Data Science Project:
Predicting Defaults on Loan Payment

Team 40, Section B

Arwen Wang, Kewei Jiang, Kieu Anh Nguyen, Taru Dharra and Vedant Sahay

I - Introduction

The dataset we have chosen focuses on the non-banking financial industry. There are 122 columns in the Main_DataSet describing the factors which could have an impact on the default by a customer on his loan payments. Some of the variables are gender, family status, contract type, annuity amount, income type, income amount etc. The datasets used in this project has been retrieved from Kaggle and is a Home Credit Group dataset. It is labeled as 'Home Credit Default Risk'. The types of variables included in the dataset are both categorical and numerical.

II - Business Understanding

Home Credit is an international non-bank financial institution and focuses on lending primarily to people with little or no credit history. A non-banking financial institution (NBFI) is a financial institution that does not have a full banking license or is not supervised by a national or international banking regulatory agency. According to Global Monitoring Report on Non-Bank Financial Intermediation 2018, globally the narrow measure (bank-like financial stability risks) grew by 8.5%, to $51.6 trillion in 2017. Since NBFI borrowers are unable to provide collateral for their loans, removing one of the main tools used by traditional financial institutions for mitigating credit risk, there is a high risk of default. So, our core task is to use this Main_DataSet to classify the potential applicants (which are present in application_test dataset) as high risk of default or not high risk of default. Apart from this, we also want to cluster the applicants in order to explore the characteristics of different applicant segments and provide recommendations for relevant product offerings.

III - Data Understanding

The Main_DataSet dataset consists of 122 variables and 307,511 observations. We have one dependent categorical variable, 121 independent variables (categorical=50 and numerical=71). Each row represents a single customer, and each column contains customer's information. The data definitions are found in HomeCredit_columns_description excel file attached along with the datasets.

Also, we will use application_test dataset (the dataset of potential applicants) which has same variables as Main_DataSet except "Target" variable which we will try to predict using the final model derived after our analysis.

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

IV - Data Cleaning and Imputation

1. Main_DataSet data cleaning and imputation

After examining the Main_DataSet data we were given, we found the following issues with the data that needs to be addressed:

- 61 variables out of 122 variables have missing values
- Date variables are coded in days
- Null values in different variables are coded differently (coded as a blank, XNA or correctly as Null)

To address these problems we decided to choose the 15% threshold of missing value to eliminate variables that have a large percentage of missing values that might impact the integrity of the analysis. As a result, we removed 53 variables that were missing more than 15% of the values. For variables remaining after this removal, we unified the way null values are coded for all variables. Then, for variables that were missing less than 10 values, we eliminated these rows from the dataset. For the remaining variables that were missing less than 15% of the data but more than 10 values, we used classification tree and regression

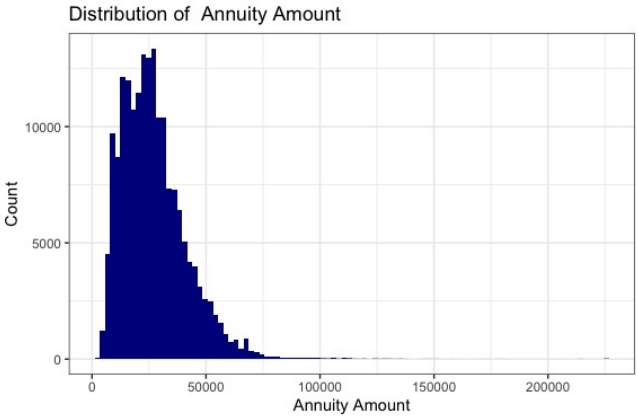
trees to impute the values. In a subsequent order, we imputed 14 variables using classification trees and regression trees. For date variables coded as days, we converted these variables into years.

