An Efficient Approach for Predicting the state of Financial Inclusion by Addressing Class Imbalance using the Oversampling Technique

Abdoulaye Balde
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
abdoulayegnbalde@gmail.com

Sabin Adhikari
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
adhikarisabin258@gmail.com

Adarsh Singh
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
adarshsingh1652@gmail.com

Nomaswati Princess Mabuela
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
lswati.mabusela@gmail.com

Ankur Choudhary
Department of Computer Science and Engineering
Sharda University
Greater Noida, India
ankur.tomer@gmail.com

***Abstract*—The development of any nation can be improved by addressing financial inclusion issues. In many nations, there are challenges in getting people financially included in the banking system. As an example, Africa faces enormous challenges that affect development and livelihood due to poor financial inclusion reported in the Zindi financial inclusion dataset. The literature revealed that various machine learning techniques have been utilized to predict the state of financial inclusion but the performance of those approaches needs to be improvised. In this paper, we have proposed an efficient approach for predicting the state of financial inclusion by addressing class unbalancing issues in the Zindi dataset. So, the class oversampling has been done using Synthetic Minority Oversampling Technique. Further, we have utilized Random Forest, XGBoost, Decision Tree, Logistic Regression, Support Vector Machine , K-Nearest Neighbors, Gradient Boosting, Histogram Gradient boosting, and Adaptive Boosting for predicting how likely a person will have a bank account. The holdout cross-validation method has been used to split the training dataset randomly into 90% train and 10% test set. The result was collected and further compared with some existing approaches, which indicates that our approach outperforms the existing state-of-the-art approach. The Random Forest was optimal among adapted models based on F1-score, precision, recall, and accuracy metrics which are 91.556%, 91.357%, 91.736%, and 91.541% respectively.**

Keywords—Financial Inclusion, Prediction, Machine Learning, SMOTE, Oversampling, Scaling.

I. INTRODUCTION

Financial Inclusion is the ability for individuals to have access to financial products and services that can effectively meet their specific needs. It is one of the noteworthy factors of growth in the finance industry as well as in the economy of the country. It is measured by how many people own and use financial services in their day-to-day life. These financial services include saving money, making payments, having access to credit and insurance.

In 2015, a pledge was taken in the World Bank Group-IMF Spring Meeting to help achieve the Universal Financial Access 2020 (UFA2020) initiative. UFA2020's main objective was simply to bank the unbanked, i.e., all adults, irrespective of their gender, have access to a transaction account to store money, send and receive payments (World Bank Group 2021) [1].

Many countries have witnessed quite a positive growth in financial inclusion since then, but still, a lot of people remain without banks. These unbanked citizens rely on alternative financial services for their needs, which can often be expensive and dangerous. With access to formal financial services, every citizen, especially low-level earners and farmers would avail the benefits of banking facilities and financial services that include loans, credit, payments, etc., they can take loans from the banks and also from the government through banking. This can improve the overall quality of their lives as they will develop a habit of saving

978-1-6654-5834-4/22/$31.00 ©20XX IEEE

money which can later help in times of emergency, consequently promoting the country's economic growth and development.

Our work addresses data imbalance using oversampling and uses machine learning models to predict if a person has a bank account based on certain features.

II. LITERATURE SURVEY

The literature shows machine learning has unfolded different ways to elevate Financial Inclusion and companies to come out as a way to increase its solutions [2]. Ismail et al, [3] have adopted various techniques and models to predict who owns a bank in which XGBoost came on top with 89.23% of accuracy on the same data. The author [4] compared many models' performances in predicting the insurance uptake using the 2016 Kenya FinAccess datasets using oversampling and under-sampling which shows the power of balancing the class label in which XGBoost came in top with 0.8655% of F1-score. Graph neural networks were utilized by [5] to support financial inclusion.

Mhlanga in the review [6] investigated the determinants of financial inclusion in Southern Africa. Using logistic regression, the study discovered that financial inclusions were driven by many factors including age, education level, income, race, gender, and marital status. In the reference the author illustrates many advantages of improving financial inclusion and investigates the major factors affecting access to financial services [7]. In another review, the author emphasizes the goal of achieving financial inclusion for sustainable development goals [8]. Another experimental application of machine learning on financial inclusion data for Governance in Eswatini shows 69.4% with SVM and 63.4% using logistic regression, it emphasizes that models' performance depends on the number of samples and the quality of data the model is trained on [9].

