# MICROFINANCE SERVICES

## Using Machine Learning to Predict Feasibility in Kenya

Giulia Maria Petrilli – 1155190793
03/01/2023
ECON4810 – Instructor: Dr. Vinci Chow

**Abstract**

***This paper aims to investigate usage of informal Kenyan microfinance investment groups, known as chamas, and based on this, to analyse the feasibility of implementing an application to assist in chama organisation and management. A population survey provides information on current usage of these groups, which was used to train predictive models. Two logistic regression models, one selecting its variables with a hyperparameter tuning and the other one by p-value significance, and a neural network using variables with p-value significance, were trained on the survey results to predict chama usage. Model performance was evaluated and compared using confusion matrices and accuracy. The results show that the first logistic regression is very accurate, with a test accuracy of 1 and a train accuracy 0.95. Such a high train accuracy suggests that the model performs well on unseen data and may not be overfitting. The second logistic regression does not perform as well, with a test accuracy score of 0.37. To investigate such a low accuracy, the paper uses a neural network analysis on the same variables. As expected, the neural network performs much better, with an accuracy score of 0.84. The study demonstrates the potential of using advanced analytics to improve the understanding and use of chama microfinance services in Kenya. The paper focuses on an unrepresented and vastly unbanked part of potential microfinance clients, the Muslim population. The study shows that belonging to the Christian religion may influence the decision of joining an informal microfinance chama group, but Muslim religion appears to be less significant with a p-value of 0.067. Understanding these factors could help to optimise the financial performance of microfinance initiatives in Kenya and similar contexts. Conclusively, digitalizing the processes of microfinance in Kenya through a digital platform for managing savings groups is worth exploring further, as it has the potential to make microfinance services more accessible and to create revenue. Making such an app for a sharia compliant population is also worth exploring, as the p-value for the Muslim religion variable is barely not significant.***

**Introduction and Background**

Informal microfinance initiatives, such as chama groups, are a type of financial service that is provided outside of the formal banking sector (Adede, 2007). These groups are typically informal organisations that are formed and run by community members, and they offer small loans and other financial services to individuals who may not have access to traditional banking options (Adede, 2007). Informal microfinance initiatives, such as chama groups, can play a vital role in improving financial inclusion and supporting economic development in communities (Kimani et al, 2012). However, there is still a lack of understanding about the feasibility and sustainability of these initiatives, and more research is needed to identify the factors that contribute to their success (Muiruri, 2013).

Stokvels, a popular form of savings in South Africa, have been shown to offer benefits such as peer pressure to save and the ability to earn better returns on pooled money (Mashigo & Schoeman, 2012). A South African digital platform for managing savings groups, Stokfella, has digitalized and hence eased the administrative work involved in running a stokvel, helping savings clubs to manage their payments digitally (Rodima-Taylor, 2022). As of early 2020, the online platform Stokfella had approximately 14,000 users and conducted more than one million transactions annually (Mathe, 2020).

As far as preliminary research goes, such a digital environment has not been established in Kenya. Nevertheless, the country vastly makes use of microfinance services (Ali, 2015). Furthermore, Kenya has a numerous Muslim population which constituted the majority in three of the eight provinces in the country - the Coast, the Eastern

Provinces and the Northeastern Province (Ayubi & Mohyuddin,1994). Both pieces of information show that there is a potential opportunity for an app for community members who use chama groups, with a further chance to tap into the Muslim population.

This paper uses logistic regression models to understand better which features are most likely to predict chama usage and to predict which demographic groups a future app or online chama aid should target in the context of the Kenyan market. Additionally, the logistic regression model performances will be compared with those of a neural network to seek the best performing model.

Interestingly, the exclusion of individuals from the financial system based on their adherence to Shariah principles has long been a challenge in microfinance (El-Komi & Croson, 2013). People who follow the religion of Islam often cannot access them because Sharia law does not allow loans, leaving many people unbanked (Burgum, 2018). Nevertheless, studies establish a Kenyan demand for a sharia compliant microfinance service (Khan, 2008). Consequently, as a sub question, this paper will predict feasibility of sharia compliant microfinance services based on current Muslim members' reasons for joining, or not joining, microfinance groups.

