PREDICT WHICH INDIVIDUALS ARE MOST LIKELY TO HAVE OR USE A BANK ACCOUNT USING TOOLS OF DATASCIENCE

FINANCIAL INCLUSION IN AFRICA

CONTECH

CONTEMPORARY TECHNOLOGY UNIVERSITY

Tools & Techniques for Data Science

Supervisor: Sergiy shevcenko

Student:

Ayanfe Isaiah Olabamiji

Date: 13/03/2023

CONTEMPORARY TECHNOLOGY UNIVERSITY

INTRODUCTION

Financial inclusion remains one of the main obstacles to economic and human development in Africa. For example, across Kenya, Rwanda, Tanzania, and Uganda only 9.1 million adults (or 14% of adults) have access to or use a commercial bank account.

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 creditworthiness and improve their access to loans, insurance, and related services. Therefore, access to bank accounts is an essential contributor to long-term economic growth.

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.

PROBLEMS AND CHALLENGES

The business problem here is to predict which Individuals are most likely to have a Bank Account i.e., create a model that can predict customers with an active bank account. Some of the challenges faced in building a model using python are explained below

CONTEMPORARY TECHNOLOGY UNIVERSITY

1. The selection of a suitable churn modelling approach

There is no single methodology to build a predictive model that can work in most situations. Machine learning techniques are mostly used by businesses due to their efficiency and ability to categorize and manipulate complex data sets. The approach of survival analysis, on the other hand, uses survival and hazard functions to predict which individual own a bank account for a particular period. So, the best solution to deal with this challenge is to compare the performance of several models and identify the most effective method for your business.

2. Exploratory analysis

Businesses face several roadblocks and risks at this stage of building predictive models such as lack of information, target leakage, and the need for optimal feature transformations. Along with domain knowledge, businesses must also have the required skills and creativity to build robust predictive models. Therefore, it is important that companies execute careful exploratory analysis and build auxiliary models before embarking on building an overall prediction model. Exploratory analysis can also help in revealing reciprocity, irregularities, outliers, and relationships between different functions, which wouldn't be possible with domain knowledge alone.

METHODOLOGY

The followings are steps to be taken in course of analysing and predicting Individuals that have a bank Account

1. Descriptive statistics of the dataset
2. Exploratory Data Analysis (univariate and multivariate analysis)
3. Data processing i.e., data cleaning, data normalization, scaling, handling missing values and removal of outliers
4. Describe on which predictive method to use e.g., Linear Regression
5. Build and evaluate the models
6. Compare their performance
7. Interpret the models

DATASCIENCE TOOLS EMPLOYED

This analysis employed the tools and techniques of Data science as relevant to the prediction model.

- **Python programming** was used to analyze the dataset while importing the necessary libraries such as Numpy, Pandas, Seaborn and Matplotlib.
- Dataset was downloaded from Zindi Competition website so there was no need for **web scrapping**. In addition, scrapping customer information from website seems unethical.
- **Visual Studio Code** as a tool was used to push the code to a repository in Github which would make it easier to make adjustment where necessary, it also encourage collaboration
- A New **repositor**y was created **on Github** to store the project.

CONTEMPORARY TECHNOLOGY UNIVERSITY

SUMMARY OF THE ANALYSIS

Data Column Description

- ✓ country: Country interviewee is in.
- ✓ year: Year survey was done in.
- ✓ uniqueid: Unique identifier for each interviewee
- ✓ location_type: Type of location: Rural, Urban
- ✓ cell phone access: If interviewee has access to a cell phone: 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 interviewees relationship with the head of the house: Head of Household, Spouse, Child, Parent, Other relative, Other non-relatives, don't know
- ✓ marital status: The martial status of the interviewee: Married/Living together, Divorced/Separated, 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/Don't 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, Don't Know/Refuse to answer

Importing Necessary Libraries

It is a good practice to import all the necessary libraries in one place so that we can modify them quickly.

Importing Dataset

Importing the dataset is pretty much simple. You can use pandas module in python to import it.

Data Processing and Descriptive Statistics

From the visualisations above it can be inferred that.

- ✓ Most of the respondent are from Rwanda.
- ✓ Majority of the respondent are from the rural area.
- ✓ Most of the respondent has access to Cell phone
- ✓ The female folk respond more to the survey and most of them are the head of their household
- ✓ Majority of the respondents are married followed by the singles respectively.
- ✓ Majority of the respondents had only Primary education followed by those with no formal education and secondary education respectively.
- ✓ Most of the respondent's survey were collected in the year 2016, then 2018 and 2017 respectively.
- ✓ Majority of the respondent has a total household size of 3 to 4.
- ✓ The average age of respondents are between 35 to 50.

Multivariate Analysis

- ✓ It is safe to assume that majority of the respondents don't have a bank account
- ✓ Household size and Age has high outliers

Correlation Analysis

- ✓ Correlation explains the relationship between the variables and from the analysis we can deduce that there no correlation between the features

Model Building

- ✓ CatBoost and XGBoost model was employed to analyse this data.
- ✓ They both gave a error of 0.30% which means an accuracy of over 70%
- ✓ Finally, an ensemble method was used to strengthen the predictive capability of the model by combining both models to give better accuracy and less mean error.

CONCLUSION

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 creditworthiness and improve their access to loans, insurance, and related services.

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.