NOVA IMS
Information Management School

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

SCALING FINANCIAL INCLUSION TO DRIVE BANKS' ANNUAL DEPOSIT INCREASE IN SUB-SAHARAN AFRICA

Improving the percentage of Banked Adult Population using Nigeria's Financial Access Data Modelling

Akinyemi Sadeeq Akintola

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

SCALING FINANCIAL INCLUSION TO DRIVE BANKS' ANNUAL DEPOSIT INCREASE IN SUB-SAHARAN AFRICA

(Improving the percentage of Banked Adult Population using Nigeria's Financial Access Data Modelling)

by

Akinyemi Sadeeq Akintola

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science.

Supervisor: Prof. Doutor Flávio Luis Portas Pinheiro

Co Supervisor: Dr. Afshin Ashofteh

October 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

13th October 2023
Lisbon, Portugal

Akinyemi Sadeeq Akintola

ACKNOWLEDGEMENTS

First and foremost, I'd want to express my gratitude to Prof. Doutor Flávio Luis Portas Pinheiro, my project supervisor. Thanks to your advice, comprehensive assessments, and honest critique, my project developed into something of which I can be proud. Many thanks! Dr. Afshin Ashofteh, my co-Supervisor, is another incredible blessing in my life. There is no one else I would have wanted. Since my early days as a NOVA IMS student, you have been an incredible motivator and model of greatness for me to aspire to. I was also able to learn a lot from your years of dedicated study and research in this sector. Thank you very much!

Prof. Leonardo Vanneschi, you did a fantastic job leading our MAA programme, and I want to thank you as well. Your classes were among my favourites. Please accept my sincere gratitude for NOVA IMS's provision of, commitment to, and maintenance of our excellent Master's degree programme.

My time at NOVA IMS, and by extension in Portugal, would not have been the same without my amazing classmates and the few people I count as close friends. The memories we made together are ones I will never forget. My dear friends in Lisbon, including Diyaolu Lekan, Ola Orisaya, Khadija, Haliyah, Aminu, and many more, have become like family to me, and for that I am eternally grateful.

I owe a great debt of gratitude to my gorgeous and understanding wife, Jelilat, for being the rock of my family while I was gone and the pillar of support when I needed it most. And especially to my handsome son, Rayyan. Thank you for being patient with me as I pursued my Master's degree, which I began when you were just two years old. And to my most precious gem, Aqeelah: your 'contribution' to Daddy's thesis has not gone unappreciated. Thanks so much to my mom and my siblings for always being there for me and supporting me no matter what.

At the end of the day, I have Allah Almighty to thank for ensuring that I completed this degree. In the face of personal and professional setbacks, I contemplated throwing in the towel on several occasions, but each time, the Lord kept me going.

“To Allâh belongs the sovereignty of the heavens and the earth and whatever lies in them, and He is the Possessor of full power to do all that He will”. (Quran: Chapter 5, Verse 120).

ABSTRACT

Banks and other financial institutions play a critical role in any nation's economy, primarily by offering services such as cash deposits, loans, advances, payments, transfers, overdrafts, and discounting bills and credits. The percentage of the adult population with access to these financial services significantly contributes to the nation Financial Inclusion (FI) index. The world has about 1.7 billion unbanked adults, meaning people who do not own an account with a bank or any other financial institution providing similar financial services. Nigeria is one of the seven nations that contributes to half of the world's unbanked population. Only 60% of the 99.6 million adult population in Nigeria are banked. Unfortunately, this percentage tends to drop further as we experience significant population growth every year. People in rural areas need more access to banks and financial information as the country's economy remains largely cash-driven. As a result, it is up to banks to meet people where they are (homes, farms, markets, religious institutions, and village square meetings) and try to onboard them. However, this strategy has always delivered modest outcomes. Therefore, this thesis seeks to determine the best mix of data point attributes to profile prospective bank clients by building and evaluating four predictive analytics models. The aim is to select the best accurate model while differentiating between banked and unbanked profiles. This enables banks to swiftly enrol customers. The models were trained using decade-long Access to Financial Services (A2F) Data from Enhancing Financial Innovation and Access (EFInA), containing over 100,000 survey respondents' replies from 2008 to 2018. This study reviews the profile patterns of banked people and helps understand unbanked adults for Financial Inclusion planning. The results revealed that the Level of Education, Age Group, Gender, State of Residence, Locality (Rural or Urban), and Type of Job all contribute to good classification prediction accuracy. Additionally, the study shows that, for an appreciable number of unbanked individuals to crossover as banked, deliberate policy alignment, government investment, legislation, or efforts in certain core areas of grassroots development, especially by improving the number of primary and secondary school graduates.

KEYWORDS

Nigeria; Financial Inclusion; Banks; Data Model; Binary Classification; Logistic Regression; XGBoost; Naive Bayes; Neural Networks; TensorFlow; Feature Engineering; Model Explainability.

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

A2F	Access to Financial Services
AI	Artificial intelligence
ATM	Automated Teller Machines
CBN	Central Bank of Nigeria
EFInA	Enhancing Financial Innovation and Access
FCT	Federal Capital Territory
FI	Financial Inclusion
GDP	Gross Domestic Product
ML	Machine Learning
NBS	National Bureau of Statistics
POS	Point of Sales
SME	Small and Medium-sized Enterprises
SQL	Structured Query Language
UI	User Interface

1. INTRODUCTION

The World Bank defines Financial Inclusion (FI) as *"the access to useful and affordable financial products and services — transactions, payments, savings, credit and insurance — that meets individuals and businesses needs in a responsible and sustainable way"* [1]. In other words, when people can, at the very least, open a bank account, make deposits, and withdraw, save money into their accounts, send, and receive money through banking channels, and have access to loans and other simple financial instruments, they are said to be financially included [2]. However, FI is difficult to achieve in many developing countries and rural areas because of the lack of access to digital services or the distance to the closest bank's physical location. In that sense, financial institutions have made efforts to onboard such populations, which they also see as untapped assets. An effort that in the end contributes positively to a country's overall FI indicator.

In general, there are three (3) main indicators of FI: Access, Usage and Quality. An access Indicator is a measure of how available and accessible financial services are to people in a particular area. Banking halls, banking apps, point-of-sale (PoS) terminals, and Cash or Automated Teller Machines (ATMs) are examples of access points. On the other hand, the usage indicator shows what products and services people consume. It measures the frequency and extent of the use of financial services rendered to people who now have access. The Usage indicator helps us to understand how inclusive the financial system is, user behavior, and what can be improved [3]. The Quality indicator is a measure of how satisfying the financial services are to the intended customer. It reflects greatly on the service provider's understanding of the needs of customer clusters, with unique characteristics and challenges, and then develops tailored solutions for such clusters. Unlike the usage indicator, the best way to hypothesize and measure quality indicators is through information gathering from end-users, such as surveys [4].

In this research, we use the Access to Financial (A2F) Services in Nigeria data curated by the FI research institute Enhancing Financial Innovation and Access (EFInA). The data contain respondents' answers to six (6) biennial survey questionnaires from 2008 to 2018, with more than 18,000 records each year. The combined curated data are then used to explore the feasibility of deploying Machine Learning models — Logistic Regression, Naive Bayes Algorithm, extreme Gradient Boosting (XGBoost), and Neural Network (NN) — to help predict the likelihood of an individual's financial inclusion.

We find that the Level of Education is the highest correlating predictor of the "banked." The more the educational degree completed after High School, the greater the likelihood of holding a bank account. The results also showed a strong positive correlation between a high level of education and other important inclusion predictors, such as Access to Mobile Phones, Access to Television, Age, Occupation, and type of locality (rural or urban). Based on these models, people who live in rural areas, those who do not have access to the Internet or email, or those who do not have a stable income are much more likely to be financially excluded.

With this model, financial institutions can now spend less time and resources on separating those who are financial inclusion-ready from those who are not, while quickly and easily onboarding the former.

2. LITERATURE REVIEW

Like every other business, banks always seek to expand their reach and scale their customer base because acquiring more customers can result in more revenue. The first business-scaling part is achieved by reaching out to potential customers *en masse* and converting them into active customers who engage with and benefit from the banking business. Owing to the nature and importance of bank services, every adult member of the population is a potential customer. However, banks must strategically and methodically target potential customers to maximize the conversion rate. The Second part of the scaling customer base is the deepening of product and service offerings to already onboarded customers, based on their profile and potential needs. This section reviews previous studies on retaining existing banking customers and attracting new ones. It also examines how this was done prior to, and in, the age of Artificial Intelligence and Machine Learning (AI/ML) technologies to target customers within and outside the banking industry.

2.1. FINANCIAL INCLUSION LANDSCAPE IN NIGERIA

While there seems to be consistent year-on-year progress with the Financial Inclusion numbers in Nigeria over the past decade, the financial landscape still does not look like an even distribution. Despite having one of the largest economies in Africa, the country still has more than half of its population living in poverty. This is a major enabler of financial exclusion. The exclusion numbers were made up of people from poor and rural neighborhoods, with little or no education, and women. According to EFInA's Landscape report, *"Today, banking leads the way to Financial Inclusion, with three in ten adults (29%) having bank accounts. Three of 100 adults (3%) had mobile money accounts and the same number (3%) had nonbank financial accounts"* [5].

Table 2.1 – The banked population increased from 36.9 million in 2016 to 39.5 million in 2018 [6]

(Total Population: 99.6 million)	**2016**	**2018**	**Growth / Decline**
Banked population	**38.3%**	**39.7%**	**+1.4%**
Remittances (send and receive)	24.2%	22.4%	-2.2%
Savings with a bank	27.7%	21.0%	-6.7%
Payments	12.1%	15.5%	+3.4%
Receive income	8.4%	9.7%	+1.3%
Loan with a bank	1.3%	1.3%	Stable
Banking agents	2.6%	3.3%	+0.6%

There are several governmental and non-governmental bodies that have led to Financial Inclusion efforts in Nigeria. One of them is EFInA. It is unsurprising to see that as a developing economy, most of the Financial Inclusion initiatives that have made huge impacts have been introduced and implemented by, or in collaboration with, the Central Bank.

2.2. About EFInA

Enhancing Financial Innovation and Access (EFInA) is at the forefront of improving Financial Inclusion in Nigeria. They identify themselves as a Nigerian financial sector development group that promotes Financial Inclusion. Before EFInA, there were no usable data to measure the country's inclusion level and how the country ranked in the global Financial Inclusion index. EFInA was formed in 2007 with the aim of being *"the leader in assisting the establishment of an all-inclusive and growth-promoting financial system"* [6]. The following is how EFInA described Nigeria's Financial Inclusion landscape.

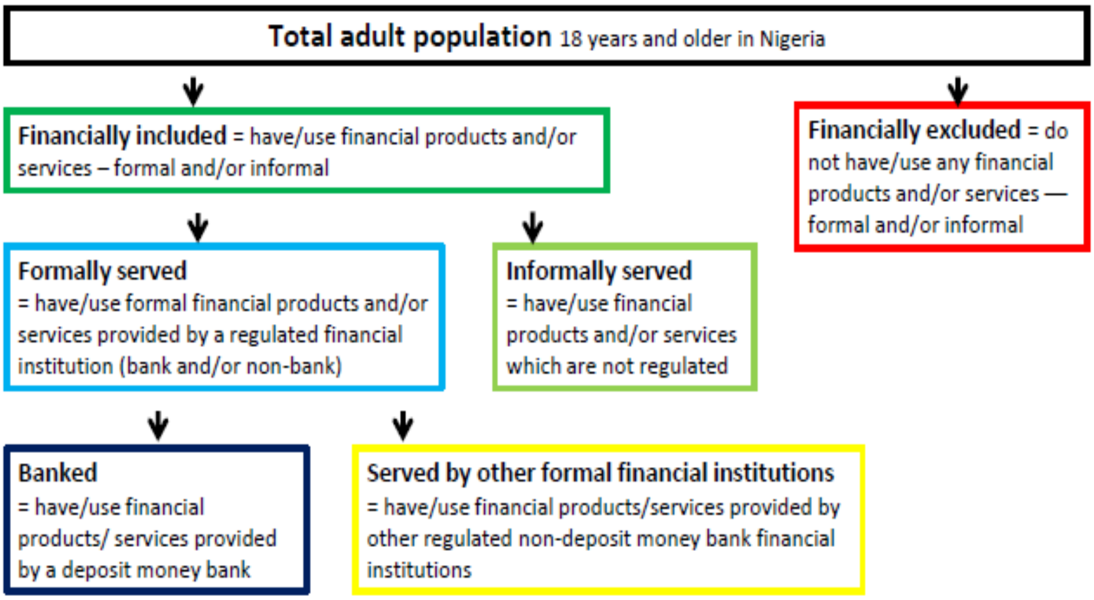


Figure 2.1– Defining Financial Inclusion in Nigeria [7]

2.3. The role of Banks in Financial Inclusion

In a country like Nigeria, which has a remarkably high percentage of people living in rural areas, the most optimal available means to financially include unbanked and underbanked people is through commercial banks. This view is also shared by Nkuna et al. [8]. Financial services provided by banks, such as loans and credit, directly benefit people and businesses. Today, some people have bank accounts solely to receive funds. Such an example would be seen in university students, who are away from home but receive monthly allowances for food and other upkeep from their parents or guardians.

Studies have been conducted on the influence of banks on Financial Inclusion over the years. One study [9] tested this idea in the Indian economy and found that Financial Inclusion is a crucial instrument for a developing nation's economic progress. In this environment, India is launching several programs to include the unbanked to promote inclusive development.

The significance that banks play in connecting people and companies to financial services cannot be overstated. Through bank networks of branches and ATMs as well as their interconnectedness with those of other banks, these outlets are the sole point of contact for new and current clients, particularly in rural regions, thus being a realistic entrance to Financial Inclusion in such places.

Banks use Agent Banking to promote rural community penetration by hiring agents. These agents function on behalf of banks and act as intermediaries. This strategy has been effective in Malawi [8].

The study found that banks aided the unserved via agent banking. Banks perceive branch location, Know Your Customer (KYC) processes, and poor literacy as barriers to financial access. Thus, by providing financial literature, a low or flexible account opening balance, and removing other obstacles, they observed a significant increase in rural onboarding.

2.4. RELATED WORKS

With the percentage of adults using formal financial institutions rising from 36.3% in 2010 to 50.5% in 2023, Nigeria's financial inclusion can be said to have improved over the past ten years [10]. However, the country fell short of its 2020 National Financial Inclusion Strategy target of 70% of Nigerians using formal financial services, as it was only able to achieve a 51% inclusion rate [11]. This indicates that more work needs to be done to include more citizens.

Digital financial services in Nigeria are, however, not non-existent, and have been instrumental in moving the FI needle, contributing to the growth of financial inclusion by making banking options available to more people and businesses in the country's underserved areas. These services include:

- **Mobile Money services** such as Paga, Airtel Money, and MTN Mobile Money enable users to conduct a range of financial transactions using their mobile devices. These transactions include sending and receiving funds, buying mobile phone credits, paying for subscription on local services, settling debts, buying mobile phone credits, and using many other financial services through mobile devices [12].
- **Unstructured Supplementary Service Data (USSD) banking** – Major mobile network providers have enabled users to conveniently access fundamental financial services using their mobile phones without the need for Internet connectivity [12]. Subscribers can engage in various financial activities, such as verifying their account balances, moving money, and purchasing airtime, by dialling designated USSD codes on mobile phones.
- **Agent banking** is a financial service model that employs authorized agents to provide banking services on behalf of banks and other financial organizations. These agents, often situated in rural or underserved regions, facilitate people's ability to make cash deposits and withdrawals, initiate account openings, and avail themselves of various financial services within their immediate localities [13].
- **Digital payment systems** such as Interswitch, Paystack and Flutterwave provide safe and simple online payment options to both individuals and businesses. Their services include peer-to-peer transfers, bill payments, and e-commerce payments [12].
- **Online lending platforms** such as Branch, Renmoney, and Carbon provide expedient and easily accessible financial services to both individuals and small enterprises by using digital technology to optimize loan application procedures and facilitate the distribution of cash, thereby enhancing the accessibility of credit to marginalized groups [14].
- **Several** banks in Nigeria provide Internet Banking services, enabling users to conveniently access their accounts and execute online banking operations, such as cash transfers, bill payments, and other related activities [12]. Internet banking offers a range of benefits, including enhanced ease and accessibility, which empowers users to effectively oversee their financial affairs from any location equipped with an internet connection.

Financial inclusion plays a crucial role in fostering equitable economic development by delivering policy measures aimed at eliminating obstacles that prevent people from accessing financial markets [15]. This measure's implementation can provide opportunities for individuals, families, and enterprises. Moreover, it serves as a potent mechanism to address economic disparities and alleviating poverty [11]. It facilitates economic growth by fostering greater levels of general consumption, promoting more lucrative investments, mitigating the issue of excessive money supply via the provision of financial services, and encouraging a transition from savings based on wages to those based on investment returns [15]. A recent study examined the influence of financial inclusion on the financial status of small- and medium-sized firms in the southwestern region of Nigeria. The research revealed that the inclusion of financial services has a beneficial effect on the financial state of small and medium-sized firms (SMEs) [16]. Additionally, it was observed that the link between financial inclusion and the financial condition of SMEs is mediated by factors such as financial self-efficacy and financial literacy.

Research shows that several factors, in varying proportions, contribute to the financial inclusion of an individual. While factors such as education or literacy level, employment or economic status, and type of environment of residence are within the control of the individual, other factors such as proximity to financial service outlets (banks, ATMs, PoS), availability, accessibility, and ease of use of these services are certainly not theirs but of the service providers. Governments also play a huge role in policy and in creating enabling environments for such financial service businesses to thrive and secure.

Prabhakar and Satyanarayana [17] agree with this position in their research, which focuses on measuring Financial Inclusion dimensions from the perspective of beneficiaries (demand-side). This paper describes Financial Inclusion as a *multidimensional construct* that would require a variety of measures to gauge financial access and usage of various financial services, while identifying the inclusion dimensions for their, and similar research as geographical proximity, availability, ease of access, affordability, and usage. As a result, recipients' socioeconomic position was found to be significantly affected by these diverse aspects of Financial Inclusion. The study concluded that *"Ease of Access"* is the most significant determining factor for Financial Inclusion, among others [17]. Prabhakar et al's [18] research on the factors influencing Financial Inclusion using banking services, considered even broader spectrum of factors. Multiple linear regression analysis was used to analyze the relationship between a single dependent variable (frequency of using banking services) and multiple independent variables (purpose of opening bank accounts, ease of use of bank products, convenience, and physical accessibility of bank branches).

The bar for any adult to be financially included is quite low in terms of modern times. Anyone who can at least have a bank is financially included. However, Financial inclusion initiatives in Sub-Saharan Africa have always faced several challenges, which can be broadly categorized into supply- and demand-side factors.

The term "supply-side factors" pertains to the obstacles encountered by financial inclusion programmes in relation to the accessibility and delivery of financial services. One prominent determinant is geographical distance and related transaction costs, which often impede access to financial services in rural areas. This limitation arises from the considerable physical separation between financial institutions and rural communities [19] as well as the expenses incurred during financial transactions. Chitimira and Ncube also mentioned that the insufficiency of the legal framework for financial inclusion in many sub-Saharan African nations is also a significant obstacle impeding the comprehensive achievement of financial inclusion for those with little income and resources [20]. In these nations, there is a notable lack of technological accessibility, particularly in rural regions where the required network and Internet infrastructure to provide digital financial services are frequently absent. This deficiency hinders the potential of such services to address the challenges posed by geographical closeness [21].

However, Demand-side factors refer to the challenges faced by financial inclusion initiatives in Sub-Saharan Africa related to the demand for financial services. Unemployment and low income, financial illiteracy, misperceptions of risk and distrust, and the absence of national identification documents all play a role. For instance, in Nigeria, where unemployment is particularly high and wages are often low in rural regions, it is challenging for individuals to obtain and use financial services. As a result, many individuals in remote areas of Sub-Saharan Africa avoid utilizing banking and financial services due to their lack of knowledge and skills necessary to do so [22] or because they are afraid of being defrauded or losing money. From the banks' point of view, because many Nigerians may lack the identification documents required to use formal financial services, there may not be a way to hold customers accountable if they breach the terms of service. As a result, financial institutions have a tougher time providing ground-breaking digital products to consumers.

