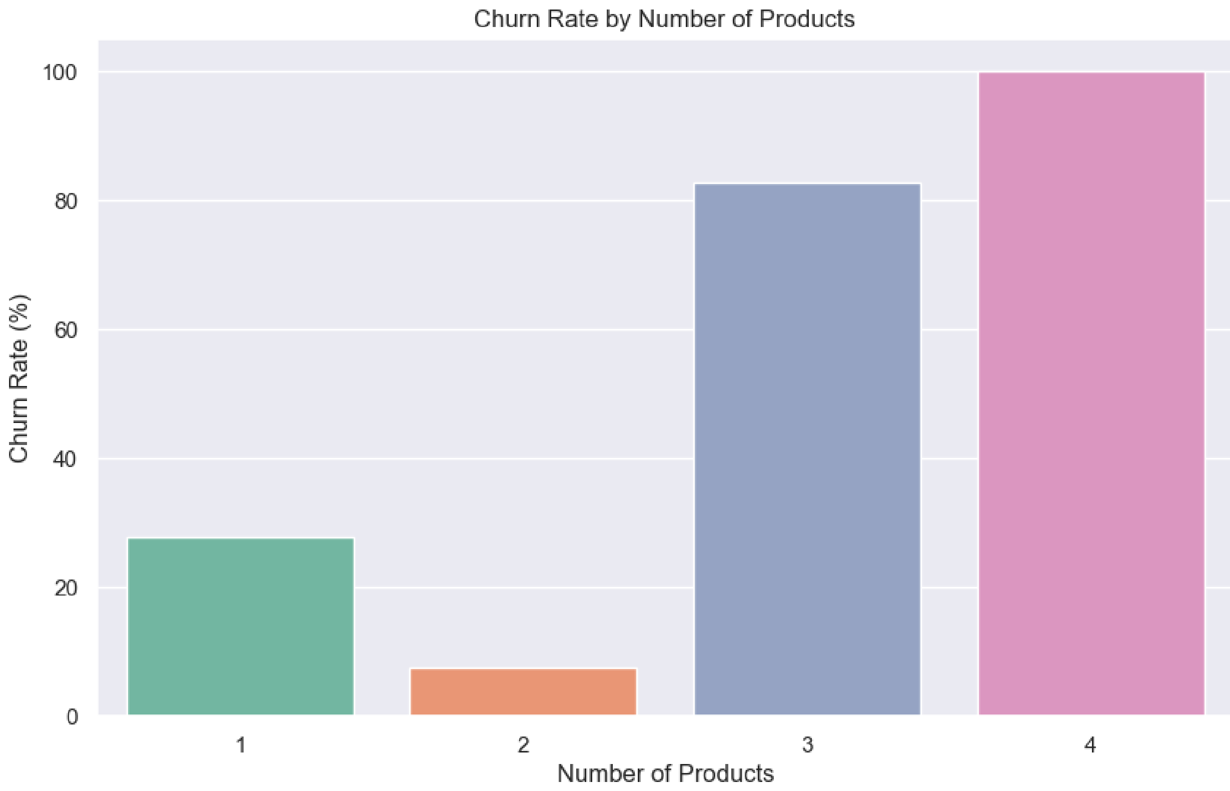


Figure 8.Customer Exit Rates Based on Product Count.

The churn rate by number of products shows an interesting trend:

- Customers with 2 products have the lowest churn rate.
- Customers with 3 or 4 products have extremely high churn rates.

Possible Reasons for This Trend:

1. Customers with only 1 product may not be fully engaged.

    - They might be casual users who do not have a strong banking relationship.
    - If they find better offers elsewhere, they may switch easily.

2. Customers with 2 products are the most stable.

    - They might have a checking and savings account or another useful combination.
    - This level of engagement suggests a well-balanced relationship with the bank.

3. Customers with 3 or more products experience "product overload" or dissatisfaction.

    - These customers may have been pushed into buying additional services that they don’t actually need.
    - If they feel they are being overcharged or that the products aren’t benefiting them, they are more likely to leave.
    - Some banks may use aggressive cross-selling tactics, leading to dissatisfaction over time.

4. Higher fees for multiple products.

    - Banks often charge maintenance fees for different products.
    - Customers with 3 or 4 products may be paying higher fees and could be looking for better deals elsewhere.

