Predicting Client's Repayment Ability

Chengyu Liang, Dongzhe Zhang, Kunpeng Huang, Haolan Ma, Yihan Jiang, Meiling Zhang

Introduction

For Home Credit, the most concerning question is how risky is the borrower? In this capstone project, our main goal is to build a machine learning model to predict whether or not the applicant will default on the loan in the future for Home Credit Group. On one hand, it is important to identify those who are unable to repay the loan to prevent business losses for Home Credit. On the other hand, it will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful. The business of Home Credit Group is currently present in China, India, Indonesia, Vietnam, Philippines, Russia, Kazakhstan, USA, Czech Republic and Slovakia. By applying our model, the company would expand its business to other countries.

Workflow

1 Observe Data

Manually look at 5-10 application records across all tables, detect systematic missing value.

Apply data into a decision tree to find the importance of NA for each variable

2 Data Cleaning

Fill NA with: 0, mean, or create model to predict what missing value should be.

Remove duplicate data

3 Features

1.Create time series parameters from bureau_balance, cash_balance and credit_balance.

2. Extract financial term parameters from dataset to indicate personal repay ability.

4 Modeling

Create a high CPU-performance compute engine on GCP

Use the data to build supervised regression models to predict default risk of new application

5 Optimization

1. Optimize linear regression model with Logistic approach.

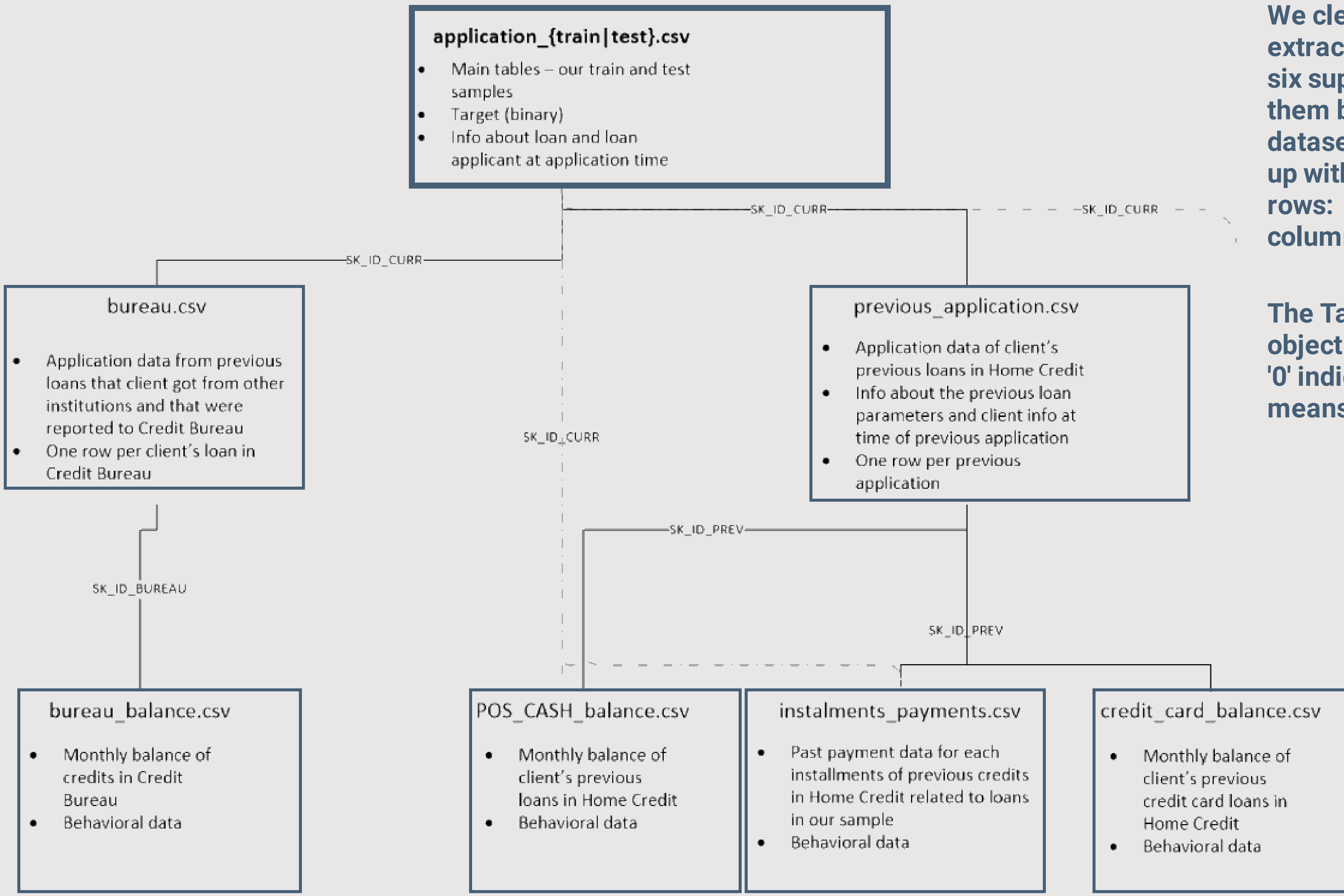
2. Optimize random forest models and XGBoost by tuning models' parameters.

