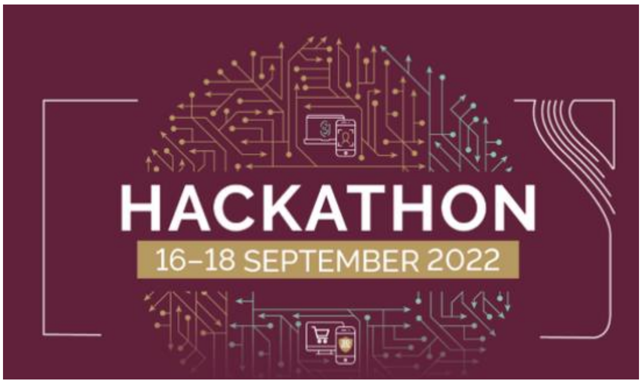

OVERVIEW

Capitec's vision is to simplify banking for our clients to live better. One way in which we're trying to do this is by ensuring that any interactions with our clients are customised resulting in a personalised client experience. The data science challenge will focus on identifying the most relevant next step that should be communicated to clients.

DATA SCIENCE CHALLENGE

We want to know which of the potential interactions are the most relevant to clients. In the dataset, nine specific interactions with clients have been identified. To simplify the challenge, the overlap between groups have been reduced, so most clients in the dataset will only have one main interaction into which they can fall.

The challenge will be to identify the right interaction for different clients. The effectiveness of how well any solution can predict the right interaction will be evaluated against the Evaluation dataset which represents the general population rates.

THE DATA

Two separate datasets are provided, the *ClientBehaviourTraining* dataset contains a balanced dataset to simplify model/process building. This file should be used for the first round of training and testing of any process. The second file, *ClientBehaviourEvaluation*, contains the normal (unsampled) data which represents the normal rates of behaviour change, this should be used as the final evaluation of how well models perform. The evaluation dataset has OOT added to the behaviour type name.

As an example, in the training dataset for clients taking up the app or not there will roughly be an equal amount of entries (50%), however the normal take-up rate is only 12% (3,067/25,000).

For each behaviour type there are four different groups which are split across the two different files. The training dataset contains entries without the OOT label added:

- *Training dataset - AppTake-up*: clients who successfully changed behaviour. This will be the target (y=1) variable. Sampling has been applied in order to simplify the building part of any model/process.
- *Training dataset - AppNoTake-up*: clients who did not change behaviour. This will be the target (y=0) variable. Sampling has been applied in order to simplify the building part of any model/process.
- *Evaluation dataset - AppTake-upOOT*: Model performance testing group. Clients who changed behaviour, but no sampling applied.

- *Evaluation dataset - AppNoTake-upOOT:* Model performance testing group. Clients who did not change behaviour, but no sampling applied.

BehaviourType	Take-up	NoTake-up	Take-up OOT	NoTake-up OOT
App	25,000	25,000	3,067	21,933
Banking	25,000	25,000	1,146	23,854
CardUser	25,000	25,000	3,029	21,971
CreditCard	24,998	24,962	1,057	24,837
DebitOrder	23,684	25,000	1,100	23,766
FinancialEducation	24,928	24,927	2,182	24,683
Loan	24,527	24,793	2,059	22,669
Saving	24,914	24,999	1,369	24,829
VirtualCard	24,969	24,854	1,208	28,745

BehaviourType Descriptions

- **App** – Clients did not have the Capitec app, based on different interactions with the bank some clients did take it up and start using it (take-up group), while some took it up, but haven't used it, or didn't download the app (control).
- **Banking** – Moved bank account to Capitec.
- **CardUser** - Clients who started swiping their Capitec cards, previously very little/no card swiping behvaiour.
- **CreditCard** – Clients who took up the Capitec Credit card compared to those who did not.
- **DebitOrder** - Clients who moved debit orders (e.g. Insurance deductions) to their Capitec account compared to those who did not.
- **FinancialEducation** – Clients identified as requiring support with financial education, e.g. through the Live Better Academy.
- **Loan** - Clients who took up a Capitec Loan offer compared to those who did not. Only clients with a stable income and bureau profile are selected to be in this group.
- **Saving** - Clients who started using specific (fixed) savings accounts with Capitec.
- **VirtualCard** – Bank safety interactions to position the benefits of using a virtual card for online purchases compared to a debit/credit card.

MEASURING SUCCESS

While some errors do carry a higher cost/impact, e.g. recommending a credit card if we should have sent information on financial education is worse than recommending the credit card rather than the loan product. However, to simplify the challenge all prediction errors will be viewed as equal.

Below a great read on understanding different metrics for multiclass problems.
https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd

Data Dictionary in excel document