# Machine Learning Engineer Nanodegree

## Capstone Proposal

Ashique Mahmood
30 July 2018

# Proposal

## Domain Background

Determining the risk profile of consumers is paramount to lending operations for financial institutions. Credit defaults affect both these institutions and consumers adversely. While there are many well established methods for modelling credit risks of corporate loans, predicting the risk profile of consumers is more challenging, especially the ones who belong to the lower-income households and informal enterprises group. Such challenges arise from the lack of prior credit history and financial services data. This leaves most of the unbanked population at a disadvantage.

## Problem Statement

Home Credit is a global consumer credit solution provider which focuses on lending to people with little to no credit history. Through their Home Credit Default Risk competition on Kaggle, we aim to predict credit default risk utilizing financial data from multiple sources. This is a binary classification problem, predicting whether a consumer will repay the loan or default. Developing a robust model for predicting this would help people in 11 countries get the loan they deserve.

## Datasets and Inputs

The data is provided by Home Credit through the Kaggle competition. The full description of each feature in the entire data set can be found here. Here's a summary of the datasets:

| **Dataset** | **No. of Features** | **Description** |
|---|---:|---|
| application_{train\|test}.csv | 122 | Static data for all loan applications. |
| bureau.csv | 17 | Credit data from other financial institutions reported to Credit Bureau |

| | | |
|---|---|---|
| bureau_balance.csv | 3 | Monthly balances of previous credits in Credit Bureau |
| POS_CASH_balance.csv | 8 | Monthly balance snapshots of previous POS and cash loans with Home Credit |
| credit_card_balance.csv | 23 | Monthly balance snapshots of previous credit cards with Home Credit |
| previous_application.csv | 38 | All previous applications for Home Credit loans |
| installments_payments.csv | 8 | Repayment history for the previously disbursed credits in Home Credit |

## Solution Statement

We are required to predict the probability of the consumer belonging in one of the two classes, repayment of loan on time or defaulting. Earlier research by Khandani et al 2010 [1] makes use of a generalized classification and regression trees (CART)-like algorithm to solve the credit risk problem. More recent research by Charpignon et al [2] has shown that advanced tree based algorithms such as Random Forests and GBTs outperform CART and logit models previously used. Although earlier attempts were made at using neural networks for solving such a problem, recent work by Peter Addo et al [3] revealed that tree-based models are more stable than the models based on multilayer artificial neural networks. Hence an attempt will be made to solving this by implementing advanced tree based algorithms such as XGBoost and LightGBM.

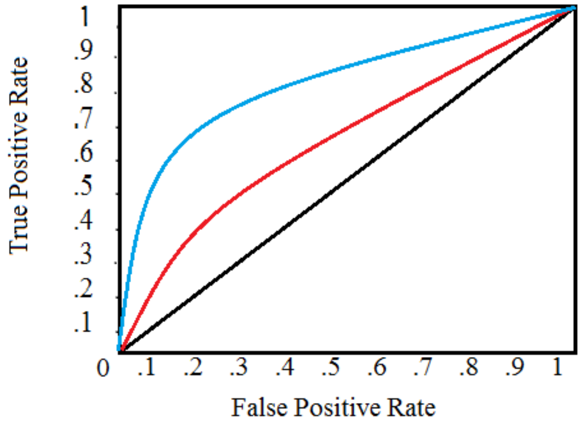
## Benchmark Model

Research and earlier work suggests tree based models as the most suitable for problems like this. Since we intend to explore advanced tree based models as a part of the solution, it will be a fair decision to compare the performance against a decision tree model. The same evaluation metric will be used for both the benchmark and the suggested solution models.

## Evaluation Metrics

The model performance will be measured using the area under the ROC curve between the predicted probability and the observed target. The ROC curve is a plot of the true positive rate against the false positive rate.

*Figure 1: ROC Curve*



$True\ Positive\ Rate = \frac{TP}{TP+FN}$ $False\ Positive\ Rate = \frac{FP}{TP+FN}$

Since the problem asks us to predict the probability of the class, using these two metrics, the ROC curve helps us understand how well the classifier separates the two classes.

## Project Design

We have a very clearly defined problem in our hands – given various data about consumers, build a model which can predict the probability of the consumer falling in one of two classes: one repaying the loan on time and the other unable to do so.

The project will be divided into the following stages:

### Exploratory Data Analysis

The goal of this step is to understand the data we will be working with and learn about the preprocessing required. We will look at the distribution of the target variables. This will be important for deciding parameters for our models later. Distribution of some features will also be investigated to learn their importance. The number of missing values for each feature will be investigated to learn what we need to impute later.

### Data Wrangling

The data will be processed to prepare for consumption by the model by using the insights gained in the previous stage. Imputing missing values, transforming, scaling, encoding categorical values will be done as needed.

### Feature Extraction

The overall number of features at our disposal for this problem checks in at X. New features will be created from the vast number of them available to us for the project, both manually and through PCA. An attempt at creating domain related features will be made as well.

### Baseline Modelling

The data will be modeled using a CART model to set the benchmark. The model performance will be evaluated. The effect of using engineered features will also be evaluated.

### Improved Modelling

A training and testing pipeline will be developed to try out different models. After evaluating their performance, the best suited model will be further improved upon through hyperparameter tuning.

Through analyzing the results, we will iterate the whole process to improve feature selection/engineering, model selection and parameter tuning and overall model performance.

## References

[1] Khandani, Amir and Kim, Adlar J. and Lo, Andrew W., Consumer Credit Risk Models Via Machine-Learning Algorithms (March 11, 2010). AFA 2011 Denver Meetings Paper. Available at SSRN: https://ssrn.com/abstract=1568864 or http://dx.doi.org/10.2139/ssrn.1568864

[2] Charpignon, M. (2014). Prediction of consumer credit risk.

[3] Peter Addo, Dominique Guegan, Bertrand Hassani. Credit Risk Analysis using Machine and Deep Learning models. Documents de travail du Centre d'Economie de la Sorbonne 2018.03 - ISSN : 1955-611X. 2018. 〈halshs-01719983〉