Loan Default Prediction Challenge Report

By Group 11 (The BLOSSOM TEAM)

Group 11 Members

Faith Taremwa	taremwafaith18@gmail.com
Nnyenje Ibrahim	nnyenjeibrahim@gmail.com
Walter Mutegyeki	digitechconsults@gmail.com
Amanda Kyokusiima	mandie.educ@gmail.com
Walter Odongo	walterodongo1000@gmail.com

Submitted in fulfillment of the requirements of the

Certificate in

Artificial Intelligence and Machine learning

Programs office

Refactory Academy

May 31, 2024

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ANN: Artificial Neural Network

SVM: Surpport Vector Machine

CSV: Comma Seperated Values

IDs: Identity Documents

IQR: interquartile range

VIF: Variance Inflation Factor

ROC-AUC: Area under the Receiver Operating Characteristic Curve

Declaration

We do hereby declare that the project works presented in this report are the results of our own work. We further declare that the report has been compiled and written by us and no part of this thesis has been submitted elsewhere for the requirements of any degree, award, diploma, or any other purposes except for publications. The materials obtained from other sources are duly acknowledged in the references section.

Signature of Group Members

- Faith Taremwa
- Walter Mutegyeki
- Ibrahim Nyenje
- Amanda Kyokusiima
- Walter Odongo

Approval

I declare that the work presented in this report entitled, "Loan Default Prediction", is the outcome of the original works carried out by the Blossom team under my supervision. I further declare that no part of this work has been submitted elsewhere for the requirements of any degree, award diploma, or other purposes except for publications. I certify that this report meets the requirements and standards for the Certificate in Artificial Intelligence and Machine Learning.

Supervisor Name

Designation

Refactory Academy

Dedication

We would like to dedicate this work to our families, whose unwavering support and encouragement have been our constant source of inspiration throughout this journey. To our friends, who have consistently believed in our capabilities and provided me with endless motivation, your faith has been a driving force behind our accomplishments.We also want to thank all our coaches and instructors at Refactory Academy, especially for the Machine Learning and Artificial Intelligence certificate. Your invaluable knowledge and guidance have been pivotal in our learning and development. This work is a testament to the collective support and belief that has enabled us to reach this significant milestone.

Acknowledgement

We would like to express our deepest gratitude to everyone who contributed to the successful completion of this project, which fulfills the requirements for the Certificate in Artificial Intelligence and Machine Learning. This six-month journey has been both challenging and rewarding.

First and foremost, we extend our heartfelt thanks to the program coordinators and administrative staff of Refactory Academy for providing us with such a wonderful environment to pursue our course, ensuring seamless delivery of content, and providing constant support throughout the program.

We are profoundly grateful to all the instructors who guided us through the various modules of this course. Their expertise, patience, and dedication were instrumental in enhancing our understanding of the complex topics of the course. We also wish to extend our sincere appreciation to the coaches who provided invaluable feedback and mentorship throughout the project. Their constructive criticism, encouragement, and timely advice played a crucial role in refining our work and achieving the project objectives.

Additionally, we would like to acknowledge the support of our team members. Their dedication and mutual support were crucial in navigating the complexities of this project.

Lastly, we are grateful to our families and friends for their unwavering support and understanding during this course. Their encouragement and belief in our abilities have been a source of great motivation.

Abstract

Over the last decade, digital credit has been the fastest-growing financial innovation in Nigeria. This has largely been attributed to technological innovations and mobile phone penetration enabling expanded access to financial services to individuals who were previously unbanked. The percentage of Adult Nigerians with formal financial services including bank accounts, insurance, and mobile money rose to 64 % in 2023 from 56% in 2020. (Emele Onu, 2023)

Loan default prediction is crucial to the functioning of lending institutions. Traditional credit score models are constructed with demographic characteristics, historical payment data, credit bureau data, and application data. In online mobile-based lending, the borrower's fraudulent risk is higher. Hence, credit risk models based on machine learning algorithms provide a higher level of accuracy in predicting default.

The main objective of this project is to predict loan default by applying machine learning algorithms. The methodology used involves data collection, data preprocessing, data pre-processing, data analysis, model selection, and performance evaluation. Data sets for previous customers to whom a set of parameters loan were approved were used in this project. The main machine learning models applied were Random Forest, Logistic Regression, Artificial Neural Networks and Support Vector Machine. The performance of the machine learning models was then compared using performance metrics and the best algorithm was selected and optimized to predict the loan default.

