BUILDING A SEMI SUPERVISED FRAMEWORK TO IMPROVE CREDIT RISK PREDICTION

SRS SUBMITTED TO THE FACULTY OF SCIENCE, TECHNOLOGY AND INNOVATION DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY IN PARTIAL FULFILMENT OF THE AWARD OF BACHELOR OF SCIENCE DEGREE IN DATA SCIENCE

SUBMITTED BY

SAMUEL CHIZUMA(BSDS0821)

ELIZABETH MFUNE (BSDS1922)

QUEEN SOSOLA (BSDS2822)

JEREMIA NKOSI (BSDS2422)

TAMANDANI YONA(3022)

MAXWELL MWALA (BSDS2322)

SUPERVISOR

MR. REUBEN MOYO

Abstract:

The scarcity of comprehensive labelled data in African financial markets have led to the development of techniques to counter this gap. This particular project aims to utilize semisupervised learning techniques namely self-training and co-training to improve model predictive power in predicting credit risk for Malawi's Higher Education Students Loans and Grants Borad (HELGB) hence create a semisupervised learning framework that can be used to improve similar models in diverse markets. The project will utilize student loan data to achieve this. The dataset will undergo cleaning, preprocessing, EDA as well as feature engineering before it is used to train various classifier models in self-training, co-training, the final machine learning model and benchmarks. The project also aims to produce a machine learning model that predicts credit risk with an accuracy score above 83% including an interface through which stakeholders will be able to utilize the model and visualize results of predictions whereby the goal is to help them make better financial and lending decisions.

Table of Contents

Introduction 4

Data requirements 4

Data collection: 4

Data description: 5

Exploratory data analysis: 5

NOTE: 6

Data cleaning and preprocessing: 8

Feature Engineering: Creating and Selecting Features 9

Pre-processed Data Summary 10

Technical requirements 10

Software and tool 10

Hardware requirements 11

Ethics and privacy 11

Appendices 11

Glossary 11

References 12

Introduction

Assessing credit risk is a crucial responsibility within the financial sector, as it determines a borrower's likelihood of repaying a loan. Accurate credit risk assessment enables financial institutions to make informed lending decisions, mitigate losses and maintain a healthy portfolio. Traditional risk models in financial institutions rely heavily on supervised learning, which demands extensive labelled datasets. However, the acquisition of such dataset is often hindered by high costs as Labelling financial data requires significant resources, including expertise and time which can be costly, lengthy process of labelling because collecting, cleaning and labelling large datasets can be a time-consuming process, delaying the development and deployment of credit risk models as well as privacy concerns since financial data is sensitive and subject to strict regulations, making it challenging to collect and share labelled datasets. This project investigates the potential of semisupervised learning to enhance credit risk prediction, leveraging both labelled and unlabelled data to improve model accuracy and efficiency. By integrating a smaller labelled dataset with a larger pool of unlabelled data, this method offers a practical approach to addressing the limitations of traditional models. The goal is to develop an efficient credit risk prediction framework that reduces dependency on labelled data while maintaining high accuracy.

Data requirements

Data collection:

Relevant data is available but limited in size which poses a significant challenge. This makes the dataset less robust for training accurate machine learning models. To address this challenge, we will employ data augmentation techniques to synthetically generate additional data points, increasing overall volume and diversity of the dataset. Specifically, SMOTE will be used to create new samples that resembles the existing data. This approach will help expand our dataset providing a more comprehensive foundation for model training. SMOTE helps to balance the dataset, helps reduce the risk of overfitting which in turn helps accuracy and reliability of our credit risk predictions.

The synthetic dataset and the original dataset will be combined for better generalisation. It also provides a comprehensive understanding of credit risk patterns ena bling models to better capture complex relationships

Data description:

The initial dataset is an excel spreadsheet which follows the .xlsx format containing 355 rows and 30 columns. The initial dataset is of size 355 by 30 containing data from the HIGHER EDUCATION LOANS and GRANTS BOARD (HELGB) as well as loan beneficiary responses to loan application and repayment process related questions. The majority of features contain categorical data as marked by more than 25 features containing qualitative values. The dataset to be used for modelling however will contain synthesized data created from the initial dataset using oversampling, generative modelling and random sampling. The dataset will potentially comprise of not less than 1000 rows and it might contain less or more features depending on the results of preprocessing and feature engineering as well as feature selection. In the initial dataset, the feature "repayment_status" was regarded as the target variable. It represents whether a beneficiary repaid the loan or not. The dataset also contains various predictor variables for example, age_range and total_loan which represent various age ranges of beneficiaries as well as the amount of loan applied for respectively.