III. RESEARCH METHODOLOGY

In this paper, we have experimented on a dataset that was obtained from 4 different countries in East Africa with the following sample Tanzania 6620, Rwanda 8735, Uganda 2101, and Kenya 6068 observations. We have experimented with state-of-the-art machine learning algorithms to find the best model for predicting financial inclusion.

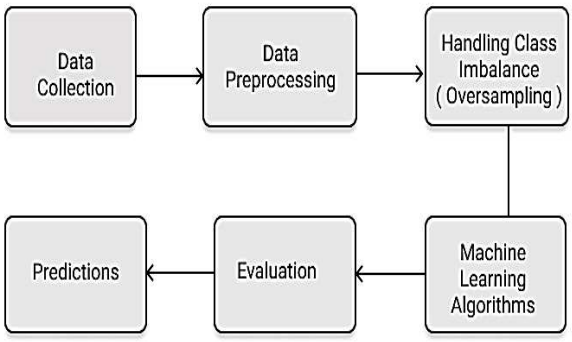

Fig. 1. Flowchart of Methodology

A. Data Collection

This data used for the research was retrieved from Zindi [10] which was extracted from various Finscope surveys ranging from 2016 to 2018. The dataset was in CSV (comma separated variable) format and it had 23524 observations and 13 columns including the target variable.

B. Features Engineering And Selection

Feature selection is one of the most important and relevant steps for mining data. To give a comprehensive view of the unbalanced data problem, we have started with exploring the target variable on the dataset [11]. The analysis shows that location, cell phone access, and gender were highly correlated. Further, the underlying statistical distribution of some features has been studied, and also a correlation between the features has been calculated. As a result, we have found that there was multicollinearity between some independent features, which means that those features are somehow independent of one another. For this reason, a feature Engineering technique has been applied to combine some features to solve the problem of multicollinearity and to pick the foremost important features.

C. Handling Class Imbalance

The problem of imbalanced data appears when the proportion of one class (majority class) has a higher ratio

than another (minority class) [11]. This could be resolved by using both under-sampling and oversampling. The data set used in this research was highly imbalanced. The ratio of those who had a bank account to those who did not have a bank account was 20212:3312 which is 6.10: 1 so the data was unbalanced with the minority class being 15% of the data with the majority class being 85% of the data. Fig 2 shows the imbalanced proportion of each category in the target variable. Since under-sampling of the data reduces the majority class to make it equal to the minority class which may lead to loss of valuable data, to prevent this situation we used the syntactic minority oversampling technique (SMOTE) [12]. Fig 3 shows the proportion of each category after oversampling.