This study addresses this problem for two reasons: first, to promote inclusion and accessibility for this group, as Islamic finance is constructed to have a real impact on an economy through job creation and entrepreneurial development (Alsonso, 2015), and second, to explore the potential for the market of a new digital platform for managing savings groups among the unbanked population in Kenya. By using advanced analytic methods, logistic regression, and neural networks evaluated using confusion matrices, this study seeks to improve our understanding of the feasibility of microfinance services in Kenya, with attention to sharia compliant ones. The combination of these methods has the potential to be a powerful tool for understanding and predicting the feasibility of informal microfinance services in Kenya.

**Data**

In 2018, the researcher Riane Kinuthia published a dataset on Kaggle titled "Islamic Microfinance Services Feasibility Study". The data was collected in 2017 and consists of information on the feasibility of implementing microfinance services in various regions in Kenya. The dataset, 506 respondents for 110 variables, includes information on the economic and cultural factors that may influence the success of such services, as well as data on the demand for and availability of financial products in different regions. The dependent variable in this study is Chama group belonging, which shows how many people in the dataset belong in the chama microfinance groups. The predictive value of this research lies in the ability to predict what kind of target population is likely to belong to a chama group, to understand feasibility and implications for a microfinance app in the future. Data pre-processing is performed using iloc and loc methods are used to remove unnecessary columns from the data frame, such as qualitative answers on the choice of a chama group instead of another. These methods allow to split the data frame by specifying rows and columns to keep or remove. In this case, the columns whose names begin with a specified string are removed. After removing the columns, the column names are changed to remove spaces, slashes, and other special characters to make them more suitable for use as variables. This is done with the str.replace method for the columns attribute of the data frame. Next, missing values in the column "On_average__how_much_are_you_able_to_save_monthly_" are replaced with the average value of this column using the SimpleImputer class from the sklearn.impute library. Further pre-processing is performed to convert categorical variables to numeric representations using one-hot encoding. This operation is performed because a logistic regression requires binary values. This is done using the get_dummies method from the Pandas library. Finally, certain hot-one encoded columns are removed from the data frame. In general, a dataset with 506 observations and 87 variables is relatively small, especially if the variables are highly correlated as it may be the case for this dataset. This may make it more difficult for the

model to learn meaningful patterns in the data and to generalize to new, unseen data.

**Data exploration**

After selecting predictors and target variables, the data frame is scaled. Then, the data set is divided into training and testing data sets. Finally, the training data are oversampled using SMOTE, which allows the chama group belonging respondents to be represented accurately in the analysis. The results are presented in figures 1 and 2, which describes the impact of data balancing on members of a Chama group (the target group).


*Figure 1: Chama group belonging before oversampling*

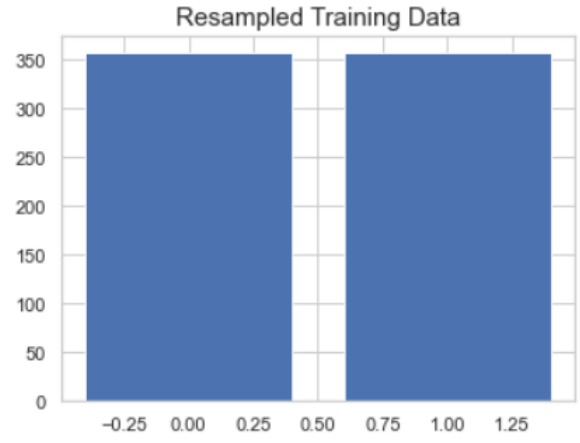


*Figure 2: Chama group belonging after oversampling*

Such an attempt was performed on the variables of religion and gender as well, but this did not improve the model, hence, this change was not conducted. Figure 3 shows that most Muslims in the questionnaire chose not to choose to belong to a chama group, which fits the literature. The plot also shows that most of the people in the survey is Muslim, which is a characteristic to be wary of since it might lead to biases in the findings.