The International Monetary Fund (IMF) conducted a study whereby they devised a digital financial inclusion index. This index was established by analyzing payment data from 52 developing nations between 2014 and 2017. This study includes both access and usage dimensions of digital financial services (DFSs). Research reveals that the integration of financial technology (FinTech) has played a significant role in promoting financial inclusion, particularly in Africa and the Asia-Pacific region, where the most notable advancements have been observed [23]. A comparative analysis of financial inclusion in 31 sub-Saharan African nations was also conducted by Asuming et al. (2018) using data from the Global Findex Database. This research discovered that many factors at the individual level, such as age, education, gender, and wealth, as well as macroeconomic variables such as the GDP growth rate and the existence of financial institutions, together with business freedom, were statistically significant predictors of financial inclusion [24]. Based on the findings of these studies, it can be inferred that rural communities characterized by inadequate contemporary infrastructure or economic disadvantage may have difficulties in adopting existing banking practices aimed at providing financial services to those without access to banking facilities. To remain competitive, banks will be required to adopt a customer-centric approach like that of FinTech companies.

When it comes to expanding financial inclusion to the unbanked, Nigeria and other Sub-Saharan African nations have much to learn from India. The Pradhan Mantri Jan-Dhan Yojana (PMJDY) is the largest financial inclusion effort in the world, with the goal of ensuring that all Indian people can afford to use basic banking services including deposit and savings accounts, as well as credit, insurance, and retirement plans. Based on a survey of 210 million households, Indian banks have established 115 million accounts as of January 17, 2015, compared to the expected aim of creating bank accounts for 75 million unbanked households in the nation. Scheduled Commercial Banks (SCBs) and Regional Rural Banks (RRBs) have made strides towards financial inclusion by increasing the number of rural banking locations, the number of basic savings bank deposit accounts (BSBDAs), the use of Kisan Credit Cards (KCCs) and General-Purpose Credit Cards (GPCs), and the use of ICT-based banking via Business Correspondents (BCs) [25].

The adoption of artificial intelligence and machine learning technologies for enhancing banking solutions has witnessed significant growth in recent decades. This trend can be attributed not only to the widespread interest in AI/ML technologies and related domains but also to their remarkable potential for disrupting entire industries and rendering businesses obsolete if they fail to engage with these technologies. The banking sector has recognised the importance of leveraging AI/ML to retain existing customers and attract new ones, thereby driving the increased adoption of these technologies. The use of artificial intelligence and machine learning technologies for client acquisition has been a prevalent approach adopted by several firms to maintain competitiveness. This trend is also seen within the banking industry. The major difference across industries is the dataset used to build the ML models. Apart from these, the strategies and goals are similar.

Machine learning algorithms have been used for the purpose of forecasting the level of financial inclusion, enhancing the precision of geographic targeting for financial inclusion initiatives, mitigating risks associated with digital financial inclusion, evaluating credit risk, and facilitating the advancement of financial inclusion. ML algorithms have been used in Africa, particularly in regions characterised by a substantial unbanked population, to forecast the level of financial inclusion. In their study, Ismail, Qusai et al. [26] employed various machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, Bagging, AdaBoosting, Voting Ensemble, K-Nearest Neighbours (KNN), Stack, and XGBoosting Classifiers. These algorithms were utilised to forecast whether an individual possesses a bank account or not. The researchers conducted experiments using these methodologies on a publicly available dataset sourced from African banks in order to forecast the level of financial inclusion in Africa. The accuracy score of the XGBoost model was determined to be 89.23%, indicating its better performance.

Also, Hersh, Jonathan et al. conducted a research study on the application of financial inclusion geographic targeting in Belize [27]. The study employed a machine learning-based small-area estimation technique to generate detailed estimates of financial inclusion at the finest geographical level referred to as Enumeration Districts (ED), which were previously lacking for Belize. To have a more comprehensive picture of the financial characteristics of the population at the ED level, a set of five metrics were constructed to assess the extent of accessibility to banking and financial services. The study further examined the variables that impact the utilisation of financial services and instruments in order to suggest suitable modifications in the tactics performed by authorities in each specific geographic region. This is one of the most creative ways of using ML to target the unbanked in a specific and familiar local area.

The analysis of many data points pertaining to participants, inhibitors, and enhancers of financial inclusion necessitates the use of ML techniques. This is crucial to effectively address the complexity of this undertaking. According to a literature study undertaken by Mhlanga [28], it has been observed that AI has a significant impact on the domain of digital financial inclusion, particularly in the realms of risk identification, assessment, and management. Additionally, the research revealed that AI has the potential to mitigate information asymmetry and enhance loan accessibility for marginalised communities. Furthermore, the research conducted by the Organisation for Economic Co-operation and Development (OECD) [29] emphasises the capacity of ML models to enhance predictability and performance in the field of finance by leveraging large-scale datasets. The research highlights the potential of ML models in discerning patterns and trends within financial data, so enabling financial institutions to enhance decision-making processes and optimise service provision.

In another paper, the International Monetary Fund (IMF) examines the use of AI/ML inside the financial industry, along with the corresponding policy considerations. The paper acknowledges the potential of AI/ML in enhancing consumer understanding and service quality inside financial institutions [30]. However, it emphasises the need to implement suitable regulatory measures and supervision to guarantee responsible utilisation of these technologies. Fares et al. conducted a comprehensive literature analysis on the use of AI in the banking industry [31]. Their findings indicate that the integration of AI technologies in banking may provide benefits such as enhanced customer experience, cost reduction, and improved operational efficiency. The research highlights the potential of AI in the creation of tailored financial goods and services, hence contributing to the enhancement of financial inclusivity. Goodell et al. [32] also provided a comprehensive study that underscores the potential of artificial intelligence AI/ML in enhancing risk management within the realm of finance. The research highlights the potential of AI/ML techniques in the detection and mitigation of hazards associated with financial transactions. This further proves that utilisation of AI/ML has the capacity to instil confidence in and enhance the security and stability of financial systems.

As stated, machine learning has been widespread in enhancing the level of financial inclusion. Banks have used machine learning models to effectively provide relevant goods and services to consumers who have a propensity for churn or inactivity. The model facilitates the identification of prospective clients for financial services via the prediction of customer behaviour and the identification of customers who may need more attention. Majid Bazarbash, in his study [33], examined the potential of machine learning-based credit assessment as a means to enhance financial inclusion and surpass conventional credit scoring methods. The study explored the utilisation of non-traditional data sources to improve the evaluation of a borrower's historical performance, assess the value of collateral, forecast income prospects, and predict changes in overall conditions. As interesting as this sounds, it is however crucial to maintain the relevance of data, particularly in circumstances when there is a significant structural shift, the potential for borrowers to falsify specific indicators, and unresolved agency difficulties resulting from knowledge asymmetry. In order to mitigate the issue of digital financial exclusion and redlining, it is essential to refrain from using discriminatory criteria for the purpose of evaluating creditworthiness.

In developing economies, underbanked people have challenges in accessing conventional collateral or identification documents necessary for obtaining loans from financial institutions. ML has the potential to significantly influence credit risk assessments via the utilisation of alternative data sources, such as publicly available data. This use of ML addresses the challenges associated with information asymmetry, adverse selection, and moral hazard . The use of alternative credit scoring methods that include mobile and social footprints has the potential to enhance credit accessibility for persons who do not possess traditional credit ratings, while minimising any negative effects on default outcomes . Rural finance plays a crucial role in agricultural output worldwide, and the use of machine learning-based technology offers promising prospects for small-scale farmers, rural youth, and women farmers who have limited access to conventional banking services [35] [36].

Kenya serves as a prominent exemplar in the realm of financial inclusion within the African context, showcasing noteworthy achievements. As of 2021, the nation boasts an impressive statistic, with a banked adult population over 80%. M-PESA, a mobile money service introduced in Kenya, has garnered substantial traction, with a use rate above 70% among the Kenyan populace. Its application spans several domains such as finance, trade, education, health, and social welfare . Based on recent research, it has been shown that the implementation of M-PESA has resulted in a poverty reduction of 2% among families in Kenya [38]. According to the findings of Ottavio Del Torso's study [39], M-PESA has effectively facilitated the advancement of financial inclusion in Kenya via enhancing the accessibility, affordability, and convenience of financial services for those who were previously excluded from such opportunities. Several scholarly articles have shown that the mobile banking platform uses machine learning techniques to discern prospective clients who are inclined to show interest in its offerings. The system conducts an analysis of many data points, including client location, income, and spending behaviours, to find those who are most likely to experience advantages from using M-PESA [40].

Also, the use of ML techniques has been employed to forecast the optimal strategies and locations for the deployment of M-PESA, potentially leading to enhanced economic opportunities for marginalised communities [39]. A comprehensive analysis was conducted by a team of specialists from Vodafone to examine the extent of uptake and utilisation of the M-PESA mobile money service inside a developing country. The aforementioned information is now being used to construct an ML model that can effectively forecast the optimal strategies and locations for deploying M-PESA [39]. The implementation of M-PESA in Kenya has shown to be effective in facilitating financial inclusion, given that a mere 7.6% of the Kenyan population had access to mobile financial services in 2013 [41]. The implementation of M-PESA has significantly contributed to the enhancement of the nation's financial inclusion, resulting in a remarkable growth from a mere 26% to a current rate over 80% [42].

In general, AI/ML may be employed as a potent instrument to enhance new consumer targeting by using innovative algorithms and data analysis methods. AI/ML may improve consumer targeting, but they should never replace human judgement or ethical concerns. Using AI/ML, we can improve our targeted marketing in the following ways:

- **Data analysis:** AI/ML algorithms can sift through mountains of information on customers, such as their demographics, online habits, and purchases and social media profiles. AI/ML models may analyse this information to learn more about the customer base. As a result, banks may improve the accuracy of their customer-targeting methods and the specificity of their marketing campaigns.
- **Customer Segmentation:** Using factors like age, gender, geography, interests, and product/services consumption behaviour, AI/ML algorithms may divide consumers into discrete groups for targeted marketing. Banks may then develop segment-specific marketing strategies to reach the most relevant audiences with their messages. It is possible to boost consumer engagement and product utilization with personalised content and offers.
- **Predictive Analytics:** With the use of AI/ML models, banks can forecast their customers' future actions based on past interactions. AI/ML systems may learn to predict customer’s needs and preferences by analysing their interactions and account activities in the past. As a result, banks may more effectively market to their target demographics and increase the possibility of quickly onboarding new customers.
- **Personalization in Real-Time:** Algorithms powered by AI/ML can instantly tailor financial products, content, and promotions to each individual user. AI/ML models may provide customised ideas, proposals, and promotions by continually analysing customer’s interactions. Customers get a better experience from their banks and are more likely to become paying customers because to this real-time customization.
- **Lookalike Modelling:** By analysing the attributes of current customers, banks can develop AI/ML "lookalike" model to locate a pool of prospective customers with similar traits. Lookalike modelling is a strategy for increasing a company's customer base by advertising to people who share its current clientele's demographic characteristics, as well as their hobbies and behaviours.

- **A/B Testing and Optimisation:** AI/ML algorithms may help automate A/B testing, in which two or more variants of a bank's marketing campaign or product offerings are put to the test to see which one yields better results. Banks may increase their consumer targeting and conversion rates over time by continually optimising their marketing tactics using this AI/ML analytics method.

3. DATA & METHODS

This chapter gives the background of our case-study country – Nigeria. It highlights her history, composition, economy, dynamics and potential, relevant financial institutions and how they all affect or contribute to the country's Financial Inclusion rate. The chapter also elucidates the methodologies used in the actual implementation of this study, together with the utilisation of theoretical principles. The process encompasses many stages, starting with the acquisition of raw data, followed by preprocessing, feature engineering, and the understanding of different machine learning models used.

3.1. DATA

3.1.1. Nigeria in a Nutshell

With a population of 202 million people, Nigeria is the world's most populous black country, according to the World Bank [43]. It has the second biggest economy in Africa and is central and home to West Africa's economic activities in location and trade figures. The country is usually referred to as the Giant of Africa, not only due to its huge population and array of natural resources, but also owing to its leadership and influence across the continent and beyond.

Nigeria's arable land area is estimated to about 34 million hectares. This makes Agriculture account for almost a third of the country's GDP and is the primary source of employment [44]. Oil and Natural Gas are the country's major export, as Nigeria remains their largest exporter in Africa [43].

3.1.2. The Central Bank of Nigeria (CBN)

The CBN is the highest monetary authority in Nigeria. It derives its existence and powers from the 1958 Act of Parliament. With six (6) amendments to the Act since its inception, The Act was last amended in 2007, giving the CBN total and complete authority and administration of the Federal Government's monetary and financial sector policies. The CBN Act, stated its objective to: *"ensure monetary and price stability; issue legal tender currency in Nigeria; maintain external reserves to safeguard the international value of the legal tender currency; promote a sound financial system in Nigeria; and act as Banker and provide economic and financial advice to the Federal Government"* [45]. Nothing happens in Nigeria's banking or broadly speaking, the financial ecosystem, without the express permission, regulation, and monitoring by the CBN.

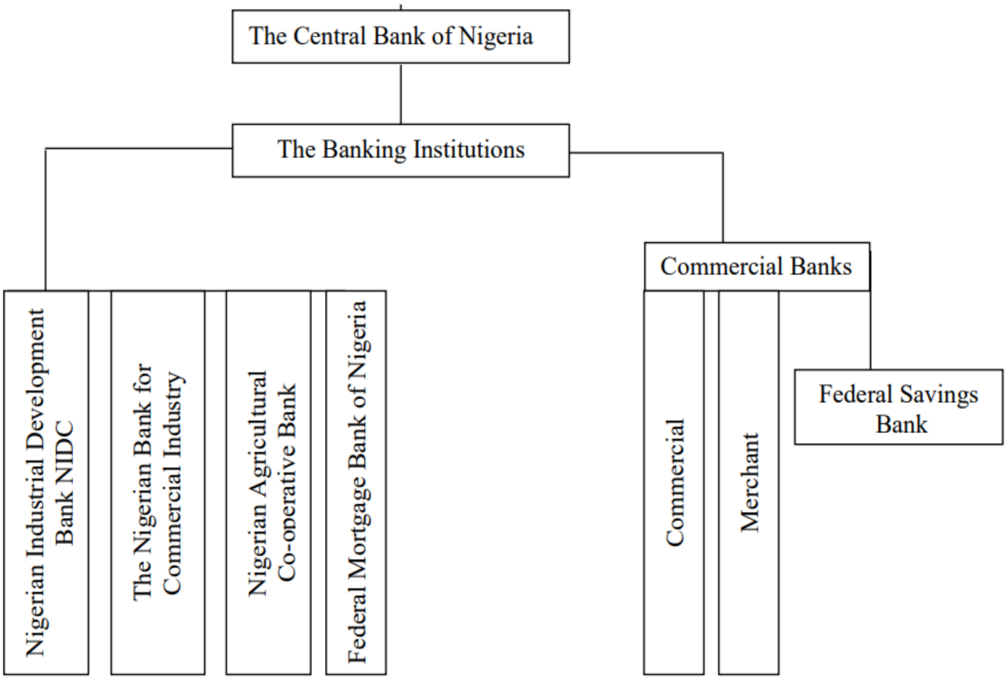


Figure 3.1 – The Nigerian Banking System [46]

3.1.3. The Data Source

We obtained the raw data from the surveys on Access to Financial (A2F) Services in Nigeria. These surveys [47] have consistently been conducted by the institution named Enhancing Financial Innovation and Access (EFInA). For this research, we obtained all six (6) biennial A2F survey data from inception in 2008 till 2018. All the data was supplied in Microsoft Excel Spreadsheet per survey year. For each year, each dataset contains at least 2,500 columns of data points with over 18,000 records each. It should be noted that the datasets were not uniform across all the years. The columns, the values and the granularity level of some data points kept evolving through the years. This was also evident from the Survey Questionnaires as the requirements of the gathered data kept evolving.

EFInA confirms dataset trustworthiness, reliability, fairness and unbiases, that the methodology employed to gather this data is *"nationally representative sample of Nigerian adults (18+) across all 36 states and FCT Abuja and that results are weighted by the National Bureau of Statistics (NBS) to provide for the total adult population and benchmarked to national population estimates for verification"* [47]

3.1.4. Feature Selection and Engineering

Most of the Feature Engineering techniques were implemented at the SQL level using SQL Query – available in the glossary section. The query was used to implement data transformation to a more ML-usable format. With most of the data columns not relevant to this research, features were selected or synthesized from the already available columns or combination of more than one column in some instances. The SQL query used to arrive at the final data set is included in the Appendix section. The columns are `Survey_Year`, `State`, `Sector`, `Residential_Area_Density`, `Gender`, `Age`, `Marital_status`, `Employment_Status`, `Monthly_Income_Rank`, `Education_Level`, `Occupation_Score`, `Access_to_Mobile_Phone`, `Access_to_Personal_Computer`, `Access_to_Internet_or_Email`, `Access_to_Television`, and the target or predicted column `Bank_Status`.

3.1.5. Preprocessed Dataset

The raw Excel data files were first converted into comma-separated value (*csv*) file formats before importing them into Google BigQuery Data warehouse for initial analysis. We observed that many of the data columns contain lots of data with either data points that were not useful to this research or null or empty data. These were the first to be eliminated in the data cleaning process. The Survey Questionnaire for each of the years provided much needed context to fully understand reasons why some details are collected and how they may be useful in deriving some patterns. A copy of the full SQL queries used in transforming the entire raw data into the usable preprocessed dataset has been attached to the Appendix section. Below is the descriptive analysis on the final dataset used:

Table 3.1 – Description of the pre-processed dataset

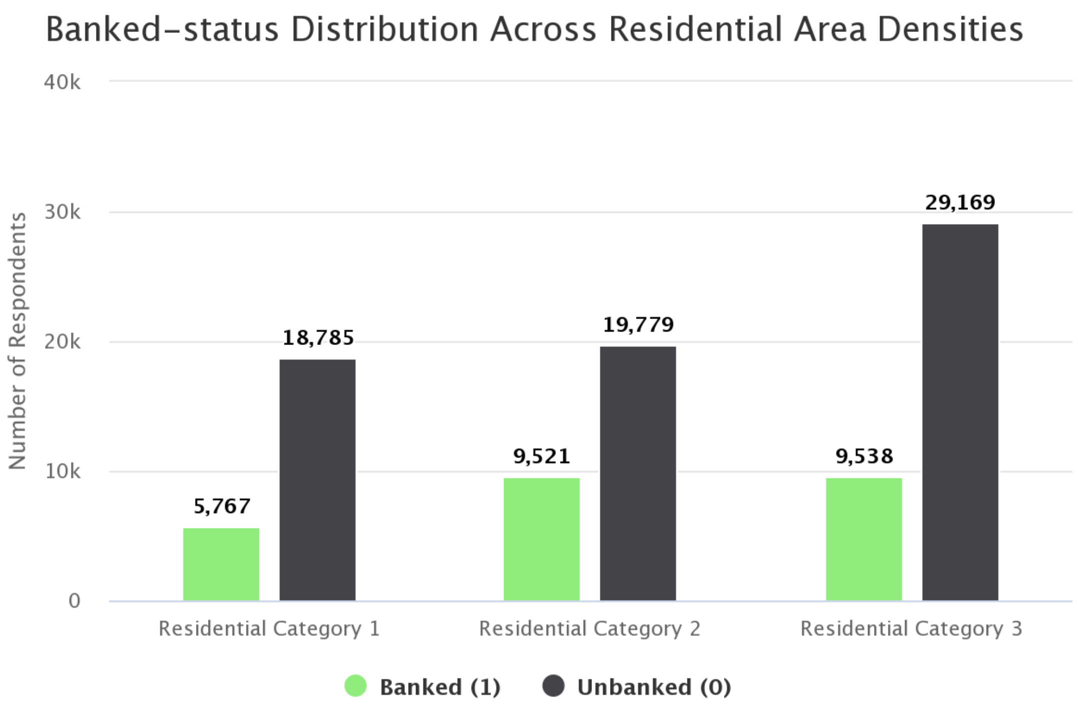
Column Name	Data type	Count	Unique	Top	Freq	Mean	Min	Max
Survey_year	Integer	92,759	N/A	N/A	N/A	N/A	2008	2018
State	String	92,759	37	Katsina	3,483	N/A	N/A	N/A
Sector	String	92,759	2	Rural	68,204	N/A	N/A	N/A
Residential_Area_Density	Integer	92,759	N/A	N/A	N/A	2.1504	1	3
Gender	String	92,759	2	Female	46,455	N/A	N/A	N/A
Age	Integer	92,759	N/A	N/A	N/A	36.5638	18	99
Marital_status	String	92,759	8	Married (Monogamy)	52,744	N/A	N/A	N/A
Employment_Status	String	92,759	9	Self-employed	44,203	N/A	N/A	N/A
Monthly_Income_Rank	Integer	92,759	N/A	N/A	N/A	4.6344	1	12
Education_Level	Integer	92,759	N/A	N/A	N/A	2.6148	0	8
Occupation_Score	Integer	92,759	N/A	N/A	N/A	1.3239	0	5
Access_to_Mobile_Phone	Integer	92,759	N/A	N/A	N/A	0.5415	0	1
Access_to_Personal_Computer	Integer	92,759	N/A	N/A	N/A	0.0387	0	1
Access_to_Internet_or_Email	Integer	92,759	N/A	N/A	N/A	0.0530	0	1
Access_to_Television	Integer	92,759	N/A	N/A	N/A	0.4882	0	1
Bank_Status	Integer	92,759	N/A	N/A	N/A	N/A	0	1

The preprocessed data consists of sixteen (16) variables, the first fifteen (15) are the features or independent variables, while ***Bank_Status*** is the target variable. The table above provides some statistics on how data is distributed in this dataset. The following is the definition of the table headers:

- **Count** is the number total number of records available for the column.
- **Unique** means total number of distinct non-number values represented in a column.
- **Top** represents the non-number value with the most occurrence in a particular column.
- **Freq** is the number of times the Top value occurred.
- **Mean** is the average of the numeric-valued columns except for the target variable.
- **Min** is the minimum value present in the numeric-valued columns.
- **Max** is the maximum value present in the numeric-valued columns.