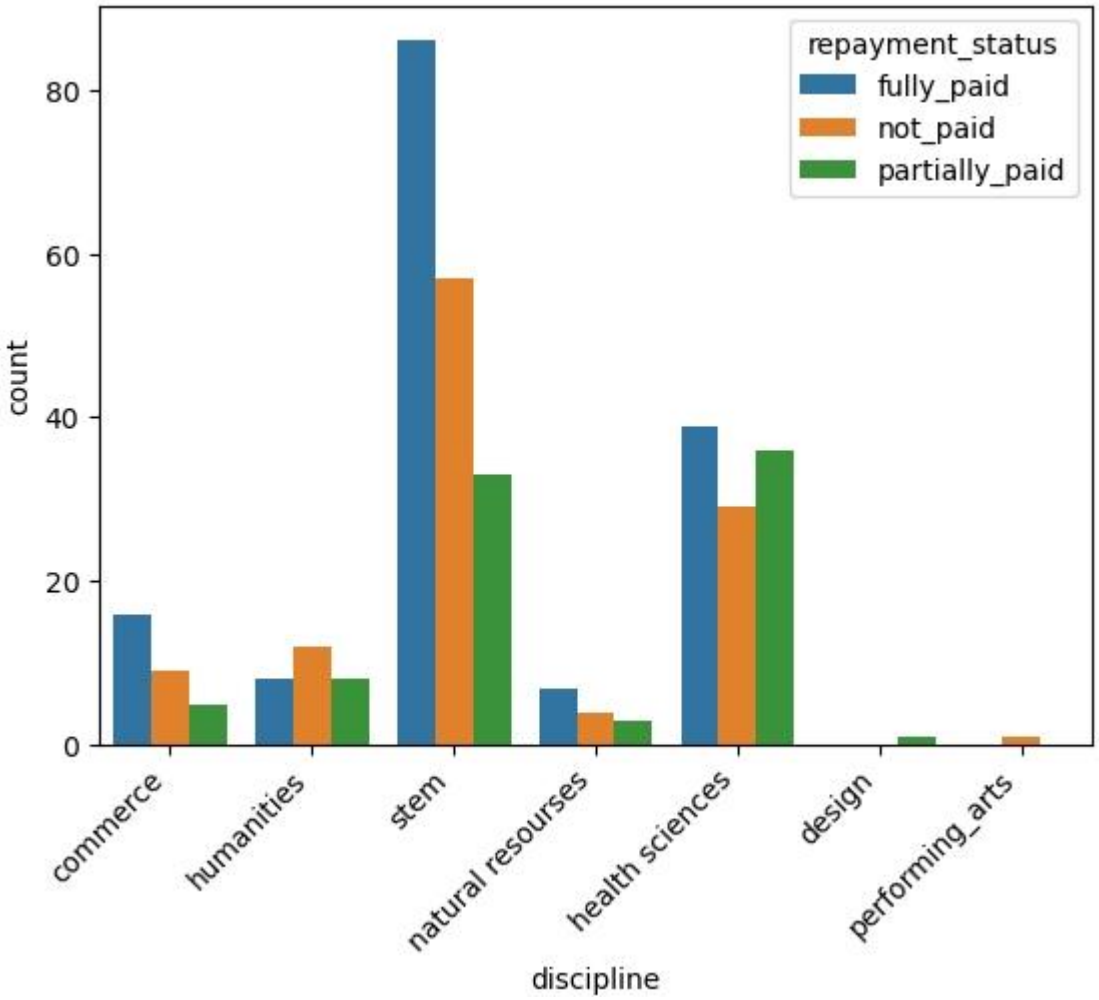
Exploratory data analysis:

1. Univariate analysis:
    a. Frequency distribution tables will be produced for all the features. These tables will be used to check the distributions of all the individual features including missing values.
    b. Bar charts as well as histograms will produced to visualize distributions. These will also be used to detect outliers based on occurrence of values.
    c. Special bar charts visualized in terms of the loan repayment status will be produced to detect outliers.
2. Multivariate analysis:
    a. Cross tabulations. Cross tabulations between the features and the target variables will be created to analyze the relationships between various features and the predictor variable.
    b. Grouping data according to different features to check weighted statistics and distributions. For example, grouping data according to gender, marital status and the loan repayment status.
    c. Scatterplots. Different features will be compared to the target variable using scatterplots.