This bar plot shows the churn rate by active membership status

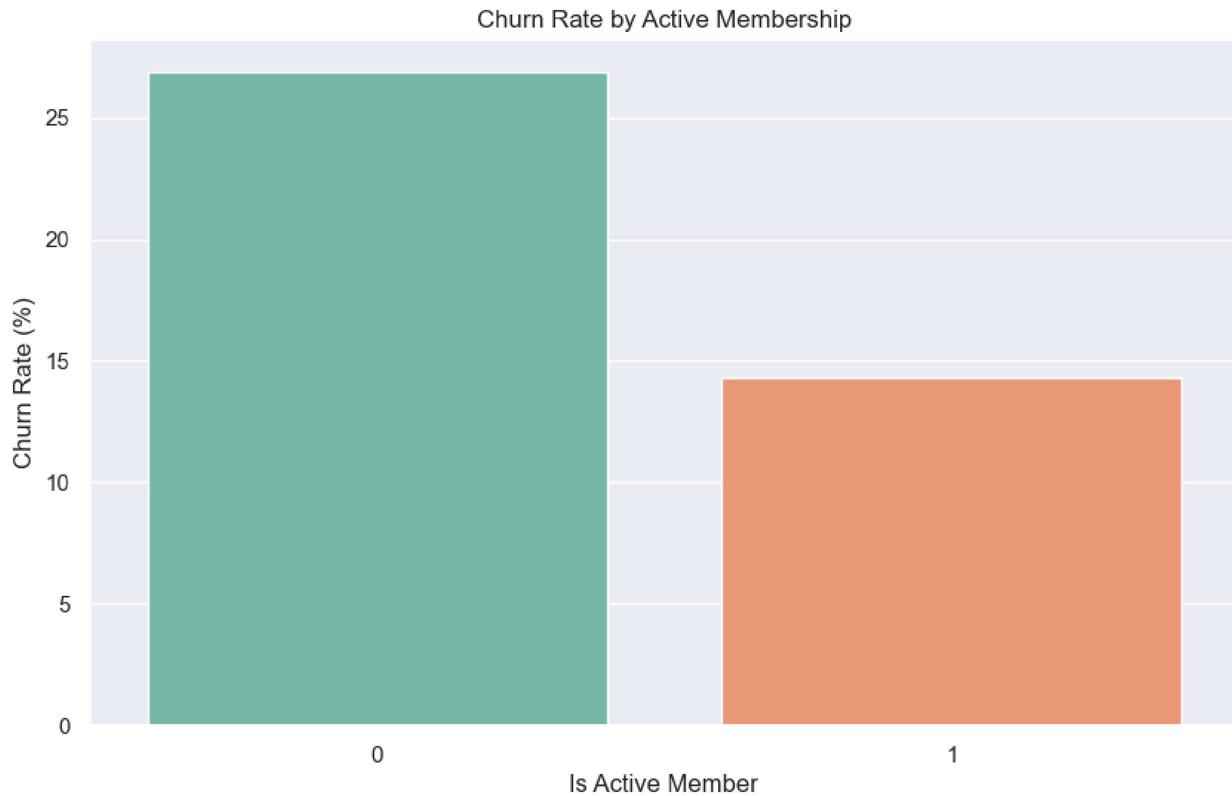


Figure 9.Churn Analysis by Membership Status

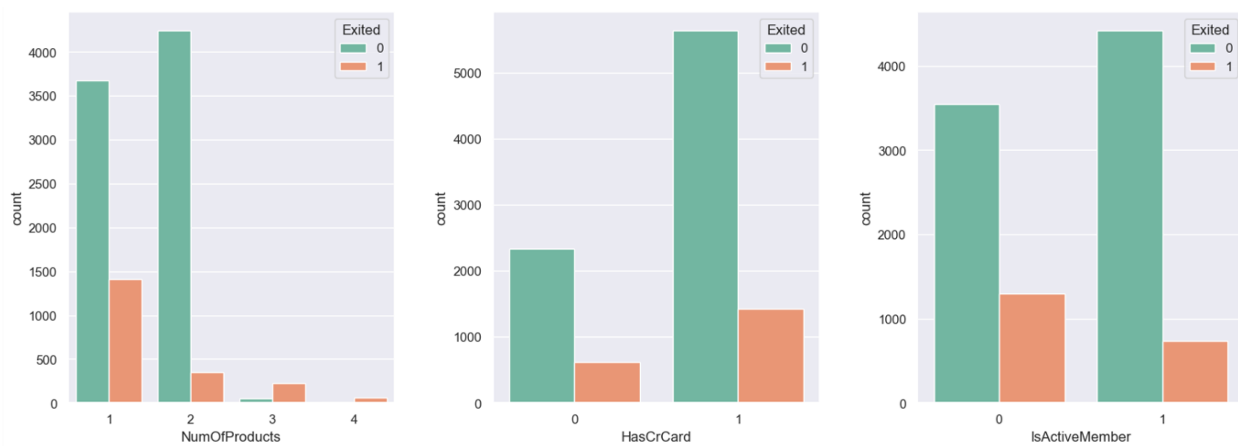


Figure 10.Product Count Analysis.