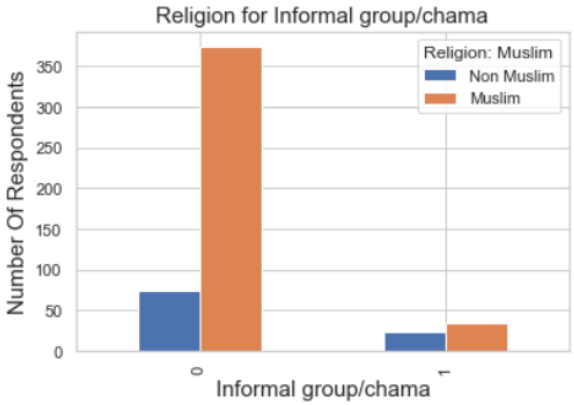


*Figure 3: Religion and chama belonging*

Gender also highly impacts chama group belonging, as seen in figure 4, with a higher proportion of in the female population choosing to belong in a chama group than the male one.

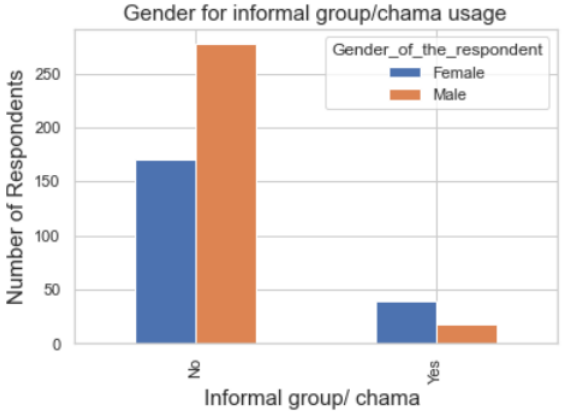


*Figure 4: Gender and chama belonging*

**First model : Logistic regression with Hyperparameter tuning**

In this logistic regression, a grid search was performed to find the best hyperparameters for the model by training and evaluating the model with a 10-fold cross-validation with different combinations of the hyperparameters 'penalty', 'C', and 'solver'. The best hyperparameters found were {'C': 1.0, 'penalty': 'l1', 'solver': 'liblinear'}, resulting in an accuracy of 1.0, which is concerningly high. The model was then trained with these hyperparameters and tested with the test data set, resulting in an accuracy of 1.0. The coefficients of the model were printed to show the significance of each feature for prediction, to

understand which features would have a stronger predictive value. The feature importance analysis revealed that the most important features are about whether the respondents had been able to make contribution through a financial provider, whether the respondent is a female, whether they are from county Kilifi and their sources of livelihood. The model performs well, hence this is trustworthy information. In fact, the logistic regression model was able to achieve perfect accuracy on the test data sets and with a 0.95 accuracy on the train one, which suggests that the model may not be overfitting. The confusion matrix confirms the model's has high accuracy, that perfectly predicted the correct and incorrect labels in most cases.

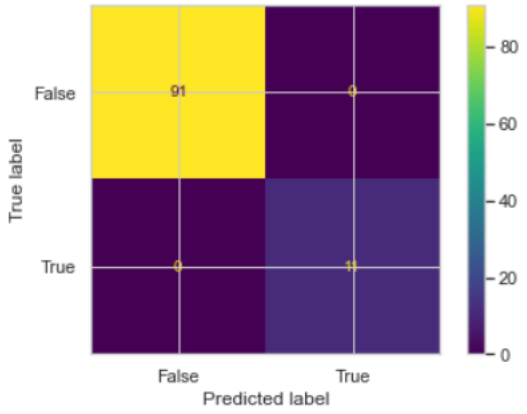


*Figure 5: Logistic regression 1*

I will perform a logistic regression that includes the characteristics with a significant p-value and use them for the second logistic regression to see whether some other useful information emerges.

