Unlocking Financial Inclusion in East Africa: A Data-Driven Approach; Data Report

Business Understanding:

In the pursuit of economic growth and individual prosperity, achieving financial inclusion is paramount. However, in East Africa, a significant gap persists in accessing banking services. Traditional approaches to expanding banking infrastructure face challenges in identifying suitable locations for new branches. Leveraging data analysis, this project aims to uncover opportunities for growth and enhance financial accessibility across the region.

Research Questions:

- What are the disparities in bank account ownership across demographic factors such as education, employment, and gender?
- How do geographic factors, such as city and state, influence the distribution of bank accounts and financial accessibility?
- Which predictive model performs best in identifying individuals without bank accounts, and what are the key factors influencing financial inclusion?

Problem Statement:
The lack of financial inclusion in East Africa poses significant obstacles to economic development. This project endeavors to address this challenge by developing a predictive model to identify individuals without bank accounts. By analyzing demographic, geographic, and socioeconomic factors, the model aims to provide insights into the barriers to financial inclusion and inform targeted interventions to promote greater access to banking services.

Objectives:

- Geographic Analysis: Mapping regions with high demand for banking services across African countries.
- Demographic Profiling: Understanding the demographic characteristics of target audiences, including age, income, and occupation.
- Accessibility Assessment: Evaluating the accessibility of banking services by considering factors such as distance to existing branches and transportation infrastructure.
- Employment Analysis: Examining employment trends in areas lacking bank branches to inform strategies for addressing economic needs and enhancing financial inclusion.
  Data Understanding:

Data Understanding:

Data Source:

The dataset utilized in this project was obtained from Zindi, a platform for data science competitions in Africa. It comprises responses from surveys conducted across East African nations, capturing information on bank account ownership, demographic characteristics, employment status, education level, and geographic location.

Data Description:

- ☐ Country: The country where the survey was conducted.
- ☐ Year: The year in which the survey was conducted.
- ☐ Unique ID: A unique identifier assigned to each respondent.
- ☐ Bank Account: Binary variable indicating whether the respondent has a bank account or not.
- ☐ Location Type: Categorical variable specifying the type of location where the respondent resides (e.g., rural or urban).
- ☐ Cellphone Access: Binary variable indicating whether the respondent has access to a cellphone or not.
- ☐ Household Size: The number of individuals living in the respondent's household.
- ☐ Age of Respondent: The age of the respondent.
- ☐ Gender of Respondent: The gender of the respondent.
- ☐ Relationship with Head: The relationship of the respondent with the head of the household.
- ☐ Marital Status: The marital status of the respondent.
- ☐ Education Level: The highest level of education attained by the respondent.
- ☐ Job Type: The type of employment of the respondent.

Data Preparation:

Cleaning:
Handling Missing Values: Any missing values in the dataset were addressed through appropriate imputation techniques. This ensured that the data used for analysis and modeling was complete and reliable.
Handling Duplicates: Duplicate entries, if any, were identified and removed from the dataset to maintain data integrity and prevent redundancy.

One-Hot Encoding:
Categorical Variables: Categorical variables in the dataset, such as 'Location Type', 'Gender of Respondent', 'Marital Status', 'Education Level', and 'Job Type', were encoded using one-hot encoding technique. This transformed categorical variables into binary vectors, enabling them to be used as features in machine learning models.
Creation of Dummy Variables: Each category within a categorical variable was converted into a separate binary variable, with a value of 1 indicating the presence of that category and 0

indicating its absence. This allowed for the inclusion of categorical variables in predictive modeling without the need for ordinal encoding.

Scaling and Data Balancing

Continuous variables were scaled to bring them to a similar range, preventing features with larger magnitudes from dominating the modeling process.

Data Balancing with SMOTE: To tackle the class imbalance, we used Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class (individuals with bank accounts), ensuring a balanced representation in the training data.

External Data Source Validation

To ensure the reliability and validity of the dataset, external source validation was conducted. The dataset was sourced from Zindi, a reputable platform known for hosting data science competitions and providing high-quality datasets for analysis and modeling. Zindi follows rigorous data collection and curation processes to ensure the datasets provided are accurate, relevant, and ethically sourced. Additionally, Zindi adheres to data privacy regulations and guidelines, safeguarding the confidentiality of individuals' information contained within the dataset. Moreover, the dataset underwent thorough scrutiny and validation by the data science community participating in Zindi competitions, where it was subjected to peer review and scrutiny, further enhancing its credibility. Furthermore, to confirm the consistency and relevance of the data, exploratory data analysis (EDA) was performed, which included examining key statistics, distributions, and patterns within the dataset. This EDA process helped validate the dataset's integrity and identify any inconsistencies or anomalies that required further investigation. Overall, the external source validation process confirmed the authenticity and suitability of the dataset for analysis and modeling, providing confidence in the results and insights derived from the data.