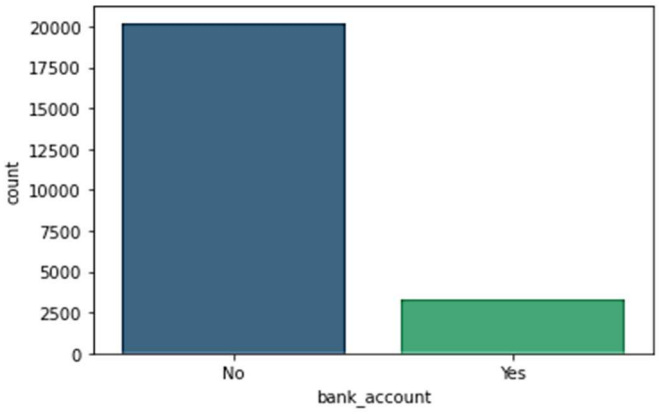

Fig. 2. Imbalanced Class

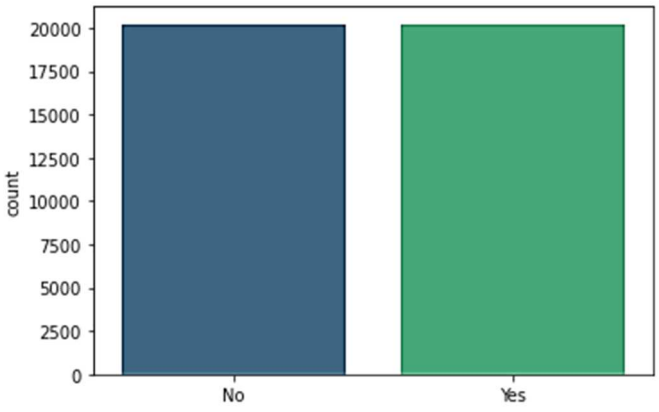

Fig. 3. Balanced Class

D. Model Building

The models used in the research were built using different classifiers. These classifiers were trained using the training data that included both the preprocessed features and the target variable. After training, the classifiers were given the test data to categorize if a person was having a bank account or not. The classifiers used for the research are Logistic Regression, Random Forest, K-,Nearest Neighbor, Gradient Boosting, Histogram-Based Gradient Boosting, Support vector machines, XGBoost, Decision Tree, and Adaptive Boosting.

E. Model Evaluation

The authors have considered 4 different performance metrics which are f1-score, accuracy, precision, and recall to determine the best-fitted model.

The confusion matrix consists of an NxN table depending on the number of classes that show the types of correct predictions and incorrect predictions made by a classifier. True Positives (TP) are those people who have a bank account and were correctly classified as having a bank account. False Positives (FP) are those people who did not have a bank account but were incorrectly classified as having a bank account. True Negatives (TN) are those people who did not have a bank account but were correctly classified as not having a bank account. False Negatives (FN) are those people who have a bank account but were incorrectly classified as not having a bank account.

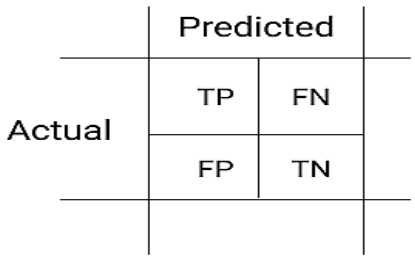

Fig. 4. Confusion Matrix

Mathematically,

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F1-score} = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

IV. Results and Discussion

From each model that has been utilized, we evaluated performance using different evaluation metrics such as F1-score, Precision, Recall, and accuracy score. A model evaluation method can be useful to quickly compare the performances of different machine learning algorithms for the same dataset. To have a fair comparison we shall use the same parameters for training each model. For example, in Random Forest and Gradient Boosting, we have given the number of estimators as 200, the learning rate as 0.1.

Table I shows the recorded performance of each model with the help of different evaluation matrices.

TABLE I. **Performance of Different Classifiers**

Algorithms	**F1-Score (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**
Random Forest	91.555	91.736	91.736	91.540
XGBoost	90.968	93.065	88.965	91.169
Hist	91.053	94.183	85.947	88.124
Gradient Boosting	89.054	92.393	85.947	89.438
KNN	90.806	90.741	90.143	90.774
SVM	89.745	95.479	84.661	90.329
Decision Tree	89.862	89.378	90.351	89.809
Logistic Regression	89.336	94.681	84.662	89.908
AdaBoost	86.641	89.545	83.918	87.064