**Second Model: Logistic regression with p-value**

The second model will perform a Recursive Feature Elimination (RFE), a feature selection method used to select a subset of the most important features from a dataset, and select features based on their ability to improve the model accuracy. This is done by recursively removing features, building a model with the remaining features, and then evaluating the performance of the model. I used RFE for my second model to improve its interpretability, to reduce the risk of overfitting, to reduce the computational cost of training the model, and to improve the generalisation performance of the model.

The RFE yielded relevant interesting findings; the variables related to sources of livelihood are relevant features to predict chama group belonging, while the variables that relate to the monthly income of the respondent are not as much. Hence, sources of livelihood may be more directly related to the respondent's decision to join a chama group. For example, if the respondent is in a profession that requires saving and investing money (e.g., farming), he or she is more likely to join a chama group for financial management purposes. The respondent's monthly income may not be a reliable predictor of Chama group membership due to a variety of factors, including income fluctuations, the respondent's spending and saving habits, or other financial considerations.

Through the RFE, Muslim religion turns out to be a relevant predictor for membership in the Chama group. However, when the variables that passed that RFE get tested for p-value significance, Muslim religion has a p-value of 0.06. This has some implications for the papers' sub-question on the feasibility of specifically Sharia-compliant microfinance services, as establishing a Sharia-compliant microfinance group may or may not be beneficial to attract Muslim clients. If religion is not a relevant characteristic for predicting membership in a chama group, it may not be necessary for the microfinance group to adopt Sharia-compliant practices to attract Muslim clients. nevertheless, the p-value is barely not significant, so further research needs to be performed to get definitive results.

The following bar plots are showing the number of features that were selected at each step of the Recursive Feature Elimination (RFE) process, based on p-value significance. The plots show that a screening was performed, but not many variables had to be given up on, to 43 after the first screening and 35 after the second one.

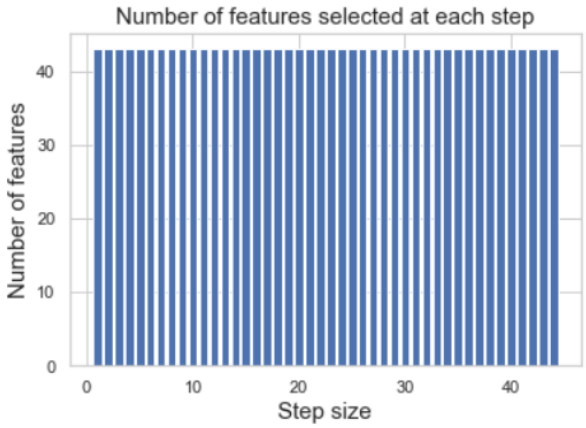


*Figure 6: first RFE*

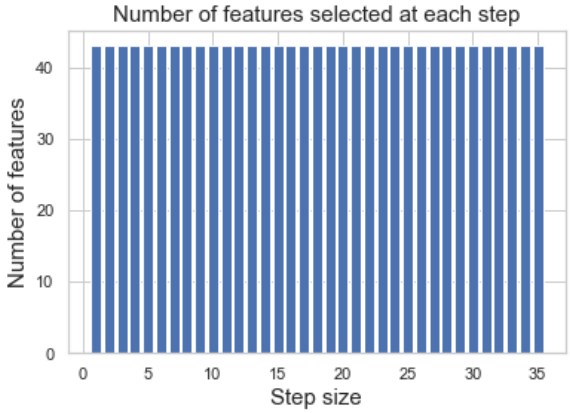


*Figure 7: second RFE*

The variables that have a p-value less than 0.05, such as respondents' sources of livelihood, gender and education level, are statistically significant predictors of belonging to the Chama group in study. This means that the relationship between these variables and Chama group membership is likely real and not due to chance. Therefore, these variables may be important in predicting Chama group membership and in designing interventions or strategies related to microfinance.

The values for R-squared and adjusted R-squared in an OLS regression indicate how well the model fits the data. A R-squared value is 0.764 and the adjusted R-squared value is also 0.752 may suggest that the model performs fairly well in predicting the outcome of interest (in this case, membership in the Chama group). The confusion matrix in figure 8, however, show that the model does not have a very high accuracy, with accuracy reaching a low 0.37. To investigate more into the p-value significant variables, the study will plug them into a neural network analysis and see if that improves the accuracy.