3.1.6. Data Distribution across Select Features

This technique was used to graphically visualize our data across select features and the target columns. It allows us to make certain preliminary inferences concerning data distribution and classification.



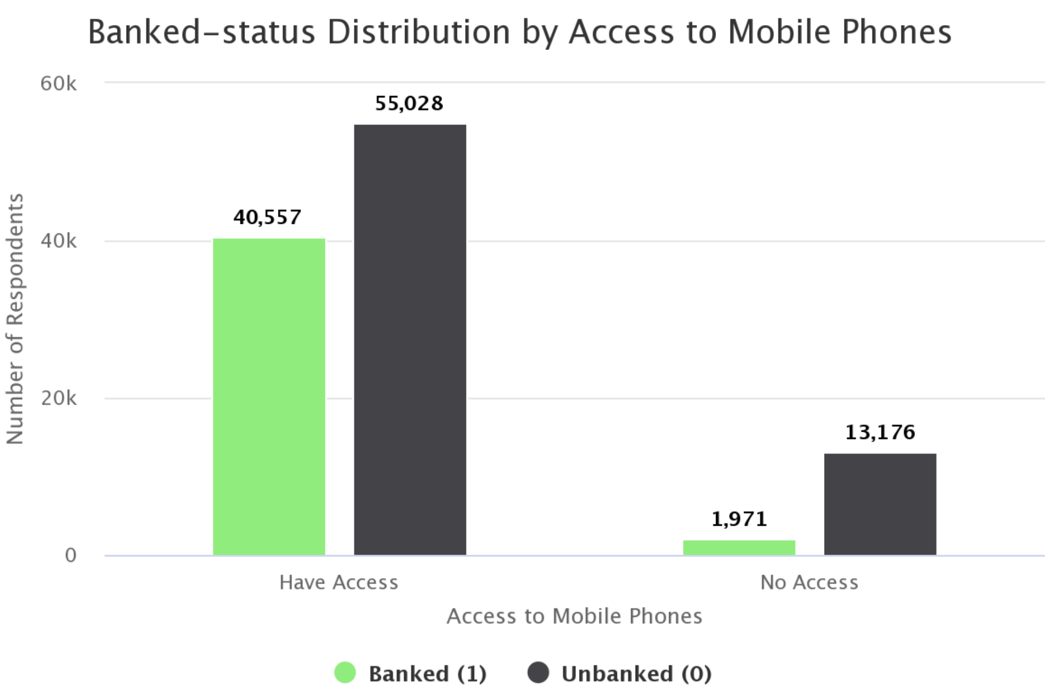


Figure 3.2 - Distribution of Features - Residential Area Density (top) and Access to Mobile Phones (bottom) against Bank Status (target)

3.1.7. Feature Importance and Correlation

We use the Correlation Matrix to understand the relationship between multiple numerical variables in the dataset. A higher positive or negative number signals a possible direct or inverse relationship and could be a strong predictor.

Figure 3.3 - Correlation Matrix of Numerical Features in dataset

3.1.8. Dealing with Categorical Variables

The final dataset contains some Categorical variables that should not be explicitly converted to numerical values using SQL, and must be preserved "in state," in order not to confuse the ML model mistaking them for ordinal variables. For each of these features, we create "dummy variables" which converts distinct values of each categorical feature into distinct columns. There are five (5) of them in our dataset - `State`, `Sector`, `Gender`, `Marital_status` and `Employment_Status`. These columns were converted using a method called One-Hot Encoding. With this approach, each category is mapped to a vector that has the values 1 and 0 to denote whether the feature is present or not. The amount of feature categories determines the number of vectors that will be created [48]

3.1.9. Imbalanced Dataset

The distribution of the features along the target variable values shows the data is highly imbalanced (see Figure 4.8). While we were able to achieve a decent accuracy score for the model, it is all due to the large amount of data in the majority (unbanked) class than the minority (banked) class. By plotting the distribution for only the Target variable (`Bank_Status`), we can observe that the dataset is imbalanced by 3:1 in favor of the unbanked (target value = 0) class. This explains why we have higher precision, recall, f1-score, and support values for the majority class – as seen in Table 4.3.

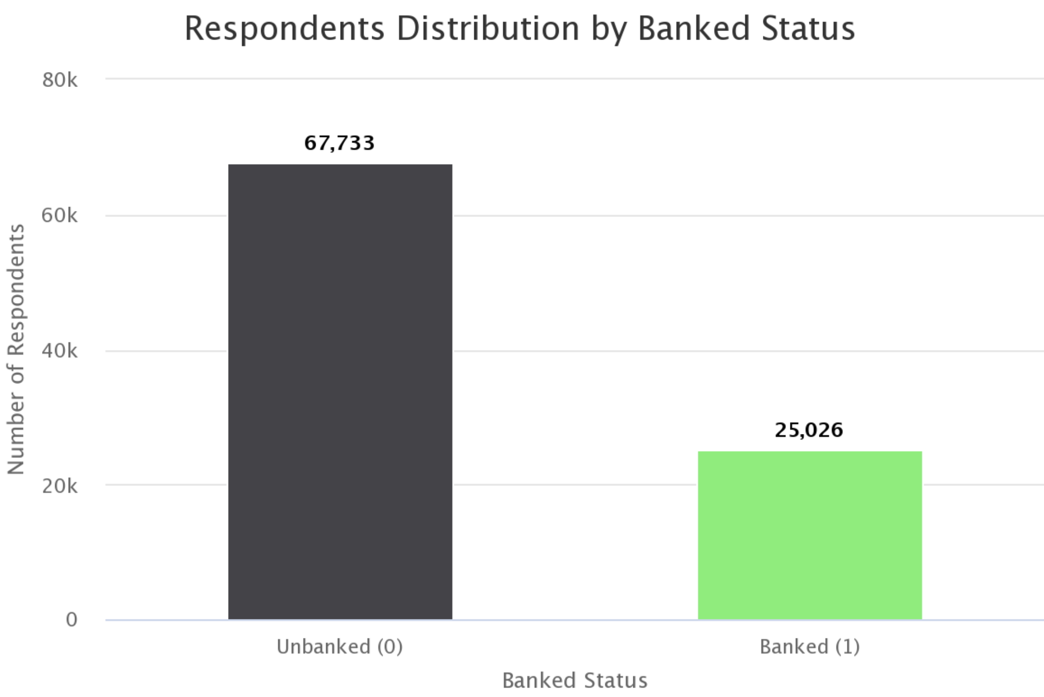


Figure 3.4 – Distribution of the Target Variable showing that the Dataset is Imbalanced

Also, from the baseline model, it can be observed that despite the model having a fair accuracy score of 0.85, most of the accuracy is tilted towards the majority (unbanked) class. This can be seen from the Recall value (0.67) for the minority (banked) class. Hence the need to balance the dataset and ensure the model is making prediction based on learning rather than guesses leaning to the majority class.

3.1.10. Handling the Imbalanced Dataset

Obviously, the majority class is about 3 times the size of the minority class. We experimented balancing the dataset using different methods to avoid overdependence on the majority class. The two (2) methods employed were:

- **Random Over-sampling:** Randomly duplicate examples in the minority class.
- **Random Under-sampling:** Randomly delete examples in the majority class.

Table 3.2 – Result of Applying Over-sampling and Under-sampling methods on the Target Variable

		Dataset Balancing Methodology	
		Random Under-sampling	**Random Over-sampling**
Classes	**0**	25026	50781
	1	25026	50781
	Total	**50,052**	**101,562**

In the end, the under-sampling method produced better results on the baseline model - Hence the reason we chose it to balance the dataset for the rest of the employed ML techniques.

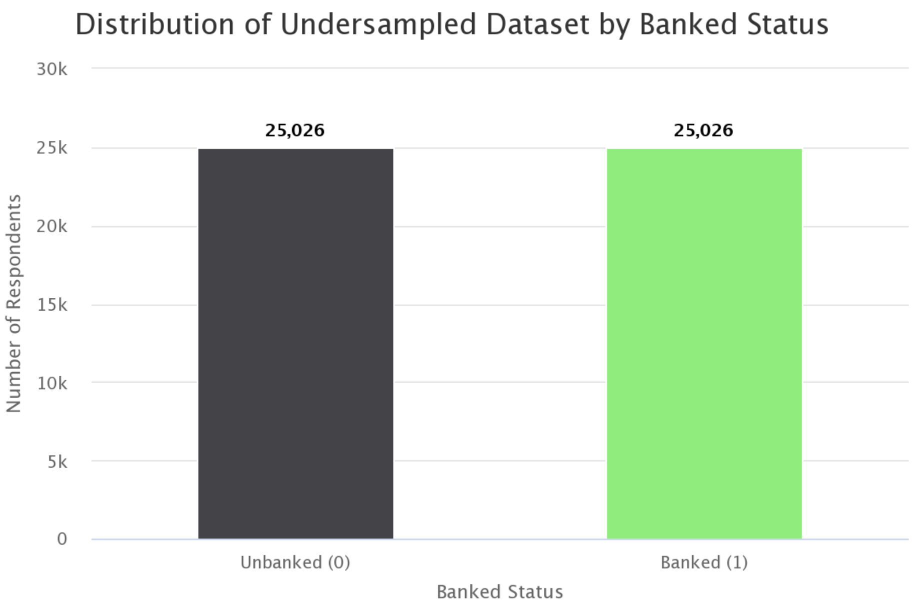


Figure 3.5 – Final Distribution of the Target Variable showing that the Dataset is now balanced (Under-sampled)

3.2. Data Exploration and Preliminary Analysis

3.2.1. Analysis on the Current Year (2018) data

The survey for 2018 revealed that Nigeria have about 63% of her adult population to be financially included. Based on Nigeria's method of FI classification, as explained in the below figure, 39.6% of the adult population have or use financial products/services provided by a deposit money bank, while 9% have or use financial products or services provided by other regulated non-deposit money bank financial institutions. The data also shows that 14.6% of the adult population consumes these same products and services, however, they are ones which are not regulated – hence the label – informally served. Nigeria has a financial exclusion rate of 36.8% [7]

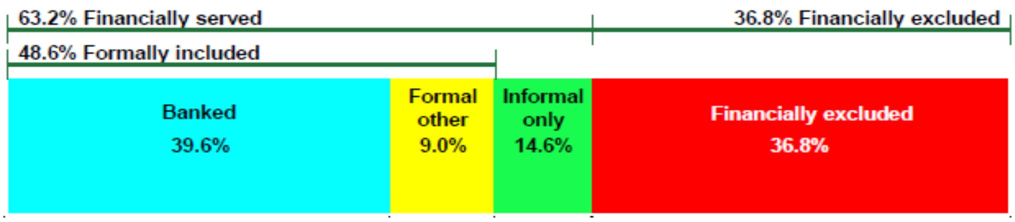


Figure 3.6 – Financial Access Strand 2018 [7]

Of the 27,470-sample size (persons whose data were gathered), about two-thirds (66.3%) live in rural regions, and over half (50.1%) are female. 56.9% of the population is under the age of 35, while 20.4% have no formal education whatsoever. Only 8% of the adult population work in Formal sector employment. 16.7% own businesses that are non-farming while 11.2% own a farming business. The survey reveals that a total of 23.4% directly or primarily rely on farming for income.

Small and individually owned businesses provided the most source of employment in 2018 to the Nigerian adults. 81.2% of the SMEs were individual entrepreneurs while the remaining 18.2% were business owners with employees. These businesses are spread across the major economic sectors: Agriculture (29.3%), Artisan services (21.2%), Retail (19.1%), Other services (12.6%) and Construction (6.0%) [47]

In 2018, affordability, institutional exclusion and peoples' attitude or perception are the greatest drawbacks to owning a bank account. The survey shows that 35% of the Unbanked population are excluded due to irregular income:

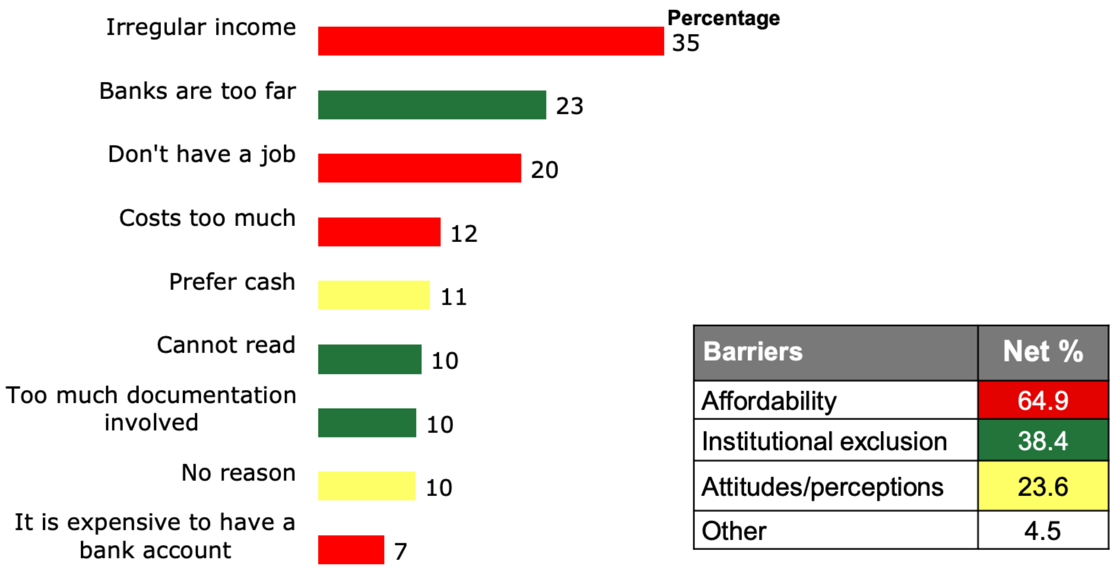


Figure 3.7 – Top 9 Obstacles of having a Bank Account for Nigerians in 2018 [7]

3.2.2. Typical Profile of the Financially Excluded Adult.

From the latest data in 2018, about 56% of adults that make up this cluster are females while 44% are male. 34% are in the age category of 18 – 25 Years old. 78.5% of them live in the Rural areas while 21.5% live in the urban areas. A little more than half of this category of people live up-North. 38.4% reside in the North-West while 18.3% live in the North-East geopolitical zones. Access to Mobile Phones is poor amongst this group with as much as 71.3% without phone access. The median monthly income of this group being as low as NGN15,000 (about USD 35) explains why 23.5% of this group earn their living through Small-scale farming and 18.3% rely on family or friends. In terms of awareness and access, about 96% of this group are unaware of Mobile Money while about 90% of them do not know who banking agents are or what they do. 76% of this group do not have access to any financial institution. Below is the groups distribution on highest level of Education.

Table 3.3 – Level of Education Distribution amongst the Unbanked (2018 EFInA Data) [47]

Highest Level of Education	Percentage Represented
No education	19.40%
Non-formal education	20.20%
Primary School	23.40%
Secondary School	34.70%
University/ OND/ HND	1.90%
Post University	0.20%

There appears to be a direct connection, or simply put, a correlation between the level of Education and Financial Inclusiveness, as it can be seen that most financially excluded adults have not been schooled beyond the Senior Secondary School level. The percentage is exclusion is quite notable with people who are well educated above the average as seen with the University and Post University numbers.

3.2.3. Financial Inclusion Trends over the years

Publicly available statistics on the adult population in Nigeria over the years were combined with the results of the surveys from the corresponding years to give an actual true number to the ratio analysis from the survey data. It can be concluded that while there has been some noticeable increase in actual Population, the Financial Inclusion numbers have also relatively improved.

Comparing the last year (2018) data to the previous year (2016), the Banked population seems to have increased by 1.4%. While Remittances (send and receive funds) reduced by 2.2%, and savings with banks reduced by 6.7%, Payments via banks witnessed some 3.4% growth. Same can be said with the percentage of adults who received income through a bank – increasing by 1.3% as previously shown in Table 2.1.

In terms of degree of usage, more banked people used their bank accounts in 2018 (68.7%) than in 2016 (59.9%). These figures indicates that the people referenced used at least one bank product one month prior to the month of the survey – These are the users in High/Active usage category. There is also an in increase in the Medium Usage catgeory from 3.2% in 2016 up to 4.6% in 2018. These are individuals who used at least one banking product during the six months prior to the survey being conducted.

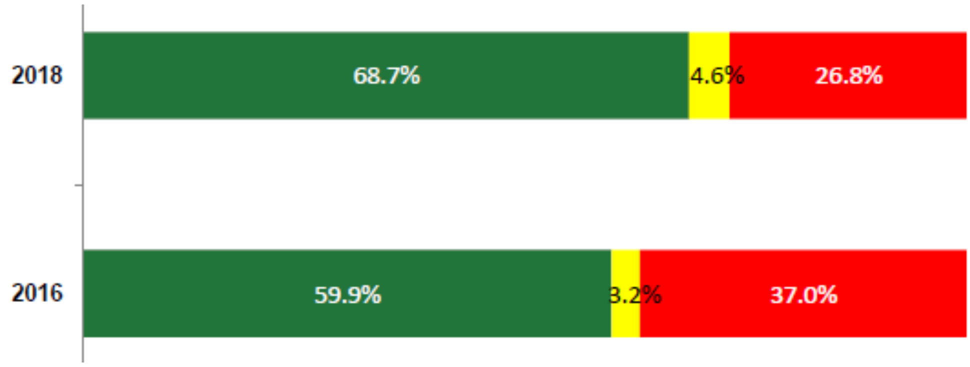


Figure 3.8 – Bank Account Usage: 2018 vs. 2016 [7]

3.3. Methods

This research seeks to determine if an individual is a right potential bank customer, ready for onboarding or not – given certain details about them. These details form the data points used as variables in the machine learning models predicting the 'Banked' status which indicates whether an individual should be followed-up on – after considering popular data points such as the ones described in section 3.1.4. Four (4) Supervised machine learning methodologies have been employed for this classification task. Below is a brief overview of the models:

3.3.1. Logistic Regression

Logistic regression is a supervised machine learning technique used for classification tasks with two possible outcomes. Logistic regression models are particularly used to analyze data whose dependent variable (the target) is categorical or binary, while the input is usually a set of independent variables [49]. According to Ursa et al, *"logistic regression classifier can use a linear combination of more than one feature value or explanatory variable as argument of the sigmoid function. The corresponding output of the sigmoid function is a number between 0 and 1. The middle value is considered as threshold to establish what belongs to class 1 and class 0. In particular, an input producing an outcome greater than 0.5 is considered to belong to class 1. Conversely, if the output is less than 0.5, then the corresponding input is classified as belonging to 0 class"* [50]. Examples of Logistic regression use cases are classification of online transaction or credit card payment as fraud/not fraud, emails as spam/not spam or any other discrete set as for / against.

Unlike Linear Regression which uses a straight line to determine the value of its output, the logistic regression uses a logistic function – an S-shaped curve which takes a number between 0 and 1. Christoph [49] defined the logistic function as:

$$\sigma(y) = \frac{1}{1+e^{-(y)}} \tag{1}$$

Where y is the linear equation given to the sigmoidal activation σ.