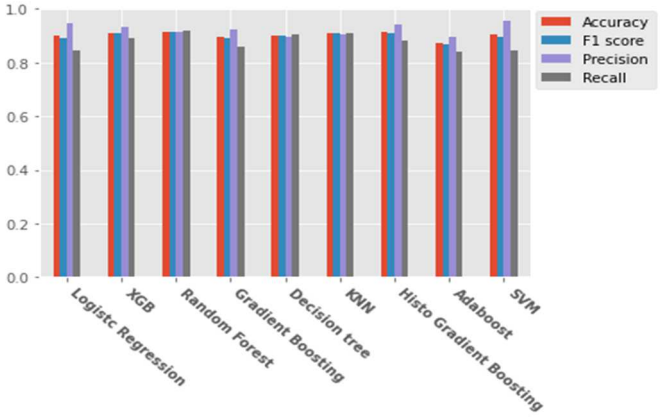

Fig. 5. Performance of Different Models

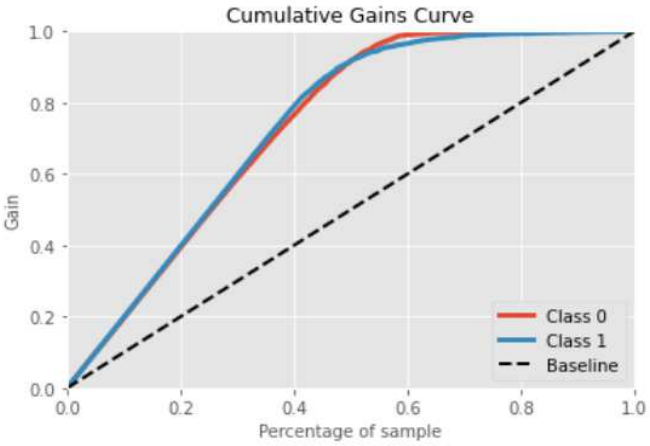

Fig. 6. Cumulative gain of Random Forest classifier

V. Conclusions

We trained all the learning algorithms using StratifiedKFold 10 fold cross-validation to preserve the percentage of samples for each class and also to avoid

model overfitting. The state of the art result we got was a big improvement of all the previous works that have been done from the papers that we went through as a result we were able to achieve a score of 91.556%, 91.375%, 91.736% and 91.541% on F1-score, Precision, Recall, and Accuracy respectively. It was noticeable that features scaling improved the speed. Moreover, we did not rely on Accuracy only because the class was unbalanced; using accuracy will always result in predicting the majority class. Future work would be trying other boosting algorithms such as Catboost, LightGBM, and a neural network might be a good try for model improvement. In addition to this, a comparison between oversampling and under-sampling will also be good work to consider.

REFERENCES

[1] "Website." https://openknowledge.worldbank.org/handle/10986/9335

[2] N. Kshetri, "The Role of Artificial Intelligence in Promoting Financial Inclusion in Developing Countries," Journal of Global Information Technology Management, vol. 24, no. 1. pp. 1–6, 2021. doi: 10.1080/1097198x.2021.1871273.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[3] 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). 2021. doi: 10.1109/icics52457.2021.9464590.

[4] N. Yego, J. Kasozi, and J. Nkrunziza, "A Comparative Analysis of Machine learning Models for Prediction of Insurance Uptake in Kenya." doi: 10.20944/preprints202010.0186.v1.

[5] L. Roa, A. Rodríguez-Rey, A. Correa-Bahnsen, and C. V. Arboleda, "Supporting Financial Inclusion with Graph Machine Learning and Super-App Alternative Data," Lecture Notes in Networks and Systems. pp. 216–230, 2022. doi: 10.1007/978-3-030-82196-8_16.

[6] Mhlanga, "Industry 4.0 in Finance: The Impact of Artificial Intelligence (AI) on Digital Financial Inclusion," International Journal of Financial Studies, vol. 8, no. 3. p. 45, 2020. doi: 10.3390/ijfs8030045.

[7] P. K. Ozili, "Impact of digital finance on financial inclusion and stability," Borsa Istanbul Review, vol. 18, no. 4. pp. 329–340, 2018. doi: 10.1016/j.bir.2017.12.003.

[8] A. Ma'ruf and F. Aryani, "Financial Inclusion and Achievements of Sustainable Development Goals (SDGs) in ASEAN," GATR Journal of Business and Economics Review (JBER) Vol. 4 (4) Oct-Dec 2019, vol. 4, no. 4. pp. 147–155, 2019. doi: 10.35609/jber.2019.4.4(1).

[9] B. A. Akinnuwesi, S. G. Fashoto, A. S. Metfula, and A. N. Akinnuwesi, "Experimental Application of Machine Learning on Financial Inclusion Data for Governance in Eswatini," Lecture Notes in Computer Science. pp. 414–425, 2020. doi: 10.1007/978-3-030-45002-1_36.

[10] "Zindi." https://zindi.africa/competitions/financial-inclusion-in-africa/data (accessed Mar. 13, 2022).

[11] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS). 2020. doi: 10.1109/icics49469.2020.239556.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16. pp. 321–357, 2002. doi: 10.1613/jair.953.