Data Scientist Nanodegree

Capstone Project - Zindi - Financial Inclusion in Tanzania

Alexandre Biebuyck

April 24th, 2019

I. Definition

From sales predictions to automatic speech creation or image recognition, Data Science has an infinite number of applications in our society and will impact our lives drastically in many ways in the short and long term. Will it be for the best or the worst? Even the greatest specialists in Data Science and Artificial intelligence have opposite visions on what the future will be[1]. It will mainly depend on how we decide to use it.

For the final project of Data Scientist Nanodegree of Udacity, I choose to bet for the optimistic option by taking part in an online "Data for Good"[2] competition, the "Mobile Money and Financial Inclusion in Tanzania Challenge" hosted by Zindi[3], an African Data Science platform.

This report includes the description of the steps followed in the realization of the project, some visualization, the main insights extracted from the data, explanations on the modelling phase and the conclusions and reflections about the project. The detail of the code is available in the Jupyter Notebook in the Github repository[4]

Project Overview

Zindi states the context and objectives of the competition:

"Only 16.7% of the population in Tanzania has a bank account. But an additional 48.6% of Tanzanians who don't have a bank account do have other types of formal financial services, primarily mobile money.

For people who have been traditionally excluded from the formal financial system in Africa and other developing markets, mobile money has become an important entry point to financial inclusion. While mobile money is a tool for transferring money among people and businesses/other institutions, it is increasingly becoming a platform for people to access a broad range of financial services, including savings, credit, and insurance."

---

[1] https://www.bbc.com/news/technology-30290540
https://www.businessinsider.com/mark-zuckerberg-shares-thoughts-elon-musks-ai-2018-5

[2] *"Data for good is a movement in which people and organizations transcend organizational boundaries to use data to improve society."* Gartner, How to Use Data for Good to Impact Society, 2018

[3] Zindi is a *"social enterprise whose mission is to build the data science ecosystem in Africa"* who *"works with companies, non-profit organizations, and government institutions to develop, curate, and prepare data-driven challenges"* .

[4] https://github.com/alexbieb/Capstone_Project_Multilabel_Classification

The objective of this competition is to create a machine learning model to predict which individuals are most likely to use mobile money and other financial services (savings, credit, and insurance). [5]

This model can help mobile money providers target new clients and markets across Tanzania more effectively, and also help financial services providers cross-sell other financial services (savings, credit, and insurance) to the existing mobile money customer base."[6]

This challenge is in line with the goals and improvement opportunities defined by the World Bank for the development of Tanzania:

- *"efforts should be made to expand access to those still not participating in financial services, including women and youth."*
- *"deepening inclusion by broadening the use of more advanced financial products and services will help Tanzania move towards a more formalized, transparent, and dynamic economy"*
- *"Tanzanians need expanded access to affordable long-term credit by lowering the cost of risk through better selection of borrowers and by reducing the disincentives on lending to the private sector…"*[7]

Zindi provides us 2 datasets and access to a Mapping Plateform to solve this challenge:

- 1) The Dataset of Demographic and Financial Service usage information of 10.000 Tanzanians[8]

	ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8_1	Q8_2	...	Q17	Q18	Q19	Latitude	Longitude	mobile_money	savings	borrowing	insurance	mobile_money
0	5086	98	2	3	1	1	2	2	0	0	...	-1	4	4	-4.460442	29.811396	0	0	0	0	
1	1258	40	1	1	3	5	1	1	1	0	...	4	1	4	-6.176438	39.244871	1	1	1	0	
2	331	18	2	4	6	3	2	1	0	0	...	-1	1	1	-6.825702	37.652798	1	0	0	0	
3	6729	50	1	1	3	1	1	1	0	0	...	-1	1	4	-3.372049	35.808307	1	0	1	0	
4	8671	34	1	1	1	1	2	1	0	1	...	-1	1	4	-7.179645	31.039095	1	1	0	1	

Figure 1 – Dataset FSDT Finscope 2017

[5] https://zindi.africa/competitions
[6] https://zindi.africa/competitions/mobile-money-and-financial-inclusion-in-tanzania-challenge
[7] Tanzania Economic Update, World Bank, April 2017
[8] Data from the FSDT Finscope 2017 survey. E.g. : House location, Education, Income sources, Sex, Age,…

- 2) The Dataset of Geospatial mapping of all cash outlets in Tanzania in 2012[9]

	region	district	ward	latitude	longitude	bank_type	bank_name	weekend_trading	yr_started	yr_started_reformatted
0	Dar es Salaam	Temeke	Mbagala	-6.92247	39.27113	Commercial Bank	Accessbank (Tanzania) Limited	Yes	16/05/11	2011-5-01
1	Shinyanga	Kahama	Kahama Mjini	-3.82858	32.60006	Commercial Bank	Accessbank (Tanzania) Limited	No	12/07/13	2013-7-01
2	Dar es Salaam	Temeke	Miburani	-6.86209	39.26233	Commercial Bank	Accessbank (Tanzania) Limited	Yes	05/03/98	1998-3-01
3	Dar es Salaam	Ilala	Kariakoo	-6.82157	39.28025	Commercial Bank	Accessbank (Tanzania) Limited	Yes	10/06/09	2009-6-01
4	Dar es Salaam	Ilala	Kariakoo	-6.81944	39.27418	Commercial Bank	Accessbank (Tanzania) Limited	Yes	18/12/13	2013-12-01

Figure 2 - Dataset FSDT - Cash Outlets

- 3) Access to Maps and regional demographic data[10]

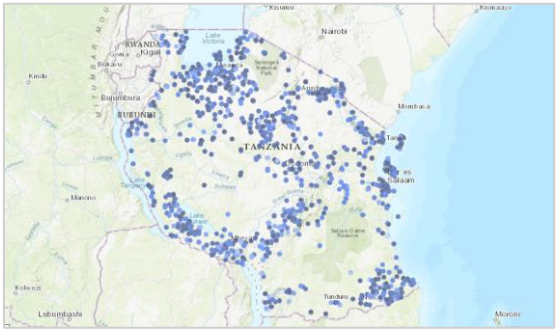


Figure 3 – Example of ArcGis Mapping

Problem Statement

Machine Learning Problem Category

The Financial Inclusion in Tanzania Competition can be defined as a Supervised Learning multiclass classification problem for the two following reasons:

- Supervised: Zindi provides us a dataset including labelled observations and a separate dataset without labels for which we have to make predictions.
- Multiclass Classification: The expected output of the model is the prediction of the probability of each individual to belong to each of the target categories.

Methodology

The resolution of this project has been done by following the Cross-Industry Standard Process for Data Mining (CRISP-DM)[11].

