Capstone Project Proposal

Loan default Prediction challenge

Group 11 Members

Name	Email	Contact
Faith Taremwa	taremwafaith18@gmail.com	0775357866
Nnyenje Ibrahim	nnyenjeibrahim@gmail.com	0780920461
Walter Mutegyeki	digitechconsults@gmail.com	0787594467
Amanda Kyokusiima	mandie.educ@gmail.com	0779339942
Walter Odongo	walterodongo1000@gmail.com	0743051849

Table 1: Group Members Information

Advisor: Mr.Solomon Nsubuga

REFACTORY ACADEMY
Certificate in Artificial Intelligence and Machine Learning Skills

May 30, 2024

1 Introduction:

The Subsharan 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 undeserved 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 networks 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.

2 Problem statement

The accuracy and reliability of loan risk assessments are paramount in the financial sector. However, there exists a significant gap 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.

2.1 Objectives

- Develop three supervised machine learning predictive models to forecast loan risk outcomes.
- Evaluate the performance of the three supervised machine learning predictive models to identify the most effective one.
- Optimize the best-performing supervised machine learning model.
- Deploy the best-performing supervised machine learning model.

2.2 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.

3 Methodology:

3.1 i) Project development environment

Data processing, feature engineering, model development, training, and testing will take place on the Google Colab platform, leveraging its extensive features that facilitate collaboration. All code for the entire model will be authored in Python, capitalizing on its robust capabilities for developing, training, and testing machine learning models. Additionally, version control will be managed using Git, ensuring efficient tracking of code changes and seamless collaboration among team members.

3.2 ii) Pre-Development methodologies

The project will adhere to a structured approach, delineated into distinct phases to ensure systematic progress and effective management. These phases encompass:

3.3 a) Data Loading:

The initial step in the project entails acquiring the requisite datasets from diverse and relevant sources. This involves identifying reputable data repositories, APIs, or proprietary databases that contain the necessary information for analysis. Careful consideration will be given to the selection of datasets to ensure they align with the project's objectives and encompass a comprehensive representation of the target domain.

3.4 b) Data Cleaning

The data cleaning methodology involves systematically addressing data quality issues to ensure the integrity and reliability of the dataset. This process begins with identifying missing values, outliers, duplicates, and inconsistencies, followed by appropriate strategies for handling each issue. Missing values shall be imputed or removed based on statistical measures, outliers evaluated and either retained, transformed, or removed, and duplicate records are identified and eliminated. Additionally, data standardization, normalization, and transformation techniques will be applied to enhance data consistency and comparability. Throughout the process, meticulous documentation will be maintained, detailing the steps taken and decisions made. Validation and quality assurance checks will be conducted to ensure that the cleaned dataset meets the desired standards for subsequent analysis and modeling endeavors.

3.5 c) Data Overview

Following data acquisition, a comprehensive examination of the dataset will be conducted to gain insights. This step will entail a thorough examination of a dataset to effectively understand its characteristics, distributions, and potential challenges. Following data acquisition and cleaning, this foundational phase will employ various exploratory data analysis techniques, including descriptive statistics and data visualization. The primary objectives will encompass gaining insights into dataset structure, detecting outliers or anomalies, assessing data quality, and identifying initial trends and patterns. Through systematic exploration, the data overview phase will provide essential groundwork for hypothesis generation and informed decision-making, ultimately facilitating actionable insights from the dataset.

3.6 d) Feature Engineering

Feature engineering will involve a systematic approach to enhancing the predictive power of our machine learning model by manipulating the dataset's features. This process entails identifying, extracting, selecting, and transforming features to better represent the underlying patterns and relationships within the data. By leveraging domain knowledge, statistical techniques, and data exploration methods, feature engineering aims to create informative features that improve model interpretability and predictive accuracy. Through careful iteration and validation, feature engineering will facilitate the creation of more robust and effective machine learning models, enabling better decision-making and actionable insights from the data.

3.7 iii) Model development

For feature scaling, the data will be standardized or normalized using libraries like scikit-learn. Train-test data will be split into a 70-30 ratio using the train_test_split function from scikit-learn. Random Forest will be trained using the RandomForestClassifier from scikit-learn, optimizing parameters like the number of trees and maximum depth. Artificial Neural Network (ANN) will be trained using Keras or TensorFlow, with considerations for the number of layers, nodes, and activation functions. Support Vector Machine (SVM) will be trained using SVC from scikit-learn, tuning parameters like the kernel type and regularization parameter.

3.8 iv) Model evaluation and selection

To evaluate model efficiency and performance metrics for Random Forest, ANN, and SVM models, we'll utilize a confusion matrix. For each model, we'll calculate accuracy, F1 score, MSE, precision, recall, sensitivity, and specificity. Accuracy represents the proportion of correctly classified instances, while F1 score balances precision and recall. MSE measures the average squared difference between predicted and actual values. Precision is the ratio of correctly predicted positive observations to the total predicted positives. Recall, also known as sensitivity, measures the proportion of actual positives correctly identified. Specificity is the ratio of correctly predicted negative observations to the total predicted negatives.

3.9 v) Model optimization

To optimize the best model, we'll employ hyper-parameter tuning techniques such as grid search or randomized search. These methods systematically explore a range of hyper-parameters to find the combination that yields the best performance. Additionally, we can employ techniques like cross-validation to ensure the robustness of our results. By iteratively adjusting hyper-parameters and evaluating model performance, we aim to enhance the model's predictive accuracy and generalization ability. Finally, we'll select the hyper-parameters that result in the highest performance metrics and retrain the model using these optimized parameters.

3.10 v) Model deployment

For model deployment, we'll utilize Joblib to save the trained model object. In Python Django, we'll create a web application following the Model-View-Controller (MVC) architecture. The model will be integrated into the back-end logic. We'll expose the model functionality through SOAP or RESTful web services to interact with the front-end. These services will handle requests, process data, and return predictions to the user interface. Finally, the deployed model will be accessible through the web application, allowing users to input data and receive predictions in real-time.

4 Possible limitations

Study limitations for the ANN, SVM, and Random Forest models include potential biases arising from incomplete or inaccurate data, along with data privacy concerns and associated risks. Addressing these issues will involve implementing rigorous data pre-processing and identification measures tailored to each model's requirements, including robust data cleaning methods such as outlier detection and imputation to enhance input data quality. Additionally, we will ensure transparency and accountability by thoroughly documenting data processing steps and model assumptions specific to the respective models.

5 Timeline:

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

6 References

Munyua, J., Ochieng, M. (2019). *The accuracy and reliability of loan risk assessments in the financial sector.* Retrieved from Finance Strategists. Accessed on

Muthoni, M. I. (2020). Credit management practices and loan performance: empirical evidence from commercial banks in Kenya. International Journal of Current Aspects in Finance, Banking and Accounting, 2(1), 51-63.