**MVP: Home Credit Loan Project**

*By Jasmine Zeng*

**Project Goal:**

To build a binary classification model to predict whether a client can repay the loan for Home Credit bank.

**Data Process:**

1. Download data from Kaggle: initial dataset has a dimension of (307511, 122)
2. Delete features with null values more than 50,000
3. Replace null values with either mean or mode
4. Look into categorical features to see whether to keep or engineer or remove that feature from the dataset:

* ‘NAME TYPE SUITE’ seems to be an immaterial feature since the default rate is quite similar across different groups, thus I remove this feature:

Chart, bar chart

Description automatically generated

* For ‘INCOME TYPE’, default rate is higher in group ‘maternity leave’ and ‘unemployed’. I keep this two groups and bag other groups into ‘Other income type’ and transfer the feature into dummy features.Chart, histogram

  Description automatically generated
* For 'EDUCATION\_TYPE', default rate seems to increase as educational level becomes lower. Hence, I transfer this feature into ordinal ‘int’ type feature.

Chart, bar chart

Description automatically generated

**Baseline model:**

I used logistic regression as my baseline model. The model’s F1 score drop from 0.57 to 0.19 after switching from train data to test data, indicating overfit. Below is an overview of confusion matrix on both in-sample data and out-of-sample data:

A picture containing graphical user interface

Description automatically generated

Graphical user interface, text

Description automatically generated

**Going forward:**

Moving forward, the goal is to deal with overfit issue, and to select a champion model out of 4-5 candidates. Further hyperparameter tuning will be performed on champion model through cross validation practice.