[9] Data from the FSDT. E.g. :Commercial banks, community banks, ATMs, microfinance institutions, mobile money agents, bus stations and post offices
[10] Data from the Esri's ArcGIS Technology
[11] https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

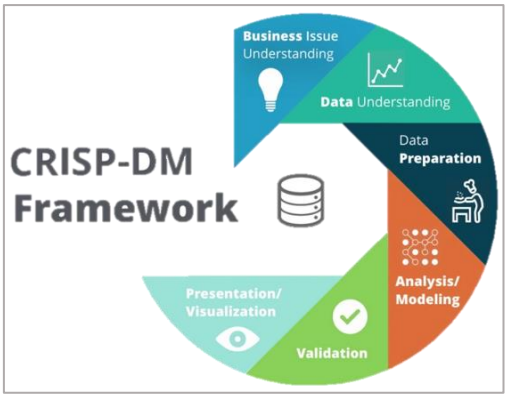


Figure 4 - CRISP-DM Framework

Therefore, the following tasks have been performed and detailed accordingly in the corresponding sections of this document:

- Real-World Issue Understanding (I. Definition)
  - Reading and analysis on the Economic situation in Tanzania
  - Obtaining statistics on the Financial inclusion and Mobile Money market
  - Visualization of the infrastructures and demography of the country (Mapping)
- Data Understanding (II. Analysis)
  - Measuring statistics of the provided datasets
  - Identification of the data types, structure of the datasets, outliers, …
  - Visualization of the correlations between variables
- Data Preparation (III. Methodology)
  - Calculating and extracting additional data from ArcGis
  - Cleaning of the outliers and missing data
  - Merging the datasets
  - Reshaping the features and creating additional features
  - Splitting the training and testing sets
  - Automatizing the pre-processing in a function
- Modelling (III. Methodology)
  - Defining a baseline model
  - Selecting classification models
  - Optimize the models (GridSearch and TPOT)
  - Training and predicting using Supervised Learning (Scikit-learn)
  - Deep Learning (Keras)
- Validation (IV. Results)
  - Designing graphs and metrics for model evaluation
  - Automatizing the creation of the output file
- Visualization (IV. Results)
  - Evaluating the score and ranking on the Zindi plateform

Expected Results

In order to properly train and validate the results of our model we will divide the 7094 observations (individuals) of the labelled data set provided by Zindi into a **training** set and **validation** set. This will allow us to train the model on labelled data (training set) and then

measure how well the model is performing on unseen data (validation set) before making predictions (test set).

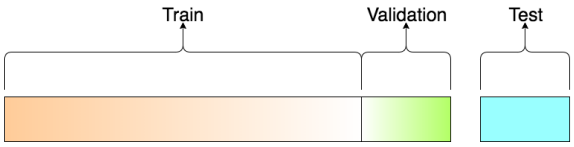


Figure 5 - Training, Testing and Validation[12]

The unlabeled **test** set provided by Zindi contains 2365 observations (individuals). For each of these observations we will have to predict the probability of being in each of the 4 target classes.

	A	B	C	D	E
1	ID	no_financi	other_only	mm_only	mm_plus
2	2352	0.0	0.0	0.2568	0.7432
3	8208	0.0129	0.0439	0.1717	0.7715
4	2785	0.0174	0.0217	0.1706	0.7903
5	2967	0.3642	0.4233	0.0526	0.1599

Figure 6 - Prediction File

The expected output of the machine learning model is a classification of the population in 4 different target classes:

Labels	Other Financial Services	
Mobile Money	NO	YES
NO	0	1
YES	2	3

- 0: Doesn't use mobile money and doesn't use financial service (saving, borrowing, insurance)
- 1: Doesn't use mobile money but use at least one other financial service
- 2: Use only mobile money
- 3: Use mobile money and at least one other financial service

Metrics

Challenge metrics

For this competition, Zindi will rank the models uploaded by the participants using Logloss, a common metrics for multiclass classification problems.

Logloss Explanation

Logloss (logarithmic loss), also called cross-entropy, is measured on the probabilities of the classifications.

[12] https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

$$\text{log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i)$$

Figure 7 - Logloss Formula

"p_i is the probability that the ith data point belongs to class 1, as judged by the classifier. y_i is the true label and is either 0 or 1. Since y_i is either 0 or 1, the formula essentially "selects" either the left or the right summand. The minimum is 0, which happens when the prediction and the true label match up. (We follow the convention that defines 0 log 0 = 0.)" [13]

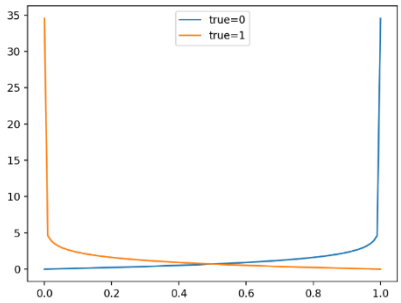


Figure 8- Line Plot of Evaluating Predictions with Log Loss[14]

Additional Internal Metrics

Accuracy, Recall or other metrics can't capture the same level of detail on how good did the model predicted the probabilities because they only take into account the class predicted versus the actual class and not the associated probability versus the actual class. However, they can provide us insights on the type of errors made in our classification model[15] and are not highly impacted by the confidence level of the predictions as is the Logloss metric[16]. Therefore, for internal evaluation of our models we will also use Accuracy and Confusion Matrices.

[13] https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/3/evaluation-metrics
[14] https://machinelearningmastery.com/how-to-score-probability-predictions-in-python/
[15] https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/
[16] https://datawookie.netlify.com/blog/2015/12/making-sense-of-logarithmic-loss/

II. Analysis

External Sources

In order to have a good initial understanding of the problematic and the current situation in Tanzania, we searched and identified reliable and pertinent sources of information:

- The World Bank Tanzania Economic Update 2017
- The Finscope Survey 2017 Report
- Market analysis from a mobile money provider[17]
- GIS Census 2014 - Financial Access Map

Visualizations on Financial Inclusion in Tanzania

They provide us with a lot of information to understand the problematic we are facing, the factors influencing the financial inclusion and the structure of the Tanzanian population and financial market. Some of the main insights taken from these reports are that:

- Financial inclusion increased over the years:

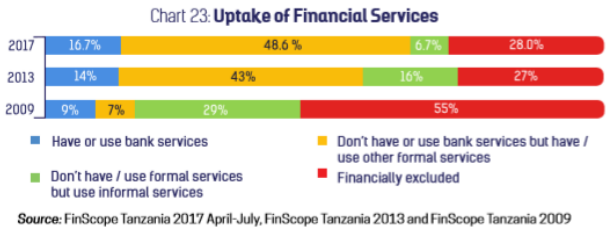


Figure 9 - Financial Inclusion Evolution - Finscope 2017

- Mobile Money is the most used financial service:

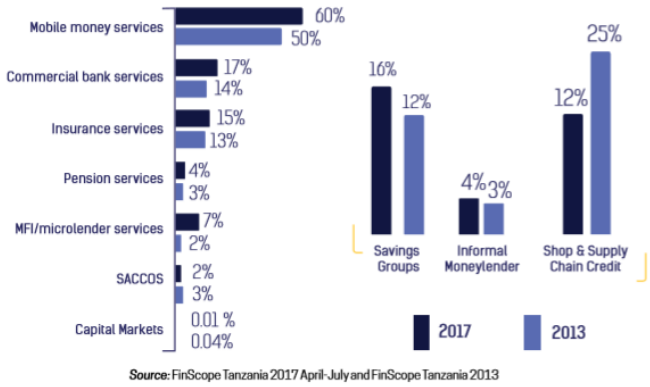
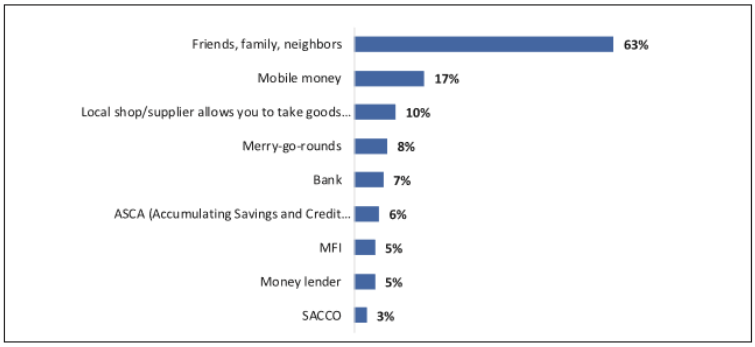