Sayed [51] explains that a linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g., predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

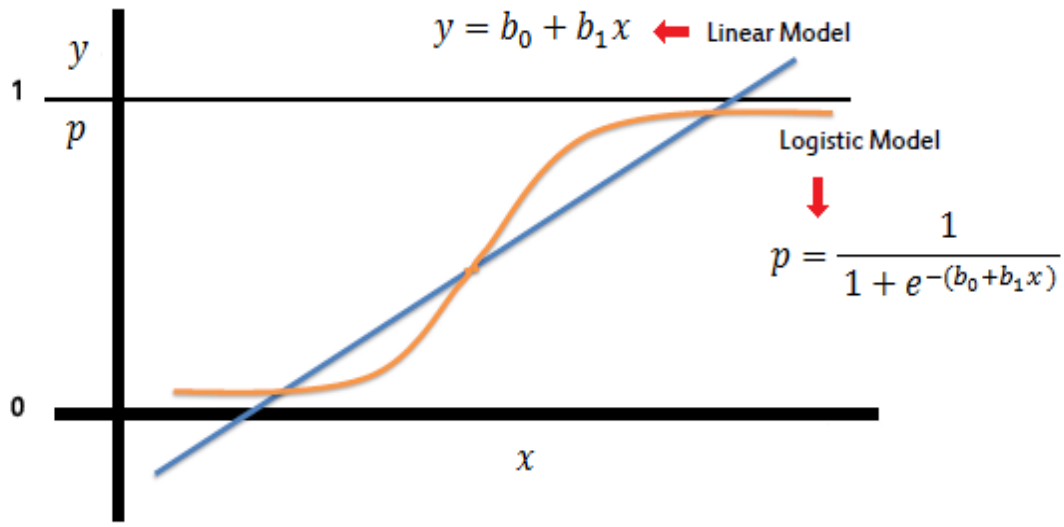


Figure 3.9 – The Linear and Logistic Regression Curves [51]

3.3.2. Naïve Bayes Algorithm for Classification

The Naïve Bayes classifier is a probabilistic machine learning model based on the Bayes theorem – a peculiar method of calculating conditional probability of events [51]. It uses a naïve assumption of conditional independence of all random variables. Classifiers based on the Naïve Bayes algorithm need a small amount of input data to be trained, but they can work with high-dimensional data points, in addition to being quick and highly scalable [52].

Naïve Bayes models are effective by predicting the probability of different classes based on available attributes. Examples can be found in binary or multiclass classification use-cases such as in spam filtering, sentiment analysis, text categorization, recommendation systems and medical diagnosis classification.

Sayed [51] represented the algorithm's conditional independence in detail as thus: *"Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naïve Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors"*:

$$P(c/x) = \frac{P(x|c)\,P(c)}{P(x)} \tag{2}$$

$$P(c/x) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c) \tag{3}$$

Where:

- *P(c|x)* is the posterior probability of class (target) given predictor (attribute).
- *P(c)* is the prior probability of class.
- *P(x|c)* is the likelihood which is the probability of predictor given class.
- *P(x)* is the prior probability of predictor.

3.3.3. eXtreme Gradient Boosting (XGBoost)

In general, Gradient Boosting is a form of ensemble learning used to solve both classification and regression problems. It is a machine learning approach that generates a prediction model in the form of a collection of weak prediction models, most often decision trees. By permitting optimization of any differentiable loss function, it generalizes previous boosting approaches that construct the model step-by-step.

XGBoost is a supervised ML algorithm that implements the gradient boosting process to produce even more accurate models by providing parallel tree boosting. Some of the features that makes XGBoost outstanding are its incredible optimization on computational speed, as well as model performance. It is also capable of some advanced features. An XGBoost model implementation has support for any of the gradient boosting – (normal) gradient boosting, Stochastic or Regularized gradient boosting. The Regularized model uses both the L1 and L2 regularization to minimize overfitting and optimize performance, according to the algorithm's creator, Tianqi Chen [53].

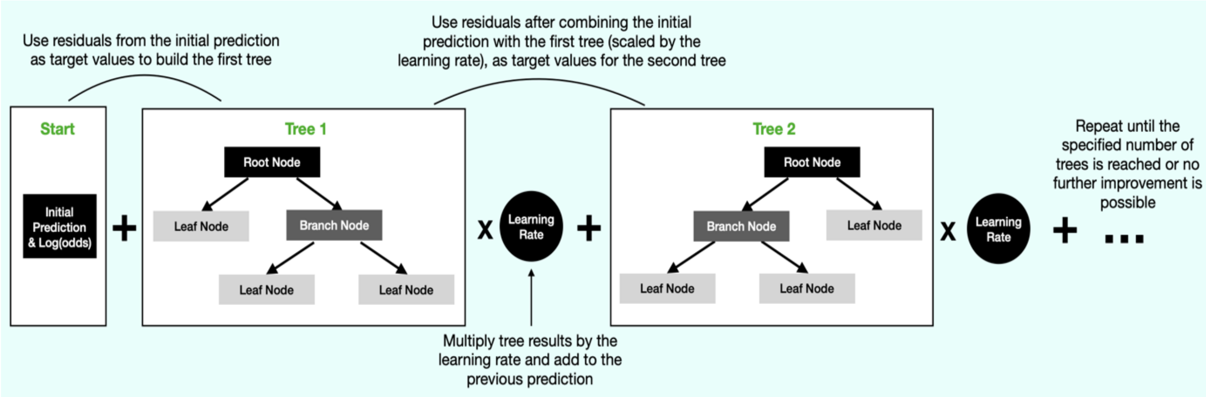


Figure 3.10 – The Process Map for Gradient Boosting and XGBoost algorithms [54]

Dobilas [54] explains the distinction between the regular and extreme gradient boosting methods. According to the article, "*regular Gradient Boosting uses a standard method to build regression trees, where a typical metric such as MSE (Mean Squared Error) or a similar one is used to determine the best split for the tree. The algorithm calculates MSE for every possible node split and then picks the one with the lowest MSE as the one to use in the tree*":

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Observed_i - Predicted_i)^2 \tag{4}$$

This is not the case with XGBoost, which employs a unique tree-building algorithm based on the Similarity Score and Gain:

$$Similarity\ Score = \frac{(\sum_{i=1}^{n} Residual_i)^2}{(\sum_{i=1}^{n}[Previous\ Probability_i * (1 - Previous\ Probability_i)] + \lambda} \tag{5}$$

Where:

- ***Residual is*** *actual (observed) value — predicted value.*
- ***Previous probability*** *is the probability of an event calculated at a previous step. The initial probability is assumed to be 0.5 for every observation, which is used to build the first tree. For any subsequent trees, the previous probability is recalculated based on initial prediction and predictions from all prior trees, as shown in the process map.*
- ***Lambda*** *is a regularization parameter. Increasing lambda disproportionately reduces the influence of small leaves (the ones with few observations) while having only a minor impact on larger leaves (the ones with many observations).*

Each leaf has a Similarity Score, and we use this score to determine the Gain.:

$$Gain = Left\ leaf_{similarity} + Right\ leaf_{similarity} - Root\ leaf_{similarity} \tag{6}$$

The optimal node split for the tree is then determined to be the one that results in the **maximum Gain**.

3.3.4. Neural Networks

There are several types of neural networks, and one of the most popular is the artificial neural networks (ANNs). Neural network's inspiration can be traced to core biological sciences concepts on how the human brain system works, which is in a complex formation and interrelation of units, regarded as a circuit of biological neurons known as the human neural networks. In other words, the Neural networks mimic the structure and behavior of the human brain.

An ANN is a collection of artificial neurons called nodes – which are like the neurons in the human brain. Hence, the entire ANN contains layers of interconnected nodes. It is through the connection that data (otherwise known as signal), flows from one node to another. There are three kinds of neurons: **input nodes**, **hidden nodes**, and **output nodes** [55].

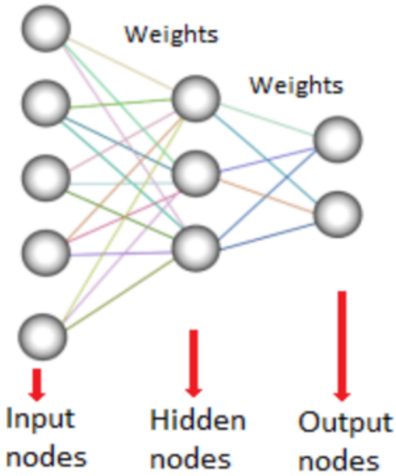


Figure 3.11 – Basic Representation of an Artificial Neural Network [51]

The input node units receive data in the form of real numbers. A crucial quantity known as a weight is assigned between two connected nodes based on the strength of the connection used to describe the connections that exist between the two units. This weight may be positive (indicating that one unit excites another unit) or negative (indicating that one unit depresses another one). When inputs are sent across connections through a network, their values are multiplied by the weights of each connection they have passed through. The network then compiles the resulting products into a single number by adding them together. If the number is below a certain limit, the node will not send any data on to the subsequent tier. If the number is more than the threshold value, the node "fires." This limit is calculated from an equation known as the activation function [51]. This kind of neural network is known as a feedforward network, and it gets its name from the process of transmitting input from one layer to the next.

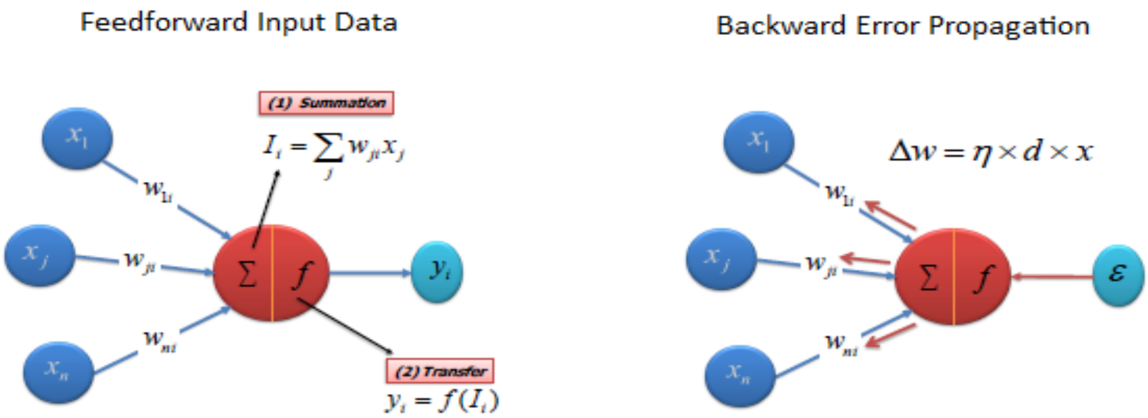


Figure 3.12 – Feedforward and Back Propagation Representation of a Neural Network [51]

The activation travels through the network, passing through any hidden layers along the way, until it reaches the nodes that are responsible for output. The error, which is the difference between the projected value and the actual value, will be propagated backward by allocating it to the weights of each node in proportion to the amount of this mistake that each node is accountable for [55].

Popular examples of activation functions used in ANNs include Linear Function, Tangent Hyperbolic function (Tanh Function), Rectified linear unit – RELU, Softmax Function, Sigmoid and Gaussian Function.

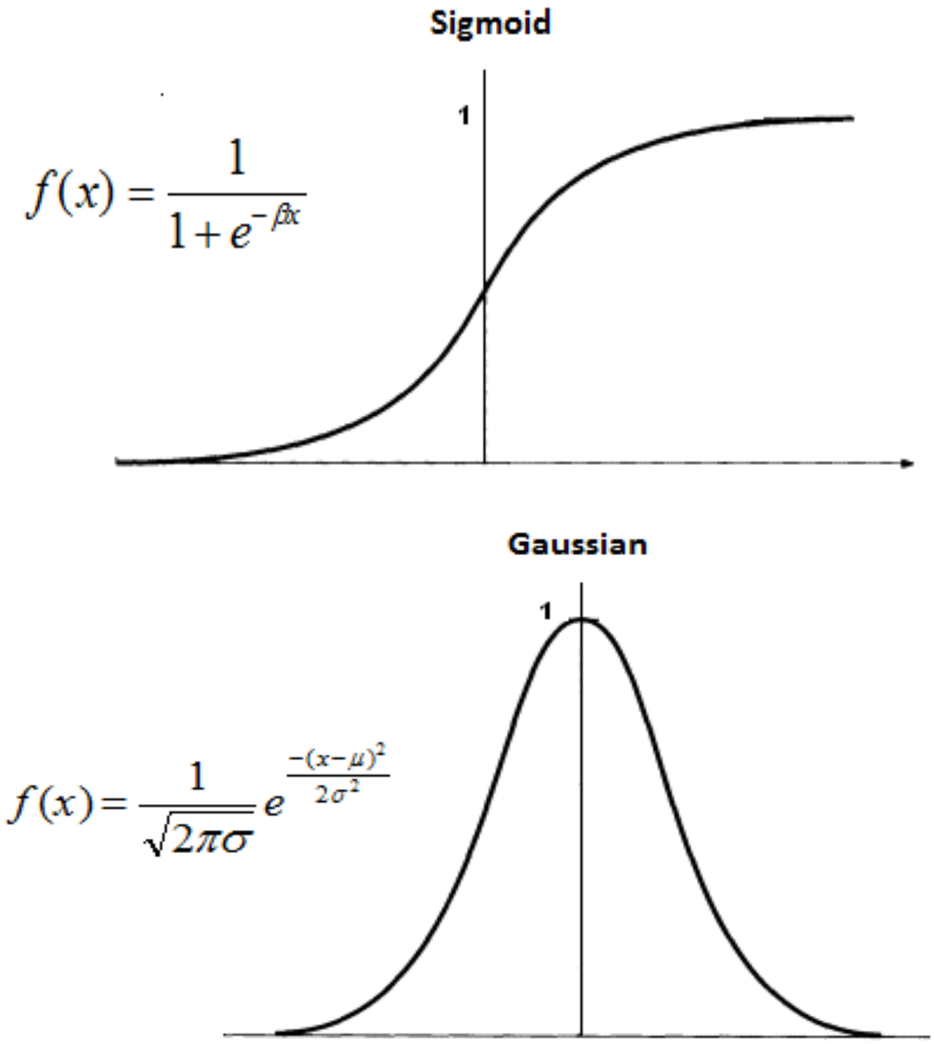


Figure 3.13 – Sample Activation Functions: Sigmoid (Top) and Gaussian (Bottom) [51]

3.4. ML Models Evaluation and Metrics

In this section we will be looking into the processes or tasks that must happen before, during and after our Machine Learning models' development. Evaluating the model is quite important because it shows the current state of the model and what area needs to be improved on. Often, there are usually specific properties of a model that makes it acceptable – For example – while a binary classification model might have an overall accuracy (term explained below) of say 95%, the accuracy on the minority class might be – say 65%. Another model of overall accuracy of 90% with both minority and majority class accuracy of above 85% would be preferred as it generalizes well the on the entire dataset. Decisions like these can only be made when we have model evaluation and metrics. It is important to define and explain the metric terms below as they would be well referred to in the next chapter.

3.4.1. Confusion Matrix

A confusion matrix is a tabular representation of certain metrics used to measure how well the ML model generalizes during prediction. It is also called the Error Matrix. It is an A x P matrix where A represents the Actual class and P represents the Predicted class. The information provided by the matrix helps to interpret what kind of error the model is making and what area to concentrate on, over the next training iteration [56]. Through the table, we observe the instances – both correctly and incorrectly classified – against their respective target classes.

For every predicted outcome in the test set, there is a corresponding actual value. Based on these two (2) values, we define certain components that form the basis of the confusion matrix. They are defined below:

- **True Positive:** when actual label is "True," and model correctly classifies as "True"
- **True Negative:** when actual label is "False," and model correctly classifies as "False"
- **False Positive:** when actual label is "False," but model wrongly classifies as "True"
- **False Negative:** when actual label is "True," but model wrongly classifies as "False"

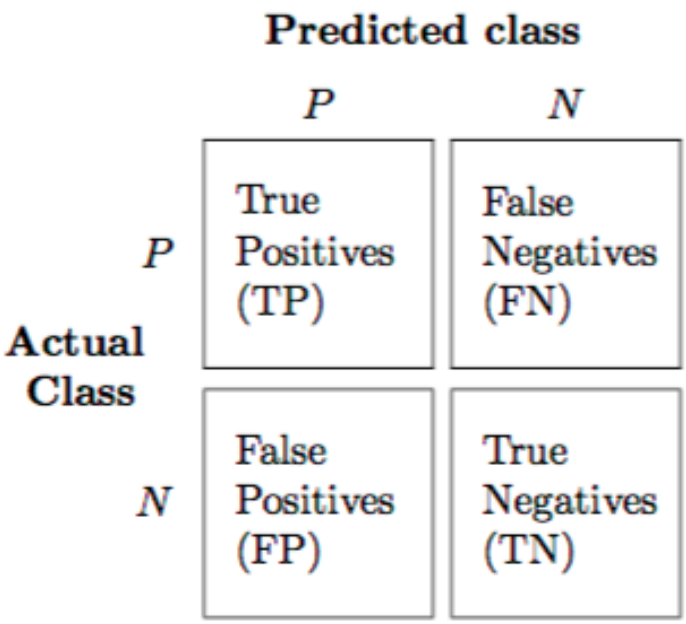


Figure 3.14 – The Confusion Matrix

The importance of the confusion matrix in classification model performance evaluation cannot be overemphasized. It is the foundation upon which other important measures are established. The four components – True Positive, True Negative, False Positive and False Negative – are pivotal in synthesizing the following measures: Accuracy, Precision, Recall, F1-Score as well as the AUC-ROC curve. These measures are explained below:

3.4.2. Accuracy

Accuracy is the measure of how many times our model was right throughout prediction. It is calculated as the fraction of correct predictions to the total number of predictions made.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{7}$$

Following the representation in the confusion matrix:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

3.4.3. Precision

Precision focuses on rate of positives – indicating what proportion of positive predictions were actually correct:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

3.4.4. Recall

While recall also focusses on the positives, it checks the actual True Positives against the universal set of relevant outcomes. It is also referred to as the 'True Positive rate' or 'Sensitivity.' It is used to tell how many times the model returns positive when the target is actually positive.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

3.4.5. F1 Score

This is the harmonic mean of precision and recall. It is often called F1 Measure because the precision and recall are evenly weighted and is used when the two measures are close.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{11}$$

3.4.6. AUC ROC score

Sometimes, numbers do not tell all the stories or present the facts at a glance. This is where visuals come in. The Receiver Operating Characteristic (ROC) curve is a graph used for visually assessing the prediction accuracy of a classification model by plotting its True Positive Rate (TPR) on the Y axis, against the False Positive Rate (FPR) on the X axis. Using these values to plot the plot our graph:

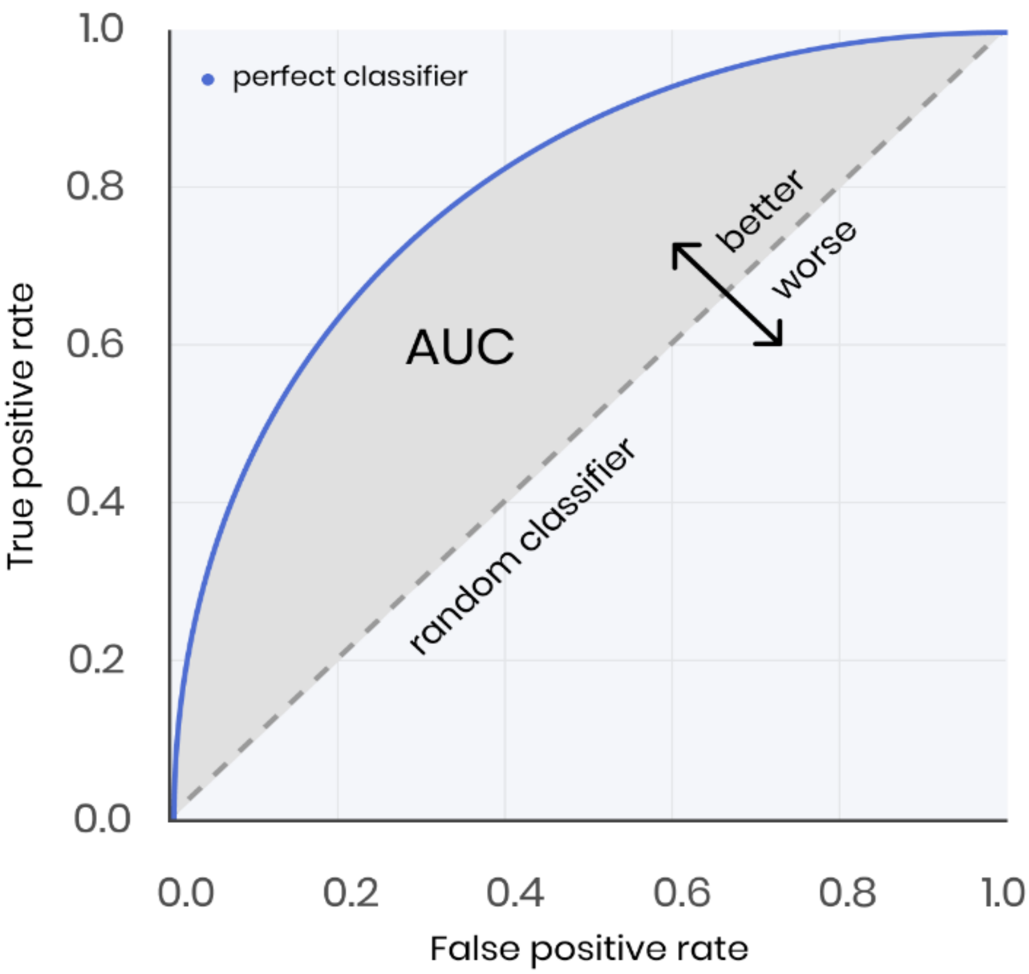


Figure 3.15 – The Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) [57]

On the graph, the model's ability to differentiate between positive and negative classes improves with increasing AUC value. So, a perfect classifier (AUC = 1) is one having a True Positive Rate of 1 and False Positive Rate of 0. But there is no such thing as a perfect classifier – which essentially means that there will be some area covered by (under) the ROC curve – AUC ROC.

When the AUC = 0.5, it is represented as straight diagonal line (from 0,0) across the graph, and it means the model is just guessing at 50-50 chances. This is not a performing model, and any value below this score (AUC < 0.5) indicates that the model is worse. As data scientists make decisions on final model selection, the AUC presents one of the best visual tools showing the trade-off between the TPR and the FPR.

4. RESULTS AND DISCUSSION

In this chapter, we present the outcome of the baseline model as well as the final models. The following were the ML models implemented and remodeled through multiple iterations to obtain and use the best set of parameters that would produce the most optimal model accuracy. They are:

4.1.1. The Baseline Model (Logistic Regression)

A baseline model is the simple model that was created using the dataset input as it is, with minimum configuration or optimization. It provides a minimum predictive ability and benchmark which subsequent models can be built upon. We used a simple Logistic Regression model to achieve the result:

Table 4.1 – Baseline Model Classification Report

		Measures			
		Precision	**Recall**	**F1-Score**	**Support**
Prediction Label	0	0.88	0.92	0.90	16,952
	1	0.75	0.66	0.70	6,238
Model Accuracy				0.85	23,190
Macro Average		0.81	0.79	0.80	23,190
Weighted Average		0.84	0.85	0.85	23,190

The measures – Precision, Recall, F1-Score and Accuracy – have already been defined. However, the term 'Support' here indicates the actual number of occurrences of the predicted class.

Table 4.2 – Contingency Table showing Actual vs Baseline Model Predicted Values

		Predicted Value		
		Unbanked (0)	**Banked (1)**	**Total**
Actual Value	**Unbanked (0)**	15,593	1,359	**16,952**
	Banked (1)	2,152	4,086	**6,238**
	Total	**17,745**	**5,445**	**23,190**

4.1.2. Logistic Regression

The final Logistic Regression model saw a considerable improvement over the initial figures from the baseline version. Though the overall accuracy dipped from 0.85 to 0.80, there accuracy, precision, recall as well as the f1-score all improved – for the minority class in the new iteration due to more balance to the data, and parameter tuning.

Table 4.3 – Logistic Regression Model Classification Report and Performance Measure

		Measures					
		Precision	**Recall**	**F1-Score**	**Support**	**Metrics**	
Prediction Label	0	0.74	0.93	0.82	6,286	**True Positive Rate:**	0.85
	1	0.90	0.67	0.77	6,227	**True Negative Rate:**	0.81
Model Accuracy				0.80	12,513	**False Positive Rate:**	0.19
Macro Average		0.82	0.80	0.79	12,513	**False Negative Rate:**	0.15
Weighted Average		0.82	0.80	0.79	12,513	**Positive Predictive Value:**	0.62

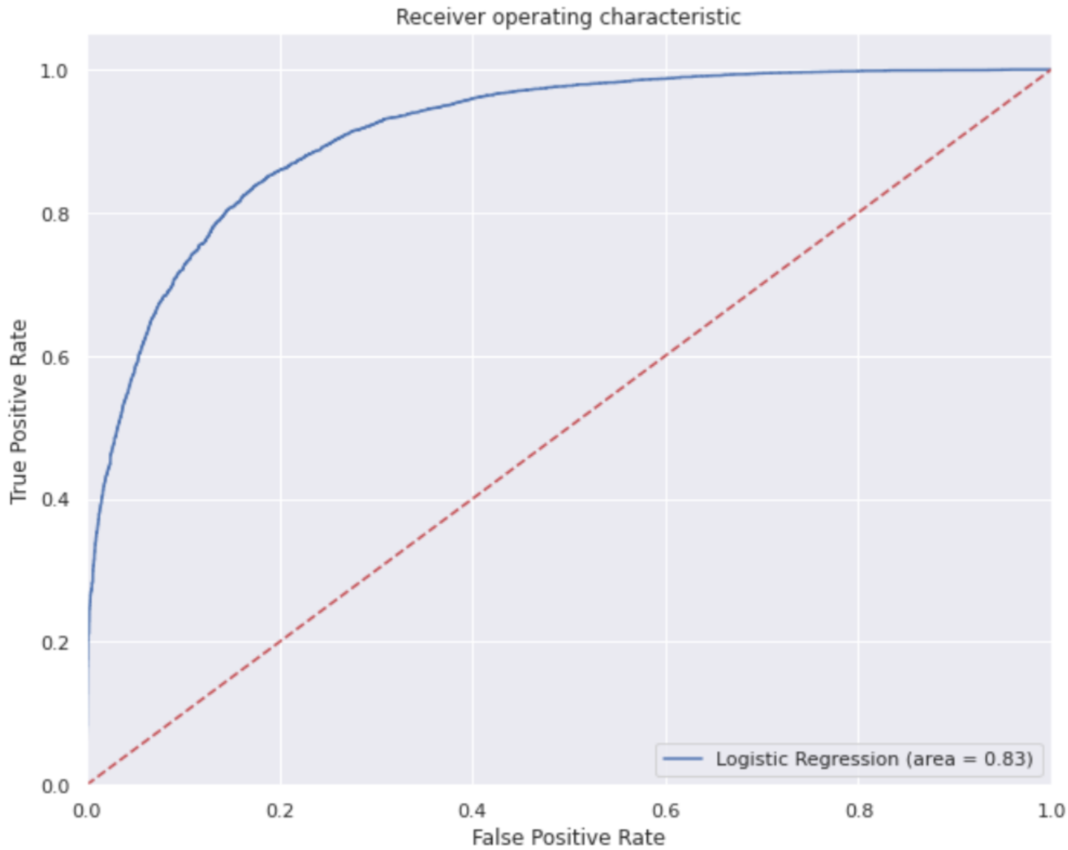


Figure 4.1 – Logistic Regression Model ROC Curve

Table 4.4 – Contingency Table for Logistic Regression Model Predicted Values

		Predicted Value	
		Unbanked (0)	**Banked (1)**
Actual Value	**Unbanked (0)**	5,829	457
	Banked (1)	2,085	4,142

4.1.3. Naive Bayes Algorithm for Classification

With an overall accuracy of 0.75, the result from the Naive Bayes model was not as good as what was obtained in the previous model iterations. It appears to be the least performing of all the models with results falling behind even the baseline logistic regression model.

Table 4.5 – Naive Bayes Model Classification Report and Performance Measure

Measures							
		Precision	**Recall**	**F1-Score**	**Support**	**Metrics**	
Prediction Label	0	0.73	0.80	0.76	6,286	**True Positive Rate:**	0.69
	1	0.78	0.69	0.73	6,227	**True Negative Rate:**	0.81
Model Accuracy				0.75	12,513	**False Positive Rate:**	0.20
Macro Average		0.75	0.75	0.75	12,513	**False Negative Rate:**	0.31
Weighted Average		0.75	0.75	0.75	12,513	**Positive Predictive Value:**	0.78

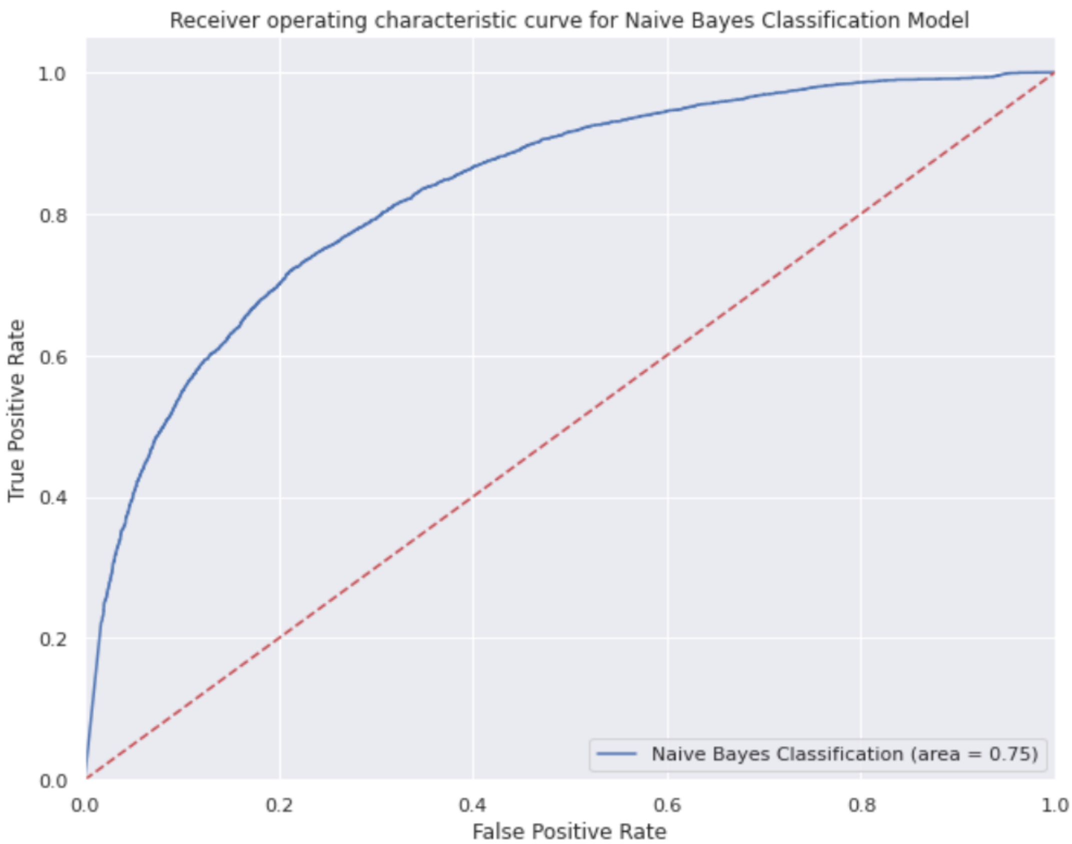


Figure 4.2 – Naive Bayes Model ROC Curve

Table 4.6 – Contingency Table for Naive Bayes Model Predicted Values

		Predicted Value	
		Unbanked (0)	**Banked (1)**
Actual Value	**Unbanked (0)**	5,021	1,265
	Banked (1)	1,921	4,306

4.1.4. eXtreme Gradient Boosting (XGBoost)

The final XGBoost model produced the best result so far with an accuracy score of 0.83. It also has rather good scores for precision, recall and f1-score.

Table 4.7 – XGBoost Model Classification Report and Performance Measure

		Measures					
		Precision	**Recall**	**F1-Score**	**Support**	**Metrics**	
Prediction Label	0	0.85	0.81	0.83	6,286	**True Positive Rate:**	0.86
	1	0.82	0.86	0.84	6,227	**True Negative Rate:**	0.81
Model Accuracy				0.83	12,513	**False Positive Rate:**	0.19
Macro Average		0.83	0.83	0.83	12,513	**False Negative Rate:**	0.14
Weighted Average		0.83	0.83	0.83	12,513	**Positive Predictive Value:**	0.82

With the use of GridSearch, we were able to perform multiple iterations of on the XGBoost Model Neural Networks. The search provided the most optimal set of parameters that would provide the best possible model. Looking at the confusion matrix, one cannot but notice the sharp contrast regions both the blue and yellow-colored regions, depicting the strengths of our True Positive and True Negative values.

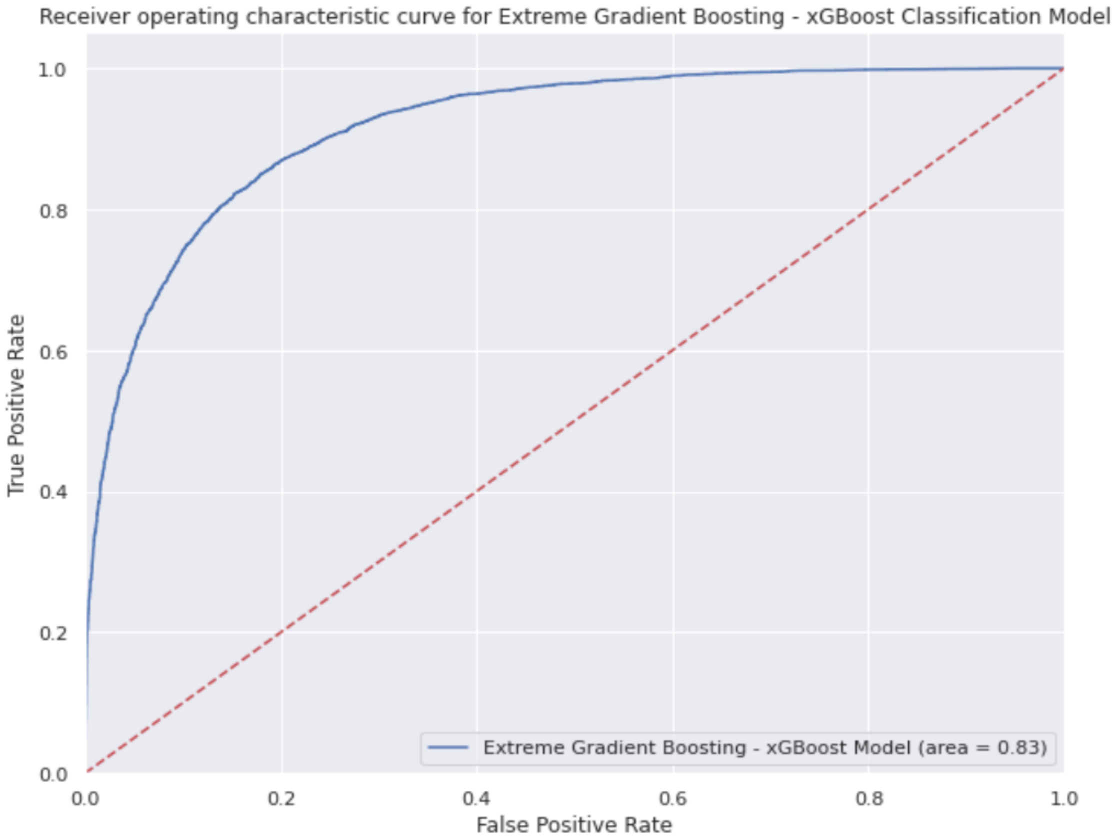


Figure 4.3 – XGBoost Model ROC Curve

Table 4.8 – Contingency Table for XGBoost Model Predicted Values

		Predicted Value	
		Unbanked (0)	**Banked (1)**
Actual Value	**Unbanked (0)**	5,101	1,185
	Banked (1)	890	5,337

4.1.5. Neural Network (NN) Model using TensorFlow

We used ***TensorFlow*** (with ***Keras***) to build our Neural Network. TensorFlow is Google's open-source software library for machine learning and artificial intelligence particularly used in building deep neural network models. Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. The NN model used a StandardScaler function from the sklearn module, to normalize the input features by setting the mean to 0 and standard deviation to 1.

Several iterations of the model were tested but the one which gave an optimal result was composed of three layers – two dense and one dropout layers with total trainable parameters of 1025.

Table 4.9 – Neural Network Model Summary

Layer (Type)	Output Shape	Number of Parameters
dense_6 (Dense)	(None, 16)	1,008
dropout_3 (Dropout)	(None, 16)	0
dense_7 (Dense)	(None, 1)	17

The model converged well after the 16th epoch with no notable change in the loss value – hence a good place to "early stop" the running iterations.

Table 4.10 – TensorFlow Neural Network Model Classification Report and Performance Measure

Measures							
		Precision	**Recall**	**F1-Score**	**Support**	**Metrics**	
Prediction Label	0	0.85	0.77	0.81	5,019	**True Positive Rate:**	0.86
	1	0.79	0.86	0.82	4,992	**True Negative Rate:**	0.77
Model Accuracy				0.82	10,011	**False Positive Rate:**	0.23
Macro Average		0.82	0.82	0.82	10,011	**False Negative Rate:**	0.14
Weighted Average		0.82	0.82	0.82	10,011	**Positive Predictive Value:**	0.79

Since the NN Model predictions are floating values ranging from 0.00 to 1.00, we set the decision threshold to 0.5. So, any predicted value equal to or more than 0.5 will be 1, else 0. With this, we achieved an accuracy of 0.82 with the model as seen in the Classification report.

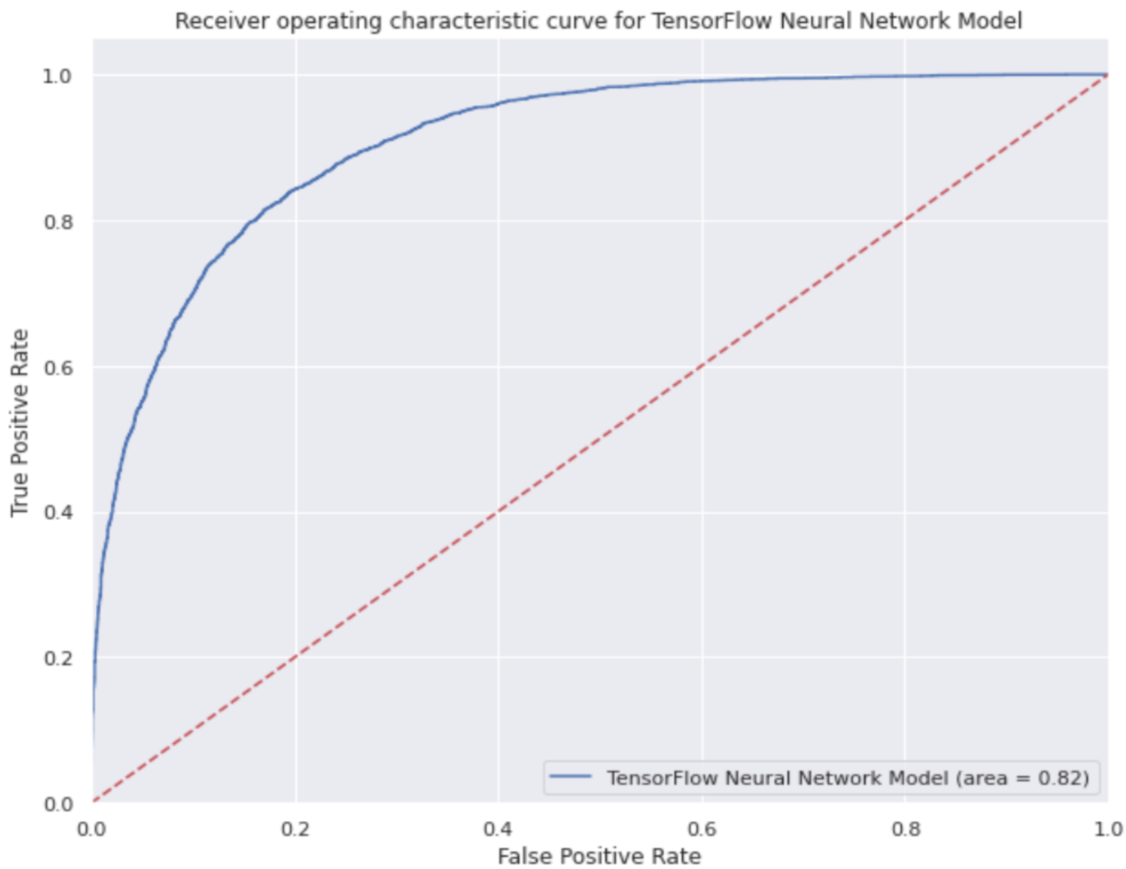


Figure 4.4 – TensorFlow Neural Network Model ROC Curve

Table 4.11 – Contingency Table for TensorFlow Neural Network Model Predicted Values

		Predicted Value	
		Unbanked (0)	**Banked (1)**
Actual Value	**Unbanked (0)**	3,886	1,133
	Banked (1)	694	4,298

4.1.6. Result Comparison

Looking at the result table, it is obvious that xGBoost Model outperforms every other model across all the measures, except for the Logistic Regression Model on the Precision score for the 1 target-class and a Recall score for the 0 target-class. While these specific values were better, it shows that the logistic regression model is not the best over-all across all other measures compared.

