**Capstone Project 1: Project Proposal**

The goal of this project is to build a classification model to predict the probability of default on a new loan, based on customer's bureau information, income, previous application, credit card balance, install payment and so on. This model will provide lenders a predicted outcome of default or non-default to guide them making decisions to approve or decline the loan application.

The data of this project comes from Home Credit, which is a non-banking financial institution founded in 1997 in the Czech Republic. The company operates in 14 countries and focuses on lending primarily to people with little or no credit history, which will either fail to be approved for loans or became victims of untrustworthy lenders. The data comes from a variety of sources with more than 200 variables in total, and in some instances goes back to as far as 8 years of monthly balances, posing great challenges in data aggregation, wrangling, missing value imputation and feature engineering. In addition, due to the imbalanced nature of the target variable, we will use both stratified sampling and down sampling methods and compare their performances on independently held out test sets.

In this project, we will be implementing the basic data aggregation by keeping only one or two of the summary statistics such as average, sum, max or min of each group having the same previous application ID. Categorical variables will be treated using one hot encoding. For variables with high cardinality of categories, some categories with low frequencies will be grouped before one hot encoding, and aggregation will be done after one hot encoding. After all data sources are combined into one lead application file, further data exploration will be conducted to treat categorical variables, perform EDA and correlation analysis on numeric variables. Finally, all missing values are imputed using the median value.

We will implement 6 classification models and report their AUC values. For each model, first a 5-fold CV will be run to select the best combination of parameters. Then each model will be fit using the chosen best parameters and performances such as precision, recall, F1 score, ROC and AUC will be reported on the test set. For tree-based models such as Random Forest, XGBoost and LightGBM, the top 10 features will be plotted to help lenders understand the most important variables in distinguishing default vs. non default applications. Note that all metrics, for example AUC, precision and recall, are reported using 0.5 decision threshold. Lenders can also choose their own threshold to come up with the tradeoff between precision and recall rate which lie within the business risk appetite and tolerance, and eventually maximize the financial benefit.