Figure 10 - Method of inclusion in the financial system - Finscope 2017

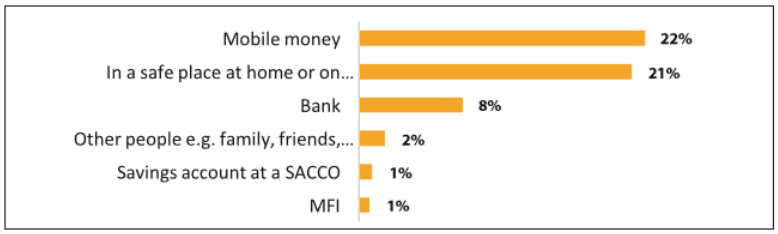
- The main borrowing and savings channels are family and mobile money:

[17] What Makes a Successful Mobile Money Implementation? - GMSA



Source: InterMedia Tanzania FII Tracker Survey Wave 3 (September-October 2015).

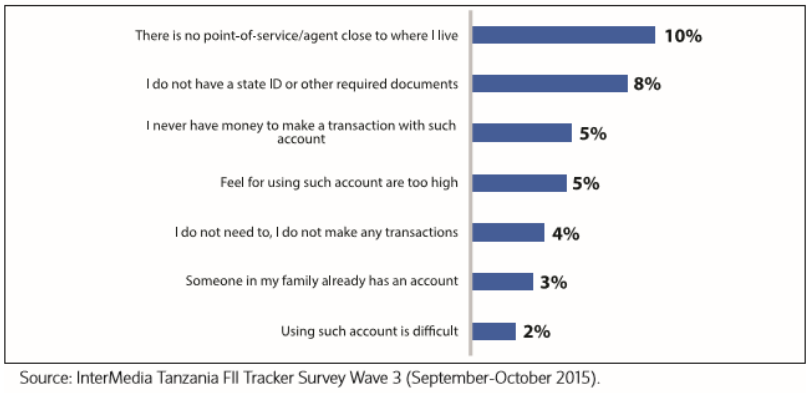
Figure 11 - Borrowing Channels - World Bank



Source: InterMedia Tanzania FII Tracker Survey Wave 3 (September-October 2015).

Figure 12 - Savings Channels - World Bank

- The main reason for not having a mobile money account is not having an agent or a point of service nearby:



Source: InterMedia Tanzania FII Tracker Survey Wave 3 (September-October 2015).

Figure 13 - Reasons for no Mobile Money - World Bank

Overview of Financial Inclusion in Tanzania

The actual situation of the Tanzanian financial sector is described by the world bank as *"Increasingly bifurcated, with a very dynamic mobile financial services sector competing against a very concentrated banking sector".*

They also define that measures should be taken to *"expand access to those still not participating in financial services. […] the range of services available to those who are already active in the financial sector should be expanded."*

And precise the type of population that is most excluded from this financial system: *"Greater attention should also be directed towards bridging the gender gap in access to. Finance in rural areas where women involved in agricultural activities suffer significant limitations. Although the proportion of women with access to an account has increased, a significant gender gap, particularly in the rural areas, still negatively impacts women's access to financial services."*.

Data Exploration & Exploratory Visualization

As we saw earlier, the input data provided by Zindi is composed of the following:

- Dataset Finscope Survey 2017
- Dataset FSDT Financial Access Mapping
- Access to the ArcGis plateform

Finscope Survey 2017: Demographic and Financial Service usage information of 10.000 Tanzanians[18]

Finscope Survey 2017 - Description

The labelled dataset, that will be split into training and validation set, from the Finscope Survey includes 7094 observations over 32 features and 5 target columns[19].

The main features provided for the modelling are:

- Continuous:
  - Age (Q1)
  - Latitude and Longitude
- Binary:
  - Gender (Q2)
  - Phone Ownership (Q7)
  - Income Sources (Q8_1 to Q8_11)
  - Do send money (Q12) – Do you receive money (Q15)
- Categorical Multilabel:
  - Marital Status (Q3)
  - Type of occupation (Q9, Q10 and Q11)
- Categorical Ordered:
  - Ownership of Land (Q5)
  - Education Level (Q4)
  - Frequency of sending (Q13) or receiving money (Q15)
  - Frequency of use of mobile money (Q16 and Q17)
  - Literacy in Kiswhahili (Q18) and Literacy in English (Q19)

The 5 target columns are:

[18] Data from the FSDT Finscope 2017 survey. E.g. : House location, Education, Income sources, Sex, Age,…
[19] C.f. Annex 1: FSDT Finscope 2017

- The multiclass output that we will have to predict (Mobile_money_classification):

Labels	Other Financial Services	
Mobile Money	NO	YES
NO	0	1
YES	2	3

- The corresponding decomposition of the output in the following 4 sub-labels:
  - Use mobile money: Yes / No
  - Has savings: Yes / No
  - Has borrowing: Yes / No
  - Has an Insurance: Yes / No

The unlabeled dataset, or test set, includes 2365 observations over the same 32 features columns, without the target columns.

Finscope Survey 2017 – Visualization and Insights

The features of the dataset have been analyzed to identify data types, outliers, missing data, correlation between features,… The details of these analysis are available in the Jupyter Notebook[20]

Some key insights on the features have extracted from this exploration phase:

- There is no missing data and all the data registered in the dataset is reasonable (e.g. no impossible ages, no unknow categories,…)
- Most of the features are categoricals[21]. (e.g. level of literacy in English)
- No important correlations between features have been identified. This is mainly due to the fact that the majority of the features are categorical.
- Some features are unbalanced and have categories with few observations. We will thus apply some feature-engineering in the pre-processing phase.