The Neural Network performance also came quite close, except that there is a wide difference in the Precision and Recall scores for the individual classes. This suggests that one class benefits from the model than the other – hence the imbalance. We have therefore chosen ***xGBoost as our winning model with a balance 83% accuracy score***.

Table 4.12 – Result Table of Model Comparison

	MEASURES										
	Precision			**Recall**			**F1-Score**				
	Target Class			**Target Class**			**Target Class**				
MODEL	***0***	***1***	***Overall***	***0***	***1***	***Overall***	***0***	***1***	***Overall***	**Accuracy**	**ROC_AUC**
Logistic Regression	0.74	0.91	0.82	0.93	0.67	0.80	0.82	0.77	0.80	0.80	0.83
Naive Bayes	0.72	0.78	0.75	0.80	0.69	0.75	0.76	0.73	0.75	0.75	0.75
xGBoost	0.85	0.82	0.84	0.81	0.86	0.83	0.83	0.84	0.83	0.83	0.83
Neural Networks	0.85	0.79	0.82	0.77	0.86	0.82	0.81	0.82	0.82	0.82	0.82

4.2. MODEL EXPLAINABILITY

Most machine learning models do not provide, by default, the means for a human to understand how the behavior and decision mechanism of the model. So, they are seen as a "Black-box" which just receives input, does a decent job, and churns out output [58]. An example is a Random Forest model which is inherently made up of multiple Decision Trees – which puts into consideration all the individual trees when making a final prediction. It will be impossible to fully understand such models. Because humans are always at the receiving end of these decisions, there is need to understand not only why certain conclusions have been reached, but also what adjustments can be made in form of input data, to make it better subsequently.

Model Explainability and Interpretability are two (2) similar and interchangeable terms on the surface, but there are differences that should be observed in their usage. For example, interpretability is about how easily one can see the link between a system's causes and effects. It is the ability to forecast what will happen if the input or computational settings were to be changed.

On the other hand, explainability measures how well a machine learning model's inner workings can be communicated in human words. The difference between being able to deduce the mechanics of what is occurring and being able to describe what is happening is called explainability [59]

Rosenfeld and Richardson [60] visually explained the relationship between interpretability, its methods and explainability. Interpretability is a means for providing explainability. Interpretability can be achieved using one, or a combination of more tools and methods. The most used in this toolbox is the Visualization tool, which can help explain the entire model as a global solution or as a localized interpretable element for specific outcomes.

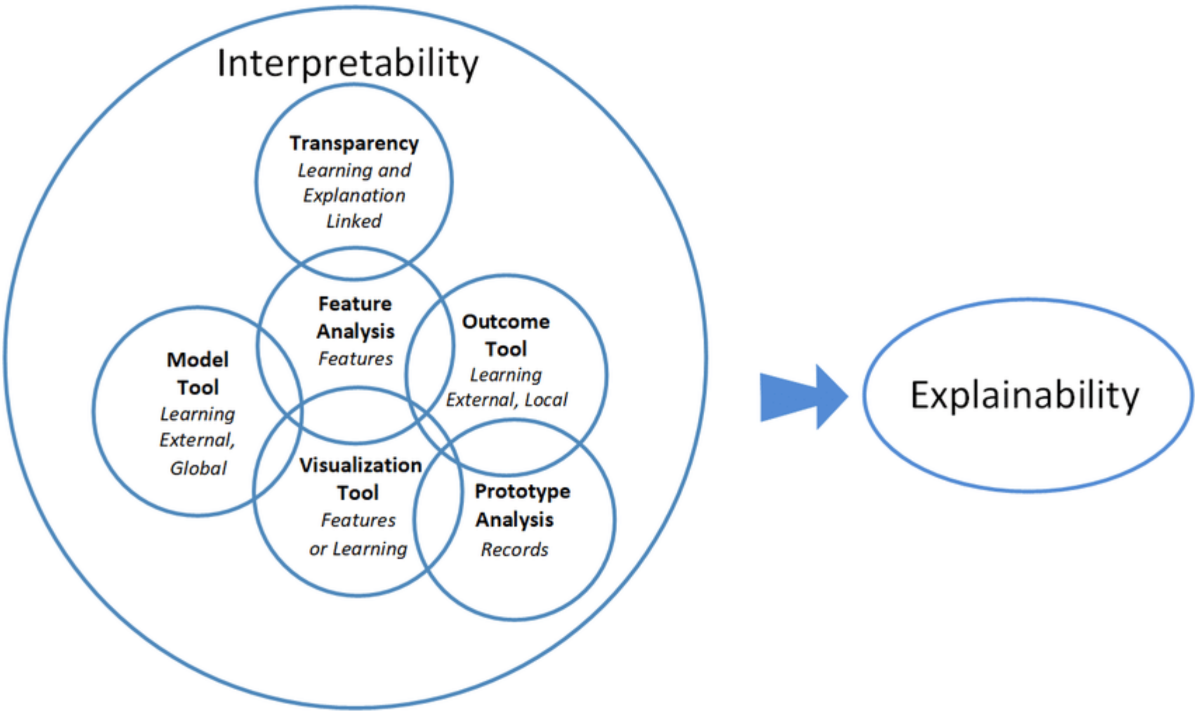


Figure 4.5 – Venn Diagram of relationship between Interpretability, its methods & Explainability [61]

There is an explicit requirement for the black box nature of AI models to be explainable, both from responsible business practitioners and from regulatory bodies. There is no industry standard on how to achieve this, but there have been some significant developments in improving the transparency and explainability of models [61].

Until very recently, a Machine Learning Model has usually been a decision black box that takes in input and give outputs, but with no human being able to understand what actually goes on inside the box to the point of explaining how certain conclusions were reached by the model. With the wide use of AI/ML in several critical domains, being able to explain our ML model is not only ethically encouraged, but also helps all stakeholders understand the strength, the weaknesses, and the reasoning behind our models' decisions.

Some ML schools of thought [59] say there is usually a trade-off between model simplicity and predictive ability. Simple models, such as calculation – or score-based models, are easy to explain and predict, but their predictive power is not great.

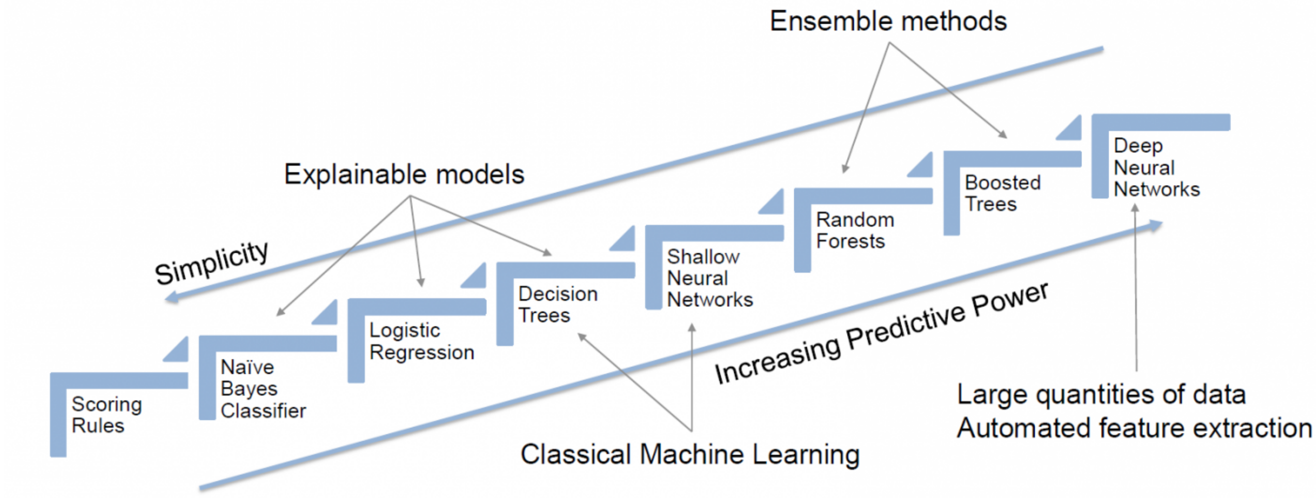


Figure 4.6 – Trade-off between model simplicity and predictive power [61]

However, complex models on the other extreme end such as Deep Neural Networks offer a high predictive power, but essentially a black-box model with opaque feature extraction methods and poor explainability. There are now several techniques used in explaining an ML model. The popular ones include Accumulated Local Effects (ALE), **Anchors,** Individual Condition Expectation (ICE) plots, Local Interpretable Model-Agnostic Explanations (LIME), Leave One Column Out (LOCO), Partial Dependence Plot (PDP) and SHapley Additive exPlanations (SHAP).

4.3. Prediction Explainability

Using the *dalex* python package [62] we were able to achieve model explainability in two key areas: Overall Variable Importance and Individual Prediction evaluation. For the latter, we examined both possible results of the prediction.

4.3.1. Overall Variable Importance

The variable importance explains in general, how each individual features contribute to the strength (or lack) through all predictions.

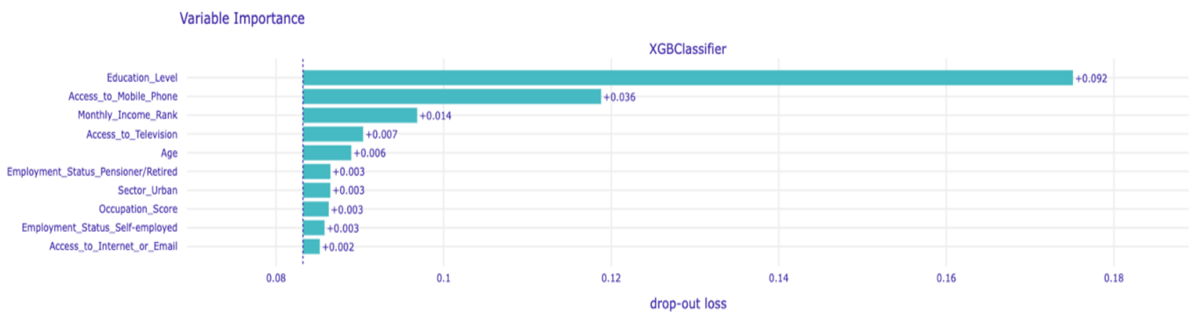


Figure 4.7 – Overall Variable Importance (Top 10)

The figure shows that `Education Level` is by far the most important contributor to the prediction. It has more than double the strength of the next best indicator – `Access_to_Mobile_Phone`. It is unsurprising to see `Access_to_Television, Age, Sector_Urban` and `Occupation_Score` also in the top 10 predictors of a Financially Includable individual.

4.3.2. Explaining the Negative Target Class Prediction

We examined the result of the individual on index 13 whose prediction was 0. From the Break Down, we can see most of the features that contribute positively to the final prediction score. Having a `Monthly_Income_Rank` of 8 is a quite high contributor. `Access_to_Television` (= 1) also certainly helps. The indicators shown in green are the positive contributors for each prediction instance and have been ordered based on levels of impact.

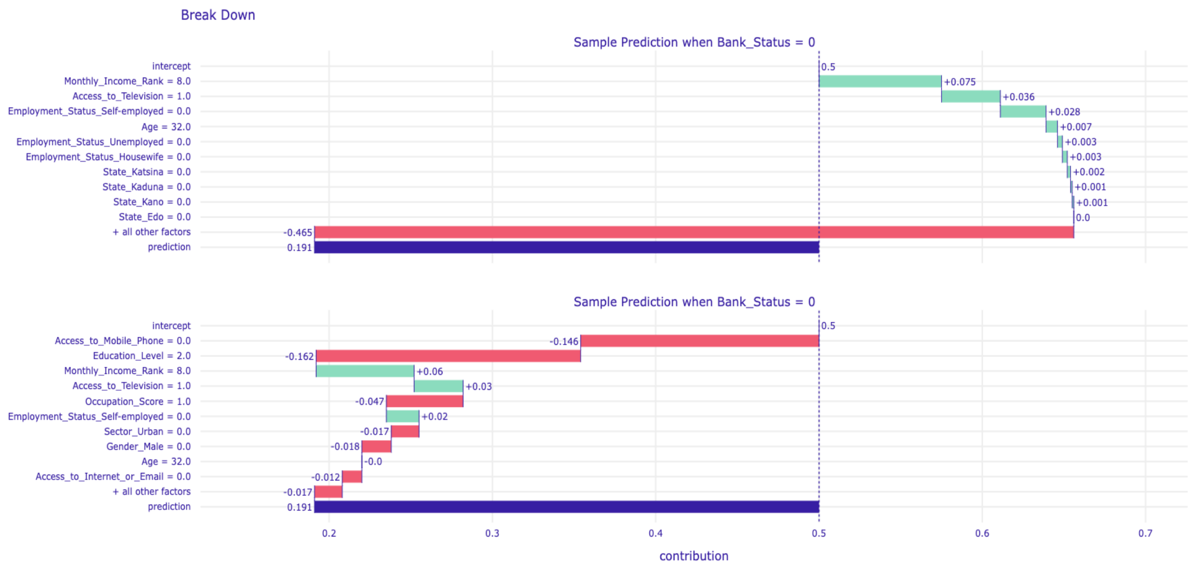


Figure 4.8 – Plots showing "Break Down" and "Contribution" of each Variable to sample Negative Prediction

However, the lower part (contribution) of the report gave more insights into why such individual is not financially included, despite the number of "green" indicators. This is because there were even stronger indicators on the *"red"* side. For example – No `Access_to_Mobile_Phone` (= 0) and a significantly low `Education_Level` (= 2), as well as a significantly low `Occupation_Score` (= 1) tend to drastically reduce the overall score of the prediction towards 0 (Actual score is: `0.19063824`). We created a *Shapley plot* to further explain these "directions of support."

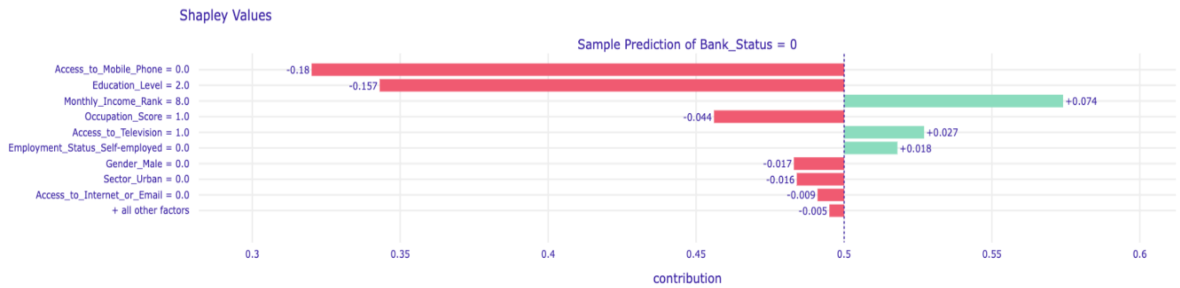


Figure 4.9 – Plot showing "Shapley Values" of each Variable to sample Negative Prediction

4.3.3. Explaining the Positive Target Class Prediction

On the contrary, the positive prediction (with target value = 1) had so many indicators contributing strongly positively to the net positive outcome. The example below was for the individual on index 50007. Obviously, the `Education_Level` of 6 the biggest impact. `Access_to_Mobile_Phone` and being in the Urban area (`Sector_Urban` = 1) are other strong contributors. It is not surprising to see the feature `State_FCT_Abuja` (= 1) supporting this strong positive prediction because *Abuja* is the country's political capital – housing the Headquarters of all major international companies.

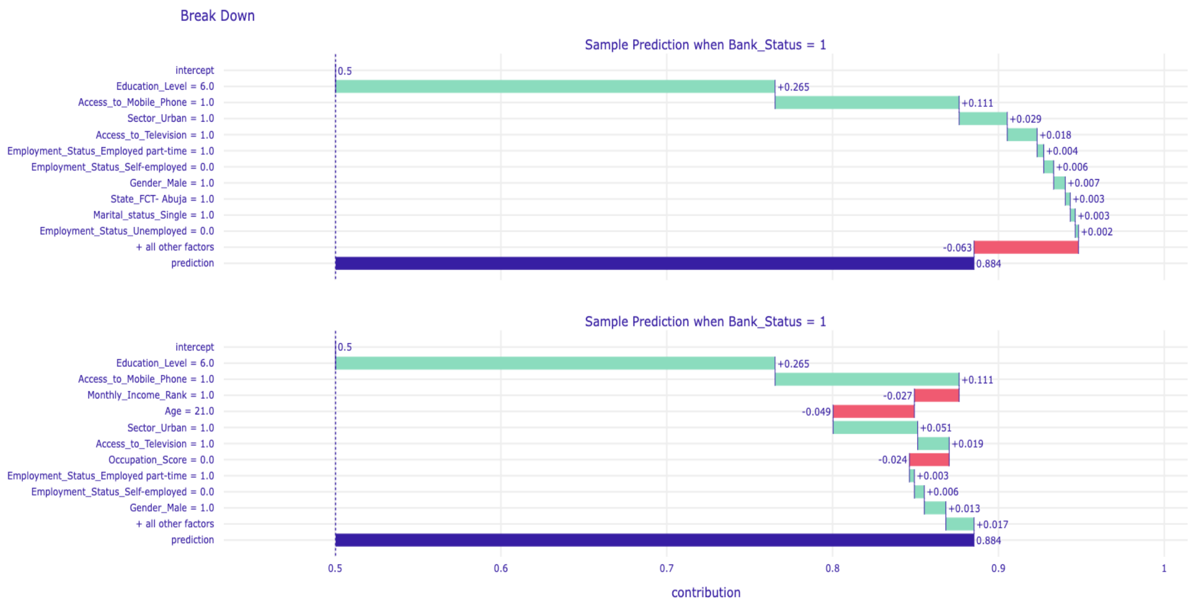


Figure 4.10 – Plots showing "Break Down" and "Contribution" of each Variable to sample Positive Prediction

Again, we used the *Shapley plot* to understand how much individually, each feature is adding to, or subtracting from the overall score. The plot below shows the top 10 most contributing features and the direction of support.

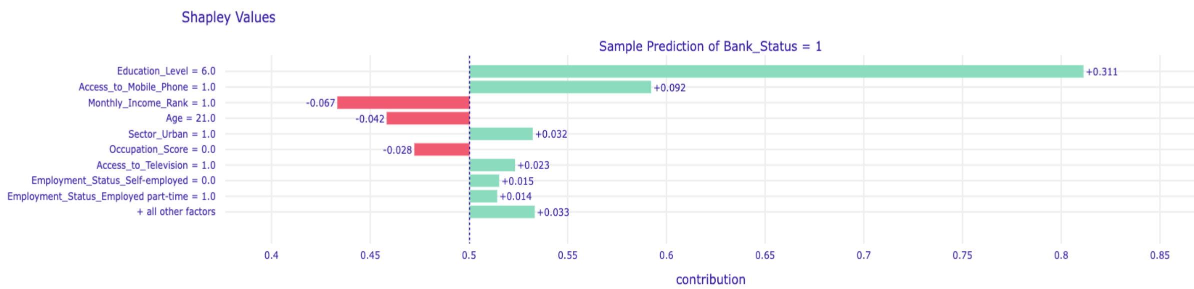


Figure 4.11 – Plot showing "Shapley Values" of each Variable to sample Positive Prediction

5. CONCLUSION

Just as the No-free-lunch theorem states that "No one algorithm is perfect for all Machine Learning Problems," we believe some algorithms would perform better than others given specific scenarios. We have used different ML models to classify this Financial Inclusion data and the best for this task amongst the tested is the eXtreme Gradient Boosting algorithm. This may however change as we have more usable data or re-engineer the data points.