Frequency distribution table for the feature 'highest_education'

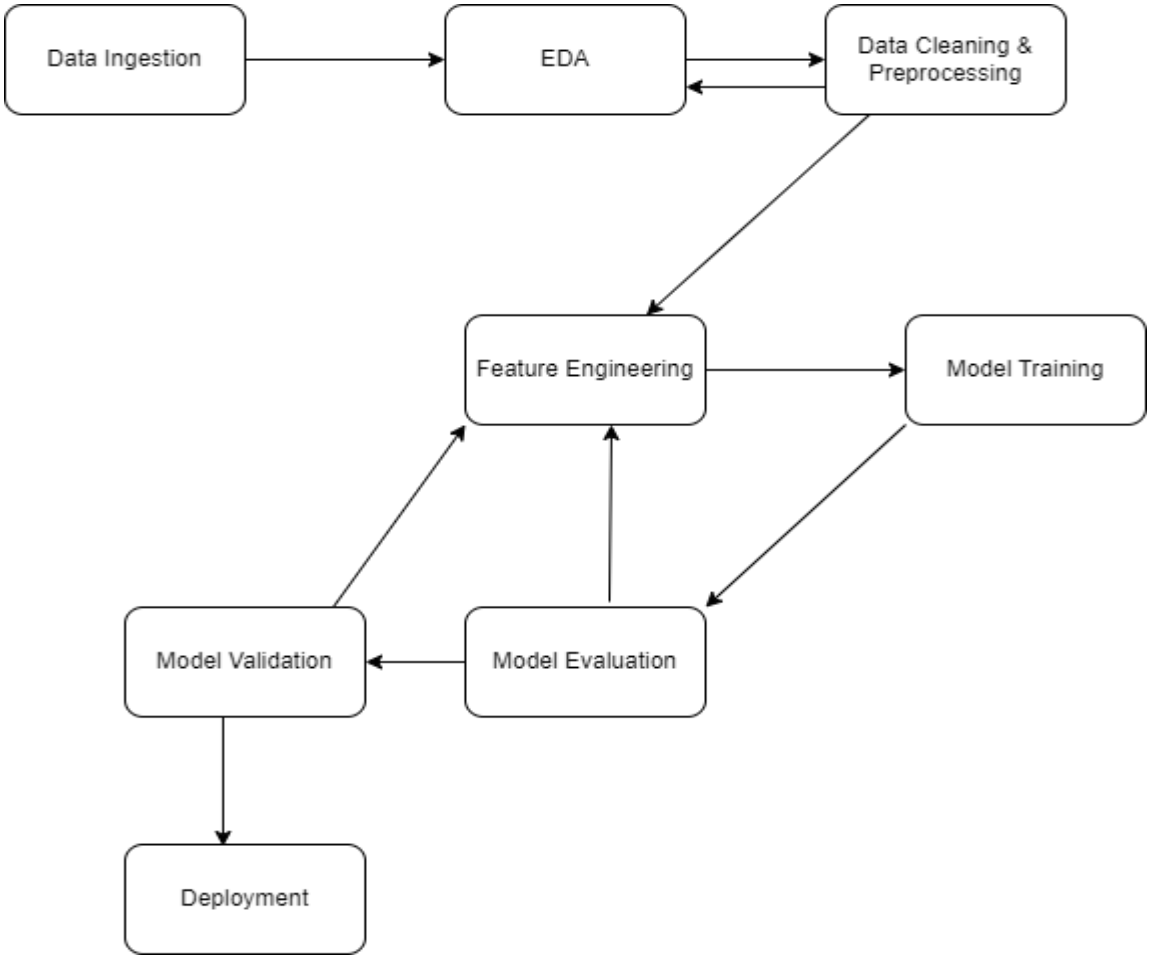
Response	Frequency	Percentage(%)
Bachelor	231	65.44
Master	79	22.38
Doctorate	42	11.90
Diploma	1	0.28

