Home Credit Default Risk - Final Report

Data Wizard Team - Ibrokhim , Kabita, Jiamin, Ritika, Namrata

2019-11-10

Contents

Chapter 1

About Home Credit

In 2017, the World Bank reported a global population of 1.7 billion unbanked adults, or those who do not own any form of bank account from any financial institution, with a majority residing in developing economies like China and India. Universally, this population has become an easy target for many institutions that tend to charge for financial services at a higher fee. To address this issue, Home Credit is dedicated to creating positive loan experiences for the unbanked around the world and promote financial inclusion. In this project, we will use the available data from Home Credit to explore various factors related to repayment abilities to ensure available loaning options to the right clients. Our goal is to find out what are major factors directly related to repayment ability of a customer.

Chapter 2

Background Research

In an overview, the word "unbanked" defines people who do not own any form of bank account, while "underbanked" characterizes those who have a bank account but do not have access to mainstream bank services such as credit and loans. In a survey conducted by the Federal Deposit Insurance Corporation in 2017, 53% of the unbanked population in the US reported insufficient financial means to maintain a bank account, while 30% cited "Don't trust banks". The lack of credit score results in inability to access common bank services such as loans and credit cards. As a result, many resorts to more expensive options when seeking financial services as simple as cashing out a check. When caught in a financial emergency, people are forced to take out at unaffordable high interest rates and therefore the cycle worsens. In expanding financial inclusion, Home Credit aims to create lines of credit specifically for the unbanked population. The main company objective is to not only build trust with clients, but also have good risk management strategies. To minimize risks of credit default, our objective is to create an algorithm on loan application approval predicting potential of repayment. However, the main challenge to this task is that 70% of Home Credit clients are first-time borrowers. As a result, there is limited record on client credit history, which serves as an important indicator in conventional credit underwriting. In a finance summit, Home Credit has stated that the key to risk management in this client pool rests in two aspects. One is client's repayment capability and understanding how much of an instalment he or she can afford. The other is to identify fraudsters who intend to loan with no intentions of repaying.

Chapter 3

Objective

In broader sense, our objective of this project is to come up with machine learning models to predict clients loan repayment capability. In order to build predictive models, we are going to utilize that dataset available in Kaggle: https://www.kaggle.com/c/home-credit-default-risk/data Home credit is using several machine learning algorithms to predict their clients' repayment capability. They also try to build repayment plan such that their clients achieve their financial goal with ease. Unlike other financial institutions that rely on credit history to check applicant loan repayment ability, home credit targets customers who has very limited or no previous credit history available. Because home credit applicants have very little (or not at all) previous loan history available, we will try to incorporate as many datasets as possible provided by Kaggle for this project.

Chapter 4

Data Overview

There are mainly seven tables among which Application (train and test) is the master table. The entities of these dataset are deidentified (replaced with ID) for security reasons. The application table contains records for each loan application with Home Credit. Other tables contain records of clients' previous loan application with Home credit, other credits which are reported to credit bureau and instalment payment records. In below table, we have mentioned some basic overview of data tables that are available to us.

Table Name	Description
Application {train|Test}.csv	This is the main table, broken into two files for Train. The Train file contains additional column TARGET which tells us the customer has repaid loan or not. The test dataset does not have TARGET column. This table also contains client's demographic and important information about client which includes whether the applicant owns a house/car, number of family members etc.
bureau.csv	This table contains other credit information of the borrower. Loans provided by other financial institutions that are reported to credit bureau are stored in this table. This table contains all other loans taken by client before the application date.
bureau balance.csv	This table contains monthly balance of previous bureau of credits
POS CASH balance.csv	Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit. This table contains month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our system.
credit card balance.csv	It contains monthly balance snapshot of previous credit cards.
previous application.csv	All previous applications for Home Credit loans of clients who have loans in our sample. There is one row for each previous application related to loans in our data sample.
installments_payments.csv	Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample. There is a) one row for every payment that was made plus b) one row each for missed payment. One row is equivalent to one payment of one instalment OR one instalment corresponding to one payment of one previous Home Credit related to loans in our sample

