BOSTON UNIVERSITY

Predicting Client’s Repayment Ability

Cohort B Team7:Chengyu Liang, Dongzhe Zhang, Kunpeng Huang, Haolan Ma, Yihan Jiang, Meiling Zhang

HOME CREDIT

INTRODUCTION

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data - including telco and transactional information - to predict their clients' repayment abilities.

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.

OBJECTIVE

- Applying various statistical and machine learning methods to make prediction on applicant's default rate.
- Identifying key features that might impact the applicant's default rate.
- Using the prediction result to segment clients and assign them different interest rates based on their default possibility.

DATA PREPARATION

- There are 7 tables in the dataset in total, including 1 main application tables with all the information of loan applications. The rest six tables are supplementary tables with applicants’ historical records.
- We cleaned and extracted important features that might related to client’s application from the supplementary tables and merge these features back to the main table with their unique identifiers.
- When dealing with NAs, we use KMeans Clusters to segment the customers into 2 groups and replace the NA's with the value from the median of the cluster that the customer belongs to. Because we thought the model might pick up any characteristics that could differentiate the default and non-default customers.

METHODOLOGY

1.Split 80% train data, 20% test data

Predict binary result on (default/not default)

2. Models

- Random Forest Model
- XGBoost Model
- Logistic Regression Model

3. Process

- 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
  - SMOTE + undersampling

- Tune the resampling parameters on AUC measure in 5-fold CV.
- Choose the optimize resample model, tune the model parameters.
- Find the cut-off point which can maximize F1-SCORE.

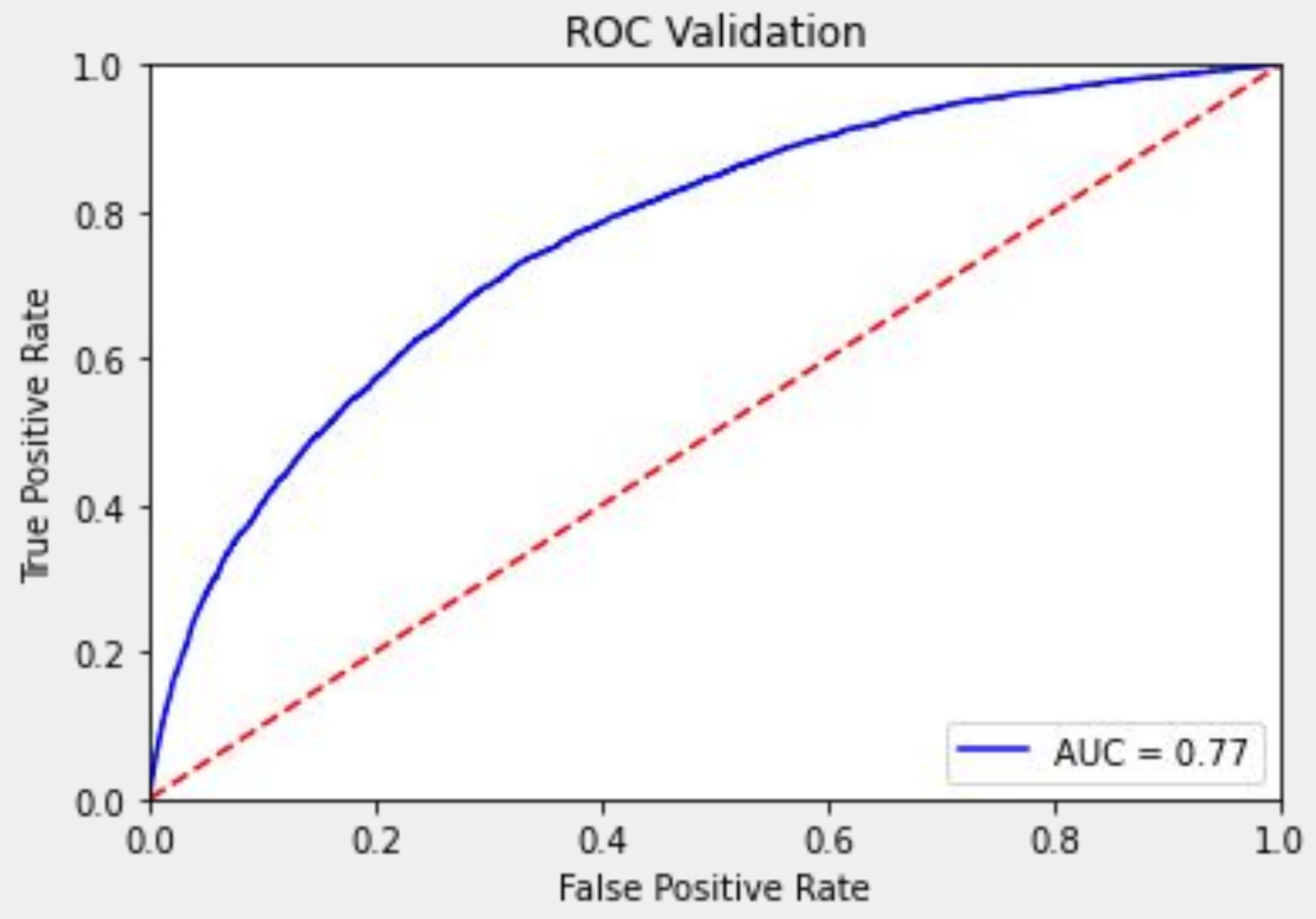
$$f1 = \frac{2 * recall * precision}{precision + recall}$$

