Home Credit Default Risk

Group 03
Zhengyuan Ding , Chenqin Yang
Jiayao Liu, Ziyu Lei, Zian Chen

Agenda

- ➔ Background
- ➔ Dataset Introduction
- ➔ Problem Statement
- ➔ Methodology
- ➔ Performance Improvements
- ➔ Schedule

Background

Home Credit tries to provide loans to underprivileged people who do not have sufficient credit histories. They used a variety of variables including telco and transactional information to predict the repayment ability of their clients.

This project will use the statistical and machine learning methods to help Home Credit fully unlock the potential of the data it collects from the clients.

Problem Statement:

Can you predict how capable each applicant is of repaying a loan?

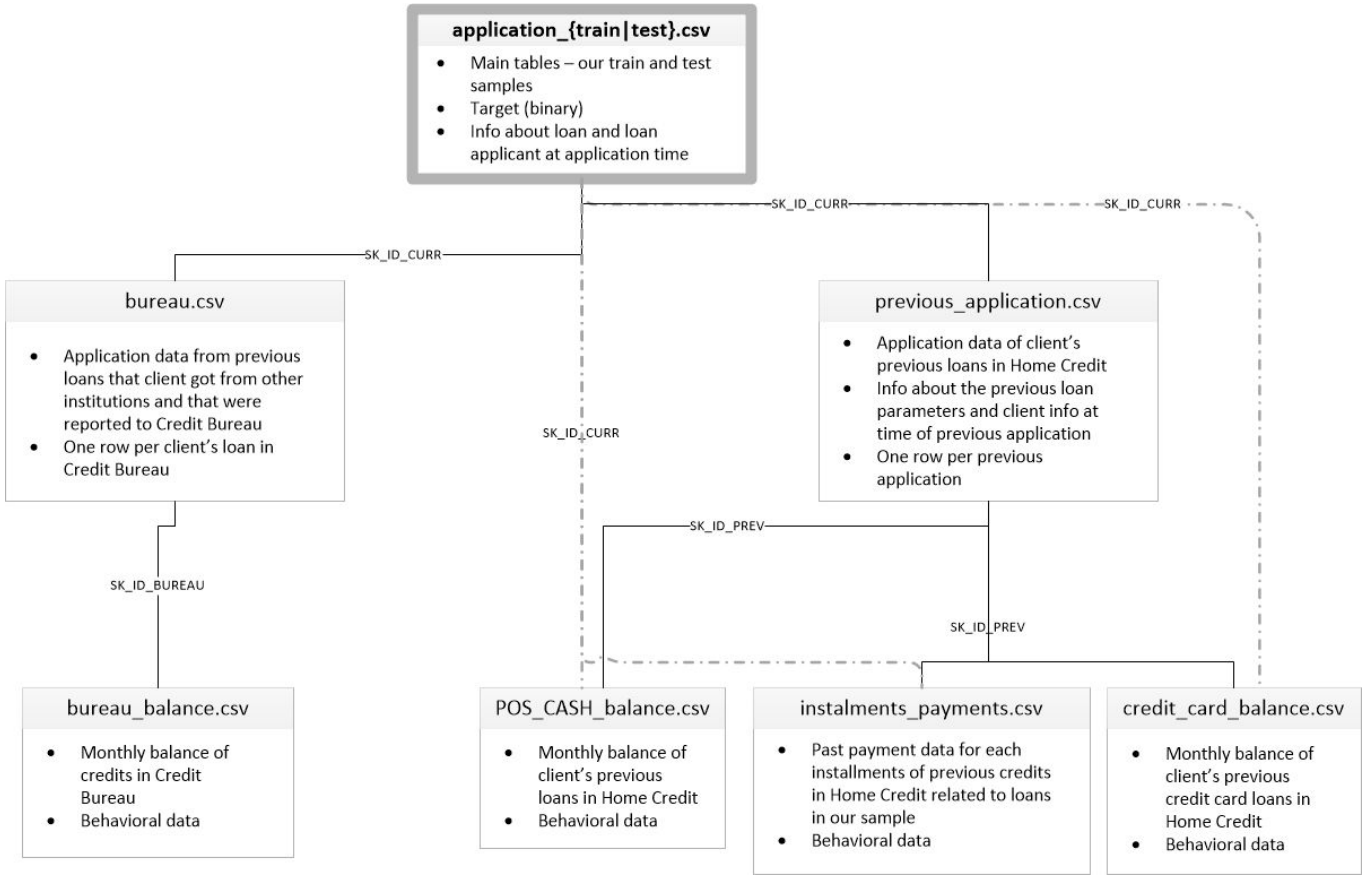
Dataset

- Application.csv
  - Main table
  - One row represents one loan in our data sample.
- Bureau.csv
  - All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample)
- Bureau_balance.csv
  - Monthly balances of previous credits in Credit Bureau
- POS_CASH_balance.csv
  - Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit

Data Source: https://www.kaggle.com/c/home-credit-default-risk/data

Dataset (Cont'd)