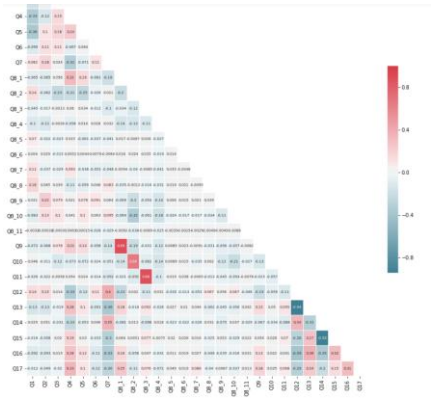


Figure 14 - Correlation Matrix

[20] https://github.com/alexbieb/Capstone_Project_Multilabel_Classification

[21] Cf. Finscope Survey 2017 - Description

The targets of the data set have also been analyzed to identify distribution characteristics, correlation with particular features, geographical patterns, …

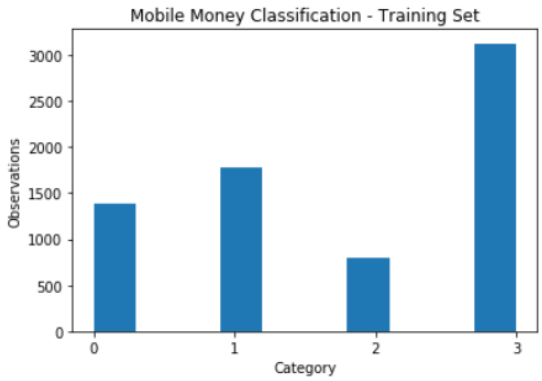


Figure 15 - Targets in the Training Set

- By plotting the distribution of the targets in the training set we identified that they are unbalanced. The class 3 (access to both mobile money and another financial service) is overrepresented with 44% while the class 2 (mobile money but no other financial service) accounts only for 11% of the observations. This may need some actions (e.g. apply resampling techniques) in the modelling phase to avoid the model to overfit to the overrepresented classes[22].

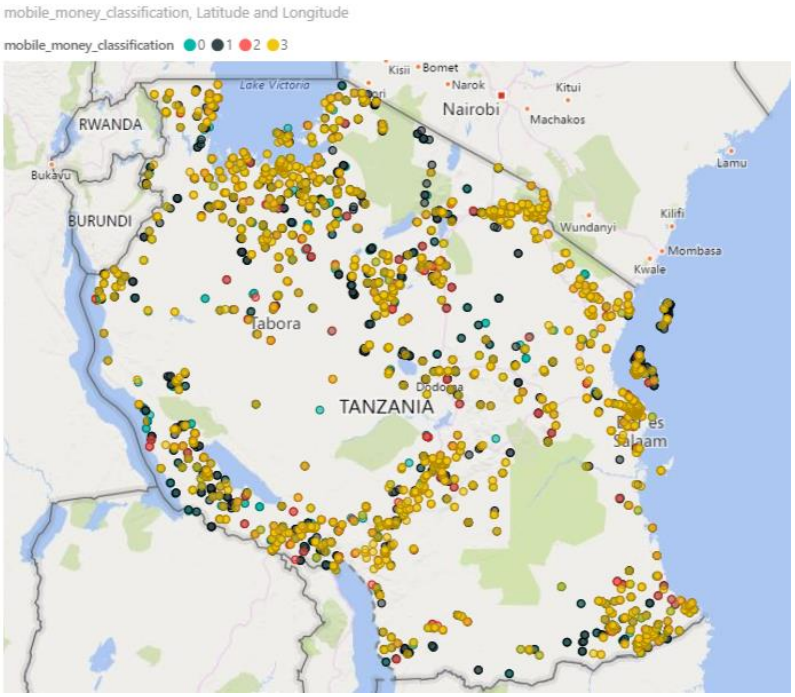


Figure 16 - Classification on Map - Power BI

- By plotting the targets on the Tanzanian map we can have a visual understanding of the classification problem and have a first sense of where the people do have financial access and where they don't.

[22] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

FSDT Financial Access Mapping: Geospatial mapping of all cash outlets in Tanzania in 2012[23]

FSDT Financial Access Mapping - Description

In addition to the dataset on individuals, and in order to enable us to understand the financial structure available in Tanzania, Zindi provides us 9 files containing the location and characteristics of the different points of sales of financial products (commercial banks, community banks, ATMs, microfinance institutions, mobile money agents, bus stations and post offices). We will mainly use them to identify the distances between the individuals and the closest point of sales of the financial products. Further analysis could be done to improve the input data of the model.

FSDT Financial Access Mapping - Visualization and Insights

- After consolidation of the 9 datasets, we analyzed their composition. The datasets contain information about:
    - 45,430 Mobile Money Sellers
    - 976 ATMs
    - 616 Commercial Banks
    - 294 Micro-Finance institutions

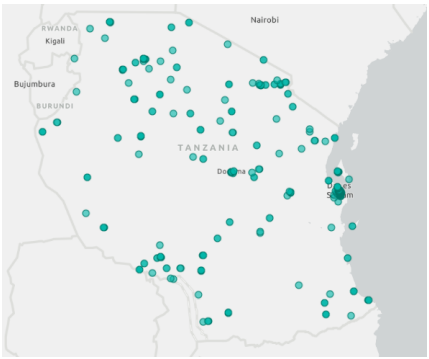


Figure 17 - Bank locations in Tanzania

- After mapping and analyzing the Financial Access datasets, we observe that the commercial banks, ATMs and Micro-Finance institution are very concentrated in the urbanized areas. On the contrary, the Mobile Money Sellers are way more distributed in all the country.

ArcGis platform: a tool to use maps and obtain regional demographic data[24]

ArcGis platform - Description

[23] Data from the FSDT. E.g. :Commercial banks, community banks, ATMs, microfinance institutions, mobile money agents, bus stations and post offices

[24] Data from the Esri's ArcGIS Technology

Zindi works in collaboration with the ArcGis platform and provides a limited access to the participants of the competition to enable them to gain insights on the Financial and demographic structure of Tanzania. The platform allows for example to calculate distances, make clusters, add layers of economic or geographic information,…

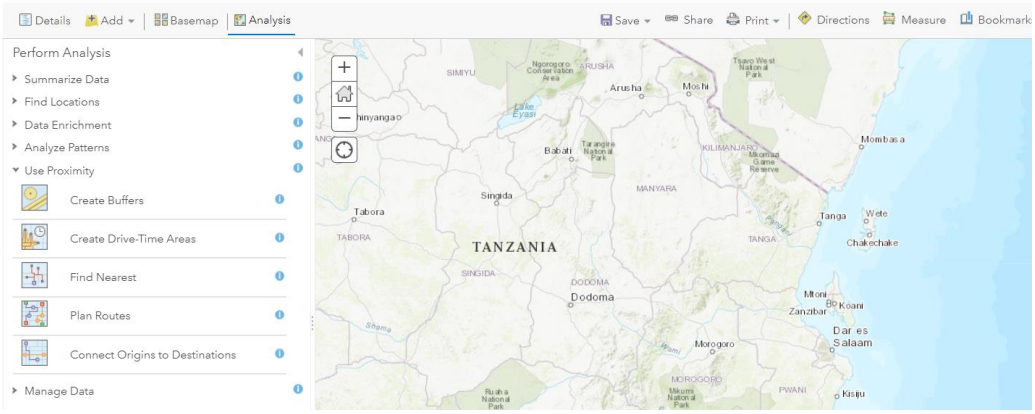


Figure 18 - ArcGis Platform

We decided to use the credits available on the ArcGis platform to:

- Measure the distances between each individual in the datasets and their closest mobile money retailer and closest commercial bank. Those additional features will be added to the training, testing and validation sets assuming they could impact directly the probability of having or not access to the financial system.
- Identify the region of each of the individuals in the dataset, based on their latitude and longitude, in order to add regional information related to the financial inclusion (percentage of access to mobile money, average income,…)

More information could have been measured and extracted from the ArcGis platform but it requires additional accesses provided for the professional accounts.