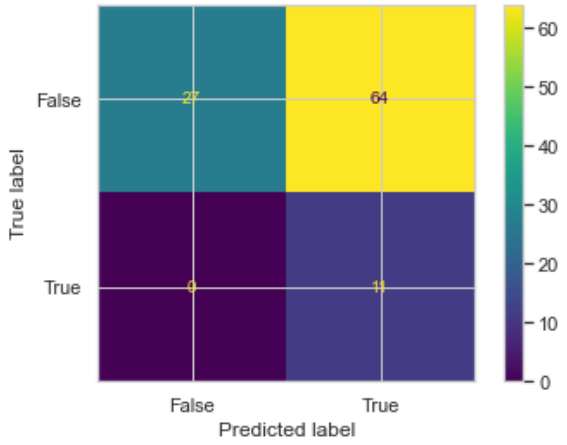


*Figure 8: Confusion matrix*

**Neural Network**

The code for this section creates and trains a neural network model using the Sequential model in the Keras library. The model consists of two fully connected layers with 64 units and an activation function of 'relu' for the first layer, and a single unit with an activation function of 'sigmoid' for the second (output) layer. The model is then compiled with a loss function of 'binary_crossentropy' and an optimizer of 'adam' . The model is then trained on the resampled training data for 50 epochs using a validation set from the original test data. After training, the model is evaluated on the original test data and the test loss and test accuracy are printed. The test loss of 0.34 and test accuracy of 0.84 indicate that the model performs well.

It appears that the model for loss (figure 9) might be overfitting the data. If the training loss continues to decrease, but the validation loss becomes horizontal or starts to increase, it is a sign of overfitting.

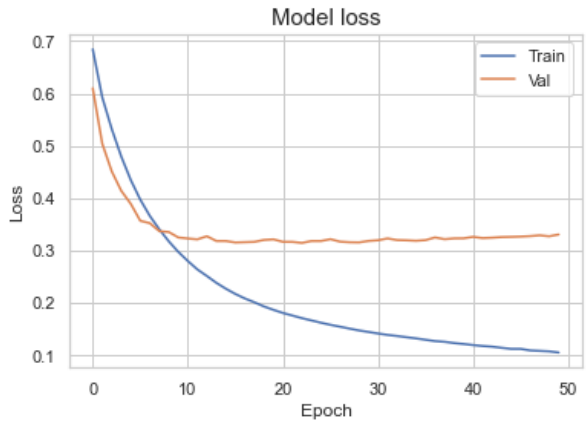


*Figure 9: loss*

The model loss plot suggests that the model is learning and improving on the training data, but it may be overfitting or not generalising well to the validation data. The validation loss becoming horizontal indicates that the model is not making significant progress on the validation set and may not be improving in accuracy. The confusion matrix in plot 11 confirms a high enough accuracy of 0.74. Accuracy for this model is higher than the logistic model one for the same variables, so a neural network performs better than a logistic regression to predict chama group belonging based on p-value.