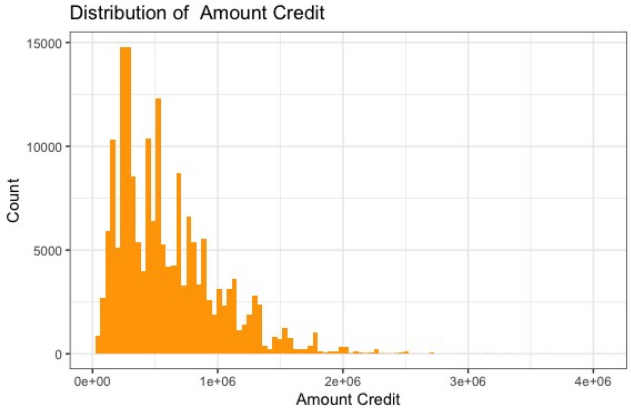
2. Application_test data cleaning

In order to match the dimensions of Main_DataSet and application_test, we converted days variables to years and dropped the columns that were also dropped in Main_DataSet.

V - Exploratory Data Analysis

To have a deeper understanding of the Main_DataSet (from now on referred as "dataset" otherwise mentioned), we first looked into the distribution of the total credit amount and annuity applicants paid. We noticed that the majority of the credit amount of the loans is between $0 and $2,000,000, while the majority of the annuity amount is between 1,000 and 80,000. Additionally, the distributions of both credit



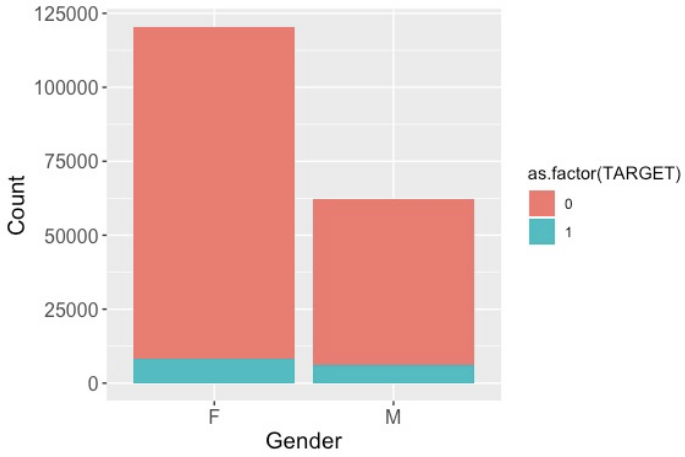


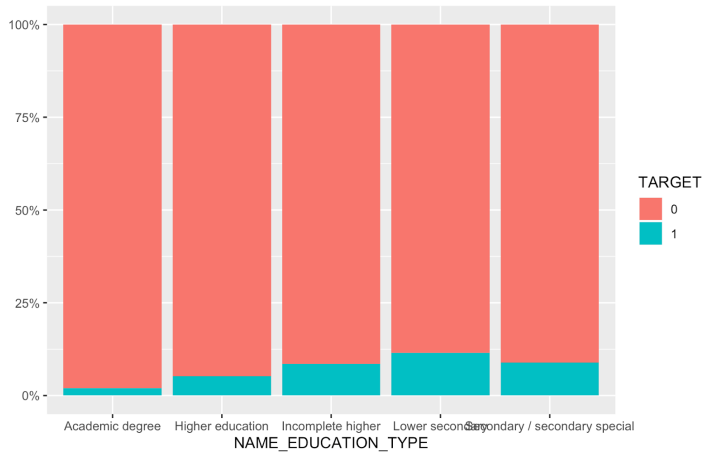
amount and annuity amount are highly right-skewed.

Considering the fact that the majority of our independent variables are categorical, we used bar plots to look at those variables that may highly influence applicant's behavior.

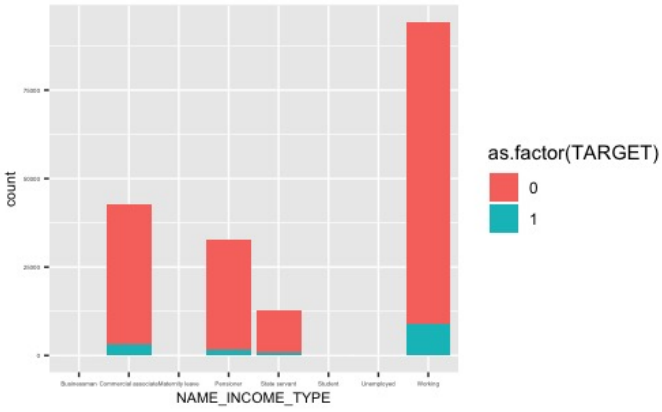
For gender, we noticed that in the dataset the number of female applicants is one time larger than the total number of male applicants, but males are more likely to default compared to females. After looking into