ArcGis platform - Visualization and Insights

After calculation on the ArcGis platform, we extracted the new datasets containing the distances and the regional information and used them to enrich our training, validation and testing set.

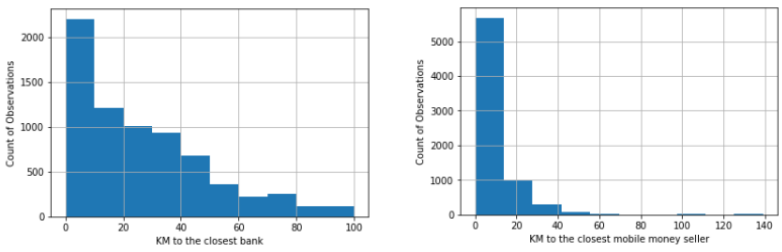


Figure 19 - Distances to banks and mobile money

By plotting the distribution of the distances between each individuals and the closest bank and mobile money reseller we can see that the large majority lives within 10km of a mobile

money reseller while more than 50% of the population lives further than 10km of a commercial bank. These new continuous features will be transformed into categorical variables in the pre-processing phase.

Algorithms and Techniques

Supervised Learning – Scikit-Learn

The Python library "Scikit-learn" provides a large list of pre-implemented algorithms for multiple types of problems.

We will select algorithms applicable for supervised learning multilabel classification[25]:

- Decision Tree
- Support Vector Classifier
- Random Forest
- Logistic Regression
- …

All of these algorithms can perform multilabel classification and we will se during the modelling and validation phase which ones perform well with our datasets.

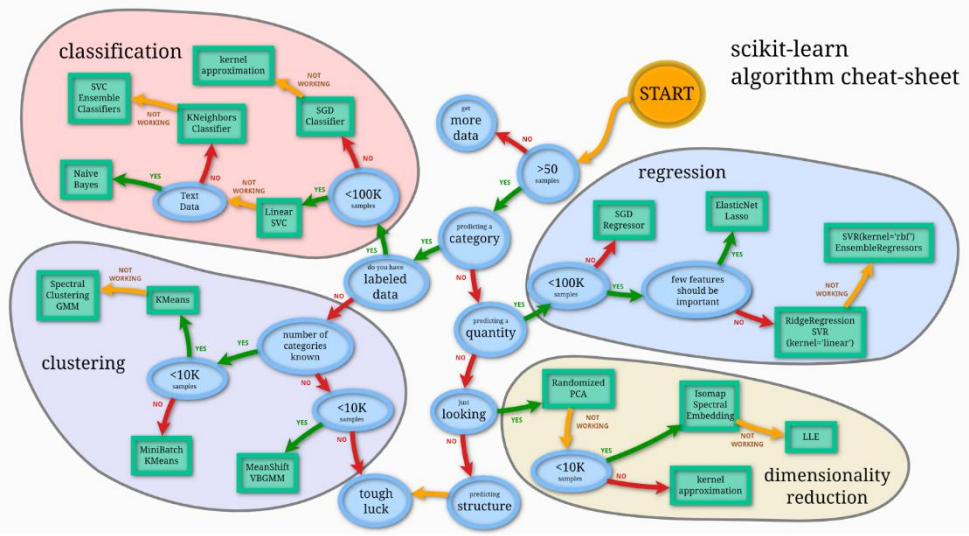


Figure 20 – Scikit-learn Algorithm cheat-sheet

Deep Learning – Keras

After implementation and optimization of supervised learning algorithms we will compare their results with a simple deep-learning model using the "Keras" library[26].

[25] https://scikit-learn.org/stable/modules/multiclass.html
[26] https://www.tensorflow.org/tutorials/keras/basic_classification

Benchmark

In order to have a reference point before designing any model, we start by uploading pseudo-random predictions. This will allow us to ensure that the models we train afterwards have reasonable scores.

- Baseline 1 : 0.25 probability for all 4 classes
- Baseline 2: Probability of each class in the training set

	Logloss	
Baseline	Training Set	Test Set
1	1.386	2.249
2	1.700	2.127

The Baseline 2 for on the test set obtains a Logloss value of 2.127.

Figure 21 - Zindi Scoring on Baseline 2

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

Before entering the modeling phase, we will preprocess our datasets in order to improve the information they contain and format it in a way that will enable the models to use it properly as input data.

In order to enrich the data and modify some features according to the observations of the exploration phase, the following steps have been done:

- Create new features based on the initial datasets and ArcGis Plateform:
    - Calculation of the distance between each individual and the closest bank and Calculation of the distance between each individual and the closest mobile-money reseller.
    - Identification of the region of each individual and addition of regional statistics such as average purchasing power, household size, density of population, …
- Aggregate the observations that were in bins with few observations, in order to reduce the effects of minor observation errors[27]. (E.g.: The Land-ownership feature transformed from 6 categories with few observations to 2 main categories (1: owns a land, 0: doesn't own a land))
- Transform the continuous variables to categorical ones by grouping them in the quartiles they belongs to[28]. (E.g.: Age and distance)
- Drop Features[29] that make no or low sense for the training of the algorithms. (E.g.: ID number, Latitude and Longitude)
- Test Dimensionality Reduction – PCA[30] to visualize the dataset and evaluate if the model trains better on a lower dimension with less features and thus less redundancy between them[31].

[27] https://en.wikipedia.org/wiki/Data_binning, https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63
[28] https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b
[29] https://scikit-learn.org/stable/modules/feature_selection.html
[30] https://en.wikipedia.org/wiki/Principal_component_analysis
[31] https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-classification-outcome-6e9ba21f0a32

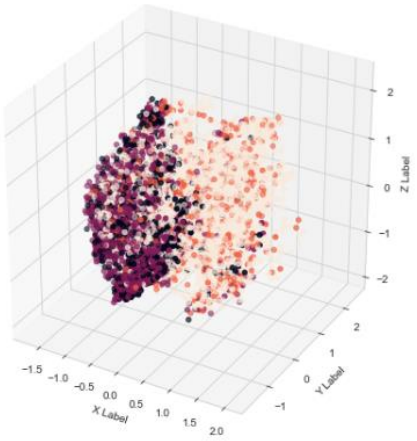


Figure 22 - Graph of training set reduced to 3 dimensions with PCA

In order to have a clean data set, that the algorithms can use for the training and prediction, the following steps have been done:

- Transform the categorical variables to multiple columns with binary information (Get_dummies[32])
- Scale[33] the data of some features to avoid having feature with too much impact on the classification for algorithms using distance classification.
- Splitting the datasets into training, validation and test sets.
- Splitting the training and validation sets into features and targets.
- Uniformize the training set and the test set columns.
- Apply the preprocessing steps to the test set.

Implementation & Refinement

After the upload of the baseline, we started the modelling phase of the project. Therefore, we implemented and refined different models and used a variety of techniques on the dataset to improve our predictions. We tested multiple supervised learning models, we then optimized them using GridSearch, we used a pipeline optimizer and finally implemented a Deep-Learning model.

Unoptimized Model Evaluation

We train a large selection of classification algorithms with their default parameters (defined by scikit-learn) and evaluate their predictions accuracy on our validation set. This give us a first idea of which algorithms are performing correctly on our dataset.

Model	Mean Accuracy	Std Dev Accuracy
LogisticRegressionCV()	66,38%	1,29%
LogisticRegression()	66,20%	1,42%
RidgeClassifier()	65,90%	1,23%
RidgeClassifierCV()	65,90%	1,33%

[32] https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html
[33] https://en.wikipedia.org/wiki/Feature_scaling

LinearDiscriminantAnalysis()	64,97%	2,00%
BernoulliNB()	62,92%	1,65%
RandomForestClassifier()	62,80%	1,78%
ExtraTreesClassifier()	62,20%	1,91%
DecisionTreeClassifier()	56,11%	1,64%
ExtraTreeClassifier()	55,05%	1,64%
GaussianNB()	53,98%	2,10%
SVC()	44,51%	1,55%
KNeighborsClassifier()	39,66%	1,63%
LinearSVC()	39,48%	14,12%

After this first evaluation we decided to keep a limited number of models for further parametrization and only trained the 7 bests of this first round.

By submitting the results of the predictions of these not optimized algorithms we obtain a best prediction with Random Forest. The Random Forest model with its default parameters improves our score to a Logloss of 1.456 on the testing set.

The random forest algorithm is a machine learning algorithm that can be used for both regression and classification and that *"builds multiple decision trees and merges them together to get a more accurate and stable prediction".* [34]

FILE	COMMENT	SCORE
predictions_6_rf_1.csv	6 - First Rf model	1.45619891898039

Figure 23 -Zindi Scoring on First Models – RandomForest

Model optimization - GridSearch[35]

In order to optimize the parameters optimization of the selected models, Scikit-learn provide us an automated way of optimizing them: GridSearch. GridSearch is a tool that will iterate through all the parameters chosen and evaluate the performance of the model on each of their combination. After the optimization of the selected models, we obtain a better score on a SVC model.

SVC (Support Vector Classification) is a type of Support Vector Machine. Support Vector Machines are algorithms that classify into separate categories by trying to keep a gap between categories as wide as possible[36]. The parameters selected by GridSearch for our SVC[37] model are the following: {'C' (Penalty): 10, 'gamma' (Kernel coefficient): 0.01, 'kernel' (Kernel type): 'rbf'}.

[34] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd
[35] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[36] https://en.wikipedia.org/wiki/Support-vector_machine
[37] https://scikit-learn.org/stable/modules/svm.html

The submission of the predictions of the SVC model with parameters optimized by GridSearch improves our score to a Logloss of 1.372 on the testing set.

Figure 24 - Zindi Scoring on a model Optimized with GridSearch

Pipeline optimization - TPOT[38]

We used an Automatized Pipeline creation to go further in the model optimization. A python library is designed to go through different pipeline structure and optimize it: TPOT[39]

TPOT will iterate through an extensive list of algorithms and techniques and combine them to evaluate the ones that perfom the best on the provided dataset.

```
Best pipeline: LogisticRegression(MultinomialNB(SelectFwe(ZeroCount(VarianceThreshold(DecisionTreeClassifier(MinMaxScaler(in
put_matrix), criterion=gini, max_depth=5, min_samples_leaf=14, min_samples_split=15), threshold=0.001)), alpha=0.039), alpha
=0.001, fit_prior=True), C=0.1, dual=False, penalty=l2)
```

Figure 25 - Best pipeline TPOT

	Logloss	
TPOT	Training Set	Validation Set
LR - NB - DT	0.7141	0.658

The submission of the predictions an optimized pipeline through TPOT improves our score to a Logloss of 1.368 on the testing set.

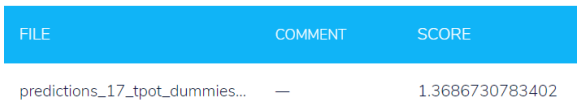


Figure 26 - Zindi scoring on a TPOT pipeline

Deep Learning:

After some basic transformation of the datasets to be able to be trained with Deep Learning, we modeled a Deep Learning structure with Keras[40]. We trained it on the Kaggle plateform to use their online GPUs[41].

Deep Learning Model with one hidden layer[42]:

[38] http://epistasislab.github.io/tpot/using/
[39] https://towardsdatascience.com/tpot-automated-machine-learning-in-python-4c063b3e5de9
[40] https://keras.io/
[41] https://hackernoon.com/learn-deep-learning-with-gpu-enabled-kaggle-kernels-and-fastai-mooc-72fee41bb4b5
[42] https://machinelearningmastery.com/multi-class-classification-tutorial-keras-deep-learning-library/

```
# Deep Learnig Function
def deepml_model():
    # Model Creation
    deepml = Sequential()
    deepml.add(Dense(8, input_dim=87, activation='relu'))
    deepml.add(Dense(8, input_dim=8, activation='relu'))
    deepml.add(Dense(4, activation='softmax'))
    # Model Compilation
    deepml.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
    return deepml
```

```
estimate = KerasClassifier(build_fn=deepml_model, epochs=200, batch_size=5, verbose=0)
```

Figure 27 - Deep-learning model

The submission of the predictions of the Deep Learning model improves our score to a Logloss of 1.362 on the testing set.

FILE	COMMENT	SCORE
predictions_deeplearning.csv	—	1.36229994429636

Figure 28 - Zindi Scoring on a simple Deeplearning structure

Additional techniques

During the training phase, many other algorithms and parameters have been tried, the above-mentioned algorithms are only the ones that impacted the most the improvement of the score on the Zindi platform. Here after are some of the other techniques that have been applied without having significant improvement on the LogLoss score:

- Dimensionality reduction
- Reducing the number of input features
- Modifying the target variables using hierarchical classification (e.g. predicting first if they have mobile money or not and then predict if they have access to another financial service or not)
- Calibration of the prediction probabilities by applying CalibratedClassifer.[43]
- Setting parameters of algorithms to make them do classifications using One-vs-Rest or One-vs-All.[44]

[43] https://scikit-learn.org/stable/modules/calibration.html
[44] https://en.wikipedia.org/wiki/Multiclass_classification

IV. Results

Model Evaluation and Validation

Confusion Matrix

To evaluate the models during the training phase we calculated the precision and the Logloss and we also plotted confusion matrix in order to visualize more in details the correct and incorrect predictions.

