**Data Science Project Protocol**

*Author(s):*

*Nimrod Ken Dror*

*Lev Solodkin*

# 

# **Introduction**

In today’s loan industry, there is a huge gap between 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, 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 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***

**General**

The data used in our project is a csv file that contains the details of loans taken from Home Credit Group, which we obtained from a [Kaggle](https://www.kaggle.com/c/home-credit-default-risk/data) competition.

The shape of our dataset before any changes made is 307,511 rows and 122 columns.

In this dataset, we can see details regarding different aspects of each loan taker, such as - age, annual income, number of cars and many more. Additionally, the data tells us regarding each loan if this loan defaulted or was returned on time - this will be our final outcome variable, where 0 will mean that the loan was returned on time, and 1 will mean that it defauted.

**Treating imbalance**

A major issue that we will be facing is the imbalance of the dataset. As expected, most of the bank’s data is comprised of entries of loans that were returned on time (otherwise the bank would go bankrupt…), therefore more than 90% of the dataset is comrpised of non-defaulted entries, and less than 10% shows defaulted loans. Such an imbalance may cause severe distortion to our results, and will require some handling through the data itself or through model generation.

**Data exploration strategy**

We plan to explore variables that we believe are key variables that affect the outcome of whether the loan defaulted or not. Exploration will bedone using Python’s matplotlib and Seaborn libraries, focusing on pie charts, histograms and boxplots to identify key trends in our data, possible outliers and also problematic values that might require our attention.

We will encapsulate some of the code with custom-made functions that will render a clean and reusable code.

**Treating the data**

Our data contains multiple variables with missing values. We will be using techniques to identify and quantify these variables and the % of missing data in each of them. After finding out the proportion of missingness and also the way the missing data behaves - Missing at random, Completely at random or Not at random - we will decide whether we delete the rows, the columns or perform some feature engineering to represent the missingness.

We will be using techniques such as OLS to find out the connection of missing data to non missing data within the dataset to determine the missingness types.

Please refer to Github for the Data Retreival Protocol for our dataset

## ***Models***

Our data will be divided into train and test, on a proportion of 80% train and 20% test.

We will keep our Target variable in the same proportions in the train and test datasets (stratified) in order to avoid imbalances in the data which will severely affect our models.

As our Target variable is 0 or 1, we will be using logistic models such as Xgboost, LightGBM, Randomforest etc. The models will be trained using train sets that will be manipulated by balancing techniques such as oversampling, under sampling, SMOTE, and also by using some of the internal model features that deal with imbalanced data.

Our winning features will be chosen by using feature importances in Random Forest, Xgboost and etc.

Each of our proposed models will then be run using basic parameters and the AUC (Area under Curve) for each model will be checked and compared. We will take the winning model and then continue tweaking it through its hyperparameters to receive a higher AUC result.

# **Results**

Below is a summary of our work and the results received from running the various models in order to predict our target variable:

The final amount of data used in our file is 305,545 rows and 354 columns.

We have removed all rows of data where missingness amounted to less than 5% of the variable.

As we identified that the missing values were strongly related to the rest of the dataset, we have chosen to keep the columns and not remove them. Therefore, where missingness was higher than 5%, we divided each numerical data to quartiles and created a “missing” option to such variable, and then used one hot encoding to encode all missing columns - thus the high amount of columns in comparison to the beginning.

We have identified a significant outlier in the “Days employed” variable (indicating that some people had 1000 years of work), and removed it from our data.

After running the various logistic models, LightGBM was chosen as the favorite model, demonstrating higher AUC than the others.

We then improved the model by tuning its hyperparameters, reaching a final AUC of 65% .

# **Conclusion**

This project has demonstrated to us the importance of using an organized workflow and methodology when attending such data science projects.

We learned the importance of treating missing values and balancing of our datasets, and also - how important it is to explore the data prior to modelling stage.

Considering the fact the Kaggle competition winners reached ~80%, we believe that if we had more computational resources, it would have enabled us to reach a better score than our current score.

We are looking forward for our next project and believe that many of the tools acquired on this project will serve us well in the projects to come.