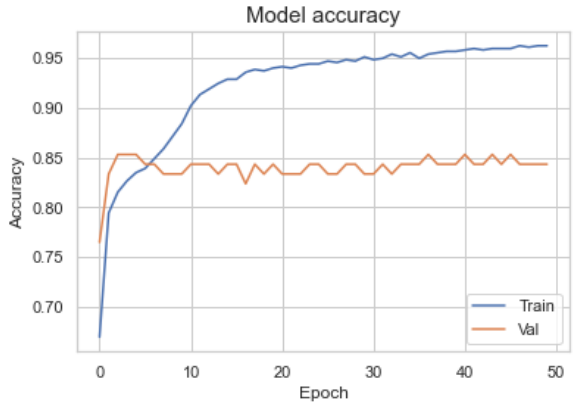


*Figure 10: Accuracy*

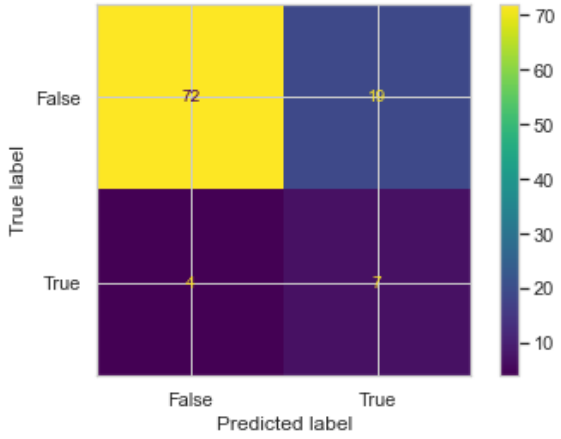


*Figure 11: NN confusion matrix*

**Discussion and Conclusion**

In summary, the results of this study suggest that the first logistic regression based on hyperparameter tuning and the neural network models developed to predict Chama group belonging in Kenya have good predictive value, with some overfitting to account for in further research. Some relevant characteristics that future research should focus on our ability to contribute to financial service providers, gender, and sources of livelihood. Further research could also explore other financial factors that could influence the feasibility of Chama groups, such as respondents' level of debt or the sustainability of the Chama business model in the long run. In addition, collecting a larger sample can improve the accuracy of the models and better understand the potential for targeting the Shariah-compliant group. It should be noted that the results may be biased due to the disproportionate number of Muslim respondents in the survey. Further research should consider the role of digitization and the potential for a conventional microfinance organisation with Shariah-compliant options, and the results for the religion variables are unclear.

**References**

Adede, A. A. (2007). *Merry Go Round Concept and Informal Financial Markets in Kenya* (Doctoral dissertation, University of Nairobi.).

Ali, A. E. E. S. (2015). The regulatory and supervision framework of Microfinance in Kenya. *Int'l J. Soc. Sci. Stud.*, *3*, 123.

Alonso, I. M. (2015). Crowdfunding in Islamic Finance and microfinance: A case study of Egypt. *Access to Finance and Human Development—Essays on Zakah, Awqaf and Microfinance*, 85.

Ayubi, S., & Mohyuddin, S. (1994). Muslims in Kenya: an overview. *Journal Institute of Muslim Minority Affairs*, *15*(1-2), 1441 156.

Begum, H., Alam, A. F., Mia, M. A., Bhuiyan, F., & Ghani, A. B. A. (2018). Development of Islamic microfinance: a sustainable poverty reduction approach. *Journal of Economic and Administrative Sciences*.

El-Komi, M., & Croson, R. (2013). Experiments in Islamic microfinance. *Journal of Economic Behavior & Organization*, *95*, 252-269.

Kimani, J. K., Ettarh, R., Kyobutungi, C., Mberu, B., & Muindi, K. (2012). Determinants for participation in a public health insurance program among residents of urban slums in Nairobi, Kenya: results from a cross-sectional survey. *BMC health services research*, *12*(1), 1-11.

Kinuthia, R. (2018). Islamic Microfinance Services Feasibility Study. Retrieved from https://www.kaggle.com/datasets/rkinuthia/islamic-microfinance-services-feasibility-study

Mashigo, P., & Schoeman, C. (2012). Stokvels as an instrument and channel to extend credit to poor households in South Africa. *Journal of Economic and Financial Sciences*, *5*(1), 49-62.

Mathe, T,2020. The age-old stokvel moves into a digital era. Available at:https://mg.co.za/article/2020-01-17-the-age-old-stokvel-moves-into-the-digital-era/[Accessed 22 December 2022]

Muiruri, P. (2013). Linking informal social arrangements, social protection and poverty reduction in the urban slums of Nairobi, Kenya. *Informal and Formal Social Protection Systems in Sub-Saharan Africa*, 43.

Rodima-Taylor, D. (2022). Platformizing Ubuntu? FinTech, Inclusion, and Mutual Help in Africa. *Journal of Cultural Economy*, 1-20.