Maximizing Fairness with Synthetic data: Access to Emergency Fund in Sub-Saharan Africa

Charavee Basnet Chettri

and
Vivian Wei

and
Betty Pu

and
Ziyue Yang *
Department of Statistical and Data Sciences, Smith College, Northampton MA

March 5, 2024

Abstract

Financial inclusion is paramount for economic stability and resilience, particularly in diverse regions like Sub-Saharan Africa, spanning low to high-income countries and encompassing both resource-intensive and non-resource-intensive economies. This study focuses on a crucial aspect of financial resilience: the accessibility of emergency funds, defined as having access to 1/20 of Gross National Income (GNI) per capita in local currency within 30 days. Leveraging previous colleagues' exploratory work on the Global Financial Inclusion Database 2021, our objective is to mitigate inherent gender biases in the dataset by rebalancing it with synthetic data, thereby enhancing fairness in predicting emergency fund accessibility. Through predictive machine learning modeling, we aim to contribute to the economic empowerment of individuals in Sub-Saharan Africa, ultimately fostering resilience and reducing disparities in access to essential financial resources.

Keywords: Fairness; predictive machine learning; emergency funds; financial inclusion; ethics; gender bias; debiasing; synthetic data; Sub-Saharan Africa; fairness metrics; Global Financial Inclusion Database

*We are grateful to the Women at the Table team and Sofia Kypraiou for the project inspiration and suggesting the path forward

1 Introduction

In the realm of financial inclusion, the accessibility of emergency funds plays a pivotal role in determining an individual's financial stability and resilience, especially in developing countries.(**?**) In this project, our goal is to predict the possibility for people in Sub-Saharan African countries to come up with emergency funds, defined as 1/20 of GNI per capita in local currency, within a 30-day period(**?**). This prediction serves as a crucial factor for establishing future public financial policies, such as determining the eligibility of individuals for loans and financial assistance. The significance of this problem lies within its direct impact on the economic well-being and empowerment of individuals in developing regions. According to the Global Financial Inclusion (Global Findex) Database 2021 published by the World Bank, only a little over half of people over 15 years of age in developing economies could access extra funds within 30 days if faced with unexpected expenses (**?**). Therefore, there is a pressing need to understand the factors influencing this accessibility and eliminate inherent bias in the dataset. By delving into this issue, we not only contribute to enhancing financial inclusion but also aid in mitigating the different effects of financial shocks on vulnerable populations.

Defining fairness is essential since the concept itself is relative among different people. In the data science discourse, fairness encompasses three key aspects: individual fairness, group fairness, and causal fairness(**?**). Individual fairness focuses on preventing discrimination against individuals with similar relevant characteristics. This means ensuring that individuals in similar situations receive similar outcomes from the model, regardless of irrelevant factors(**?**). Group fairness aims to prevent disparities in outcomes for different groups. This ensures equal opportunities for all groups, regardless of their membership(**?**). Causal fairness goes beyond simply observing disparities and delves into understanding

their underlying causes. It seeks to mitigate these root causes to achieve fair outcomes within and across groups(**?**).

Our previous colleagues conducted analysis using the <AI & Equality> Human Rights Toolbox and used a Decision Tree Classifier machine learning model implemented via Python to predict access to emergency funds with 68% accuracy. Their work laid a solid foundation by exploring demographic and financial variables within the dataset, and assessed the fairness of the decision tree classifier, particularly concerning gender bias, and then applied various processing techniques to enhance the fairness of the model(**?**).

Based on their work, our group aims to incorporate synthetic data to rebalance the dataset, ensuring equitable representations amongst both genders in the dataset. Our approach aims to consider a broader selection of machine learning algorithms and mitigate the disparities the previous colleagues found in the decision tree's classifier's predictions and enhance fairness with synthetic data, ultimately better predict access to emergency funds in south-Saharan Africa countries(**?**).