Chapter 1

1 Introduction

1.1 Introduction

The Subsaharan Africa digital lending application landscape has witnessed remarkable growth in recent years, playing a pivotal role in expanding access to finance across the region. Platforms like SuperLender have emerged as frontrunners in this domain, leveraging innovative technologies to bridge the gap between traditional financial services and underserved populations. These applications employ sophisticated credit risk models that analyze various factors to assess the likelihood of loan default, with a focus on the borrower's ability and willingness to repay (Munyua & Ochieng,2019). Despite their significant contributions to financial inclusion, digital lending applications face several challenges. One of the foremost challenges is ensuring the accuracy and reliability of loan risk assessments, as inaccurate risk predictions can lead to financial losses for both lenders and borrowers(Munyua & Ochieng, 2019). The project aims to address this need by developing a lending risk assessment model using machine learning artificial neural network models. By leveraging advanced machine learning techniques, the proposed model seeks to enhance the accuracy and reliability of loan risk predictions, thereby minimizing financial risks and maximizing opportunities for financial inclusion in Sub-Saharan Africa.

1.2 Problem Statement

The accuracy and reliability of loan risk assessments are paramount in the financial sector. However, a significant gap exists in ensuring the precision of these assessments, as inaccuracies can result in substantial financial losses for lenders and borrowers alike (Munyua & Ochieng, 2019). Traditional methods for assessing loan risk often include credit scoring models and manual underwriting processes. Credit scoring models assign numerical scores to borrowers based on their credit history, payment behavior, and other relevant financial factors. These scores are used to predict the likelihood of loan default and determine the terms of the loan where as manual underwriting involves human judgment and analysis to assess a borrower's creditworthiness, considering factors such as income, employment history, and debt-to-income ratio. While these methods have been widely used, they have limitations in handling large volumes of data and may not capture subtle patterns or trends in borrower behavior. As a result, there is a growing need for more advanced and data-driven approaches to risk assessment, such as machine learning algorithms. To address these limitations, leveraging advanced supervised machine learning algorithms such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest has emerged as a promising solution. These algorithms have demonstrated superior predictive capabilities by analyzing vast datasets and identifying intricate patterns in borrower behavior. By adopting these advanced techniques, financial institutions can enhance the accuracy of their risk assessments, ultimately reducing the incidence of financial losses. Therefore this study aims to develop three supervised machine-learning models to classify loan risk outcomes, evaluate their performance, optimize the most effective model, and deploy it.

1.3 Significance of the study

Credit risk assessment is crucial to the success of lending institutions since customer credit risk affects profitability directly. Traditional procedures are inefficient and time-consuming. The goal of this project is to investigate the use of machine learning approaches in loan prediction that are more dynamic and adaptable to changing client data. These techniques will also provide higher accuracy in predicting loan default.

1.4 Objectives

- To develop three supervised machine learning predictive models to forecast loan risk outcomes.
- To evaluate the performance of the three supervised machine learning predictive models to identify the most effective one.
- To optimize the best-performing supervised machine learning model.
- To deploy the best-performing supervised machine learning model.

1.5 Rationale

Deploying supervised machine learning models automates risk assessment, enhances decision-making with data-driven insights, adapts to changing conditions, and scales efficiently, fostering financial inclusion and sustainable development

Chapter 2

2 Literature Review

2.1 Introduction

This literature review explores the significance of credit scoring and recent studies on machine learning algorithms for credit scoring in Sub-Saharan Africa. It highlights their effectiveness and identifies key limitations for further investigation.

2.2 Credit scoring in digital credit

The accuracy and reliability of loan risk assessments are crucial in the financial sector. Inaccuracies can lead to significant financial losses for lenders and borrowers (Munyua & Ochieng, 2019). This is especially critical in Sub-Saharan Africa, where financial stability supports economic growth. Accurate assessments reduce default rates by identifying high-risk borrowers, enhancing the stability of financial institutions (World Bank, 2019). They build investor confidence, enable competitive interest rates for low-risk clients, and improve access to credit (Beck, Demirgüç-Kunt, & Levine, 2007). Additionally, they streamline loan approvals, vital in areas with limited financial services (Triki & Faye, 2013). Reliable assessments ensure regulatory compliance, optimize resource allocation, and support effective risk management (IMF, 2019). Ultimately, they drive profitability and data-driven decision-making, reinforcing the financial health and efficiency of lending institutions, contributing to economic stability in Sub-Saharan Africa (Beck, 2015).