Exploratory Data Analysis (EDA)

EDA was conducted to gain insights into the patterns and trends present in the dataset regarding financial inclusion in East Africa. Initial observations revealed disparities in bank account ownership across different demographic groups and geographic regions. Visualizations such as histograms, bar charts, and heatmaps were utilized to explore relationships between variables. For instance, the distribution of bank account ownership was analyzed based on factors such as education level, employment status, and household size. Additionally, geographic maps were employed to visualize the spatial distribution of bank account holders across East African countries, highlighting areas with higher or lower levels of financial inclusion.

Modeling

In our analysis, we employed various machine learning algorithms to predict bank account ownership based on demographic and socioeconomic features. Specifically, we utilized K-Nearest Neighbors (KNN), Decision Trees, and Logistic Regression models. Below is a detailed overview of our modeling approach and results:

K-Nearest Neighbors (KNN):

Initially, we implemented a basic KNN model without hyperparameter tuning.
The model achieved an accuracy of approximately 61% on unseen data.
To enhance performance, we conducted hyperparameter tuning using GridSearchCV to find the optimal combination of hyperparameters.
The best hyperparameters found were {'n_neighbors': 11, 'p': 1, 'weights': 'uniform'}.
With the tuned model, the accuracy improved to approximately 88% on the test set.
Further evaluation was performed using a classification report, revealing precision, recall, and F1-score for both classes (bank account holders and non-holders).

Decision Trees:

A Decision Tree classifier was trained on the preprocessed data.
The model demonstrated an accuracy of approximately 89% on the test set.
Hyperparameter tuning was conducted to optimize the model's performance.
The best hyperparameters obtained were {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}.
With the tuned model, the accuracy slightly improved to approximately 89% on the test set.
Additionally, a Receiver Operating Characteristic (ROC) curve was plotted to visualize the model's performance.

Logistic Regression:

Logistic Regression was applied after scaling the features using StandardScaler.
The model achieved an accuracy of approximately 87% on the test set.
Model evaluation included precision, recall, accuracy, and F1-score metrics for both the training and testing sets.
A confusion matrix was plotted to visualize the true positive, true negative, false positive, and false negative predictions.
Finally, the model was tested on new data (test dataset) to make predictions on bank account ownership.

Our analysis demonstrates the effectiveness of machine learning algorithms in predicting bank account ownership based on demographic and socioeconomic factors.
Decision Trees and Logistic Regression models exhibited promising performance, achieving accuracies of around 89% and 87%, respectively, on the test set. These models can be valuable

tools for financial institutions and policymakers in understanding and promoting financial inclusion in East Africa.

Conclusion

Financial Inclusion Discrepancy: The analysis illuminates a significant disparity in financial inclusion levels across East Africa, with a predominant portion of respondents indicating an absence of bank account ownership. This underscores the imperative for concerted efforts aimed at dismantling existing barriers to access and fostering a more inclusive financial landscape.

Educational Influence: Intriguingly, completion of primary or secondary education does not seem to substantially elevate the likelihood of possessing a bank account. This suggests that additional factors beyond educational attainment, such as accessibility to banking infrastructure and financial literacy, may wield greater influence over one's banking status.

Employment Dynamics: A notable trend emerges regarding the impact of employment status on bank account ownership, with self-employed individuals comprising the largest share of respondents possessing bank accounts. This underscores the potential efficacy of tailored financial services and targeted support mechanisms geared towards various employment sectors in enhancing financial inclusion.

Gender Disparity: There exists a palpable gender gap in bank account ownership, with males representing the majority of respondents with active bank accounts. Addressing gender-specific obstacles to financial access and implementing policies that prioritize gender equality in financial inclusion initiatives are paramount for fostering a more balanced financial landscape.

Model Performance: The Logistic Regression model emerges as the best-performing model, exhibiting commendable precision and recall rates above 88%. Despite both models demonstrating robust predictive capabilities, the Logistic Regression model's slightly superior performance underscores its suitability for predicting bank account ownership in this context.

Recommendations

- Targeted Expansion: Utilize the geographic analysis to identify regions with high demand for banking services and prioritize expansion efforts in these areas. This can help in increasing accessibility to banking services where they are most needed.
- Demographic Targeting: Use demographic profiling to tailor financial products and services to the specific needs and characteristics of different demographic groups. This can include age-appropriate financial education programs and customized banking solutions.
- Accessibility Improvement: Based on the accessibility assessment, consider investing in infrastructure improvements such as building new bank branches or enhancing transportation networks to make banking services more accessible to underserved communities.

- Employment-Based Strategies: Develop strategies to target different employment sectors identified in the analysis. For example, offering specialized financial products for self-employed individuals or providing financial literacy programs for farmers and fishermen.
- Gender-Specific Initiatives: Given the gender disparity in bank account ownership revealed in the analysis, implement targeted initiatives to promote financial inclusion among women. This could involve providing women with access to microfinance loans or creating women-focused financial empowerment programs.