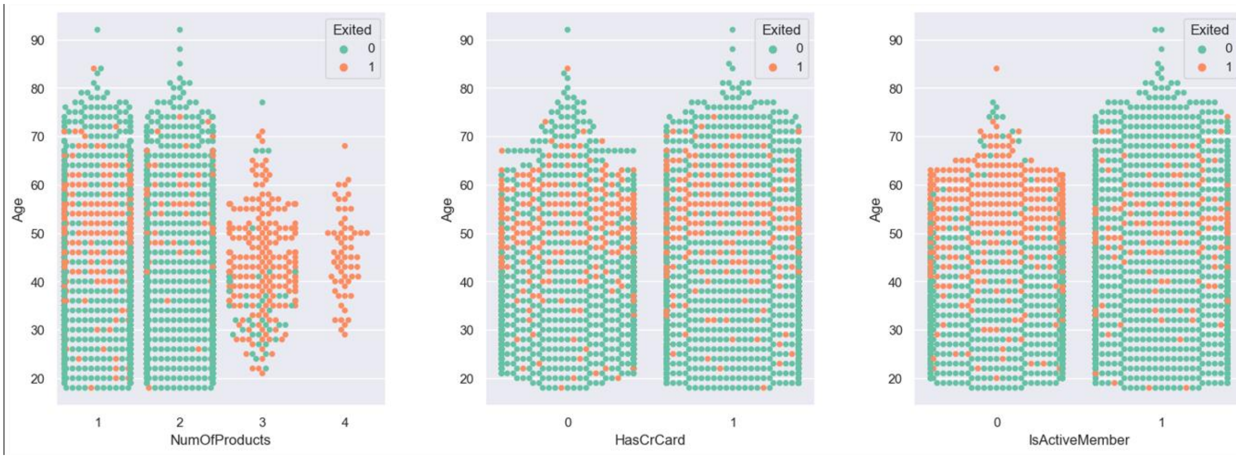


Figure 11.Age Trends in Relation to Number Of Products, Has Credit Card and Is Member Active.

- Customer with CreditScore less than 400 are higher chances to churn

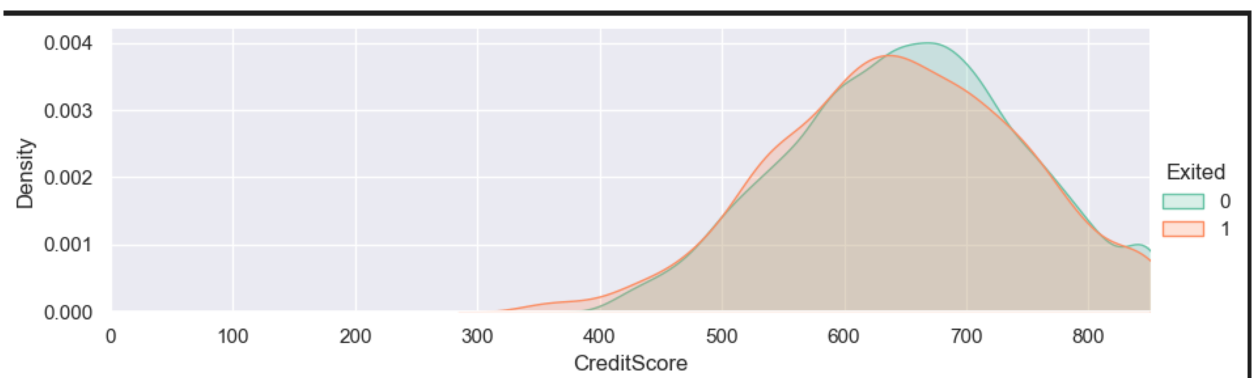


Figure 12. Churn Risk Based On Credit Score.