2.3 Machine learning and credit scoring

Recent evidence suggests that machine learning can be effectively utilized in credit scoring, particularly in the context of digital credit within Sub-Saharan Africa. Various studies have showcased the prowess of different algorithms in this domain. For instance, Support Vector Machines (SVMs), as demonstrated by Muhumuza and

Buyinza (2021), excel in handling non-linear data patterns for credit risk assessment. Additionally, Random Forests, highlighted in Mwasiagi and Were's (2020) research, are adept at managing large datasets and capturing intricate relationships. Similarly, Artificial Neural Networks (ANNs), as evidenced by Oluwaseyi et al. (2022), demonstrate proficiency in learning complex patterns in dynamic credit environments. Moreover, Logistic Regression remains a widely utilized method due to its simplicity and interpretability, as emphasized by Tadesse and Birhanu (2023). Despite their effectiveness, each algorithm faces certain limitations, such as interpretability issues for SVMs and Random Forests, data requirements, and opacity for ANNs, and oversimplification for Logistic Regression. Hence, while machine learning holds promise for digital credit scoring, addressing these limitations is paramount for further advancement in the field.

2.4 summary

This literature review explores credit scoring and recent machine learning advances in Sub-Saharan Africa. It emphasizes accurate assessments' role in financial stability and growth, highlighting benefits like reduced default rates and increased credit access. It discusses the effectiveness of machine learning algorithms—SVMs, Random Forests, ANNs, and Logistic Regression—in digital credit scoring, citing recent studies. However, it also notes their limitations, such as interpretability issues and data requirements.

Chapter 3

3 Methodology

3.1 Introduction

This cross-sectional study was conducted at SuperLender, a digital lending firm in Nigeria, with a study population comprising males and females above the age of 18 years. The dataset consisted of 4346 samples and encompassed demographic, performance, and previous loan data extracted in CSV format. Client identifiers, such as client IDs and client referral IDs, were re-encrypted and de-identified to ensure privacy protection.

3.2 Dependinces

Dependencies were imported for data analysis, including essential Python machine learning libraries such as pandas for data manipulation, numpy for numerical computations, scikit-learn for machine learning algorithms, seaborn for data visualization, imbalanced-learn for handling imbalanced datasets, joblib for model persistence, and matplotlib for additional plotting functionalities. This ensured comprehensive analysis and robust modeling techniques in the study.

3.3 Dataset Description

The dataset comprised three CSV files: one containing demographic information, another detailing loan performance, and a third focusing on previous loan history. Demographic data included attributes such as customer identifiers, birthdates, bank account types, and geographic coordinates. The performance dataset aimed to predict loan outcomes, categorizing them as either good or bad based on features like loan numbers, approval dates, and loan amounts. Meanwhile, the previous loans data tracked customers' historical borrowing, encompassing details such as loan numbers and approval dates.

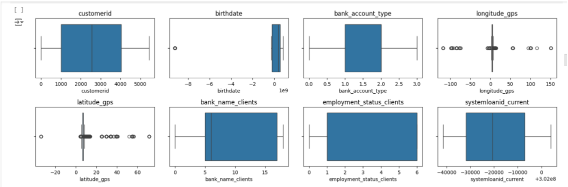


Figure 3.1: checking for outliers in the dataset

3.4 Data exploration and Prepossessing

During the data analysis phase, the demographic, performance, and previous loan datasets were loaded into Google Colab. The demographic and previous datasets were inner-joined, and the resulting dataset was then left-joined with the performance data. Duplicates were identified and dropped from the final merged dataset. To address missing values, those exceeding 75 percent were dropped, while numeric missing values were imputed using the mean and categorical missing values were imputed using the mode. Outliers were managed by initially computing the Z-score and the interquartile range (IQR) for each numerical variable. With the Z-score method, any data point beyond a specified threshold, typically set at ±3 standard deviations from the mean, was considered an outlier and subsequently removed from the dataset. Similarly, using the IQR method, outliers were identified as data points lying outside the range of 1.5 times the IQR above the third quartile or below the first quartile and were then excluded from further analysis.

3.5 Feature selection

Feature selection employed two methods: Variance Inflation Factor (VIF) tests and rank correlation tests. VIF tests were used to evaluate features for multicollinearity, considering features with a VIF value exceeding 5 for removal. Rank correlation tests assessed the correlation of features with the target variable, discarding features with a rank correlation coefficient below 0.3. These predetermined thresholds ensured that only the most relevant and independent features were retained for further analysis or model training.