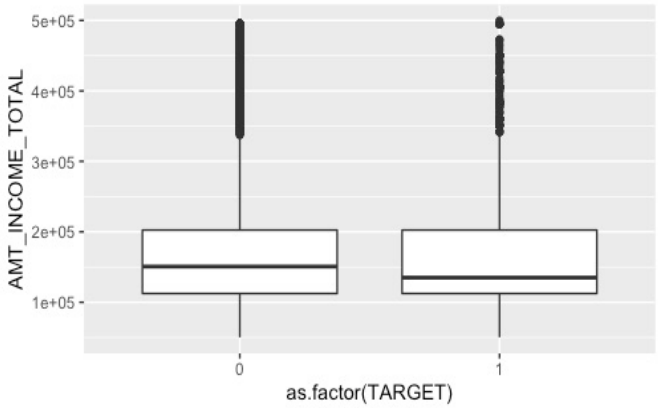
the education levels of applicants, we found that applicants who have a higher education level are less likely to default as the plot indicates.





As for income, a majority of the applicants earn their income from working, commercial associate jobs and pension. When looking into the total income of the applicants, the boxplot indicates that applicants who did not default have a higher median income compared with those who defaulted.





VI - Segmentation of Application using K-means

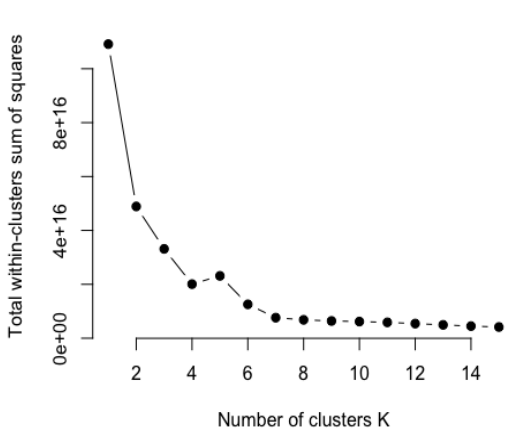
In order to provide more tailored products and rates for Home Credit's clients, we implemented the K-means clustering on the applicants' train dataset.

First, we created the sub dataset with the selected variables: AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE and REGION_POPULATION_RELATIVE. We chose these

variables because we thought that these variables can help us to better segment applicants for the Home Credit Default products.

Then we used elbow method to define the optimal number of clusters, and we found that the optimal number of clusters is 6.

After we standardized the dataset and run the K-means, we got the result below:




```
K-means clustering with 6 clusters of sizes 48172, 12708, 84576, 92400, 14136, 55500

Cluster means:
  AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE REGION_POPULATION_RELATIVE
1      0.163621106  1.2150057   1.0116606       1.2165681                -0.1635642
2      0.592381506  2.7014775   2.3031916       2.7811103                 0.3990571
3     -0.166925464 -0.8403191  -0.7981227      -0.8288675                -0.6038351
4     -0.003494494  0.1213907   0.1792794       0.0933742                -0.2981399
5      0.399794922  0.4023590   0.6313030       0.3702063                 2.8637615
6     -0.119291203 -0.6971721  -0.6484714      -0.6793798                 0.7377287
```

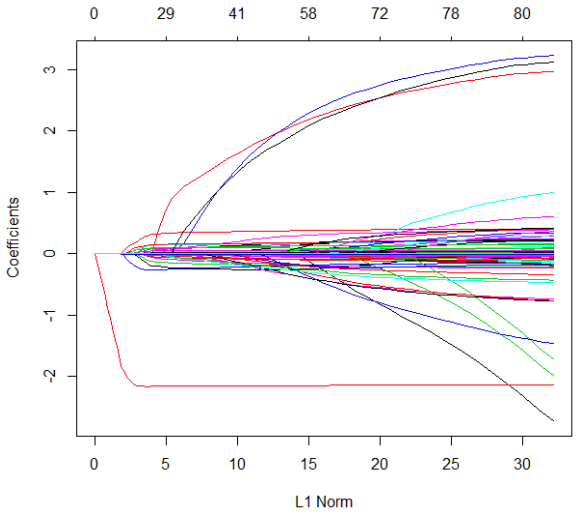
In these results, we clustered data into 6 clusters based on the initial partition that was specified. Cluster 1 contains 48172 observations and represents the clients who have slightly above-average income, relatively high credit amount of the loan, high loan annuity, relatively high price of goods in consumer loans and live in a below-average populated area; Cluster 2 contains 12708 observations and represents clients who have the highest income, highest credit amount of the loan, loan annuity and price of goods in consumer loans; Cluster 3 and Cluster 4 contain the most clients who have below-average income, the

