**Project Proposal CS6220 Summer 2018**

**TEAM MEMBERS**  Sharyu Deshmukh, Somya Bhargava, Mounica Subraman, Chia-Yi Liaw

**SUMMARY**

[Home Credit](http://www.homecredit.net/about-us.aspx), the host of the competition, is a finance provider that focuses on serving the unbanked population. Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. So, [Home Credit](http://www.homecredit.net/about-us.aspx) tries to broaden financial inclusions for unbanked population. The [Home Credit Default Risk c](https://www.kaggle.com/c/home-credit-default-risk)hallenge is a standard supervised machine learning task where the goal is to use historical loan application data to predict their clients' repayment abilities based on datasets provided.

**DATA EXPLORATION**

* *application training/testing*: contains the loan applicants which with the labels included,  a binary 0 - if the loan was repaid and a 1 - if the loan was not repaid ),
* Bureau.csv: application data from other credit institution with historical record
* Bureau\_balance.csv : Monthly balances of previous credits in Credit Bureau
* *previous\_applications* : previous loans information at Home Credit by same clients
* *Credit\_card\_balance.csv:* Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
* POS\_CASH\_balance.csv : cash loans that the applicant had with Home Credit
* Installments\_payments.csv : Repayment history for the previously disbursed credits in Home Credit

**PROPOSED PLAN OF APPROACH**

**Preprocessing :** handle missing value with deletion or imputation and apply over sampling method on the unbalance distribution, encode categorical variable as well as feature selection using dimension reduction approach such as random forest, PCA and gradient boosting.

**Modeling :** Apply multiple machine learning models on the dataset, such as random forest, SVM , XGBoost and logistic regression achieve the in-depth data analysis, classification and prediction on client repayment abilities. Perform cross validation includes setting hyperparameter and model training.

**DESIRED OUTCOME**

**Evaluation :** Model evaluation , comparison and discussion, includes accuracy score and Receiver Operating Characteristic Area Under the Curve. In addition , Interpret for influential predictor.

**Visualizations :** Utilize Tableau and Python to deliver data visualization of model evaluation and prediction results.

**REFERENCE**

1. <https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction>
2. <https://www.kaggle.com/codename007/home-credit-complete-eda-feature-importance/notebook>
3. Optimization? (Run Bayesian Optimization)
4. Modeling :

* Set Hyperparam values
* Perform cross validation
* Model Training (CV train dataset)
* Model Evaluation (Validation dataset), Calculate AUC
* XGBoost
* KNN

1. Model fitting (full train data)
2. Predict (using test data)

Modeling phase: During training, we provide our model with the features — the variables describing a loan application — and the label — a binary 0 if the loan was repaid and a 1 if the loan was not repaid — and the model learns a mapping from the features to the label. Then, during testing, we feed the model the features for a new batch of applications and ask it to predict the label.

*One to three paragraphs clearly outlining*

*o Your name and your project partners’ names*

*o what you plan to do,*

*o how you plan to approach it, and*

*o what your final deliverable will be.*

# Data[¶](https://www.kaggleusercontent.com/kf/3914089/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..179QMq6JGD3_XTA-OIs5Dw.4g_PUClR41Fm8xF1mGSh5pkTYtsIIqmWIdqbu19_7nNSBH_Kp3TpF2O-2aKuKIePxq8XCo0FfoXXgmjfNaG3thkphF0cfQX_IiPo8zeY9eyn7jOay1_QozL0e_b09rOl3qb7t77UP3ll-p-nXbYmN8XCzad0wUo4cMm6Z9XKT9Y.GuJoUIf8d4ZzDFUZJAWVkA/__results__.html#Data)

The data is provided by [Home Credit](http://www.homecredit.net/about-us.aspx), a service dedicated to provided lines of credit (loans) to the unbanked population. Predicting whether or not a client will repay a loan or have difficulty is a critical business need, and Home Credit is hosting this competition on Kaggle to see what sort of models the machine learning community can develop to help them in this task.

There are 7 different sources of data:

* application\_train/application\_test: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK\_ID\_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.
* bureau: data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
* bureau\_balance: monthly data about the previous credits in bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
* previous\_application: previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK\_ID\_PREV.
* POS\_CASH\_BALANCE: monthly data about previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
* credit\_card\_balance: monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
* installments\_payment: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.