In conclusion, this study, like other studies before it has revealed that – with the right targeting strategy, more adults could be included financially, and the ripple positive effect could be massive for the economy. As more people save in banks, the net loanable fund with the banks increases. This empowers the banks to do even more loans for individuals and businesses. Whilst the businesses to which the banks have loaned money expand and grow, the banks record more revenue through interests and investment partnerships. Economic outlook for the government would greatly improve as more would be earned through taxes.

All the models we have developed have shown that certain policies and infrastructures must be in place or greatly improved, for any economy to benefit from a scaled and intentional Financial Inclusion efforts. The models well identified the key drivers of Financial Inclusion and clearly suggest an unambiguous direct relationship between improved positive factors as much as it suggests the inverse relationship between negative factors and improved Financial Inclusion.

In clearer terms, key stakeholders in the societies must in improving education not only for the young, but also for the middle-aged and old people, provide jobs as well as an enabling environment for personal businesses to thrive. Make modern technology such as Television, Computers, Mobile Phones and Internet easily accessible and affordable to citizens as they have proven to be strong determinants for Financial Inclusion. Women and people living in rural areas make up a sizable number of the excluded groups. Governments and other concerned groups must work together on each specific group-based needs to address the wide gaps in exclusion compared to the other groups.

5.1. LIMITATIONS AND CHALLENGES WORKING WITH THIS DATASET

1. Too many data points - the dataset is a general-purpose dataset that could be used for multiple purposes hence the reason for about 2500 columns for each individual record.
2. The structure of the questionnaire evolved over the years since 2008 with the addition, removal, rearrangement, redesign of the questionnaire for the following year. This made consolidating data from multiple years together as one giant dataset difficult.

5.2. RECOMMENDATIONS FOR IMPROVEMENT / FURTHER WORKS

This thesis looked at only the Onboarding phase of Financial Inclusion, the dataset also include data that can be used to predict Deepening Financial Inclusion and Bank Product/Services recommendations. There should be follow-up research to address not how to deepen Financial Inclusion.

6. BIBLIOGRAPHY

[1] World Bank, “Financial Inclusion Overview,” World Bank. Accessed: Sep. 26, 2020. [Online]. Available: https://www.worldbank.org/en/topic/financialinclusion/overview#1

[2] A. Demirguc-Kunt, L. Klapper, and D. Singer, “Financial inclusion and inclusive growth: A review of recent empirical evidence,” 2017. Accessed: Sep. 27, 2020. [Online]. Available: http://econ.worldbank.org.

[3] A. Demirgüç-Kunt *et al.*, *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution*. World Bank, 2018. doi: 10.1596/978-1-4648-1259-0.

[4] ‘Femi Asu, “Nigeria’s financial inclusion falls below 2012 level -CBN,” Punch Newspapers. Accessed: Sep. 28, 2020. [Online]. Available: https://punchng.com/nigerias-financial-inclusion-falls-below-2012-level-cbn/

[5] Financial Inclusion Insights, “Nigeria · Financial Inclusion Insights from Kantar,” Finclusion.org . Accessed: Jan. 07, 2021. [Online]. Available: http://finclusion.org/country/africa/nigeria.html#dataAtAGlance

[6] EFInA, “The Landscape of Financial Inclusion and Microfinance in Nigeria,” 2018. Accessed: Jan. 07, 2021. [Online]. Available: file:///C:/Users/IMO/Desktop/Fin Descipiline/EFInA The Landscapeof Financial Inclusion and Microfinance in Nigeria.pdf

[7] EFInA, “EFInA Access to Financial Services in Nigeria 2018 Survey,” 2018.

[8] O. Nkuna, A. F. Lapukeni, P. Kaude, and G. Kabango, “The Role of Commercial Banks on Financial Inclusion in Malawi,” *Open Journal of Business and Management*, vol. 06, no. 04, pp. 813–832, Aug. 2018, doi: 10.4236/ojbm.2018.64061.

[9] M. M. Gandhi, “Role of banks in financial inclusion in India: Issues, challenges and strategies,” *International Journal of Applied Business and Economic Research*, vol. 11, no. 2, pp. 163–183, 2013, Accessed: Jan. 04, 2021. [Online]. Available: https://www.researchgate.net/publication/298642684_Role_of_banks_in_financial_inclusion_in_India_Issues_challenges_and_strategies

[10] Abraham Augustine, “The Limits of Accelerating Digital-Only Financial Inclusion,” Carnegie Endowment for International Peace. Accessed: Oct. 01, 2023. [Online]. Available: https://carnegieendowment.org/2023/07/13/limits-of-accelerating-digital-only-financial-inclusion-pub-90175

[11] British High Commission Abuja and British Deputy High Commission Lagos, “Nigeria: new data from EFInA shows financial inclusion growth,” GOV.UK. Accessed: Oct. 01, 2023. [Online]. Available: https://www.gov.uk/government/news/nigeria-new-data-from-efina-shows-financial-inclusion-growth

[12] Rashida Monguno and A. Oludamola, “UPDATE ON THE DIGITAL FINANCIAL SERVICES (DFS) AWARENESS GUIDELINES BY THE CONSUMER PROTECTION DEPARTMENT, CENTRAL BANK OF NIGERIA.,” *Finconet.org*, 2022, Accessed: Oct. 01, 2023. [Online]. Available: https://www.finconet.org/Update-DFS-Awareness-Guidelines-Nigeria.pdf

[13] Tayo Fabusiwa, “Financial Inclusion and Digital Financial Services in Nigeria.” Accessed: Oct. 01, 2023. [Online]. Available: https://www.linkedin.com/pulse/financial-inclusion-digital-services-nigeria-tayo-fabusiwa/

[14] Torsten Wezel and Jack J Ree, “Nigeria—Fostering Financial Inclusion through Digital Financial Services: Nigeria,” *International Monetary Fund*, Mar. 2023, Accessed: Oct. 01, 2023. [Online]. Available: https://www.imf.org/en/Publications/selected-issues-papers/Issues/2023/03/07/Nigeria-Fostering-Financial-Inclusion-through-Digital-Financial-Services-Nigeria-Nigeria-530633

[15] M. N. Azimi, “New insights into the impact of financial inclusion on economic growth: A global perspective,” *PLoS One*, vol. 17, no. 11, Nov. 2022, doi: 10.1371/JOURNAL.PONE.0277730.

[16] M. Bojuwon, B. R. Olaleye, and A. A. Ojebode, “Financial Inclusion and Financial Condition: The Mediating Effect of Financial Self-efficacy and Financial Literacy,” *https://doi.org/10.1177/09722629231166200*, May 2023, doi: 10.1177/09722629231166200.

[17] P. Nandru and S. Rentala, “Demand-side analysis of measuring financial inclusion: Impact on socio-economic status of primitive tribal groups (PTGs) in India,” *International Journal of Development Issues*, vol. 19, no. 1, pp. 1–24, Oct. 2019, doi: 10.1108/IJDI-06-2018-0088.

[18] Prabhakar Nandru, Byram Anand, and Satyanarayana Rentala, “Factors influencing financial inclusion using banking services,” *Journal of Contemporary Research in Management*, 2015, Accessed: Apr. 18, 2021. [Online]. Available: https://www.researchgate.net/publication/301636204_Factors_influencing_financial_inclusion_using_banking_services

[19] M. C. Simatele and L. Maciko, “Financial Inclusion in Rural South Africa: A Qualitative Approach,” *Journal of Risk and Financial Management*, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:251855969

[20] H. Chitimira and M. Ncube, “Legislative and Other Selected Challenges Affecting Financial Inclusion for the Poor and Low Income Earners in South Africa,” *J Afr Law*, vol. 64, no. 3, pp. 337–355, Oct. 2020, doi: 10.1017/S0021855320000182.

[21] M. Simatele and L. Maciko, “Financial Inclusion in Rural South Africa: A Qualitative Approach,” *Journal of Risk and Financial Management*, vol. 15, no. 9, Sep. 2022, doi: 10.3390/JRFM15090376.

[22] D. W. Ulwodi and P. W. Muriu, “Barriers of Financial Inclusion in Sub-Saharan Africa,” *Journal of economics and sustainable development*, 2017.

[23] E. Ezzahid and Z. Elouaourti, “Financial inclusion, mobile banking, informal finance and financial exclusion: micro-level evidence from Morocco,” *Int J Soc Econ*, vol. 48, no. 7, pp. 1060–1086, 2021, doi: 10.1108/IJSE-11-2020-0747.

[24] P. Khera, S. Ng, S. Ogawa, and R. Sahay, “Measuring Digital Financial Inclusion in Emerging Market and Developing Economies: A New Index,” *Asian Economic Policy Review*, vol. 17, no. 2, pp. 213–230, Jul. 2022, doi: 10.1111/AEPR.12377.

[25] B. Areas Preeti, “Financial Inclusion-Role of Indian Banks in Reaching Out to the Unbanked and Backward Areas,” *International Journal Of Business Management Available at www.ijbm.co.in ISSN*, vol. 2, no. 1, 2015, Accessed: Oct. 01, 2023. [Online]. Available: www.ijbm.co.in

[26] Q. F. Ismail, E. S. Al-Sobh, S. S. Al-Omari, T. M. Bani Yaseen, and M. A. Abdullah, “Using Machine Learning Algorithms to Predict the State of Financial Inclusion in Africa,” *2021 12th International Conference on Information and Communication Systems (ICICS)*, pp. 317–323, May 2021, doi: 10.1109/ICICS52457.2021.9464590.

[27] J. Hersh, L. Martin Rivero, and J. Leslie, “Where are the Unbanked in Belize?: Using Maching Learning Small Area Estimation to Improve Financial Inclusion Geographic Targeting,” *Where are the Unbanked in Belize?: Using Machine Learning Small Area Estimation to Improve Financial Inclusion Geographic Targeting*, Jun. 2021, doi: 10.18235/0003381.

[28] D. Mhlanga, “Industry 4.0 in Finance: The Impact of Artificial Intelligence (AI) on Digital Financial Inclusion,” *International Journal of Financial Studies 2020, Vol. 8, Page 45*, vol. 8, no. 3, p. 45, Jul. 2020, doi: 10.3390/IJFS8030045.

[29] OECD, “Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges and Implications for Policy Makers,” *Organisation for Economic Co-operation and Development*, 2021.

[30] E. B. Boukherouaa, K. AlAjmi, J. Deodoro, A. Farias, and R. Ravikumar, “Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance,” *Departmental Papers*, vol. 2021, no. 024, Oct. 2021, doi: 10.5089/9781589063952.087.A001.

[31] O. H. Fares, I. Butt, and S. H. M. Lee, “Utilization of artificial intelligence in the banking sector: a systematic literature review,” *Journal of Financial Services Marketing*, pp. 1–18, Aug. 2022, doi: 10.1057/S41264-022-00176-7/TABLES/4.

[32] J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, “Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis,” *J Behav Exp Finance*, vol. 32, p. 100577, Dec. 2021, doi: 10.1016/J.JBEF.2021.100577.

[33] M. Bazarbash, “Fintech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk,” *FinPlanRN: Other Finance Planning Fundamentals (Topic)*, vol. 2019, no. 109, p. 1, May 2019, doi: 10.5089/9781498314428.001.

[34] D. Mhlanga, “Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment,” *International Journal of Financial Studies*, vol. 9, no. 3, Sep. 2021, doi: 10.3390/IJFS9030039.

[35] S. Agarwal, S. Alok, P. Ghosh, and S. Gupta, “Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech,” *FEN: Behavioral Finance (Topic)*, Jan. 2019, doi: 10.2139/SSRN.3507827.

[36] A. Kumar, S. Sharma, and M. Mahdavi, “Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review,” *Risks*, vol. 9, no. 11, Nov. 2021, doi: 10.3390/RISKS9110192.

[37] S. Natile, “Digital Finance Inclusion and the Mobile Money,” *Historical Social Research*, 2020, doi: 10.12759/HSR.45.2020.3.74-94.

[38] S. Centellegher *et al.*, “Mobile Money: Understanding and Predicting its Adoption and Use in a Developing Economy,” vol. 18, Dec. 2018, Accessed: Oct. 01, 2023. [Online]. Available: https://arxiv.org/abs/1812.03289v1

[39] Ottavio Del Torso, “Mobile Payment Systems and Financial Inclusion: the case of M-Pesa,” 2019. doi: 10.13140/RG.2.2.10605.13280.

[40] J. N. Omanga and J. K. Dreyer, “Innovation and Financial Inclusion in Kenya: A Case Study of M-PESA,” pp. 207–227, Apr. 2017, doi: 10.4018/978-1-5225-2319-2.CH008.

[41] L. Van Hove and A. Dubus, “M-PESA and Financial Inclusion in Kenya: Of Paying Comes Saving?,” *Sustainability 2019, Vol. 11, Page 568*, vol. 11, no. 3, p. 568, Jan. 2019, doi: 10.3390/SU11030568.

[42] T. Suri and W. Jack, “The long-run poverty and gender impacts of mobile money,” *Science (1979)*, vol. 354, no. 6317, pp. 1288–1292, Dec. 2016, doi: 10.1126/SCIENCE.AAH5309.

[43] World Bank, “The World Bank In Nigeria,” The World Bank Research Observer. Accessed: Sep. 28, 2020. [Online]. Available: https://www.worldbank.org/en/country/nigeria/overview

[44] Organisation for Economic Co-operation and Development (OECD), “West Africa Gateway: Nigeria.” Accessed: Sep. 28, 2020. [Online]. Available: https://www.oecd.org/swac/publications/Nigeria_e-version_en_light.pdf

[45] Central Bank of Nigeria, “Central Bank of Nigeria | About CBN,” *Central Bank of Nigeria*. pp. 1–6, 2011. Accessed: Oct. 11, 2020. [Online]. Available: https://www.cbn.gov.ng/AboutCBN/

[46] I. Obienusi and E. A. Obienusi, “Banking Reforms and the Nigerian Economy, 1990-2007,” 2015. Accessed: Sep. 28, 2020. [Online]. Available: www.iiste.org

[47] Enhancing Financial Innovation and Access, “Access to Financial Services in Nigeria survey – EFInA.” Accessed: Aug. 22, 2021. [Online]. Available: https://efina.org.ng/our-work/research/access/

[48] Shelvi Garg, “How to Deal with Categorical Data for Machine Learning - KDnuggets,” *KD Nuggets*, May 2021. Accessed: Dec. 20, 2021. [Online]. Available: https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html

[49] Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2019. Accessed: Jun. 19, 2021. [Online]. Available: https://christophm.github.io/interpretable-ml-book/logistic.html

[50] A. Urso, A. Fiannaca, M. La Rosa, V. Ravì, and R. Rizzo, “Data mining: Prediction methods,” in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 413–430. doi: 10.1016/B978-0-12-809633-8.20462-7.

[51] Saed Sayed, “An Introduction to Data Science: Data Mining Map,” 2017, Accessed: Jun. 19, 2021. [Online]. Available: https://saedsayad.com/data_mining_map.htm

[52] P. Ambika, "Machine learning and deep learning algorithms on the Industrial Internet of Things (IIoT)," in *Advances in Computers*, vol. 117, no. 1, Academic Press Inc., 2020, pp. 321–338. doi: 10.1016/bs.adcom.2019.10.007.

[53] Kan Nishida, "Introduction to Extreme Gradient Boosting in Exploratory," *Exploratory - Learn Data Science*, Mar. 09, 2017. Accessed: Jun. 20, 2021. [Online]. Available: https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7

[54] Saul Dobilas, "XGBoost: Extreme Gradient Boosting — How to Improve on Regular Gradient Boosting," *Towards Data Science*, Mar. 21, 2021. Accessed: Jun. 20, 2021. [Online]. Available: https://towardsdatascience.com/xgboost-extreme-gradient-boosting-how-to-improve-on-regular-gradient-boosting-5c6acf66c70a

[55] Sakshi Tiwari, "Activation functions in Neural Networks," GeeksforGeeks. Accessed: Jun. 20, 2021. [Online]. Available: https://www.geeksforgeeks.org/activation-functions-neural-networks/

[56] Aniruddha Bhandari, "Confusion Matrix for Machine Learning," *Analytics Vidhya*, Apr. 17, 2020. Accessed: Dec. 05, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

[57] "Area under the ROC Curve (AUC) - Evispot." Accessed: Dec. 15, 2021. [Online]. Available: https://evispot.ai/area-under-the-roc-curve-auc/

[58] Conor O'Sullivan, "Interpretable vs Explainable Machine Learning," Towards Data Science. Accessed: Jan. 03, 2022. [Online]. Available: https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48

[59] Richard Gall, "Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI," KDnuggets. Accessed: Jan. 03, 2022. [Online]. Available: https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html

[60] A. Rosenfeld and A. Richardson, *Explainability in Human-Agent Systems*. 2019.

[61] Opti-Num Solutions, "Explainability for machine learning models in MATLAB - Opti-Num Solutions." Accessed: Jan. 02, 2022. [Online]. Available: https://optinum.co.za/explainability-for-machine-learning-models-in-matlab/

[62] "dalex: Responsible Machine Learning in Python." Accessed: Jan. 24, 2022. [Online]. Available: https://dalex.drwhy.ai/python/

7. APPENDIX

7.1. RAW DATASET

https://liveeduisegiunl-my.sharepoint.com/personal/m20170002_novaims_unl_pt/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fm20170002%5Fnovaims%5Funl%5Fpt%2FDocuments%2FMicrosoft%20Drive%2FMaster%20Thesis%20Raw%20Data&ga=1

