# Executive Summary – Direct Marketing Optimization

## 1.1 Summary

The purpose of this report is to establish the optimal targeting strategy for the placement of three financial products (a Mutual Fund (MF), a Credit Card (CC), and a Consumer Loan (CL)) to a set of KBC customers, constrained to 100 final targets.

The optimal mix, given the limitations to be discussed in subsequent sections, entails targeting 45 customers for CL, 37 for CC, and 18 for MF. This is done with the consideration that CL and CC enjoy both higher average revenues and predictive accuracy. Estimated expected revenues, weighting average purchase per customer by the effectiveness of the model at prediction, amount to 887€.

## 1.2 Project overview and methodology

To tackle this problem, two main variables had to be estimated:

1) The propensity of each customer to purchase a given product; and
2) The expected revenue generated by the purchase of a given product.

The first task is a classification problem, where the goal is predicting the probability of a customer buying a product (propensity). The second is a regression problem, but for reasons to be explained in the subsequent sections, was tackled by weighting the average revenues per product by the predictive power of the model used to estimate propensity.

The steps involved in this project are the following:

- Exploratory data analysis: aimed at understanding the data at hand, which features are correlated with which others. This is done to guide the subsequent steps and identify areas where the data is in need of cleaning or enrichment before feeding it to any model.
- Feature engineering: in this step, the data is cleaned of spurious observations which could bias algorithms, missing values are managed, and new variables are created by combining existing ones (this step can be iterative by measuring its interactions with the subsequent one).
- Modeling: once the data is deemed ready, the next step is developing models to tackle the problem. Multiple ones are tried and the best performing employed. To make sure the models can effectively predict the target variables, they are trained (e.g., developed and fine-tuned) on a part of the data, whereas they are tested (e.g. assessed) on a different one, so that their capability to generalize can be reliably assessed.

## 1.3 Model interpretation and core information

The models employed to scoring propensity are a set of three XG Boost classifiers. In this sort of problems, there are three main variables of interest in evaluating the predictive power of a model:

1) Accuracy: out of all the predictions of the model, how many were correct
2) Precision: out of all the **positive** predictions of the model, how many were correct
3) Recall: out of all the examples in the data of the positive class, how many did the model identify

Using all three can be important when one of the classes (e.g., whether a customer will buy a MF) is rarer than the other: in this case accuracy can be misleading (if 1 customer out of 10 purchases the product, and the classifier says “always no”, then it will be still 90% accurate, but practically useless).

To this end, precision and recall have to be taken into account and balanced one against the other. A very good classifier would be great at both, but this is not often the case in reality. Favoring one versus the other is akin to casting a fishing net: a large net (focusing on recall) will catch more fish, but also more garbage. A smaller one (focusing on precision), instead, will catch fewer fish, but also much less garbage.

In this project, a balance had to be struck. Given the constraint of targeting only 100 customers, it was clear precision was the metric of most interest, as it is preferable to have few good prospects than to have many less-reliable ones. Considering the three target variables, the models range from 87% precision (for CL), to 73% for CC and 70% for MF. Recall is much lower, but of less interest.

Expected revenues per targeted customer have been computed by multiplying the average revenues generated by each product times the precision of each model. E.g., for MF: 9.7€ * 70% = 6.8€. This was then summed for each customer targeted for each product.

Customers have been targeted by prioritizing on 1) profitability of each product (in terms of average revenues), and 2) predictive power of each model. CL was favoured on both metrics. It was assumed that all three products had to be promoted. Otherwise, a sample of 100 prospects with high reliability belonging to only, or mainly CL, could be produced. This would drive higher average revenues.

Considering targeted customers, CL & CC prospects tend to have higher current account balances than average (the inverse is true for MF). Tenure is higher for all save CC. MF & CC prospects tend to have higher credit volumes, while CL exhibits lower ones. Debit volumes are higher for all categories.

## 1.4 Limitations and possible next steps

There are two main limitations of the proposed solution, one practical and one theoretical:

- Revenue estimation: the current procedure is fairly simplistic, and a better model could potentially be created. This choice was mainly driven by timing constraints; and
- Scalability: while a "small fishing net" was an optimal solution for a small set of prospects, were the targeting capacity to be expanded, the current system would struggle to produce high-quality prospects.

In terms of next steps:

- The revenue estimation has margin for improvement which should be acted upon.
- Depending on whether the amount of direct marketing capacity could increase, the models could be improved to increase recall.
- The "profile" of high-probability prospects could be shared with the marketing department to help them identify suitable targeting strategies for these customers: it could be used for a data-driven persona definition.