# Credit modelling challenge

Congratulations on reaching this stage of the interview process!

This is our technical challenge and you have 7 days to accomplish it. Please send the results to [v.brodski@atbank.nl](mailto:v.brodski@atbank.nl).

## Description of the challenge

The context of this challenge is a charitable organisation that is trying to target people effectively through direct mail and would like to understand how to best target people who are most likely to give to charity.

From a data science point of view, it consists of 2 modelling exercises:

1. The building of a binary classification model, predicting which people are more likely to donate to a charity
2. The building of a regression model, predicting the amount of money that each person is likely to donate

We are then interested to understand how we could combine the results of those 2 models to build a list of potential people to target. The goal is to maximise the amount of money that would be collected through a direct mail activity where letters would be posted to people in order to collect their donations.

We would be interested to understand how many people we should target if each envelope was costing us $5. What would happen if each envelope only cost $1?

## Expectations

We expect the following pieces of code in a jupyter notebook or equivalent, ideally written in python:

1. A set of data quality/data processing steps designed to clean the data prior to its use for the binary and regressions models
2. A binary classification model as described above
3. A regression model as described above
4. A function that creates a list of ideal people to target based on the price of each envelope

We expect a small presentation that would highlight:

1. The data processing activities you carried before the modelling activities
2. The model building process, variable selection, performance, validation
3. The final recommendation in terms of the number of people who should be targeted in the campaign if the cost of each envelope is $5 or $1. Alongside the number of people to target, we would like to understand how much money you think the charity will be able to collect from this campaign.

What we assess:

* Analytical skills (your ability to perform a deep dive into the data set, slice and dice in order to find pockets of data that perform particularly well or badly, define top donors or those donating without patterns)
* Python proficiency
* Presentation skills (how much time and effort is required to understand the results)

## Datasets

The datasets are located here:

<https://www.kdd.org/kdd-cup/view/kdd-cup-1998/Data>

Specifically, you will want to use:

* Data dictionary: <https://www.kdd.org/cupfiles/KDDCupData/1998/cup98dic.txt>
* Training dataset: <http://kdd.org/cupfiles/KDDCupData/1998/cup98lrn.zip>
* Validation dataset: <http://kdd.org/cupfiles/KDDCupData/1998/cup98val.zip>
* Validation targets: <http://kdd.org/cupfiles/KDDCupData/1998/valtargt.txt>

## Is it credit risk?

Not really, but you can think about donation events as PDs and amounts as LGDs 😊