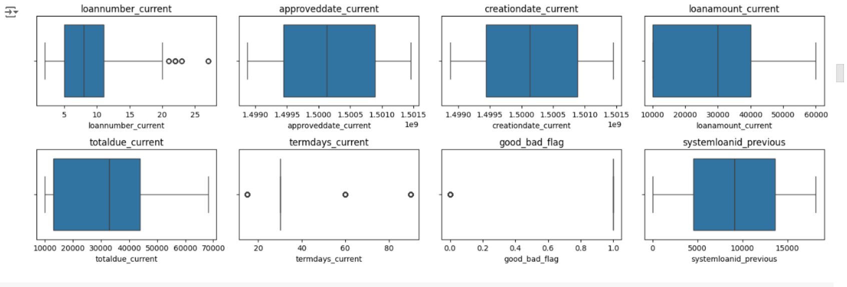
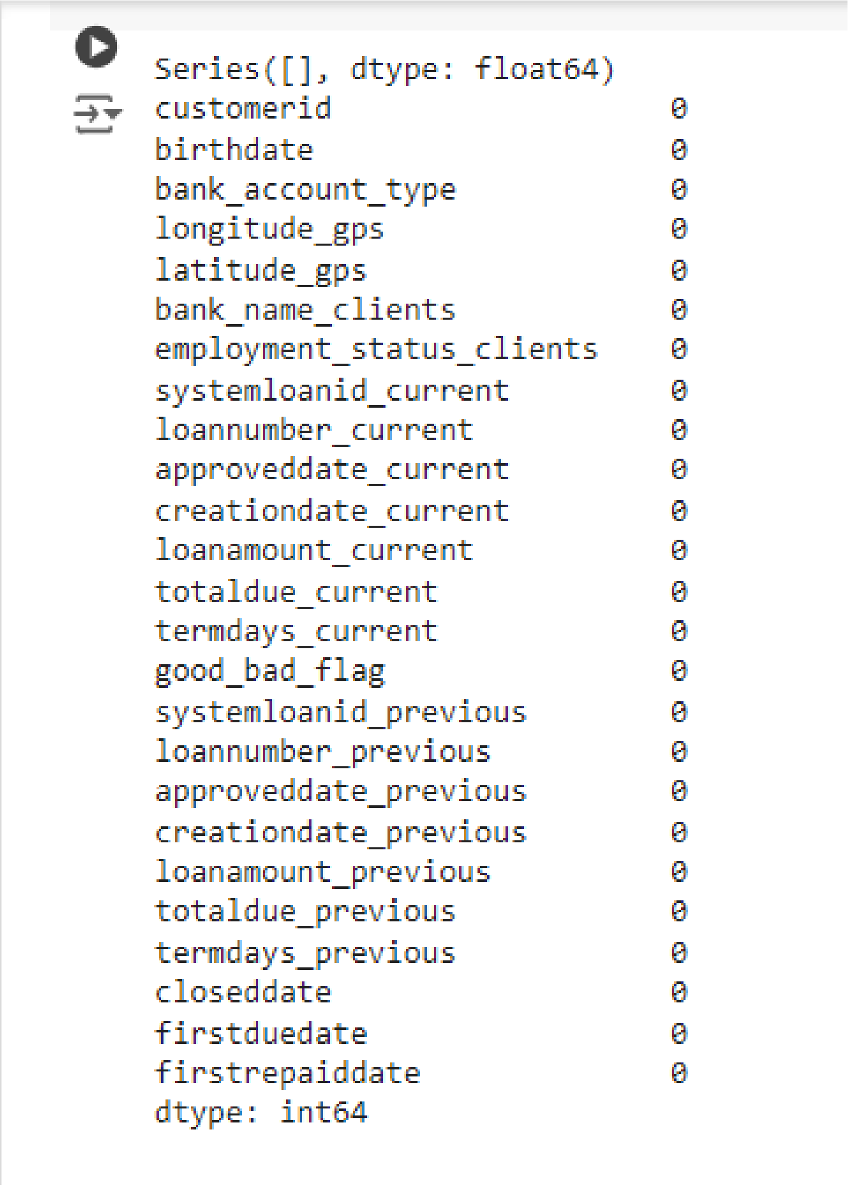
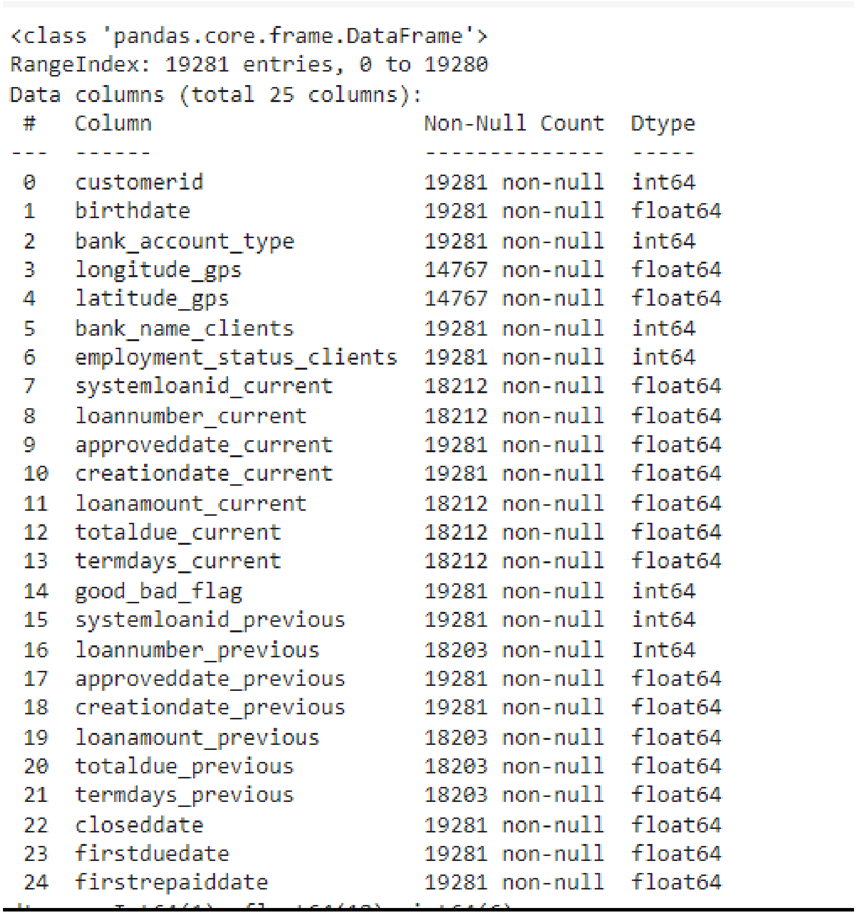