6 Business Implication

Apply model to decide whether applicant could receive money and the loan interest rate

Dataset Overview

Dataset Overview



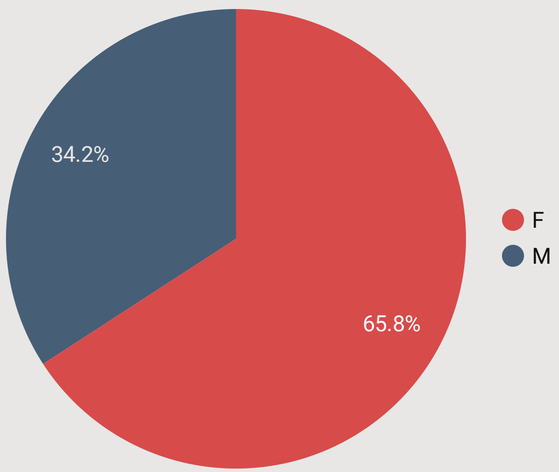
We cleaned, aggregated and extracted important variables from six supplementary datasets, and join them back to the main dataset(application_train). Which end up with
rows:
columns:

The Target variable is our prediction objectives.
'0' indicates the loan was repaid, '1' means not repaid.(Default.)

Exploratory Data Analysis

Workflow | Dataset | EDA | Methodologies | Implementation

Gender Distribution



Record Count
307,511

Default Rate
8.07%

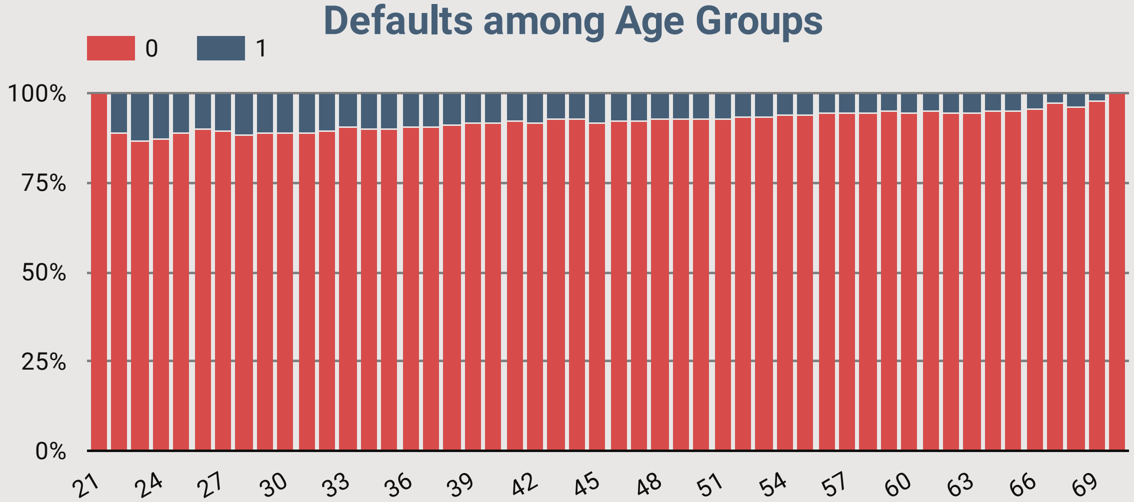
Average Income
168,797.92

Occupation and Default Rate

	OCCUPATION_TYPE	Record Count	TARGET	default ▾
1.	Low-skill Laborers	2,093	359	17.15%
2.	Drivers	18,603	2,107	11.33%
3.	Waiters/barmen staff	1,348	152	11.28%
4.	Security staff	6,721	722	10.74%
5.	Laborers	55,186	5,838	10.58%
6.	Cooking staff	5,946	621	10.44%
7.	Sales staff	32,102	3,092	9.63%
8.	Cleaning staff	4,653	447	9.61%
9.	Realty agents	751	59	7.86%
10.	Secretaries	1,305	92	7.05%

1 - 18 / 18 ‹ ›

Defaults among Age Groups



Applicants' Education vs.Default Rate

	NAME_EDUCATION_TYPE	Record Count	TARGET	Default ▾
1.	Lower secondary	3,816	417	10.93%
2.	Secondary / secondary special	218,391	19,524	8.94%
3.	Incomplete higher	10,277	872	8.48%
4.	Higher education	74,863	4,009	5.36%
5.	Academic degree	164	3	1.83%

1 - 5 / 5 ‹ ›

Methodologies

1.Split 80% train data, 20% test data
Use train data to predict binary result on default/not default of test data.

2. Choose of Models
Random Forest Model
XGBoost Model
Logistic Regression Model

3. Tune on Resample Parameters on 5-fold CV