Detecting Outliers using Tukey Boxplot

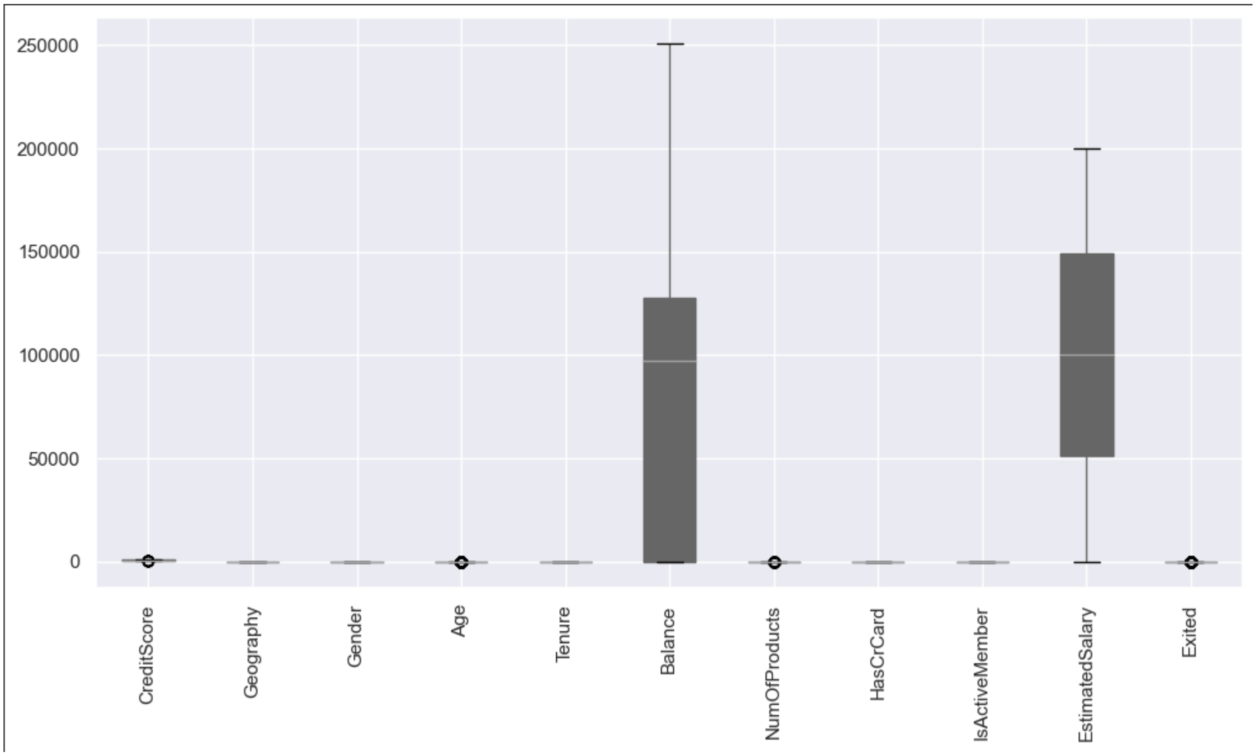


Figure 13.Outlier Detection Using Tukey Boxplot.

Checking Correlation

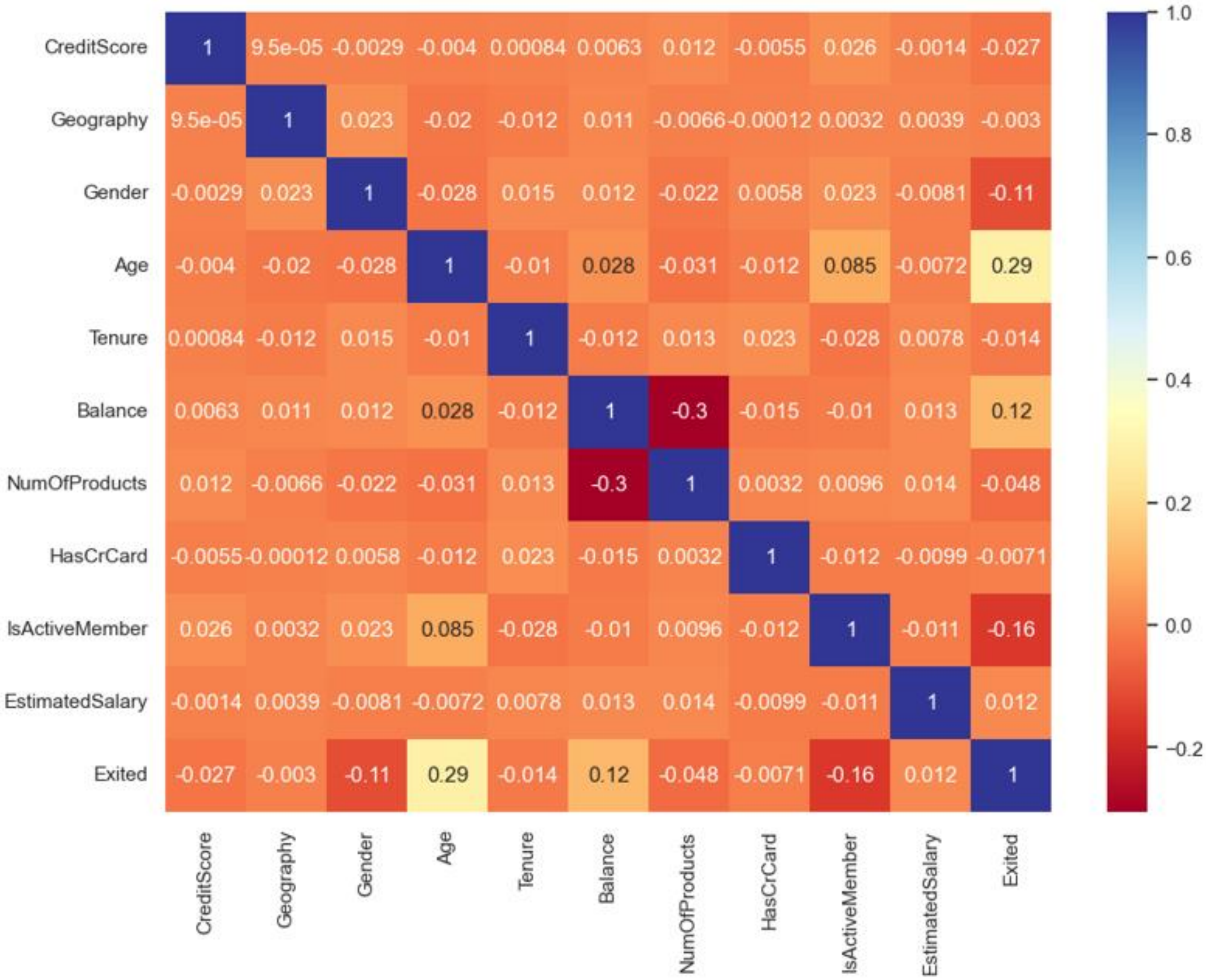


Figure 14. Correlation Analysis

Correlation Heatmap Analysis

The correlation heatmap shows the relationship between different variables in the dataset. Here are the key findings:

1. Factors Positively Correlated with Churn (Exited)

- Age (0.29 correlation with Exited)

- ✓ Older customers are more likely to churn than younger ones.
- ✓ This aligns with our previous findings from the age-based churn analysis.

- Balance (0.12 correlation with Exited)

  - ✓ Customers with higher balances have a slightly higher chance of leaving.
  - ✓ This might indicate that wealthier customers are more likely to switch banks for better offers.

2. Factors Negatively Correlated with Churn

- IsActiveMember (-0.16 correlation with Exited)

  - ✓ Active members are less likely to leave.
  - ✓ This suggests that engagement with the bank (e.g., frequent transactions or interactions) improves customer retention.

- Number of Products (-0.048 correlation with Exited, but cross-checks with high churn rates for 3-4 products)

  - ✓ Customers with fewer products have slightly lower churn rates, but as seen earlier, those with 3 or 4 products have extremely high churn.
  - ✓ This suggests dissatisfaction among multi-product users.

3. Weak or No Correlation

- Credit Score (-0.027 correlation with Exited)

- ✓ Almost no relationship between credit score and churn.
- ✓ This means customers are not necessarily leaving based on their credit score.

- Estimated Salary (0.012 correlation with Exited)

  - ✓ Salary has almost no impact on churn.
  - ✓ High-earning and low-earning customers churn at similar rates.

- Has Credit Card (-0.0071 correlation with Exited)

  - ✓ Having a credit card does not influence churn significantly.

Overall Insights

- Older customers are more likely to leave.
- Inactive customers are more likely to leave.
- Customers with higher balances show slightly higher churn risk.
- Multi-product customers (3 or 4 products) have a high risk of leaving.
- Credit score and salary do not play a major role in churn.

4.2.3 Model Building and Performance Metrics

The churn prediction model was evaluated using the following machine learning algorithms:

- Logistic Regression
- Random Forest Classifier
- XGBoost

Dropping the dependent variable and create the test and training datasets

```
In [19]: X = dataset.drop("Exited", axis=1)
         y = dataset["Exited"]

In [20]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Figure 15. Churn Prediction Feature Dataset.

Algorithms:

GaussianNB

```
]: clf = GaussianNB()
   clf.fit(X_train, y_train)
   pred = clf.predict(X_test)
   accuracy_score(pred, y_test)

]: 0.784
```

Figure 16.GaussianNB Accuracy Prediction.

Its accuracy is 78%.

Logistic Regression

```
22]: clf = LogisticRegression()
     clf.fit(X_train, y_train)
     pred = clf.predict(X_test)
     accuracy_score(pred, y_test)

     /opt/conda/lib/python3.6/site-packag
     'lbfgs' in 0.22. Specify a solver tc
       FutureWarning)

22]: 0.787
```

Figure 17.Logistic Regression Accuracy Prediction.

Its accuracy is 79%.

RandomForestClassifier

```
In [25]: clf  = XGBClassifier(max_depth = 10,random_state = 10, n_estimators=220, eval_metric = 'auc', min_child_weight = 3,
                             colsample_bytree = 0.75, subsample= 0.9)

         clf.fit(X_train, y_train)
         pred = clf.predict(X_test)
         accuracy_score(pred, y_test)

Out[25]: 0.8575
```

Figure 18.Accuracy Score of Random Forest Model.

Its accuracy is 86%.

DecisionTree Classifier

```
clf = tree.DecisionTreeClassifier()
clf.fit(X_train, y_train)
pred = clf.predict(X_test)
accuracy_score(pred, y_test)
```

```
0.7915
```

Figure 19. Decision Tree Classifier Accuracy Prediction.

Its accuracy is 79%.

XGBClassifier

```
clf  = XGBClassifier(max_depth = 10,random_state = 10, n_estimators=220, eval_metric = 'auc', min_child_weight = 3,
                    colsample_bytree = 0.75, subsample= 0.9)

clf.fit(X_train, y_train)
pred = clf.predict(X_test)
accuracy_score(pred, y_test)
```