7.2. FINAL DATASET AND PUBLISHED ML MODEL CODES

https://github.com/SadeeqAkintola/MasterThesis

7.3. FEATURE ENGINEERING SQL QUERIES

```
SELECT

2008  Survey_Year,
#CAST(State AS INT64 ) State,
CASE
    WHEN State = 1.0 then 'Abia'
    WHEN State = 2.0 then 'Adamawa'
    WHEN State = 3.0 then 'Akwa Ibom'
    WHEN State = 4.0 then 'Anambra'
    WHEN State = 5.0 then 'Bauchi'
    WHEN State = 6.0 then 'Bayelsa'
    WHEN State = 7.0 then 'Benue'
    WHEN State = 8.0 then 'Borno'
    WHEN State = 9.0 then 'Cross Rivers'
    WHEN State = 10.0 then 'Delta'
    WHEN State = 11.0 then 'Ebonyi'
    WHEN State = 12.0 then 'Edo'
    WHEN State = 13.0 then 'Ekiti'
    WHEN State = 14.0 then 'Enugu'
    WHEN State = 15.0 then 'Gombe'
    WHEN State = 16.0 then 'Imo'
    WHEN State = 17.0 then 'Jigawa'
    WHEN State = 18.0 then 'Kaduna'
    WHEN State = 19.0 then 'Kano'
    WHEN State = 20.0 then 'Katsina'
    WHEN State = 21.0 then 'Kebbi'
    WHEN State = 22.0 then 'Kogi'
    WHEN State = 23.0 then 'Kwara'
    WHEN State = 24.0 then 'Lagos'
    WHEN State = 25.0 then 'Nasarawa'
    WHEN State = 26.0 then 'Niger'
    WHEN State = 27.0 then 'Ogun'
    WHEN State = 28.0 then 'Ondo'
    WHEN State = 29.0 then 'Osun'
```

```
    WHEN State = 30.0 then 'Oyo'
    WHEN State = 31.0 then 'Plateau'
    WHEN State = 32.0 then 'Rivers'
    WHEN State = 33.0 then 'Sokoto'
    WHEN State = 34.0 then 'Taraba'
    WHEN State = 35.0 then 'Yobe'
    WHEN State = 36.0 then 'Zamfara'
    WHEN State = 37.0 then 'FCT- Abuja'
ELSE NULL END State,

CASE WHEN Sector = 1.0 then 'Urban' ELSE 'Rural' END Sector,
CASE WHEN Residential_Area = 1.0 then 3 WHEN Residential_Area = 2.0 then 2 WHEN
Residential_Area = 3.0 then 1  ELSE 0 END Residential_Area_Density,
CASE WHEN (Gender_of_Selected_Respondent = 1.0) THEN 'Male'  ELSE 'Female' END
Gender,
CAST(Age_of_Selected_Respondent AS INT64 ) Age,

CASE
    WHEN Q57__Best_Description_of_Marital_Status = 1.0 then 'Married (Monogamy)'
    WHEN Q57__Best_Description_of_Marital_Status = 2.0 then 'Married (Polygamy)'
    WHEN Q57__Best_Description_of_Marital_Status = 3.0 then 'Co-Habiting'
    WHEN Q57__Best_Description_of_Marital_Status = 4.0 then 'Divorced'
    WHEN Q57__Best_Description_of_Marital_Status = 5.0 then 'Separated'
    WHEN Q57__Best_Description_of_Marital_Status = 6.0 then 'Widowed'
    WHEN Q57__Best_Description_of_Marital_Status = 7.0 then 'Single'

ELSE NULL END Marital_status,

 CASE WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (1.0, 2.0)) THEN
'Employed full-time'
                WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (3.0,
4.0)) THEN 'Employed part-time'
                WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (5.0,
6.0)) THEN 'Unemployed'
                WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (7.0))
THEN 'Self-employed'
                WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (8.0))
THEN 'Student'
                WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (9.0))
THEN 'Pensioner/Retired'
                WHEN (Q2A_Best_Description_of_Current_Employment_Status IN (10.0))
THEN 'Housewife'
ELSE 'Others' END Employment_Status,

CASE WHEN (Q63a_Personal_Monthly_Income >= 13.0) THEN NULL  ELSE
CAST(Q63a_Personal_Monthly_Income AS INT64 ) END Monthly_Income_Rank,

CASE WHEN (Q58__Best_Description_of_Literacy_Level IN (1.0)) THEN 1
                WHEN (Q58__Best_Description_of_Literacy_Level IN (2.0)) THEN 2
                WHEN (Q58__Best_Description_of_Literacy_Level IN (3.0, 13.0, 15.0))
THEN 3
                WHEN (Q58__Best_Description_of_Literacy_Level IN (4.0, 5.0, 14.0,
16.0)) THEN 4
                WHEN (Q58__Best_Description_of_Literacy_Level IN (6.0, 7.0)) THEN 5
```

```
            WHEN (Q58__Best_Description_of_Literacy_Level IN (8.0, 10.0)) THEN 6
            WHEN (Q58__Best_Description_of_Literacy_Level IN (11.0)) THEN 7
            WHEN (Q58__Best_Description_of_Literacy_Level IN (12.0)) THEN 8
         ELSE 0 END Education_Level,

CASE WHEN (OCCUPATION IN (1.0, 2.0)) THEN 5
             WHEN (OCCUPATION IN (3.0, 5.0)) THEN 4
             WHEN (OCCUPATION IN (4.0)) THEN 3
             WHEN (OCCUPATION IN (6.0)) THEN 2
             WHEN (OCCUPATION IN (7.0, 8.0)) THEN 1
         ELSE 0 END Occupation_Score,

CASE
        WHEN
(CAST(REPLACE(CAST(q43c_Prepaid_mobile_Phone_Q43C_Personal_and_Household_Ownership_
of_Products_Services AS STRING), "-99.99", "0")AS INT64 )
        +
CAST(REPLACE(CAST(q43c_Contract_mobile_phone_Q43C_Personal_and_Household_Ownership_
of_Products_Services AS STRING), "-99.99", "0")AS INT64 )  = 0) then 0
        ELSE 1 End Access_to_Mobile_Phone,

CASE WHEN

(CAST(REPLACE(CAST(q43c_Personal_computer_at_home_Q43C_Personal_and_Household_Owner
ship_of_Products_Services AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43c__Personal_computer_at_work_Q43C_Personal_and_Household_Owner
ship_of_Products_Services AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43a_Personal_computer_at_home_Q43A_Personal_Access_to_Products_S
ervices AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43a__Personal_computer_at_work_Q43A_Personal_Access_to_Products_
Services AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43b_Personal_computer_at_home_Q43B_Regular_Usage_of_Products_Ser
vices AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43b__Personal_computer_at_work_Q43B_Regular_Usage_of_Products_Se
rvices AS STRING), "-99.99", "0")AS INT64 )
        = 0) then 0
        ELSE 1 End Access_to_Personal_Computer,

        CASE WHEN

(CAST(REPLACE(CAST(q43a_Internet_E_mail_at_home_Q43A_Personal_Access_to_Products_Se
rvices AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43a__Internet_E_mail_at_work_Q43A_Personal_Access_to_Products_Se
rvices AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43a___Internet_E_mail_at_business_centres_Q43A_Personal_Access_t
o_Products_Services AS STRING), "-99.99", "0")AS INT64 ) +
```

```
CAST(REPLACE(CAST(q43b_Internet_E_mail_at_home_Q43B_Regular_Usage_of_Products_Servi
ces AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43b__Internet_E_mail_at_work_Q43B_Regular_Usage_of_Products_Serv
ices AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43b___Internet_E_mail_at_business_centres_Q43B_Regular_Usage_of_
Products_Services AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43c_Internet_E_mail_at_home_Q43C_Personal_and_Household_Ownershi
p_of_Products_Services AS STRING), "-99.99", "0")AS INT64 )
        = 0) then 0
        ELSE 1 End Access_to_Internet_or_Email,
        CASE WHEN

(CAST(REPLACE(CAST(q43a_Television_Q43A_Personal_Access_to_Products_Services AS
STRING), "-99.99", "0")AS INT64 ) +
        CAST(REPLACE(CAST(q43b_Television_Q43B_Regular_Usage_of_Products_Services
AS STRING), "-99.99", "0")AS INT64 ) +

CAST(REPLACE(CAST(q43c_Television_Q43C_Personal_and_Household_Ownership_of_Products
_Services AS STRING), "-99.99", "0")AS INT64 )
        = 0) then 0
        ELSE 1 End Access_to_Television

, CASE WHEN Q7__Bank_Status = 1.0 THEN 1 ELSE 0 END Bank_Status

FROM `masters-thesis-316415.efina_survey_data.efina2008`


UNION ALL


SELECT
# * ,
2010  Survey_Year,
State,

Sector__Urban_Rural Sector,
CASE WHEN RESIDENTIAL_AREA_OBSERVE_= 'High Density' then 1 WHEN
RESIDENTIAL_AREA_OBSERVE_= 'Medium Density' then  2 WHEN RESIDENTIAL_AREA_OBSERVE_=
'Low Density' then  3  ELSE 0 END Residential_Area_Density,
Respondent_gender Gender,
CAST(Respondent_age AS INT64 ) Age,
DM1__Marital_status Marital_status,
C2___Employment_status Employment_Status,

CASE WHEN  DM7__Personal_monthly_income = '250 or less' then 2
     WHEN  DM7__Personal_monthly_income = '251 - 1,000' then 3
     WHEN  DM7__Personal_monthly_income = '1,001 - 3,000' then 4
     WHEN  DM7__Personal_monthly_income = '3,001 - 6,000' then 5
     WHEN  DM7__Personal_monthly_income = '6,001 - 13,000' then 6
     WHEN  DM7__Personal_monthly_income = '13,001 - 20,000' then 7
```

```
     WHEN  DM7__Personal_monthly_income = '20,001 - 40,000' then 8
     WHEN  DM7__Personal_monthly_income = '40,001 - 70,000' then 9
     WHEN  DM7__Personal_monthly_income = '70,001 - 100,000' then 10
     WHEN  DM7__Personal_monthly_income = '100,001 - 200,000' then 11
     WHEN  DM7__Personal_monthly_income = 'Above 200,000' then 12
     WHEN  DM7__Personal_monthly_income = 'No income' then 1
ELSE NULL END Monthly_Income_Rank,

CASE WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN ('Primary Incomplete')) THEN 1
                WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN ('Primary
Complete')) THEN 2
                WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN ('Secondary
Incomplete')) THEN 3
                WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN ('Secondary
Complete','Islamic College Degree', 'Vocational Training/Technical college')) THEN
4
                WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN
('University/Polytechnic Incomplete OND')) THEN 5
                 WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN
('*University/Polytechnic Complete HND')) THEN 6
                WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN ('Post University
Incomplete')) THEN 7
                 WHEN (RESPONDENT_S_HIGHEST_EDUCATIONAL_LEVEL IN ('*Post University
Complete')) THEN 8
            ELSE 0 END Education_Level,

CASE WHEN (RESPONDENT_S_OCCUPATION IN ('Senior Management/Senior Admin')) THEN 5
                WHEN (RESPONDENT_S_OCCUPATION IN ('Professional e.g Doctor, Lawyer,
Engineers')) THEN 4
                WHEN (RESPONDENT_S_OCCUPATION IN ('Manager')) THEN 3
                WHEN (RESPONDENT_S_OCCUPATION IN ('Skilled Workers',
'Retired/Pensioner')) THEN 2
                WHEN (RESPONDENT_S_OCCUPATION IN ('Housewife', 'Unskilled Worker',
'Student')) THEN 1
            ELSE 0 END Occupation_Score,


# GF1__Interest_in_financial_matters Financial_Matters_Interest_Level,

CASE WHEN TE1___Do_you_own_a_mobile_phone_ =true then 1 ELSE 0 END
Access_to_Mobile_Phone,
CASE WHEN ((qte6_07__Access_to___Personal_computer_at_work  = 'Personal computer at
work')or (qte6_06__Access_to___Personal_computer_at_home = 'Personal computer at
home')) THEN 1 ELSE 0 END Access_to_Personal_Computer,
CASE WHEN ( qte6_08__Access_to___Internet_E_mail_at_home != '#NULL!'
    OR qte6_09__Access_to___Internet_E_mail_at_work != '#NULL!'
    OR   qte6_10__Access_to___Internet_E_mail_on_mobile_phone != '#NULL!') Then 1
    ELSE 0 END Access_to_Internet_or_Email,
 CASE
       WHEN qte6_12__Access_to___Television = 'Television' Then 1
       ELSE 0 END Access_to_Television,
CASE WHEN Banked_status = 'Have/use a bank product' THEN 1 ELSE 0 END Banked_Status
FROM `masters-thesis-316415.efina_survey_data.efina2010`
```

```
UNION ALL
SELECT
# * ,

2014 Survey_Year,
 CASE
    WHEN State = 1.0 then 'Abia'
    WHEN State = 2.0 then 'Adamawa'
    WHEN State = 3.0 then 'Akwa Ibom'
    WHEN State = 4.0 then 'Anambra'
    WHEN State = 5.0 then 'Bauchi'
    WHEN State = 6.0 then 'Bayelsa'
    WHEN State = 7.0 then 'Benue'
    WHEN State = 8.0 then 'Borno'
    WHEN State = 9.0 then 'Cross Rivers'
    WHEN State = 10.0 then 'Delta'
    WHEN State = 11.0 then 'Ebonyi'
    WHEN State = 12.0 then 'Edo'
    WHEN State = 13.0 then 'Ekiti'
    WHEN State = 14.0 then 'Enugu'
    WHEN State = 15.0 then 'Gombe'
    WHEN State = 16.0 then 'Imo'
    WHEN State = 17.0 then 'Jigawa'
    WHEN State = 18.0 then 'Kaduna'
    WHEN State = 19.0 then 'Kano'
    WHEN State = 20.0 then 'Katsina'
    WHEN State = 21.0 then 'Kebbi'
    WHEN State = 22.0 then 'Kogi'
    WHEN State = 23.0 then 'Kwara'
    WHEN State = 24.0 then 'Lagos'
    WHEN State = 25.0 then 'Nasarawa'
    WHEN State = 26.0 then 'Niger'
    WHEN State = 27.0 then 'Ogun'
    WHEN State = 28.0 then 'Ondo'
    WHEN State = 29.0 then 'Osun'
    WHEN State = 30.0 then 'Oyo'
    WHEN State = 31.0 then 'Plateau'
    WHEN State = 32.0 then 'Rivers'
    WHEN State = 33.0 then 'Sokoto'
    WHEN State = 34.0 then 'Taraba'
    WHEN State = 35.0 then 'Yobe'
    WHEN State = 36.0 then 'Zamfara'
    WHEN State = 37.0 then 'FCT- Abuja'
ELSE NULL END State,
Sector,
CASE WHEN SC5= 1.0 then 3 WHEN SC5= 2.0 then 2 WHEN SC5= 3.0 then 1  ELSE 0 END
Residential_Area_Density,
CASE WHEN DM5 = 1.0 Then 'Male'
             ELSE 'Female' END Gender,
CAST(DM6 AS INT64 ) Age,
CASE
    WHEN DM1 = 1.0 then 'Married (Monogamy)'
    WHEN DM1 = 2.0 then 'Married (Polygamy)'
    WHEN DM1 = 3.0 then 'Co-Habiting'
```

```
    WHEN DM1 = 4.0 then 'Divorced'
    WHEN DM1 = 5.0 then 'Separated'
    WHEN DM1 = 6.0 then 'Widowed'
    WHEN DM1 = 7.0 then 'Single'
ELSE NULL END Marital_status,

CASE WHEN (DM9 IN (1.0)) THEN  'Employed full-time'
                WHEN (DM9 IN (2.0)) THEN 'Employed part-time'
                WHEN (DM9 IN (3.0)) THEN 'Self-employed'
                WHEN (DM9 IN (4.0)) THEN 'Unemployed'
                WHEN (DM9 IN (5.0)) THEN 'Student'
                WHEN (DM9 IN (6.0)) THEN 'NYSC'
                WHEN (DM9 IN (7.0)) THEN 'Pensioner/Retired'
                WHEN (DM9 IN (8.0)) THEN 'Housewife'
ELSE 'Others' END Employment_Status,

CASE WHEN (DM3 >= 13.0) THEN NULL  ELSE CAST(DM3 AS INT64 ) END
Monthly_Income_Rank,

CASE            WHEN (DM7 IN (1.0)) THEN 0
                WHEN (DM7 IN (2.0)) THEN 1
                WHEN (DM7 IN (3.0)) THEN 2
                WHEN (DM7 IN (4.0)) THEN 3
                WHEN (DM7 IN (5.0)) THEN 4
                WHEN (DM7 IN (6.0)) THEN 5
                 WHEN (DM7 IN (7.0)) THEN 6
                WHEN (DM7 IN (8.0)) THEN 7
                 WHEN (DM7 IN (9.0)) THEN 8
            ELSE 0 END Education_Level,

CASE WHEN (SC7 IN (1.0)) THEN 5
                WHEN (SC7 IN (2.0, 4.0)) THEN 4
                WHEN (SC7 IN (3.0)) THEN 3
                WHEN (SC7 IN (5.0, 98.0)) THEN 2
                WHEN (SC7 IN (6.0, 7.0 )) THEN 1
            ELSE 0 END Occupation_Score,

CASE WHEN TE3 = 1.0 Then 1 ELSE 0 END Access_to_Mobile_Phone,
CASE WHEN (SC1_18 = 2.0 AND SC1_19 = 2.0) THEN 0 ELSE 1 END
Access_to_Personal_Computer,
CASE WHEN TE6 = 1.0 or TE8 = 1.0 Then 1 ELSE 0 END Access_to_Internet_or_Email,
CASE WHEN (SC1_5 = 2.0 AND  SC1_11 = 2.0) THEN 0 ELSE 1 END Access_to_Television,

CASE WHEN BA3 = 1.0 THEN 1 ELSE 0 END Banked_Status

FROM `masters-thesis-316415.efina_survey_data.efina2014`



UNION ALL
SELECT
2018 Survey_Year,
 CASE
    WHEN _10__State = 1.0 then 'Abia'
```

```
    WHEN _10__State = 2.0 then 'Adamawa'
    WHEN _10__State = 3.0 then 'Akwa Ibom'
    WHEN _10__State = 4.0 then 'Anambra'
    WHEN _10__State = 5.0 then 'Bauchi'
    WHEN _10__State = 6.0 then 'Bayelsa'
    WHEN _10__State = 7.0 then 'Benue'
    WHEN _10__State = 8.0 then 'Borno'
    WHEN _10__State = 9.0 then 'Cross Rivers'
    WHEN _10__State = 10.0 then 'Delta'
    WHEN _10__State = 11.0 then 'Ebonyi'
    WHEN _10__State = 12.0 then 'Edo'
    WHEN _10__State = 13.0 then 'Ekiti'
    WHEN _10__State = 14.0 then 'Enugu'
    WHEN _10__State = 15.0 then 'Gombe'
    WHEN _10__State = 16.0 then 'Imo'
    WHEN _10__State = 17.0 then 'Jigawa'
    WHEN _10__State = 18.0 then 'Kaduna'
    WHEN _10__State = 19.0 then 'Kano'
    WHEN _10__State = 20.0 then 'Katsina'
    WHEN _10__State = 21.0 then 'Kebbi'
    WHEN _10__State = 22.0 then 'Kogi'
    WHEN _10__State = 23.0 then 'Kwara'
    WHEN _10__State = 24.0 then 'Lagos'
    WHEN _10__State = 25.0 then 'Nasarawa'
    WHEN _10__State = 26.0 then 'Niger'
    WHEN _10__State = 27.0 then 'Ogun'
    WHEN _10__State = 28.0 then 'Ondo'
    WHEN _10__State = 29.0 then 'Osun'
    WHEN _10__State = 30.0 then 'Oyo'
    WHEN _10__State = 31.0 then 'Plateau'
    WHEN _10__State = 32.0 then 'Rivers'
    WHEN _10__State = 33.0 then 'Sokoto'
    WHEN _10__State = 34.0 then 'Taraba'
    WHEN _10__State = 35.0 then 'Yobe'
    WHEN _10__State = 36.0 then 'Zamfara'
    WHEN _10__State = 37.0 then 'FCT- Abuja'
ELSE NULL END State,
CASE WHEN _11__Sector = 1 then 'Urban' ELSE 'Rural' END Sector,
NULL Residential_Area_Density,
CASE WHEN (E6__Respondent_s_gender = 1) THEN 'Male'  ELSE 'Female' END Gender,
CAST(E7__Respondent_s_exact_age AS INT64 ) Age,
CASE
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 1.0 then
'Married (Monogamy)'
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 2.0 then
'Married (Polygamy)'
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 3.0 then
'Co-Habiting'
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 4.0 then
'Divorced'
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 5.0 then
'Separated'
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 6.0 then
'Widowed'
```

```
    WHEN E4__Which_of_the_following_best_describes_your_marital_status_ = 7.0 then
'Single'

ELSE NULL END Marital_status,
NULL Employment_Status,
CASE WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual = 1 then 5
    WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual = 2 then 6
    WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual = 3 then 7
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 4 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 1000 then 3
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 1001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 3000 then 4
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 3001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 6000 then 5
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 6001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 13000 then 6
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 13001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 20000 then 7

     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 20001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 40000 then 8
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 40001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 70000 then 9
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 70001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 100000 then 10
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 100001 and
IE1a__M6a__Total_Personal_Monthly_Income_actual <= 200000 then 11
     WHEN  IE1a__M6a__Total_Personal_Monthly_Income_actual > 200000  then 12
ELSE 1 END Monthly_Income_Rank,

CASE WHEN (E8__Respondent_s_level_of_education IN (1, 10)) THEN 1
                WHEN (E8__Respondent_s_level_of_education IN (2)) THEN 2
                WHEN (E8__Respondent_s_level_of_education IN (3)) THEN 3
                WHEN (E8__Respondent_s_level_of_education IN (4, 10)) THEN 4
                WHEN (E8__Respondent_s_level_of_education IN (5)) THEN 5
                 WHEN (E8__Respondent_s_level_of_education IN (6)) THEN 6
                WHEN (E8__Respondent_s_level_of_education IN (7)) THEN 7
                 WHEN (E8__Respondent_s_level_of_education IN (8)) THEN 8
            ELSE 0 END Education_Level,

NULL Occupation_Score,
CASE WHEN TE1__Do_you_own_a_mobile_phone_ = 1 then 1 ELSE 0 END
Access_to_Mobile_Phone,
CASE WHEN ((D3__Household_asset_ownership___Computer__laptop_  = 1 ) or
(D3__Household_asset_ownership___Computer__desktop_ = 1)) THEN 1 ELSE 0 END
Access_to_Personal_Computer,
CASE WHEN F1b__Preferred_sources_of_information___Internet = 1 then 1 ELSE 0 END
Access_to_Internet_or_Email,
CASE WHEN ((D3__Household_asset_ownership___Colour_TV  = 1 ) or
(D3__Household_asset_ownership___Black___white_TV = 1)) THEN 1 ELSE 0 END
Access_to_Television, Banked Banked_Status

FROM `masters-thesis-316415.efina_survey_data.efina2018`
```

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa