Document

# Assignment 1: predictive modeling on tabular data

customers ending their subscriptions or contracts with their telecom service providers

In this assignment, you will construct a predictive model to predict churn of telco custemers. The data set is based on an old "classic" used by SPSS Modeller, with some modifications. The data set can be downloaded from here (https://seppe.net/aa/assignment1/data.zip). You will play "competitively" using this website (https://seppe.net/aa/assignment1/)

- The test set does not contain the target, so you will need to split up the train set accordingly into your own validation set. The test set supplied in the data is used to rank and assess your model on the competition leaderboard
- Your model will be evaluated using two metrics: profit @ top-20, and AUC. The reasons for this is to be in line with a more realistic setting. E.g. one can image data scientists in a team arguing to use AUC and optimize for that. However, as seen in the course, for this scenario, we also imagine management arguing that there is not enough budget (in terms of time and money) to contact a lot of people (or hand out a lot of promotions). Hence, they have come up with the following: based on the top-k would-be churners as predicted by your model, sum some proxy of "retained profitability" in case the customer was indeed a churner, or zero otherwise
- As a proxy of profitability, the feature average cost min was deemed to be a good value. Based on the size of the test set, k=20 was deemed to be a good choice. Hence, management cares about optimizing this metric
- Note that only about half of the test set is used for the "public" leaderboard. That means that the score you will see on the leaderboard is done using this part of the test only (you don't know which half). Later on through the semester, submissions are frozen and the resuls on the "hidden" part will be revealed
- Class imbalance and the peculiar evaluation metric adopted here will make for challenges to be overcome
- You will have received a password and your final group number through an email, which you need to make submissions
- The results of your latest submission are used to rank you on the leaderboard. This means it is your job to keep track of different model versions / approaches / outputs in case you'd like to go back to an earlier result
- The leaderboard will be frozen and the hidden results shown a few weeks before the deadline. You should then reflect on both results and explain accordingly in your report. E.g. if you did well on the public leaderboard but not on the hidden one, what might have caused this? The idea is not that you then step in and "fix" your model, but to learn and reflect
- Also, whilst you can definitely try, the goal is not to "win", but to help you reflect on your model's results, see how others are doing, etc.
- Your model needs to be build using R, Python (or Go, Rust, Julia or whatever you prefer as long as it involves coding). As an environment, you can use e.g. Jupyter (Notebook or Lab), RStudio, Google Colaboratory, Microsoft Azure Machine Learning Studio... and any additional library or package you want

The first part of your lab report should contain a clear overview of your whole modeling pipeline, including exploratory analysis (if any), preprocessing, construction of model, set-up of validation and results of the model: