**Data Science Project Protocol**

*Author(s):*

*Nimrod Ken Dror*

*Lev Solodkin*

# Introduction

In today’s loan industry, there is a huge gap between the what financial institutes know about their loaning customers, their credit history and their ability to repay their debts. This leads to information gap that causes inefficiency in loan providing and high loan interest rates. This also prevents many adequate and potential customers from applying for a loan, thus causing financial loses to both sides.

The goal of this project is to create a, based on historical loan data, a Machine Learning based model that will aid financial institutes in the process of screening and approving potential customers.

This model will try to predict whether a user will be able to to repay his debt within the required timeframe or not.

Based on our current knowledge and assessment, we predict that the outcome will be influenced by some or all of the following factors:

* Employment status
* Gender
* Marital status & number of kids (if any)
* Age
* Past loans
* Annual income from salary and other sources
* Assets owned by the client

Apart from the above stated factors, we believe that there may be additional factors that have to be considered in order to gain an accurate prediction regarding the possibility that a loaner will default.

In order to answer this question we will perform Machine Learning based analysis on a dataset of actual users and loan history provided by Home Credit Group, as provided by [Kaggle](https://www.kaggle.com/c/home-credit-default-risk/).

# 

# 

# 

# Methodology (Project design)

## Data

Here you have to describe how do you plan to manipulate the data. For this you have to answer to the following questions:

* Which data will be used?
  + Describe data sources
  + Describe possible external data sources that may enrich our data
  + Data for external validation?
* On which time frames periods will your project will be based on?
  + Time-frame for training
  + Time-frame for test?
* How do you define your subjects?
  + Inclusion criteria?
  + Exclusion criteria?
* Which would be your outcome variable?
* Are there confounder variables that may affect the outcome?
* Is there a possible source of bias in our data?
* Describe your data exploration strategy.
* Which techniques will be applied to enrich the data?
* How you will deal with outliers?
* How you will deal with missing values
* Add at the end of the protocol (appendix) the [Data retrieval protocol](https://docs.google.com/spreadsheets/d/1pYYjgwZ_8PS1Bcmc2kRNHTL0f_rk__GCJALLs1JHPUQ/edit#gid=0)

## Models

Here you have to describe how do you plan to develop your models:

* How do you plan to divide your data
  + Training, validation, test - proportions, techniques
* Do you need to balance your data? How?
* Do you need to stratify/subsample your data? How?
* What techniques will you apply to model your outcome?
  + Unsupervised
  + Regression
  + Classification
* Will you use cross-validation and/or bootstrap?
* Which measures you will use to train and evaluate your models? Why?
* Do you plan to use ensembling or will use your best model?

## Deployment of your model

* Who will make the QA of the project?
  + Which units will be assessed
  + Write a QA protocol for each step of the project
* Who is the final user of the predictions?
* How the prediction will be presented to the final user?
* How will the final user be trained to use and interpret the prediction?
* On which platform the predictions will be deployed?
* How frequently the model will be updated?
* What will happen in cases where the model return a null prediction (eg. incomplete data)?
* Which models were used and which were selected for the final prediction.
* Which measurements were used to evaluate the prediction.
* Which results we got from those models.

# Results

Here you will present the main results of all the process. We will describe:

* The final amount of data used (total, train, test, etc)
* The amount of outliers and the way of treating them,
* The amount of missing values and the methods used for imputing them,
* The distribution of the data (timeframes)
* The methods used to transform the data and to generate new features.

# Conclusion

Here you will write about how the project began, which were the most important challenges you had when developing the project, and how did you get the final prediction. You have to discuss also the limitations of the model, when it can be used and when not.