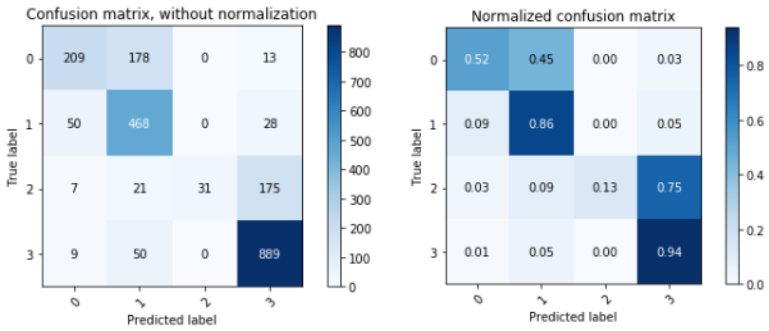


Figure 29 - Confusion Matrices on training set

As a reminder the targets are the following:

Labels	Other Financial Services	
Mobile Money	NO	YES
NO	0	1
YES	2	3

Plotting the predictions on confusion matrix provides us interesting insights about our models:

- True labels 0 or 1 (without mobile money) are barely never predicted as 2 or 3 (with mobile money) and vice-versa. This means that the models manage to identify the division between the individuals that have mobile money and those who don't but struggle to identify people having access to other financial services.
- We notice that the targets labelled 1 and 3 are well predicted with in the observed case respectively 86% and 94% of items correctly predicted. Targets 0 and 2 however are poorly predicted with respectively 52% and 13% of correct predictions. Target 0 is frequently predicted as 1 and target 2 is predicted as item 3. This means that the model tends to predict the class that is the most represented in the dataset for its classification between. It could be caused by the overrepresentation of the classes 1 and 3 in the training set.

Justification

Zindi Ranking

Our best model on the Zindi plateform is the Deep-Learning model with a Logloss of 1.362 on the visible testing set and ranked 48th of 72 ranked participants on 23rd of April 2019.

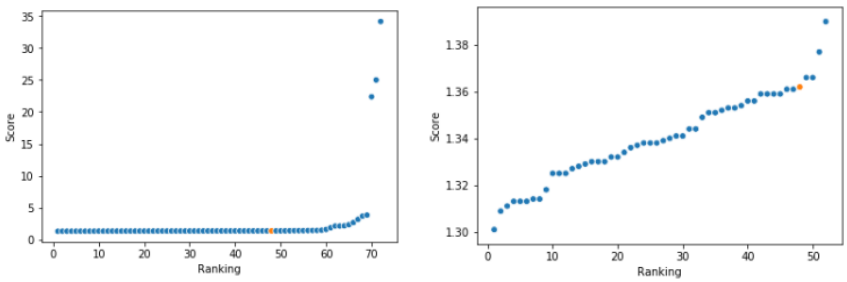


Figure 30 - Ranking and Logloss of participants

The leader of the competition has a Logloss of 1.300[45]

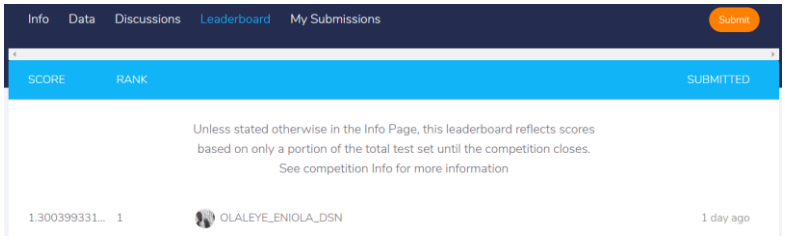


Figure 31 - Leader of the Challenge on 23rd of April

Ranking Evolution

We can evaluate our ranking in the competition with a Logloss of 1.362 as reasonable given that the Logloss values of the other participants range from 35 to 1.3. But there is a clear opportunity for improvement given that 65% of the ranked participants made better scores with their models.

Personal or Competitor	Model or Ranking	Logloss Zindi Testing Set
Competitor	72nd Model Zindi	34.159
Personal model	Baseline	2.127
Personal model	Unoptimized	1.456
Personal model	GridSearch	1.372
Personal model	TPOT	1.368
Personal model	**48th Deep Learn.**	**1.362**
Competitor	36th Model Zindi	1.352
Competitor	1st Model Zindi	1.300

Figure 32 - Scoring Improvement and Competitors scoring

[45] https://zindi.africa/competitions/mobile-money-and-financial-inclusion-in-tanzania-challenge/leaderboard

V. Conclusion

Free-Form Visualization

For the resolution of this challenge, one of the main issues has been to find features that can help us classify the individuals in their correct category. The famous "Garbage in, garbage out"[46] straightforwardly tells us that if we don't have clean and meaningful data as input of our model if will be difficult to have good predictions.

Geographical Feature

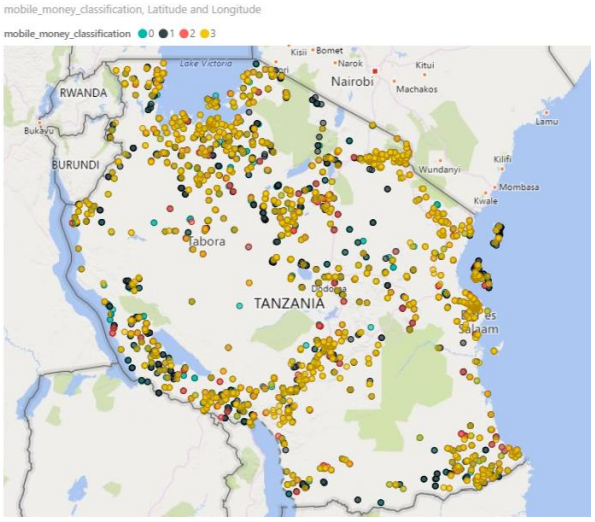


Figure 33 - Classification Targets on the Map of Tanzania

In this first visualization of the targets on the map of Tanzania we can see that there are some patterns related to the geographic location of the individuals (e.g. the areas with low density of observations have higher proportion of targets without access to financial services.) However, at first sight these patterns don't allow us to make a clear separation between the targets. That's why we performed calculation to identify the distances between each individual and finance providers.

Dimensionality Reduction

Location is far from being the only feature, after modification and creation of additional features in our preprocessing phase, we obtained more than 100 features. In order to visualize the main variations in the data set we reduced the dimensionality to 3 dimensions using PCA (Principal Composant Analysis).

[46] https://en.wikipedia.org/wiki/Garbage_in,_garbage_out

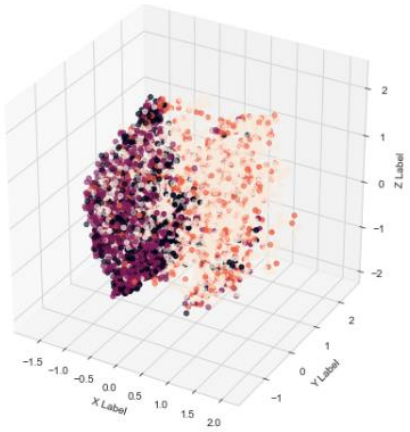


Figure 34 - Training Set features reduced to 3 dimensions - PCA

When we plot the features of the training set after application of PCA to 3 dimension (X,Y and Z) and using colors to differentiate the labels, we can identify groups of data with the same labels but some of them remains very mixed and seems difficult to separate.

Feature Selection

Finding features that help us separate clearly the dataset between labels is one of the most challenging part of the data science process.

Reflection & Improvement

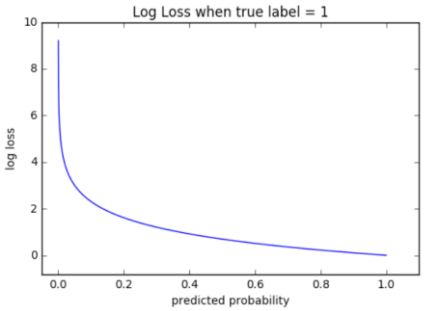
The fact that our model ranks below 65% of the participants of the competition raise the question on the areas of improvement that we could apply to improve it (what additional or different step have done the other competitors to rank better?). Some reflections and improvement opportunities have been identified:

Data Enrichment and Engineering

- The information provided by ArcGis could be used in more details and could help find other features that could improve our predictions. For example calculation the distance between each individual and the closest ATM, SACCO,… Infrastructure such as roads, mobile network coverage, bus stops,… could also be computed and integrated as features
- On the contrary some other features could have been dropped or transformed. For example we could have tried to build a feature that takes into account the profession, age, source of incomes,… to define if the person could have the need to borrow or save money.

Logloss Optimization

- While uploading our predictions to the Zindi platform we identified that the Logloss scoring can be largely impacted by too confident values[47]. We calibrated some of our predictions in order to improve the Logloss score[48] but some other techniques may be investigated to optimize the scoring on Logloss.



- Some algorithms have been optimized using a scoring function different than Logloss (e.g. Accuracy). We could have developed a home-made scoring function that matches the Logloss to train these algorithms that don't have Logloss as predefined scoring function.

Deep Learning Optimization

- In this project we focused on traditional supervised learning algorithms and at the end of the project we compared the performances with a Deep Learning prediction. It turns out that the Deep Learning model performed better so the next step will be to improve this Deep Learning model to try to reach even better results.

Conclusion

For this Data for Good Challenge we predicted the probabilities of Tanzanians to have access to the Financial system. We started by understanding the context and analyzing the structure of the datasets available. We then added new features and prepared the data to enter the Machine Learning models. Different models and techniques have been tried and at the end the Deep Learning model bet all classical supervised learning algorithms and gives results far more precise than the pseudo-random baseline.

This competition is the perfect example that Machine Learning can be used for the best and that the Data Science community is ready to collaborate and to challenge itself to bring help on real-world social issues.

[47] http://wiki.fast.ai/index.php/Log_Loss
[48] https://scikit-learn.org/stable/modules/calibration.html

Annex 1: FSDT Finscope 2017

Variable ID	Question	Values
ID	Unique respondent ID	<String>
Q1	Age	<Number>
Q2	gender	1 Male 2 Female
Q3	Marital status	1 Married 2 Divorced 3 Widowed 4 Single/never married
Q4	Highest level of education completed?	1 No formal education 2 Some primary 3 Primary completed 4 Post primary technical training 5 Some secondary 6 University or other higher education 7 Don't know
Q5	Which of the following applies to you? Read out; Single response	1 You personally own the land/plot where you live 2 You own the land/plot together with someone else 3 A household member owns the land/plot 4 The land/plot is rented 5 You don't own or rent the land 6 Don't know
Q6	Do you personally own land (other than the land you live on) that you have land certificates of ownership for?	1 Yes 2 No
Q7	Do you personally own a mobile phone?	1 Yes 2 No
Q8_1 through Q8_11	Different people have different ways of getting money, please tell me how you get the money you spend? Multiple mention possible	
Q8_1	Salaries/wages	1 Yes 0 No
Q8_2	Money from trading/selling Anything you produce/grow/raise/make/collect with the intention of selling	1 Yes 0 No
Q8_3	Money from providing a service – i.e. such as transport, hairdressing, processing, hospitality services (food & accommodation)	1 Yes 0 No
Q8_4	Piece work/Casual labor/Occasional jobs	1 Yes 0 No
Q8_5	Rental income	1 Yes 0 No
Q8_6	Interest from savings, investments, stocks, unit trusts etc.	1 Yes 0 No
Q8_7	Pension	1 Yes 0 No
Q8_8	Social welfare money/grant from Government	1 Yes 0 No
Q8_9	Rely on someone else/others to give/send me money	1 Yes 0 No
Q8_10	Don't get money – someone else pays my expenses	1 Yes 0 No
Q8_11	Other	1 Yes 0 No
Q9	Only for those who said they get a salary/wages. Who do you work for?	-1 not applicable 1 Government 2 Private company/business 3 Individual who owns his own business 4 Small scale farmer 5 Commercial farmer 6 Work for individual/household e.g. security guard, maid etc. 7 Other

Q10	Only for those who said they get money from selling things – what kind of things do you MAINLY sell (get most money from)?	-1 not applicable 1 Crops/produce I grow 2 Products I get from livestock 3 Livestock 4 Fish you catch yourself/aquaculture 5 Things you buy from others – agricultural products 6 Things you buy from others – non-agricultural products 7 Things you make (clothes, art, crafts) 8 Things you collect from nature (stones, sand, thatch, herbs) 9 Things you process (honey, dairy products, flour) 10 Other
Q11	Only for those who said they get money from providing a service – what kind of services do you MAINLY provide (get most money from)?	-1 not applicable 1 Personal services (hairdressers, massage, etc.) 2 Telecommunications/IT 3 Financial services 4 Transport 5 Hospitality – Accommodation, restaurants, etc. 6 Information/research 7 Technical – mechanic, etc. 8 Educational/child care 9 Health services – traditional healer etc. 10 Legal services 11 Security 12 Other, specify
Q12	In the past 12 months, have you sent money to someone in a different place within the country or outside of Tanzania?	1 Yes 2 No
Q13	When did you last send money?	-1 not applicable 1 Yesterday/today 2 In the past 7 days 3 In the past 30 days 4 In the past 90 days 5 More than 90 days ago but less than 6 months ago 6 6 months or longer ago
Q14	In the past 12 months, have you received money from someone in a different place within the country or from outside the country?	1 Yes 2 No
Q15	When did you last receive money?	-1 not applicable 1 Yesterday/today 2 In the past 7 days 3 In the past 30 days 4 In the past 90 days 5 More than 90 days ago but less than 6 months ago 6 6 months or longer ago
Q16	In the past 12 months, how often did you use mobile money for purchases of goods and/or services?	-1 not applicable 1 Never 2 Daily 3 Weekly 4 Monthly 5 Less often than monthly
Q17	In the past 12 months, how often did you use mobile money for paying your bills?	-1 not applicable 1 Never 2 Daily 3 Weekly 4 Monthly 5 Less often than monthly
Q18	Literacy in Kiswhahili	1 Can read and write 2 Can read only 3 Can write only 4 Can neither read nor write 5 Refused to read

Q19	Literacy in English	1 Can read and write 2 Can read only 3 Can write only 4 Can neither read nor write 5 Refused to read
Latitude	Approximate latitude	Number
Longitude	Approximate longitude	Number
Mobile_money	Do you use mobile money?	1 Yes 0 No
Savings	Do you save?	1 Yes 0 No
Borrowing	Do you borrow?	1 Yes 0 No
Insurance	Do you have insurance?	1 Yes 0 No
Mobile_money_classification		0 no mobile money and no other financial service (saving, borrowing, insurance) 1 no mobile money, but at least one other financial service 2 mobile money only 3 mobile money and at least one other financial service