Bar chart visualizing repayment status with respect to discipline of study

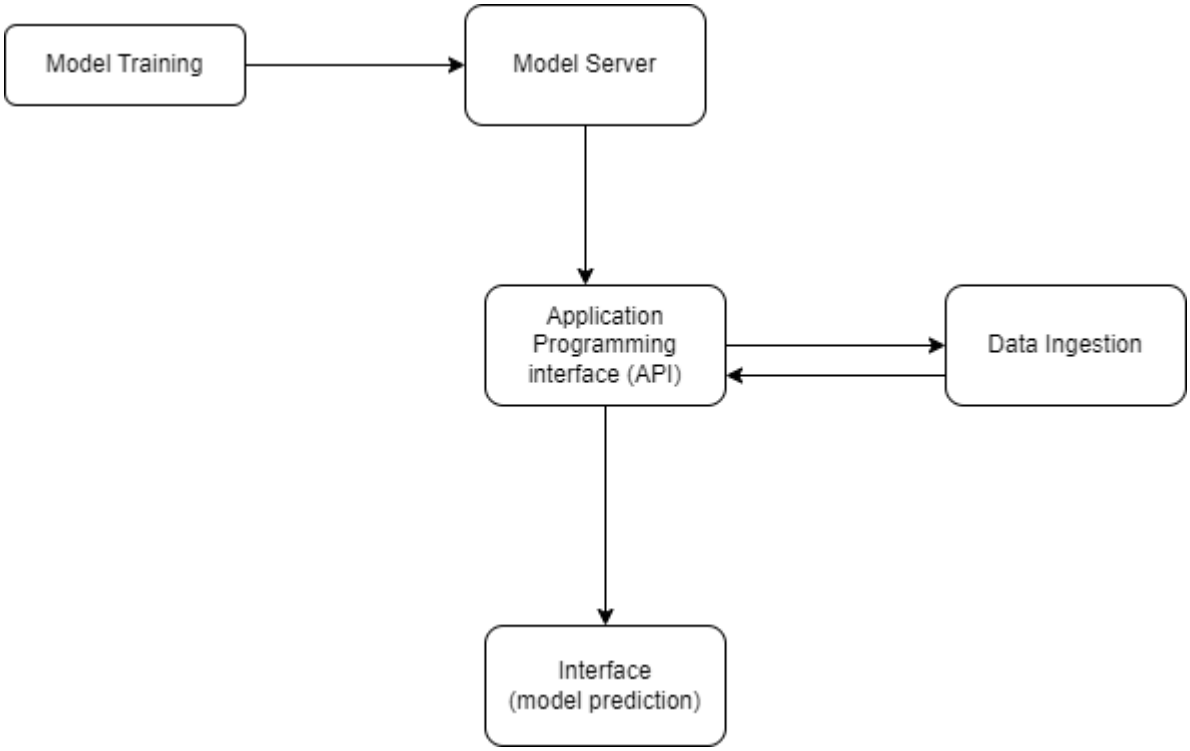


NOTE: more visual analysis will be conducted after data preprocessing since it is easier to visualize numbers than is to get the same level of visual meaning from categorical data.

Machine learning pipeline



Model deployment plan



Data cleaning and preprocessing:

1. Converting all column names to lowercase. This will be done to ease further preprocessing and reduce chances of errors due to casing.
2. Converting all the values in the dataset to lowercase. This will be performed to ease further analysis as well as reduce access errors due to mismatching casing.
3. Checking datatypes and converting them where necessary. Datatypes of the values of different features will be checked for uniformity and will be converted to follow a uniform type for the particular feature.
4. Handling outliers. The outliers identified in EDA will be handled using appropriate techniques as follows:
    a. Dropping.
    b. Replacing.
5. Identifying and handling missing values. the dataset will be checked for missing values, and if any are identified they will be handled as follows:
    a. Dropping rows
    b. Dropping columns if missing values exceed a certain threshold.
    c. Imputing using appropriate techniques including using the mode.

6. Feature transformation. Certain features will be changed in order to provide more meaning and relevance for example:
    a. Program of study will be changed to discipline whereby programs conforming to a particular discipline will be labeled as that discipline for instance pharmacy and mmbs will both be registered as medicine as a discipline.
    b. Age range will be converted to age group. the age group will contain the specific intervals of ages of the beneficiaries for instance; '38 to 48' will be converted to '38 – 40'.
    c. Values under the total loan feature will be converted to sensible intervals and will be stripped of the currency which will be annotated together with the feature itself, thus, loan_amount (MWK).
7. Dropping features which will be intuitively considered irrelevant. For example, repayment method recommendation which would be irrelevant to the final model.
8. Feature encoding. All the columns will be converted to floating point numbers and integer types using appropriate encoding techniques which are as follows:
    a. Label encoding. For ordinal types.
    b. Mean target encoding. For both ordinal and nominal types of data.
    c. One-hot encoding. For the categories which follow hierarchy.