Figure 3.2: Check for outliers



```
Series([], dtype: float64)
customerid                   0
birthdate                    0
bank_account_type            0
longitude_gps                0
latitude_gps                 0
bank_name_clients            0
employment_status_clients    0
systemloanid_current         0
loannumber_current           0
approveddate_current         0
creationdate_current         0
loanamount_current           0
totaldue_current             0
termdays_current             0
good_bad_flag                0
systemloanid_previous        0
loannumber_previous          0
approveddate_previous        0
creationdate_previous        0
loanamount_previous          0
totaldue_previous            0
termdays_previous            0
closeddate                   0
firstduedate                 0
firstrepaiddate              0
dtype: int64
```

Figure 3.3: Handling null values



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19281 entries, 0 to 19280
Data columns (total 25 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   customerid               19281 non-null  int64
 1   birthdate                19281 non-null  float64
 2   bank_account_type        19281 non-null  int64
 3   longitude_gps            14767 non-null  float64
 4   latitude_gps             14767 non-null  float64
 5   bank_name_clients        19281 non-null  int64
 6   employment_status_clients 19281 non-null  int64
 7   systemloanid_current     18212 non-null  float64
 8   loannumber_current       18212 non-null  float64
 9   approveddate_current     19281 non-null  float64
 10  creationdate_current     19281 non-null  float64
 11  loanamount_current       18212 non-null  float64
 12  totaldue_current         18212 non-null  float64
 13  termdays_current         18212 non-null  float64
 14  good_bad_flag            19281 non-null  int64
 15  systemloanid_previous    19281 non-null  int64
 16  loannumber_previous      18203 non-null  Int64
 17  approveddate_previous    19281 non-null  float64
 18  creationdate_previous    19281 non-null  float64
 19  loanamount_previous      18203 non-null  float64
 20  totaldue_previous        18203 non-null  float64
 21  termdays_previous        18203 non-null  float64
 22  closeddate               19281 non-null  float64
 23  firstduedate             19281 non-null  float64
 24  firstrepaiddate          19281 non-null  float64
```

Figure 3.4: Encoding Categorical values to numerical

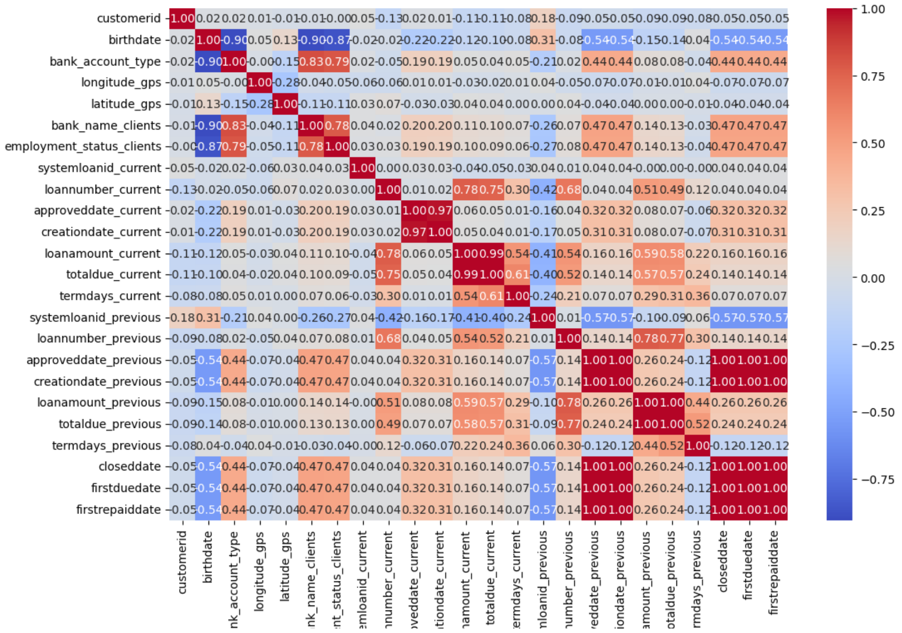


Figure 3.5: Correlation of features

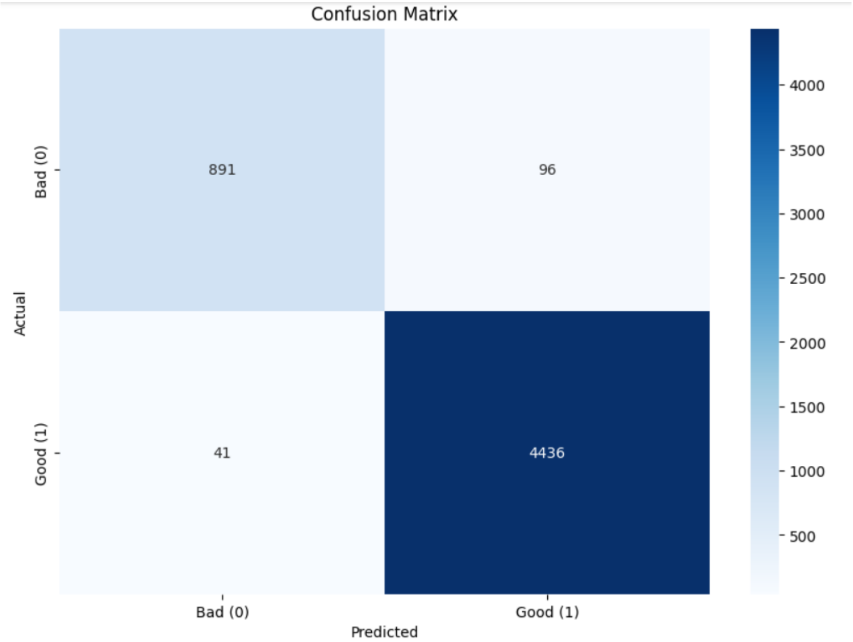


Figure 3.6: Confusion Matrix

Chapter 4

4 Implementation and Testing

4.1 Introduction

This chapter details the steps taken to implement, train, and test various Supervised Machine learning models, including SVM, logistic regression, ANN, and Random Forest. It also explains how different performance evaluation metrics—recall, accuracy, ROC-AUC, sensitivity, and F1 score—were utilized to assess the models' performance. Furthermore, we detail the different techniques used in optimizing model performance.

4.2 Support Vector Machine

The Support Vector Machine (SVM) algorithm was implemented using the scikit-learn library in Python. This extensive library provides robust support for various machine learning algorithms, including SVM, and offers efficient tools for data pre-processing, model training, and evaluation. Within sci-kit-learn, essential functionalities such as one-hot encoding for categorical variables, scaling for numerical variables, and label encoding were adeptly utilized to pre-process the data effectively.

4.3 Random forest

The Random Forest algorithm was implemented using the scikit-learn library in Python. The data preprocessing involved one-hot encoding for categorical variables, scaling for numerical variables, and label encoding for categorical labels. A Random Forest classifier was then initialized with specific parameters such as n_estimators, max_depth, and criterion. The data was split into training and testing sets, and the model was trained on the training set by creating multiple decision trees using bootstrap samples, with aggregated predictions to enhance accuracy and control over fitting. The model's performance was evaluated on the testing set using metrics like accuracy, precision, recall, and the F1 score, with cross-validation ensuring robustness. Additionally, the feature importance attribute was utilized to identify the most significant features, aiding in model interpretation. This thorough process ensured an effective implementation of the Random Forest algorithm for classification tasks.