low to average credit amount of the loan, loan annuity, the price of goods in consumer loans, and live in less populated areas.

Generally speaking, we can see that it tends to be more clients applying for small amount of loans and smaller loan annuity in less populated areas. It tends to be more variations in highly populated areas. There are clients applying for comparably larger amount of loans and also a certain amount of clients applying smaller credits amount of loan in highly populated areas. So we recommend the Home Credit company to focus on geographic segmentation.

VII - Dimension Reduction

Post visualization, we intended to have the most important variables and decided to perform a LASSO on the data set. LASSO penalized 37 variables using the best lambda value and we used the unpenalized variables to run models. With a small L1 norm, there is a lot of regularization and as L1 increases, the coefficients are non-zero and hence are included in the model.



VIII- Data Mining task

1. Model selection

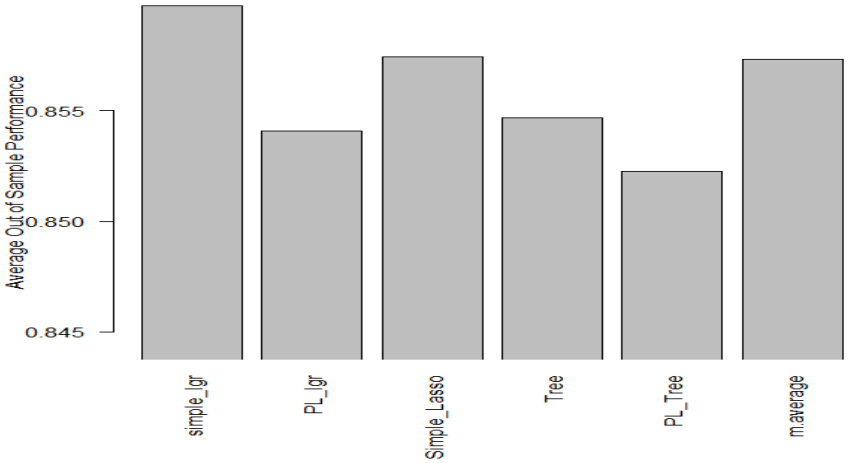
As we have defined, the core task of this project is to predict the probability that a client will be late on his/her payment and potentially default on their loans. Considering the size of our dataset, aside from

considering the interpretability, marginal impact and probability estimation ability of the models, we also have to consider whether or not we have the computation speed and capability to run the models for this task. This informed our model selection to consider the following data mining techniques and the following objectives: (Y- Model meets criteria, N- Model doesn't meet criteria)

No.	Model/Objectives	Computation speed	Interpretability	Marginal impact	Probability estimation
1	**Logistic regression**	Y	Y	Y	Y
2	**Lasso Regression**	Y	Y	Y	Y
3	**Post-Lasso Logistic Regression**	Y	Y	Y	Y
4	**Post-Lasso Logistic Regression with Interaction (Lasso selected)**	N	Y	Y	Y
5	**Classification Tree**	Y	Y	N	Y
6	**Post-Lasso Classification Tree**	Y	Y	N	Y
7	**SVM**	Y	Y	N	N
8	**Neural Network**	N	N	N	Y
9	**Ensemble model**	Y	N	N	Y
10	**Random Forest**	N	N	N	Y
11	**K Nearest Neighbor**	N	Y	N	Y