Model AUC

	Baseline	SMOTE	SMOTE(0.1) and undersampling(0.5)	SMOTE(0.2) and undersampling(0.5)
Logistic Regression Model	*56.39%*	*60.43%*	*63.08%*	*61.10%*
Random Forest	*73.16%*	*71.97%*	*74.51%*	*73.56%*
XGBoost Model	*75.66%*	*72.76%*	*75.78%*	*74.96%*

XGBoost Model Tuning

Parameter Tested	Best Parameter	AUC
Depth	*5*	*76.37%*
Weight	*3*	*76.37%*
Gamma	*0*	*76.50%*
Subsample	*0.8*	*76.65%*
Alpha	*0*	*76.79%*



RESULTS

Look Into the Best Model:

- **Top 5** important variables
  1. NAME_INCOME_TYPE.Working
  2. 'NAME_EDUCATION_TYPE.Secondary...secondary.special
  3. NAME_FAMILY_STATUS.Married
  4. OCCUPATION_TYPE.Not.Provided
  5. Previous application reject ratio

- In this project, due to the unavailability of building a profit-oriented model, we simply use optimized f1 score threshold as our final cutoff point. When applying this threshold, the confusion matrix of the test dataset is:

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

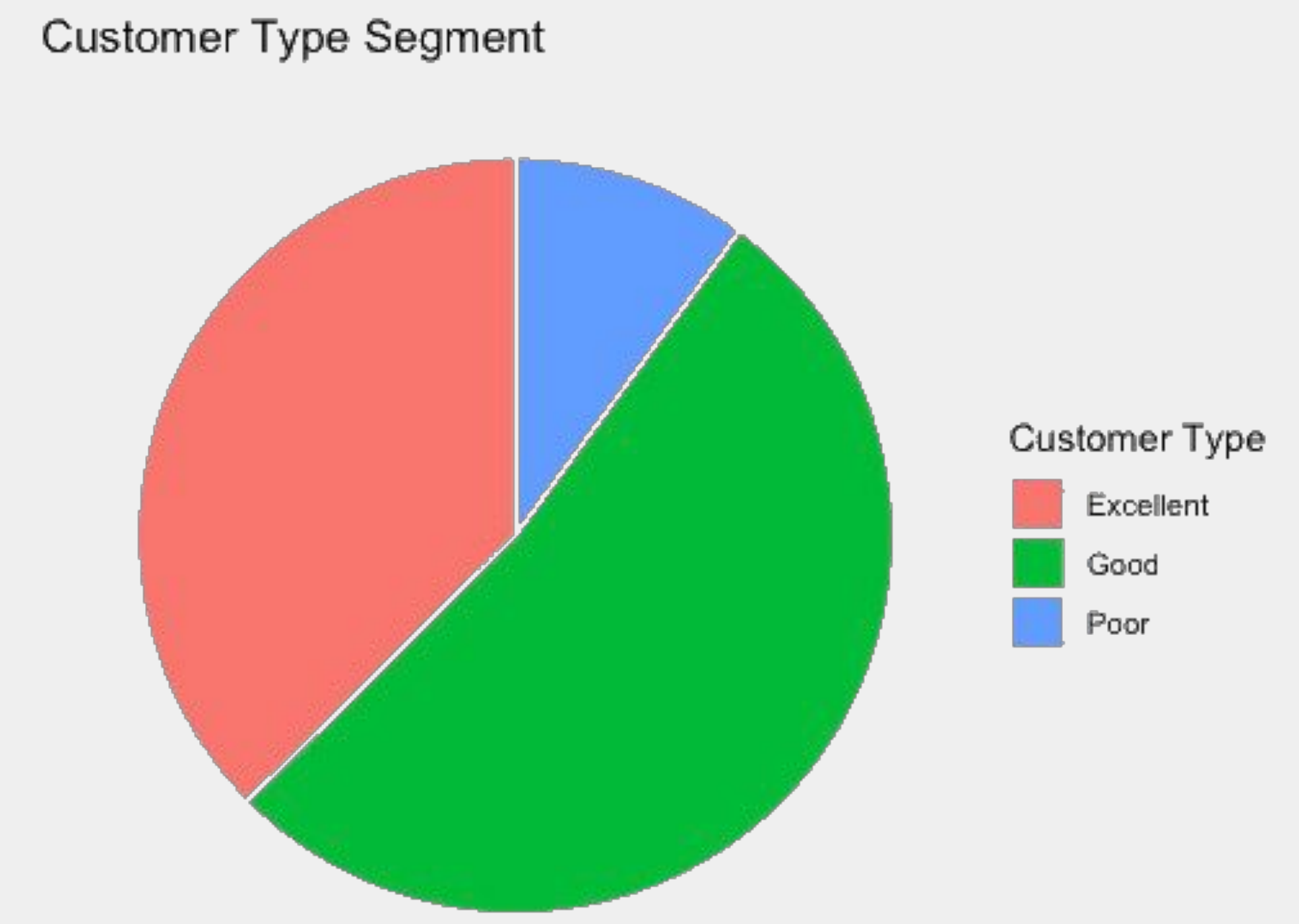
BUSINESS IMPLEMENTATION

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.



ACKNOWLEDGEMENT

We thank Professor Mohammad Soltanieh-ha (Boston University) for his teaching, feedback and support throughout this project. We thank Brock Tibert (Boston University) for his instruction to us for starting our capstone project. We thank all other Professors who lectured us and gave us help throughout this program.

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

Boston University, Questrom School of Business