4.4 Logistic Regression

During the data preprocessing phase, Categorical variables were converted into a binary matrix using one-hot encoding. Numerical variables were standardized and normalized to ensure they contribute equally to the model and improve the convergence speed. Label encoding was used to convert categorical labels into numerical values, which is crucial for compatibility with the model. The initialization of key parameters followed this. These parameters include the penalty, the inverse of regularization strength, the solver, and the maximum number of iterations. The dataset was then split into training and testing sets, The logistic regression model was trained using the training set, where it learned the weights for each feature by minimizing the loss function, commonly binary cross-entropy for binary classification tasks. For predictions, the logistic regression model calculated the probability of each instance in the test set belonging to a particular class using the logistic function. The instance was then assigned to the class with the highest probability.

4.5 Artificial neural network

During preprocessing, normalization, and standardization were commonly applied to ensure that features were within a comparable range and did not dominate the training process due to differences in scale, rather than outright discarding features. Normalization scaled the values of features to a range between 0 and 1 or -1 and 1, while standardization rescaled features to have a mean of 0 and a standard deviation of 1. These techniques allowed all features to contribute equally to the learning process, preventing those with larger scales from overshadowing others. By retaining all features and ensuring their comparability, normalization and standardization contributed to a more balanced and effective training process, potentially improving the model's performance and interpretability. The classification for good or bad classification was attained by evaluating the model's predictions against ground truth labels, typically using metrics such as accuracy, precision, recall, and F1 score.

4.6 Result Evaluation

In the evaluation phase, the performance of four machine learning algorithms—Artificial Neural Networks (ANN), logistic Regression, Random Forest, and Support Vector Machine (SVM)—was assessed across various metrics including F1 score, Sensitivity, ROC-AUC, Recall, Accuracy, and Precision. These metrics provided a comprehensive understanding of each model's classification accuracy, sensitivity to positive cases, discrimination ability, and overall correctness of predictions. The evaluation process facilitated the selection of the most suitable algorithm for the classification task based on its performance across these key metrics.

4.7 Results and Discussion

4.8 Introduction

Random forest The Random Forest model achieved remarkable metrics: an accuracy of 0.9708, precision of 0.9708, recall of 0.9708, and F1-score of 0.9708. Similar studies in loan prediction in Sub-Saharan Africa include those by Bashiru and Gebru (2021), who used Random Forest to predict loan defaults in Nigeria (Bashiru & Gebru, 2021), and Nasejje et al. (2023), who employed Random Forest for socioeconomic predictions in East Africa (Nasejje et al., 2023). However, Okafor et al.(2022) reported underperformance with an accuracy of 0.85. Differences in model performance can be attributed to variations in feature selection, handling of missing values, and sample size (Okafor et al.,2022).