Chapter 5

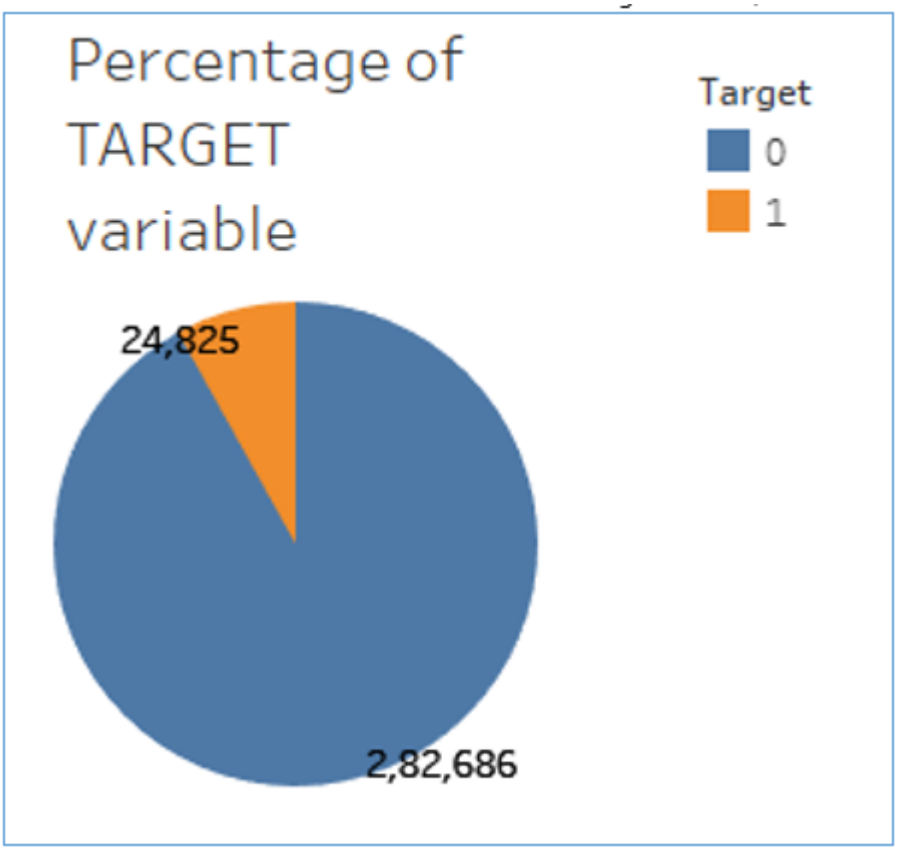
Exploration of Data

In order to better understand the risk of default, we want to use the TARGET variable from the application_train dataset and compare it various predictors. Here, a value of 0 represents a repaid loan, while a value of 1 represents a default loan. Before we proceed with modelling, we want to first understand the distribution of the target variable, as shown below:

5.1 Target

As we can see, class 0 is significantly larger than category 1. This means only few loans were not repaid compared to loans that are repaid. When classes are imbalanced, machine learning models are biased towards category with higher number of observations. The classifier will achieve a high accuracy based on the majority class. This doesn't make a good predictive model because the minority class will simply be ignored. To overcome this problem, we plan to apply oversampling on class 1 to achieve a balanced dataset.

Another important point to note is that the application_test dataset does not contain a target variable. When we are ready to run our model, we need to split the application_train dataset into training and testing subsets to train our classifier and evaluate the predicted outcomes with the actual outcomes.



5.2 Independent Variables

Chapter 6

Challenges:

6.1 Too many attributes:

The application table consists of 122 variables. We need to carefully choose variables that are needed for analysis removing less important variables in order to avoid noise while not losing valuable information. In order to avoid this problem, we can do feature selection.

6.2 Dealing with Missing values:

There are high percentage of missing values for few attributes. There are many ways to handle missing values. Most common methods are dropping records with missing values (when percentage of missing value is low). One other method is to discard columns with high number of missing values. Missing values can be handled by Imputation techniques as well. In imputation, missing values for categorical variables are replaced by either most frequently used category or a considered as third category. At this earlier stage, we will prefer imputation over dropping in order to avoid loss of information.

6.3 Imbalanced Data:

The application dataset has less than 10% values for TARGET category 1. The rest of the dataset has TARGET 0. Machine learning algorithms are highly biased with the categories with high number of observations. To overcome this situation, we want to use sampling techniques.

Chapter 7

Methodology:

The application train table has TARGET column, which is missing in application test. We need to predict TARGET column (dependent variable) with the help of machine learning. This is direct case of supervised classification problem.

7.1 Supervised:

Some observations are labeled with TARGET category 0 or 1. We need to predict the TARGET column based on the predictor variables. Model is open for continuous improvement with the help of validation techniques.

7.2 Classification:

The dependent variable (TARGET) is categorical with two levels o and 1. Through model, entire population of dataset will be labeled as either class 0 or class 1 based on predictor variables.

Chapter 8

Metrics

8.1 ROC

The most common classification metric is Receiver Operating Frequency Curve commonly known as ROC Curve. When data is highly imbalanced, accuracy is not the best measure of performance. Area Under Curve (AUC) of ROC needs to be taken into consideration. As more area (AUC) is better perfection, we want to achieve as high AUC as possible for our model.

8.2 Confusion Matrix

Another validation measure of our classification model can be confusion matrix. Confusion matrix represents four values with observations that are positive and labeled as positive (True Positives TP), positive but recorded as negative (FN), Negative but recorded as positive (FP) and Negative recoded as negative (TN). In our case, customers who are not able to repay loan (actual 1) but marked as 0 (able to repay, not default) is very sensitive and raises risk factor for Home Credit. In this case False Negative is the parameter, we should closely monitor. We need to tune our model sufficiently to minimize False negatives.

Chapter 9

References

- Apaam, Gerald, et. al "FDIC National Survey of Unbanked and Underbanked Households: Executive Summary". Federal Deposit Insurance Corporation. 20 Sep. 2019. https://www.fdic.gov/householdsurvey/2017/2017execsumm.pdf
- Demirguc-Kunt, Asli, et. al. "Measuring Financial Inclusion and the Fintech Revolution". The Global Findex Database. 20. Sep. 2019. https://globalfindex.worldbank.org/sites/globalfindex/files/2018-04/2017%20Findex%20full%20report_0.pdf
- Home Credit Default Risk: https://www.kaggle.com/c/home-credit-default-risk/data
- "Serving the Unbanked Population: a speech by Home Credit China CEO Ondrej Frydrych to the 21st Century Finance Summit of Asia". Home Credit. 10 Jan. 2017. http://www.homecredit.net/media/press-releases/2017/170110-ondrej-frydrych-serving-the-unbanked-population.aspx