- Credit_card_balance.csv
  - Monthly balance snapshots of previous credit cards that the applicant has with Home Credit
- Previous_application.csv
  - All previous applications for Home Credit loans of clients who have loans in our sample
- Installments_payments.csv
  - Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample


application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

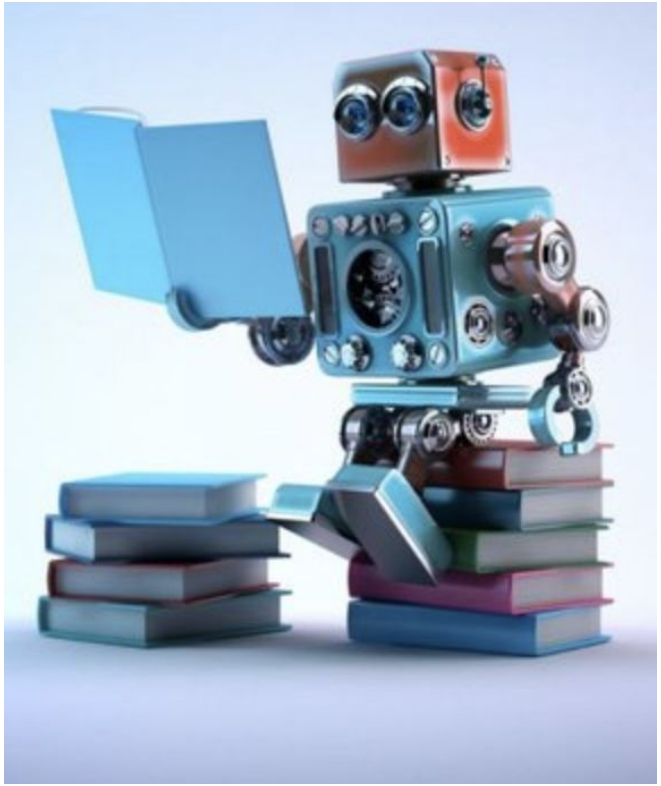
Methodology

Binary Classification Models

- Baseline model: Naive Bayes
- Logistic Regression
- SVM
- Random Forest
- Gradient Boosting Method

Evaluation Metrics

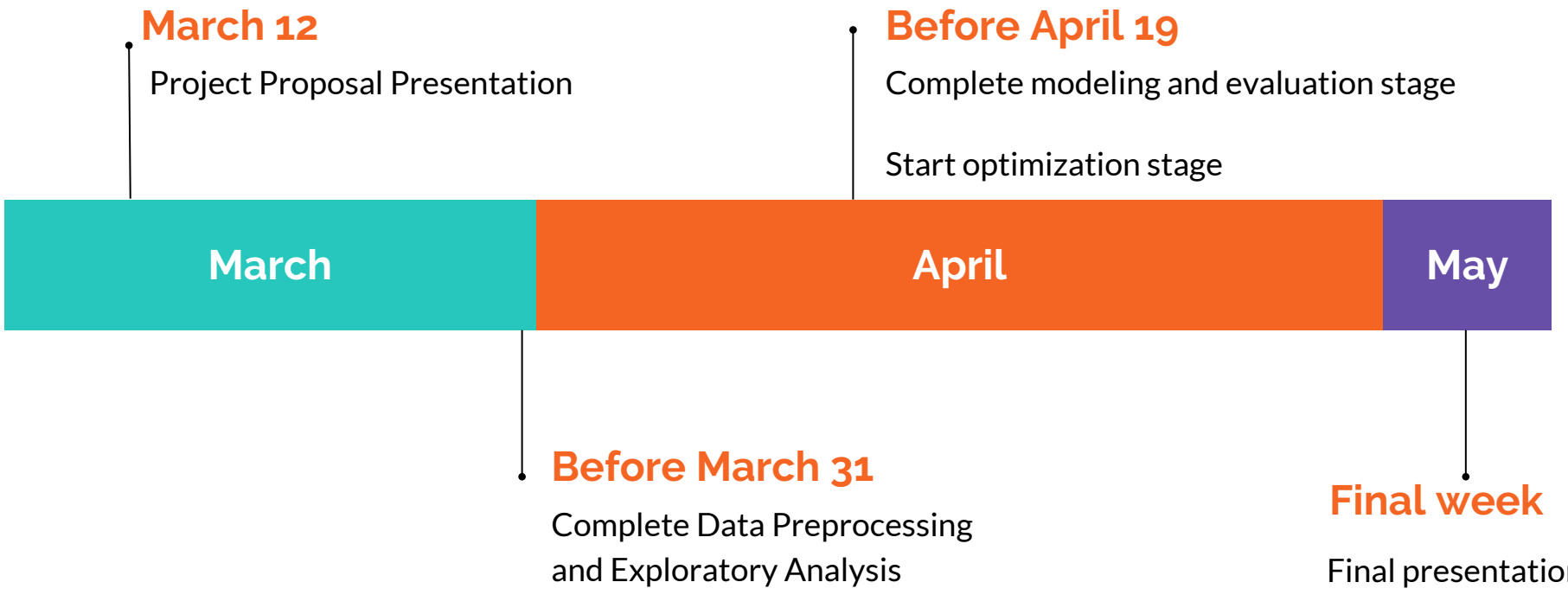
- AUC



Possible Performance Improvements

- Cython — Operations between columns (groupby count, max, min, mean)
- Check code sanity by following Python performance tips
    - Itertools, function call overhead, built-in tools…….
- Do Python performance tuning by using cProfile and line profiler
- Numba: JIT Compiling, Vectorize decorator
- PySpark

Schedule


March 12
Project Proposal Presentation
Before April 19
Complete modeling and evaluation stage
Start optimization stage
March
April
May
Before March 31
Complete Data Preprocessing
and Exploratory Analysis
Final week
Final presentation

Thank you!