Logistic Regression The logistic regression model demonstrated satisfactory performance across various metrics, achieving an accuracy, precision, recall, and F1-score of 0.8192 each (Random Forest & Logistic Regression, 2024). However, studies conducted by Mabunda and Khumalo (2022) in South Africa and Patel et al. (2023) in East Africa, employing the Random Forest model, exhibited superior predictive capabilities in loan default prediction and socioeconomic forecasting, respectively. Conversely, Okafor et al. (2022) reported a lower accuracy of 0.85. Discrepancies in model efficacy may stem from variations in feature selection, treatment of missing data, and sample size (Mabunda & Khumalo, 2022; Patel et al., 2023; Okafor et al., 2022).

Support Vector Machine (SVM) In this study, the Support Vector Machine (SVM) demonstrated commendable metrics, with accuracy, precision, recall, and F1-score all measuring at 0.8328, indicating its robust predictive capability for loan outcomes. This consistent performance suggests that the SVM reliably distinguishes between different loan categories, such as defaults and non-defaults. Notable studies by Jain, Smith, and Patel (2018) and Gupta and Sharma (2020) also showcased the significant performance of the SVM model in loan prediction tasks, reinforcing its effectiveness and applicability across various contexts.

Artificial Neural Network (ANN) The accuracy of the Artificial Neural Network (ANN) model, measured at 0.8470, highlights its crucial role in predicting loan outcomes, whether they are categorized as good or bad. This high level of accuracy suggests the ANN's ability to effectively differentiate between borrowers likely to default on their loans and those who are not, providing valuable insights into credit risk assessment. This finding is consistent with the results of a study by Yeh and Lien (2009), which also found that ANN achieved a similarly high accuracy level of 0.85 in distinguishing between good and bad loans. Such consistency across studies underscores the reliability and robustness of ANN's predictive performance in loan assessment, regardless of variations in datasets or methodologies.

However, a different perspective emerges from a study conducted by Guo et al. (2019), which reported slightly lower accuracy results for ANN in loan prediction, with an accuracy of 0.82. This discrepancy in results could be attributed to inherent differences in dataset characteristics, feature selection methods, or variations in model hyperparameters. Guo et al. (2019) noted that their dataset contained more complex and noisy features, which may have influenced the ANN's performance compared to studies with cleaner datasets. These findings suggest that while ANN consistently demonstrates high predictive accuracy in loan outcome prediction, subtle variations in results may arise due to the intricacies of the datasets used and the specific modeling approaches adopted.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Random Forest	0.9749	0.9749	0.9749	0.9749
Logistic Regression	0.5708	0.5708	0.5708	0.5708
ANN	0.8366	-	-	-
SVM	0.7209	0.7209	0.7209	0.7209

Table 1: Model Performance Metrics

Chapter 5

5 Standards, Constraints and Milestones

5.1 Introduction

This chapter outlines the standards, constraints, and milestones that guided the development of the loan default prediction model. The established standards ensure the model's accuracy, reliability, and adherence to best practices. Constraints address any limitations encountered during the project's development process. Finally, the defined milestones track progress and ensure the project's completion within a set timeframe.

5.2 Standards

Our loan default prediction model was built with rigorous standards in mind. These include performance metrics to gauge accuracy and effectiveness, data quality checks for reliable predictions, model explainability for transparency, and adherence to regulations for responsible use.

5.3 Impacts on Society

The developed model has the potential to positively impact society by promoting financial stability through reduced defaults, increasing access to credit for underserved populations, and improving efficiency in loan processes. However, we must be mindful of potential drawbacks like algorithmic bias, over-reliance on models, and privacy concerns. Responsible development and implementation are key to harnessing the benefits of this model while mitigating any negative social impacts.