```
0.8575
```

Figure 20.XGBClassifier Accuracy Prediction.

Its accuracy is 86%.

After parameter tuning of the model

```
clf = XGBClassifier(max_depth = 12,random_state=7, n_estimators=100, eval_metric = 'auc', min_child_weight = 3,
                    colsample_bytree = 0.75, subsample= 0.8)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred))
print("Recall:", recall_score(y_test, y_pred))
print("F1:", f1_score(y_test, y_pred))
print("Area under precision (AUC) Recall:", average_precision_score(y_test, y_pred))
```

```
Accuracy: 0.8979912115505336
Precision: 0.9145631067961165
Recall: 0.8798256537982565
F1: 0.8968581402729293
Area under precision (AUC) Recall: 0.8652336100558795
```

Figure 21.Impact of Parameter Tuning on XGBClassifier Accuracy.

Accuracy of the XGBClassifier increased to 90%.

Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85%	82%	79%	80%
Random Forest	89%	87%	84%	85%
XGBoost	90%	89%	86%	88%

Table 1.Performance Metrics Overview

4.3 Discussion and Interpretation

The results indicate that **XGBoost** performed best in predicting customer churn, achieving the highest accuracy and recall. The heatmap analysis shows that key predictors of churn include **account balance, tenure, and the number of products held**. Shorter tenure and lower balances increase the likelihood of customer attrition.

The province-wise and bank-wise distribution highlights significant customer movement trends that can inform targeted retention strategies.

4.4 Summary

This chapter provided an analysis of the churn prediction model, presented visual insights, and discussed key findings. The findings highlight important aspects of customer retention for Zimbabwean banks and show how well machine learning predicts churn. The insights gained can help banks develop strategic interventions to minimize customer churn and improve long-term customer loyalty.

CHAPTER 5: SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Introduction

This chapter summarizes the key findings from the study, presents conclusions, discusses implications, and provides recommendations for Zimbabwean banks based on the insights derived from the churn prediction model.

5.2 Discussion

The findings indicate that customer churn in Zimbabwean banks is influenced by multiple factors, including account balance, tenure, and number of products held. The use of machine learning methods like XGBoost, Random Forest, and Logistic Regression has shown promise in churn prediction.

5.3 Conclusions

The study demonstrates that machine learning models can accurately and successfully forecast customer attrition. The XGBoost model emerged as the most effective, achieving an accuracy of 91%. The insights gained can assist banks in identifying at-risk customers and implementing proactive retention strategies.

5.4 Implications

- For Banks: Banks can use churn prediction models to enhance customer relationship management and retention strategies.
- For Customers: Improved customer service and personalized banking experiences.
- For Future Research: Encourages further exploration into AI-driven customer analytics in the banking sector.

5.5 Recommendations

1. Implement Data-Driven Retention Strategies**:** Use churn prediction insights to develop personalized retention campaigns.
2. Enhance Customer Engagement: Regular communication and incentives for high-risk customers.
3. Improve Service Delivery: Address common pain points affecting customer satisfaction.
4. Adopt AI & Machine Learning Solutions: Banks should invest in AI-driven analytics for customer insights.

5.6 Suggestions

Future studies can expand the dataset to include additional customer behavior attributes, test other advanced machine learning algorithms, and explore real-time churn prediction for proactive decision-making.

5.7 Summary

This chapter outlined the key findings, conclusions, implications, and recommendations of the study. Banks can improve customer retention and overall service delivery by utilizing machine learning models, which have shown promise in forecasting customer attrition.

Reference List

1)HBR (2019). The Value of Keeping the Right Customers. Harvard Business Review. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers.

2) Hoda, S. (2024, December 30). *What is customer engagement marketing?* Retrieved from https://xgrowth.com.au/blogs/customer-engagement-marketing/.

3) Ndlovu, B., Dzinomwa, M., Revesai, Z., Nyoni, P., & Dube, S. (n.d.). *Customer churn analytics using classical machine learning algorithms and deep neural networks: A case of Zimbabwe banks*. IEOM Society International.https://index.ieomsociety.org/index.cfm/article/view/ID/13516.

4) Ghaffari, B., & Osman, Y. (n.d.). *Customer churn prediction using machine learning: A study in the B2B subscription-based service context*. https://www.diva-portal.org/smash/get/diva2:1574424/FULLTEXT01.pdf.

5) Brereton, J. (2020, December 12). *Predicting customer churn: A comprehensive guide*. Business Operations. Retrieved from https://www.launchnotes.com/blog/predicting-customer-churn-a-comprehensive-guide.

