# Machine Learning Model Report

## Abstract

This report presents the development and evaluation of a machine learning model designed to address Financial inclusion in Africa. The objective was to create a machine learning model using what I have learnt so far to predict the likelihood of an individual to create a bank account or not.

## 1. Introduction

**Problem Statement**: Financial inclusion is and has been a hindrance to Africa’s development. For example, across Kenya, Rwanda, Tanzania and Uganda only 9.1 million adults have access to or use commercial bank accounts.

Traditionally, access to bank accounts has been regarded as an indicator of financial inclusion. Despite the proliferation of mobile money in Africa, and the growth of innovative fintech solutions, banks still play a pivotal role in facilitating access to financial services. Access to bank accounts enable households to save and make payments while also helping businesses build up their credit-worthiness and improve their access to loans, insurance, and related services. Therefore, access to bank accounts is an essential contributor to long-term economic growth.

**Objective**:

The objective of this competition is to create a machine learning model to predict which individuals are most likely to have or use a bank account. The models and solutions developed can provide an indication of the state of financial inclusion in Kenya, Rwanda, Tanzania and Uganda, while providing insights into some of the key factors driving individuals’ financial security.

## 2. Data Overview

**Data Source**:

The source of the data is Zindi.com

**Description:** The features of the data are;

Country: Country interviewee is in.

Year: Year survey was done in.

Uniqueid: Unique identifier for each interviewee

location\_type: Type of location: Rural, Urban"

cellphone\_access: "If interviewee has access to a cellphone: Yes, No"

household\_size: Number of people living in one house

age\_of\_respondent: The age of the interviewee

gender\_of\_respondent: "Gender of interviewee: Male, Female"

relationship\_with\_head: "The interviewee’s relationship with the head of the house:Head of Household, Spouse, Child, Parent, Other relative, Other non-relatives, Dont know"

marital\_status: "The martial status of the interviewee: Married/Living together, Divorced/Seperated, Widowed, Single/Never Married, Don’t know"

education\_level: "Highest level of education: No formal education, Primary education, Secondary education, Vocational/Specialised training, Tertiary education, Other/Dont know/RTA"

job\_type: "Type of job interviewee has: Farming and Fishing, Self employed, Formally employed Government, Formally employed Private, Informally employed, Remittance Dependent, Government Dependent, Other Income, No Income, Dont Know/Refuse to answer"

**Exploratory Data Analysis (EDA)**:

Some things I noticed from the data were;

1. The data was highly imbalanced:

About only 15% of the individuals had created bank accounts and the remaining 85% had no bank accounts.

1. The data had no missing values.
2. There were some outliers in the age of respondent and household size columns.

## 3. Methodology

**Data Preprocessing**:

Some things I did in the preprocessing stage were;

1. Creating a function called processing data to encode the target variable which was initially “yes” and “ no”.
2. Encoding categorical variables with label encoder.
3. As part of the function, using minmax scaler to scale the data to fall between 0 and 1.
4. Dropping the unique id column because it had no relation with the label.

**Model Selection**:

I explored and experimented with different models such as XGB and lightGBM but in the end I choose the sklearn’s linear regression model because it met most of the expectations.

**Feature Engineering**:

The data given was already split into the train and test data so all I had to do was to work with it . After I called the function on the train and test data I assigned them to the variables, processed\_train and processed\_test respectively.

**Data Splitting**:

I used sklearn’s train\_test\_split module to split the data into train for training and val for testing the model.

## Model Training and Evaluation

**Training**

: The data was fitted into the logistic regression model I created

**Evaluation Metrics**: For evaluation metrics I used;

Accuracy: 0.8334

Precision:0.4427

Recall:0.7129

F1\_score:0.5462

ROC- AUC:0.8647

## 5. Challenges and Limitations

One challenge I faced was balancing the label data and the adjustment of the threshold as this is my first machine learning model. In the end I fixed the imbalance in the data by setting the class\_weight parameter in the logistic regression model to balanced.

## 6. Appendices

Repository link: https://github.com/shepherdmac/Financial-inclution.git