We notice that there are only 8.07% default records which cause imbalanced problem. To solve this problem, we decide to resample the train dataset with following methods:

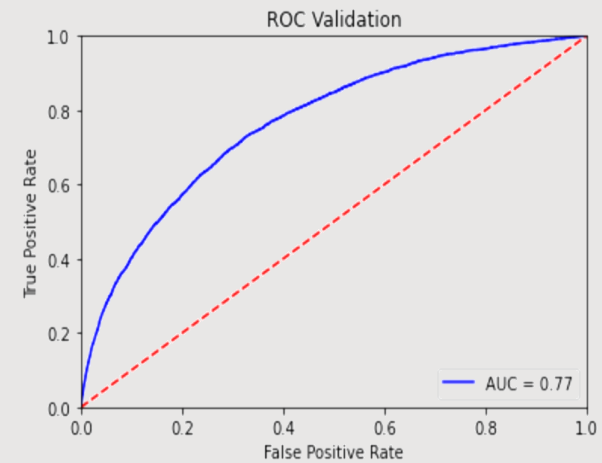
SMOTE
-- Generate synthetic samples in between minority samples and the k-nearest neighbor of all minority samples.

SMOTE + undersampling
-- Generate synthetic minority samples, at the same time, reduce majority samples.

	Model	Baseline	Oversampling	Oversampling + Undersampling
1.	Logistic Regression	56.39%	60.43%	63.08%
2.	Random Forest	73.16%	71.97%	74.51%
3.	XGBoost	75.66%	72.76%	75.78%

4.Tune the model parameters on AUC measure in 5-fold CV.
Choose the optimize resample model, tune the model parameters. The chart below shows the improvement on tuning each parameters.

	Paramete...	Best ...	AUC ▲
1.	Depth	5	76.37%
2.	Weight	3	76.37%
3.	Gamma	0	76.5%
4.	Subsample	0.8	76.65%
5.	Alpha	0	76.79%



5.Find the cut-off point which can maximize F1-SCORE.

$$f1 = \frac{2 * recall * precision}{recall + precision}$$

$$recall = \frac{TP}{TP + FN} \qquad precision = \frac{TP}{TP + FP}$$

The confusion matrix at the optimal cutoff point is:

Actual \ Predict	Non-default	Default
Non-default	*41452*	*3547*
Default	*2497*	*1399*

Implementation

Given the prediction result, we to segment the applicants into three categories:
1. Extraordinary applicants
2. Great applicants
3. Unqualified applicants

We derive a formula for generating each applicant's credit score following the FICO credit scores distribution.

$$cut\ off = optimized\ threshold * 550 + 300$$
$$credit\ score = nondefault\ probability * 550 + 300$$

Applicants who score lower than the cutoff point would be rejected, applicants who score higher than the cutoff point, we will charge them different interest rates based on customer segmentation.