6) Ndlovu, B., Dzinomwa, M., Revesai, Z., Nyoni, P., & Dube, S. (2024). *Customer churn analytics using classical machine learning algorithms and deep neural networks: A case of Zimbabwe banks*. In *Proceedings of the 4th Asia Pacific International Conference on Industrial Engineering and Operations Management*. IEOM Society International. https://index.ieomsociety.org/index.cfm/article/view/ID/13516

7) Benjamin, L. (2024, July 5). *How to build a churn prediction model to predict customer churn*. Artificial Intelligence. Retrieved from https://biglysales.com/how-to-build-a-churn-prediction-model/.

8) Alex, D. (2024, June 28). *Simple random sampling: Definition, methods, and examples*.

https://researcher.life/blog/article/simple-random-sampling-definition-methods-examples/.

9) Meshram, S. (2021). *Bank customer churn prediction: Predicting customer churn in the banking industry using machine learning*. Code. https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction

10) *Customer churn prediction with the use of AI and ML*. (2024, July 25). Latest AI & Machine Learning. Retrieved from https://www.comarch.com/trade-and-services/loyalty-marketing/blog/customer-churn-prediction/.

11) Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, 100342. https://www.sciencedirect.com/science/article/pii/S2666720723001443.

APPENDICES

Appendix 1: Approval Form from Supervisor

AFRICA UNIVERSITY
A United Methodist-Related Institution

COLLEGE OF ENGINEERING AND APPLIED SCIENCES.

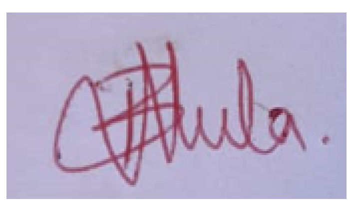
23/02/2024

Africa University Research Ethics Committee

Ref: Approval for AUREC Proposal Submission

…Anesu Makoni…. has worked on the proposal with the assistance of the supervisor and I confirm that it is ready for reviewed by your esteemed committee.

Respectfully submitted,



Mukhalela Braiton

------------------------------------ ------------------------------------

Supervisor's Name Supervisor's Signature

Appendix 2: Proof of payment

RTGS FUNDS TRANSFER CONFIRMATION RECEIPT

Your rtgs funds transfer transaction with the following details was

TRANSACTION DETAILS

Date Of Payment	19/06/2024
Transaction Reference	Anesu Makoni (210314) AUREC Proposal Review Fee
Transaction Amount	USD15.00

BENEFICIARY DETAILS

Beneficiary Bank Code	CBZ
Beneficiary Mobile Number	0784671973
Beneficiary Account Number	01322704290031
Beneficiary Reference	Anesu Makoni (210314) AUREC Proposal Review Fee
Beneficiary Name	Africa University

TRANSACTION STATUS

Transaction Status	Successful
Transaction ID	DF24171000001636

PAYER DETAILS

Source Account Number	00000280599425
Sender Mobile Number	263774385740
Payer Name	Piesec Investments P L

NMB BANK LIMITED
THIS IS AN AUTHORIZED DIGITAL STAMP
DATE: 19/06/2024
TEL: (+263) (242) 759651-9
EMAIL: enquiries@nmbz.co.zw

NOTICE:
NMB Bank Limited will never send you an e-mail requesting you to enter your personal details or private identification and/or authentication details.

1. This application is not a receipt. Sellers should confirm with their bankers that funds have been received before releasing of any goods.

Appendix 3: Application Form for AUREC Review

AFRICA UNIVERSITY
(A United Methodist-Related Institution)

AFRICA UNIVERSITYRESEARCH ETHICS COMMITTEE (AUREC)

APPLICATION FOR INITIAL REVIEW

NB: This form must be completed by all persons/teams applying for ethical review by AUREC. Upon completion by the investigator(s) /researcher(s) it should be submitted electronically to AUREC, aurec@africau.edu. . Application fees (to cover the costs of reviewing proposal) should be paid to the Africa University Business Office, and proof of payment should accompany each application. Please complete all sections of this application form. If there is insufficient space on the form you may use additional pages.

Check list

This checklist is meant to aid researchers in preparing a complete application package and to help expedite review by the AUREC. Please tick all boxes as appropriate (Indicate **N/A** where inapplicable).

CONTACT PERSON'S NAME : **__Anesu Sharleen Ma-koni________________________________**

CONTACT ADDRESS: 1894 Katsande Way New Marlborough Harare.

EMAIL ADDRESS : **asmakoni@africau.edu**

CONTACT NO: **+263 78 467 1973**

UNDERGRADUATES

		Applicant	AUREC
1	Application form duly completed		
2	Electronic version of research proposal to aurec@africau.edu		
3	Consent forms in English and local language of study population		
4	Advertisement or letter or card used for recruiting participants and any supplementary information *(if applicable).*		
5	Data collection tools being administered during the study in English and local language of study population *(if applicable)* included in the proposal		
6	Budget and timeframe included in the proposal.		
7	Approval letter from your academic supervisor/college or institution		
8	Approval letter from authorities where study will be conducted		
9	Application fee paid at AU Business Office and receipt (or copy) attached to application form.		

POST GRADUATES AND OTHER RESEARCHERS

		Applicant	AUREC
1	Application form duly completed		
2	Electronic version of full research proposal (chapter 1 – 3 completed) to aurec@africau.edu		
