# Home Credit Default Risk

## Capstone 1 Proposal, Springboard

Araks Stepanyan, 07/07/2018

Predicting whether the applicant will repay the loan is crucial and helps lending institutions make a decision of whether to give the loan or not. Nowadays machine learning exceeds the human-level performance when it comes to these kinds of problems. And in this project I am going to tackle this problem with state of the art machine learning techniques.

The data that will be used in this project comes from a Kaggle competition “Home Credit Default Risk”. The host is Home Credit Group, an international consumer finance provider with operations in 10 countries. The task is to predict how capable each applicant is of repaying a loan. Although this is a data of a single institution, the same problem is applicable to all the banks and lending institutions. So the techniques used here will not necessarily be specific to only this particular data set. The data is stored in 8 different .csv files. A simple download from Kaggle is enough to acquire the data. And the competition page, https://www.kaggle.com/c/home-credit-default-risk/data, describes in details what the common columns of the different files are. These common columns will be used to combined all 8 files into one big data set.

Here, it is important to note that credit default prediction problems are usually more concerned with a good prediction of whether the client will default and less concerned with trying to understand why the client defaulted. This means that it’s not necessary to explain the connection between the input and the prediction. Which in turn means that there is no restriction of using only those machine learning algorithms that are easy to interpret.

Considering the above, throughout the project a variety of algorithms will be used, ranging from simple Logistic Regression to complicated Deep Learning techniques. And most probably, the best performance will be reached by combining different models. But it’s important to note that when we are dealing with structured data, tree based models usually work really well. So the main focus will be on Gradient Boosting algorithms. And of course, a huge portion of the project will include feature engineering. Good feature engineering is a key to success in these kinds of problems.

To summarize, I am going to use machine learning techniques to predict whether the client will repay the loan at the time of application. A code and a slide deck will be shared on GitHub.