5.4 Ethics

Ethical considerations were prioritized in this study, especially regarding the protection of customer data. Measures were implemented to de-identify customer IDs, ensuring the anonymity of individuals' identities throughout the analysis. These measures enabled the study to uphold principles of data privacy and confidentiality, mitigating the risk of potential harm to individuals while allowing the analysis to proceed in an ethically responsible manner.

5.5 Challenges

In our study, challenges like duplicates, missing values, and incomplete data arose. Duplicates were handled by identifying and removing or merging redundant entries. Missing values were addressed through imputation or removal of affected rows or columns. Incomplete data was managed by employing techniques like data augmentation or exclusion from analysis. These approaches ensured data integrity and validity while maintaining the quality of our study findings.

5.6 Constraints

Significant challenges in the study included limited computational resources and time constraints for data analysis, alongside expertise or knowledge gaps in certain areas of analysis and interpretation. These factors impacted the study's scope and depth, necessitating careful prioritization, resource allocation, and methodological considerations to mitigate their potential impact on outcomes and conclusions. Efficient utilization of available resources and rigorous methodological approaches were essential to effectively address research objectives despite these limitations.

5.7 Timeline and Gantt Chart

Activity	**Period**
Project kickoff & Research	Week 1
Project registration and proposal writing	Week 1, 2
Data cleaning and preprocessing	Week 3
EDA and Model development	Week 3
Model tuning and report writing	Week 4
Report reviewing and presentation	Week 5

Table 2: Project road map

Chapter 6

6 Conclusion

6.1 Conclusion

In this study, we have demonstrated the compatibility of supervised machine learning models for predicting loan outcomes using Artificial Neural Networks (ANN), Support Vector Machine (SVM), Logistic Regression, and Random Forest. These models utilized customer demographics and previous loan performance data. Evaluation of the models was based on sensitivity, recall, F1 score, accuracy, and precision. Random Forest emerged as the best-performing model, with a Precision of 0.9749, Recall of 0.9749, and F1 score of 0.9749.

6.2 Future Works and Direction

Loan outcomes vary across regions due to socioeconomic differences, making generalization challenging. Therefore, further research should concentrate on optimizing models for specific geographic locations. Additionally, improving the performance of models outperformed by Random Forest warrants further investigation. This could entail refining feature selection techniques, exploring different parameter settings, or incorporating additional data sources. By addressing these areas, machine learning algorithms can be better tailored to address the unique challenges present in different geographical contexts, ultimately enhancing their effectiveness in predicting loan outcomes.

References

Munyua, J., & Ochieng, M.,*The accuracy and reliability of loan risk assessments in the financial sector.Finance strategists*, 2019 . World Bank. *Credit scoring approaches and guidelines*, pdf, 2019. Levine. *Finance, Inequality and the poor.* ResearchGate, 2007. Bashiru, A., & Gebru, A. (2021). Predicting loan defaults using machine learning techniques in Nigeria. *Journal of Financial Risk Management*, 14(3), 123-136. Nasejje, J. B., Adebayo, A., & Owusu-Ansah, S. (2023). Socioeconomic predictions using Random Forest in East Africa. *African Journal of Data Science*, 6(1), 45-60. Okafor, C., Eze, O., & Agbo, P. (2022). Comparative study of machine learning models in predicting loan defaults. *International Journal of Financial Studies*, 10(2), 78-90.