Feature Engineering: Creating and Selecting Features

The goal is to capture meaningful patterns and structures in the data, reduce dimensionality while retaining relevant information, and improve model performance and interpretability.

Steps for Feature Creation:

1. Data Exploration: Understand data types and identify potential patterns or relationships.

2. Feature Extraction: Extract relevant features, such as:

- Statistical Features: Mean, median, standard deviation, skewness, kurtosis.

- Time-Series Features: Trends, seasonality, cyclical patterns.

3. Feature Engineering: Transform and combine existing features, such as:

- Normalization/Scaling: Standardize features for equal contribution.

- Feature Interactions: Ratios, products, or differences between features.

Steps for Feature Selection:

1. Correlation Analysis: Identify and remove highly correlated features.
2. Mutual Information: Select features with high mutual information.
3. Dimensionality Reduction: Use techniques like PCA, t-SNE, or autoencoders.
4. Domain Knowledge: Leverage domain expertise to select relevant features.

Pre-processed Data Summary

The final cleaned and prepared dataset will be presented, along with relevant statistics and insights.

Dataset Overview:

- Summary Statistics: Mean, median, mode, standard deviation, and other metrics.
- Data Visualizations: Histograms, scatter plots, and other charts.

Key Statistics:

- Descriptive Statistics: Numerical variable summaries.
- Frequency Distributions: Categorical variable summaries.
- Correlation Matrices: Relationship analysis.

Visualizations:

- Histograms: Distribution analysis.
- Scatter Plots: Relationship analysis.

Technical requirements

Software and tools

- Programming language that we will is python
- Some of libraries is pandas for data manipulation, numpy mathematical calculations, sklearn for model training and lebel encoding and one hot encoding, joblib for creating pkl file that will be processed during model deployment
- Matplotlib and searbon for data visualisation

- Editors such as vs code

Hardware requirements

- Minimum of 4gb ram
- Cpu multicore is also helpful

Ethics and privacy

Ensuring informed consent by making sure users understand what they are agreeing to and why when determining borrower's likelihood of repaying the loan to assess the credit risk within the financial sector. Users will understand what data is being collected, why it's needed, and how it will be used.

Ensuring transparency by communicating the data practices clearly about the data usage and providing ways for the users to manage their data. Transparency will also be addressed by ensuring there is explanation of how data will be collected, used, and shared so as to be easily understandable for users.

Privacy will be highly considered by making sure any personal information shared is protected to ensure user privacy. It will not be made publicly available and financial institutions can only collect data that is necessary for their stated purpose without gathering excessive or irrelevant data which is unethical and increases security risks.

Data encryption which will be a fundamental technique to ensure data privacy and protect data so as data cannot be stolen, altered or compromised. This will ensure there is confidentiality and high data security on data privacy issues.

Appendices

Glossary

Term,	**Definition**
Semi-supervised Learning,	A machine learning approach that uses both labeled and unlabeled data for training.
Credit Risk Prediction,	The process of forecasting the probability of a borrower defaulting on a loan.

Self-training,	A semi-supervised method where a model trained on labeled data generates pseudo- labels for unlabeled data.
Co-training,	A technique where two classifiers train on different views of the data and label each other's data.
Consistency Regularization,	A method that leverages the assumption that model predictions should be consistent under input perturbations.
Pseudo-label,	A label generated by a model for previously unlabeled data, used to extend training data.
Baseline Model,	A simple model used as a comparison benchmark for evaluating performance improvements.
F1-Score,	The harmonic mean of precision and recall, used to assess model performance.
Data Normalization,	A preprocessing technique used to scale numerical input variables.
Class Imbalance,	A condition in which the number of observations in one class greatly outweighs the others.

References

X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1–130, 2009.

O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, Cambridge, MA: MIT Press, 2006.

J. Van Engelen and H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2020.

Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/

Python Software Foundation, "Python Language Reference," [Online]. Available: https://www.python.org/

W. McKinney, *Python for Data Analysis*, 2nd ed., O'Reilly Media, 2017.

M. Lutz, *Learning Python*, 5th ed., O'Reilly Media, 2013.

J. Brownlee, “Semi-Supervised Learning with Label Propagation,” *Machine Learning Mastery*, [Online]. Available: https://machinelearningmastery.com/semi-supervised-learning-with-label-propagation/

Dataset provided by project supervisor (unpublished).