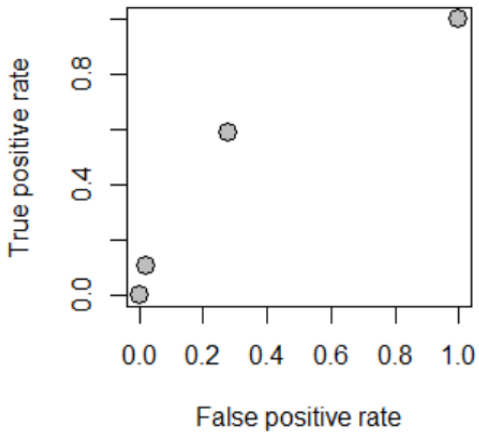
After considering the criteria mentioned, we decided to run the models we highlighted. Due to the size of the dataset, the constraints of computational power and the purpose of the data-mining task, we have put more weight on the criteria of computation speed and probability estimation. With these criteria, we decided to run the highlighted models and check for their OOS accuracy using k-fold cross validation with 10 folds.

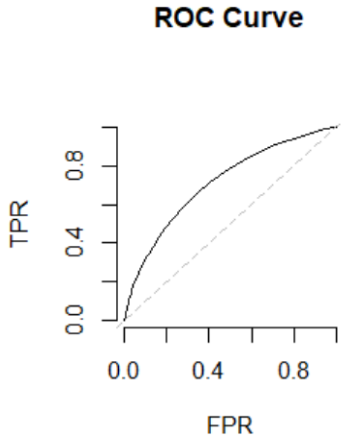
VIII - OOS Accuracy using K-Fold Cross Validation

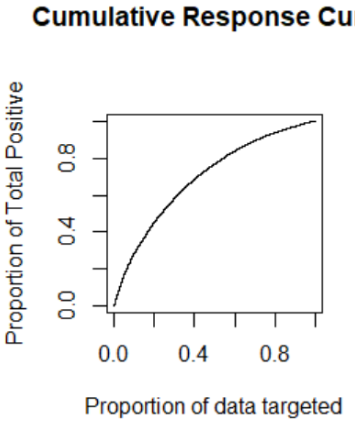


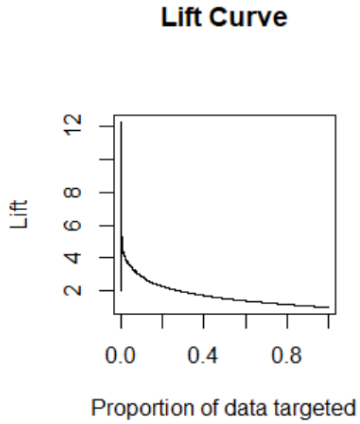
After running the different models and ignoring those models with constraints, we performed a 10-fold cross validation of 6 models namely - Simple Logistic Regression, Post LASSO Logistic Regression, LASSO, Classification Tree, Post LASSO Classification Tree and the Ensemble Model. As per the plot above, we see that Simple Logistic Regression is the best performing model amongst the five. The marginal difference between the OOS accuracy of all the models is very small indicating that the models are very comparable in terms of predictive power. Regardless, Simple Logistic Regression has the best OOS accuracy and hence we are considering Simple Logistic Regression as our best model for prediction. Our Simple Logistic Regression is as follows:

P = 1 + e^(-2.11e^{01}+CODE_GENDERM(3.57e^{-01}) + FLAG_OWN_CARY(-2.61e^{-01}) + AMT_INCOME_TOTAL(2.38e^{-07}) + AMT_CREDIT(2.57e^{-06}) + AMT_ANNUITY(8.15e^{-06}) + AMT_GOODS_PRICE(-3.06e^{-06}) + NAME_FAMILY_STATUSMARRIED(-1.66e^{-01}) + NAME_FAMILY_STATUSWidow(-1.55e^{-01}) + REGION_POPULATION_RELEVANCE(1.72) + FLAG_WORK_PHONE(2.36e^{-01}) + FLAG_PHONE(-7.63e^{-02}) + REGION_RATING_CLIENT(-1.83e^{-01}) + REGION_RATING_CLIENT_W_CITY(3.43e^{-01}) + WEEKDAY_APPR_PROCESS_STARTMonday(-6.72e^{-02}) + WEEKDAY_APPR_PROCESS_STARTSaturday(-9.15e^{-02}) + LIVE_REGION_NOT_WORK_REGION(2.94e^{-01}) + REG_CITY_NOT_LIVE(1.81e^{-01}) + EXT_SOURCE_2(-2.19) + DEF_30_CNT_SOCIAL_CIRCLE(1.71e^{-01}) + DAYS_LAST_PHONE_CHANGE(-3.89e^{-02}) + FLAG_DOCUMENT_2(2.79) + FLAG_DOCUMENT_3(5.16e^{-01}) + FLAG_DOCUMENT_5(4.88e^{-01}) + FLAG_DOCUMENT_6(4.36e^{-01}) + FLAG_DOCUMENT_8(3.02e^{-01}) + FLAG_DOCUMENT_13(-8.53e^{-01}) + FLAG_DOCUMENT_14(-8.6e^{-01}) + FLAG_DOCUMENT_16(-5.42e^{-01}) + FLAG_DOCUMENT_18(-4.93e^{-01}) + FLAG_DOCUMENT_20(1.03) + AMT_REQ_CREDIT_YEAR(3.94e^{-02}) + DAYS_REGISTRATION_YEAR(-4.91e^{-02}) + DAYS_ID_PUBLISH_YEAR(-2.22e^{-02}) + YEARS_BIRTH(-7.85e-^{03}) + YEARS_EMPLOYED(3.06e-^{02})









As we can see in the ROC graph on our hold-out data, we can see that our champion model performs better than random guessing. We looked at different thresholds of classifying the target variable in our OOS data. According to the graphs, the threshold that gave us the best TPR and FPR out of sample was 0.1. In other words, any probability generated by the model that is greater than 0.1 can be classified as 1. If we use this threshold, we have an OOS true positive rate of 53.51%, an OOS false positive rate of 23.64% and an OOS accuracy of 74.5%. According to the lift chart, if the company target approximately 40% of their customer base, the model perform approximately 2 times better than random guessing.

IX- Deployment and Business insights

Micro loan is one of the most important and fast growing markets in the financial industry now. Thus, given the fact that applicants for micro loans have no or very limited credit history, we believe it is of great importance for Home Credit to use the model we built to predict the default rate of the potential applicants. When deploying the model into practice, Home Credit should be aware that 26.01% of the applicants in the applicant_test data set without default status are flagged as having a high probability of defaulting. Using the model, Home Credit can identify more than 32,000 applicants with a high risk of defaulting on their loans out of the applicants with unknown default status from the application_test dataset.

Additionally, according to the result of the predictive model and the clustering model, we came up with some business insights for Home Credit regarding target segments and issues that the company should pay attention to. The variables giving information about applicants characteristics that highly increase the likelihood of default rate are age, gender, education level, and family status. Thus, we suggest Home Credit to regard potential applicants who are married older females with a higher education level as the target segment. For the predictive model to perform approximately twice as well as random guessing, the company can target up to 40% of their customer base.

As for concerns that Home Credit should address when issuing micro loans are:

- Filling in document 2 is of great importance to decrease the default rate of applicants given their very limited credit histories
- Given the fact that the likelihood of default is lower for applicants who apply for loans in their residential cities, Home Credit should encourage potential applicants to apply for micro loans locally
- According to clustering, it will help Home Credit to meet the needs of the potential applicants better by providing a larger number of small-amount loans in less populated area.

Appendix I

Team Contribution:

Contribution/ Team Member				
Arwen Wang	Kewei Jiang	Kieu Anh Nguyen	Taru Dharra	Vedant Sahay
Business and Data Understanding	Business and Data Understanding	Business and Data Understanding	Business and Data Understanding	Business and Data Understanding
Data Cleaning and imputation	Data Cleaning and imputation	Data Cleaning and imputation	Data Cleaning and imputation	Data Cleaning and imputation
Segmentation of Application using K means	Exploratory Data Analysis (EDA)	OOS Test using K-fold	Exploratory Data Analysis (EDA)	Dimension Reduction
Data Mining Task/Modeling	Data Mining Task/Modeling	Data Mining Task/Modeling	Data Mining Task/Modeling	Data Mining Task/Modeling
Deployment and Business insights	Deployment and Business insights	Deployment and Business insights	Deployment and Business insights	Deployment and Business insights