2 Background on Sub-Saharan Region

Sub-Saharan Africa is a region characterized by a diverse economic landscape, encompassing low, lower-middle, and upper-middle-income countries. Demographically, Sub-Saharan Africa is marked by a rich tapestry of cultures and a population exceeding 1.2 billion people. This expansive and diverse demographic landscape includes 22 countries grappling with fragility or conflict, posing unique challenges to development efforts. Additionally, 13 small states within the region are characterized by limited human capital, modest populations, and constrained land areas.

According to the World Bank's definition, middle-income countries had a per capita

gross national income of more than US$995.00 in the years 2015–17. Among the 35 countries included in our dataset, 20 are classified as low-income countries, and 15 are classified as middle-income countries. Additionally, 11 are classified as countries in Fragile and Conflict-Affected Situations, which, by definition, have experienced a peacekeeping or peace-building mission within the last three years.

Sub-Saharan African countries not only differ in terms of economic prosperity but also in economic structure and resource intensity. Resource-intensive countries include both oil-exporting nations, where net oil exports make up 30 percent or more of total exports, and commodity exporters, where nonrenewable natural resources represent 25 percent or more of total exports. The divergence between resource-intensive and non-resource-intensive countries became more entrenched following the commodity price shock of 2015[ ]. Non-resource-intensive countries have proven more resilient, supported by their more diversified economies. On the other hand, resource-intensive economies generally have a less diversified structure, making them more susceptible to external shocks. In our dataset, 6 of the countries are oil exporters, and 11 export other commodities such as iron ore, copper, cotton, coffee, and sugar. The remaining 18 countries are non-resource-intensive and their economies are not reliant on exports.

In recent years, the Sub-Saharan Africa region has grappled with significant economic challenges, including soaring inflation, pronounced exchange rate pressures, debt vulnerabilities, and widening economic disparities within the region. Therefore, addressing these structural and economic disparities is imperative for tackling developmental issues in the region.

3 Data

In order to continue assessing and enhancing the fairness of the machine learning models, we use the same Global Findex Database as our colleagues before us. The Global Findex database [ ] was first launched in 2011 by the World Bank—with funding from the Bill & Melinda Gates Foundation. It is the world's most comprehensive data set on how adults save, borrow, make payments, and manage risk. The dataset contains over 200 indicators including account ownership, payments, savings, credit, and financial resilience and has coverage over 140 nations, representing 97% of the world's population for year 2017, 2014, and 2011.

The data was constructed through a series of surveys carried out by Gallup, Inc in association with the annual Gallup World Poll. They randomly sampled 1000 individuals from each country and asked them to respond to a survey either over the phone or in-person. The target population is the civilian, non-institutionalized population 15 years and above. In consistency with the former colleagues, our sample covers only the Sub-Saharan region (35 countries total), thus there are 35,000 total observations and 105 total variables. Since the sampling was random and the sample size is large, we can assume that our sample is representative of the total population of people living in the 35 Sub-Saharan countries in the data. The data was collected directly from individuals over the 2017 calendar year and is self-contained, meaning it does not rely on any external resources.

3.1 Variable Definition and Descriptions

After basic data cleaning, we have created various visualizations to help understand who is in the data and acknowledge potential sources of bias within the data.

Outcome Variable of Interest: Access to Emergency Fund (Fin24) The variable "Fin24"

in the dataset asked participants the question: Now, imagine that you have an emergency and you need to pay [1/20 of GNI per capita in local currency]. Is it possible or not possible that you could come up with [1/20 of GNI per capita in local currency] within the NEXT MONTH? In order to make it more straightforward for interpretations, we restrict the response into a binary variable for Yes and No. The overall distribution of access to emergency funds showed 17, 599 individuals had access to emergency funds while 14342 did not. This indicates that over half of individuals represented in the data do not have access to emergency funds.

To help conceptually understand how the outcome variables might vary among different levels of the variables, we chose 50% as the benchmark proportion for checking if the possibility of coming up with emergency funds for each variable group of interest is different from a random 50/50 chance.