# 1 Introduction

This report is part of the syllabus the course Introduction to Programming in the Master of Data Science and Advanced Analytics at Nova IMS and has the purpose to demonstrate programming skills in Python in a Data Science related manner.

The data set used for hour project, further described later (cf. XXX) contains sample data of bank customers of the year Y0. One of the variables is the binary variable “exited” indicating whether a customer is or no longer is a customer after time T1.

For the purpose of this report the following fictional problem is introduced:

The marketing department of the bank has developed a campaign which prevents customers from leaving the company. This campaign is to be executed a data set from the current year Y1.

The goal for us, as aspiring Data Scientists, is to maximize the return on investment of the marketing campaign, that is predicting which customer is going to leave and which is not.

The marketing department specified the following parameters of the campaign:

* Cost of applying the campaign per customer: 150€
* Average customer value till T1: 450€,
* Average customer value from T1 till T2: 450€
* Moreover, applying the campaign to a customer which had not been leaving still has some positive effect on the customer’s likelihood to leave the bank in the future. This effect is quantified as: 10€
* Assumption at time T0: no costumer leaves till T1 (for simplicity)
* Customers who were prevented from leaving will stay till T2

Four cases can be derived from the above for the training data set:

1. If a customer leaves till T1 and no campaign is applied: loss of 450€ (money expected to be gained by T1)
2. If a customer does not leave and no campaign is applied: no loss or gain (everything happened as excepted till T1)
3. If a customer does not leave because the campaign is applied: gain of 200€ (no loss of 450€ by T1, a gain of 450€ because the customer is going to stay till T2, a loss for applying the campaign of 150€)
4. If a customer does not leave and was not going to leave but the campaign is applied: loss of 140€ (no loss of 450€ by T1, a loss for applying the campaign of 150€, a gain for the positive effect of the campaign of 10€)

Later, this information is used to evaluate the performance of different prediction models.

The remainder of this report is structured as following: first, the data set is described (cf. XXX), and different prediction models are introduced (cf. XXX). Afterwards the result of the models are compared (cf. XXX) of which the *two* best are further optimized and the best model is chosen (cf. XXX). At last a conclusion is drawn (cf. XXX).