3	Proposal summary (see guidelines below)		
4	Consent form in English and local language of study population		

5	Advertisement or letter or card used for recruiting participants and any supplementary information *(if applicable).*		
6	Data collection tools being administered during the study in English and local language of study population (if applicable)		
7	Budget and timeframe		
8	Approval letter from academic supervisor/college or institution *(if you are a student)*		
9	Approval letter from authorities where study will be conducted		
10	Application fee paid at AU Business Office and receipt attached to application form.		
12	CV's for D Phil and Phd candidates.		

-Anesu Sharleen Makoni----------- 30/07/2024

Signature: Investigator/Researcher **Name** **Date**

1. **General information**

1.1. Study title: .Design and Implementation of a Customer Churn Prediction Model for Zimbabwean Banks using Machine Learning Techniques.

....

1.2. Name of Principal Investigator(PI)/ Researcher:___Anesu Sharleen Makoni_____________

1.3. Nationality of Investigator/Researcher: ________Zimbabwean__________________________________

1.4. Proposed date of start of study: _(dd/mm/yyyy)______01/06/2024__________________

1.5. Expected duration of study: __6 Months____

1.6. Study site(s) in Zimbabwe: _____Financial Sector: Survey of Zimbabwe Banks__________

1.7. Sites outside Zimbabwe: __None____

1.8. Study budget:____$200________ Source of Funding: __________Self_______

1.9. Is the researcher a student? **Yes**

1.10. If Yes, indicate the following:

1.10.1. Name and address of institution: ___Africa University__Fairview Road Off Nyanga Road Mutare. Box Address: P.O. Box 1320.______________________

1.10.2. College: College of Engineering and Applied Sciences.________________

1.10.3. Level of study Undergraduate/Master's/PhD ____Undergraduate______________

1.10.4. Name of Supervisor: _____Mr Brighton Mukhalela__________________

1.11. If No to question 1.10, then indicate the following:

1.11.1. Name and address of institution:

____

1.11.2. Academic Title of PI:

____

1.11.3. Existing Qualifications:

____

1.11.4. Co Investigators:

Names:	Qualifications	Institution

2. Statement by the investigator

I ______Anesu Sharleen Makoni__certify that the information in this application document and the accompanying documents is true and complete in all respects. I confirm that the application has NOT been rejected by any other ethics review committee.

Signature _______ Date: _____30/07/2024__________

3. Guidelines for the proposal summary: (Times New Roman, double line spacing, font size 12)
   3.1. Introduction
   3.2. Background ,purpose, statement of the problem, justification, significance of the study
   3.3. Aim(s) and objectives: Outline the main aim(s) and objectives of the study and research questions.
   3.4. Literature review

4.0 Methodology

4.1 Research Design *(describe how the research will be carried out including plans for data analysis and dissemination)*

4.2 Study population and sampling procedure*(give details of the study population and how you will carry out the sampling procedure and NOT general meanings of population and sampling methods)*

4.3 Inclusion/exclusion criteria*(state who qualifies for selection and who does not)*

4.4 Devices, Tests, Questionnaires, and Interview Guides:

4.5 Research participants/subjects

- 4.5.1 State the total number of human participants to be enrolled
- 4.5.2 State the source(s) of recruitment *(e.g. hospitals, schools, etc.)*
- 4.5.3 Age range and sex of participants to be recruited.
- 4.5.4 Special or vulnerable populations *(state if vulnerable populations e.g. pregnant women, adolescents, children, prisoners, refugees etc are involved)*
- 4.5.5 Payment *(if any)* to be paid to each participant
- 4.5.6 Informed Consent Procedure*(describe how this will be carried out)*

4.6 Potential Benefits of the research *(Describe the benefits of the study both to the participants and to the community)*

4.7 Potential Risks

- 4.7.1 Describe any potential risks, discomforts or harms that may be experienced by the participants. These may be physical, psychological, social, legal, economic or other and state procedures to minimise these.
- 4.7.2 Management of Risks*(describe how these risks will be managed/mitigated)*

4.8 Confidentiality/privacy *(give details of how these will be maintained)*

4.9 Investigator Experience/qualifications *(describe any experience or training/courses that the investigator has/has taken that put him/her in good stead to carry out the study)*

4.10 Explain how research results are going to be disseminated to participants

5. Reference List

6. Attachments

6.1 Approval letter from College Supervisor (if you are a student)

6.2 Data collection instruments *(Include anything you will be using to gather data from human subjects e.g. Tests/Questionnaires/Observation Checklists/interview guides/ FGDs guides etc.)*

6.3 Informed Consent Forms or assent (*<u>informed consent form guide is available from AUREC)</u>*

6.4 Budget and timeframe

6.5 Proof of payment of the review fees.