



Enhancing Financial Inclusion:

A Data-Driven Approach to Analyzing and Predicting Repayment
Abilities for Unbanked Populations

Spring 2023

Professor Veronika Rockova

Diemeng Hu, Yifang Zhang, Ruixin Dai, Ke Shao, Kai Gong

*We pledge our honor that we have not violated the Honor Code during the preparation of
this assignment.*

Executive Summary

In this report, we conducted an analysis of the house loan applicants' data provided by Home Credit. Our objective was to gain insights into the characteristic features and credit histories of the applicants and use this information to predict their repayment behavior. The key findings and insights derived from our analysis are as follows:

In our initial analysis, we employed various regression techniques, namely lasso, cart tree model, and random forest, to explore the relationship between fundamental information and credit histories and the repayment behavior of loan applicants. To evaluate the performance of these models, we divided our dataset into training and testing sets. Notably, all the models exhibited reasonable accuracy on the testing dataset, with the random forest model demonstrating the highest predictive capability. Furthermore, we observed a consistent pattern across all regression techniques: the variable representing the mean of the normalized credit scores from three external data sources emerged as the most significant predictor of repayment behavior. This finding suggests that the aggregated credit scores from these sources play a crucial role in predicting applicants' likelihood of repayment.

Next, we study for the most significant 10 covariates in the full LASSO model and conduct the casual LASSO method to answer if we may conclude a causal relationship between those influential factors and the response variable. We compute the confounding effects for all other variables despite the on-examining one for each covariate and include it to fit a LASSO to see if there is still some effect for the in-studying variable. We observe there are some covariates that can be mostly explained by other variables and hence the relationship between which and TARGET should not be considered as causal, while there are some variables that we are confident to conclude a causal effect given the current controls.

Then, we employ several unsupervised methods to investigate if there exist patterns in credit histories that are related to the organization types of clients. We apply PCA to perform dimensionality reduction on the 175-dimensional non-static features, reducing the data to 75 dimensions. We conduct the K-Means algorithm on the dimensionality-reduced data to perform clustering analysis with resulting six clusters and visualize clustering results by t-SNE method. We notice that there are discernible patterns in credit history records that are associated with the organization types of clients by the clustering results.

At last, we examine the prediction performance of each model we trained previously. And we also build new models to overcome the systematic errors. Confusion matrix and ROC curve are typical tools to compare the model performance. So they are used in each implementing model, in hopes to select a high accuracy and low false-positive-rate model. However after implementing different models, the prediction performance still remains pessimistic. We then retrospect on the dataset again and found some potential structure problems of the data that might be the reasons for the poor performance.

Introduction

In recent years, many people are struggling to get loans due to the insufficient or non-existent credit histories so it's hard for them to make applications and sometimes is taken advantage of by untrustworthy lenders. Home Credit strives to broaden financial inclusion for the population by providing a positive and safe borrowing experience. In order to make sure this population has a positive loan experience, Home Credit present dataset contains historical credit behaviors for the clients and covers a great variety of static data for all applications including income of the client, level of highest education the client achieved, type of organization where clients works.etc. Accordingly, we use this data to conduct our investigation into various questions related to the future repayment behavior of the clients from application.

After merging all the data together and basic cleaning, we got 204 variables and 307507 samples in total(Please refer to Appendix I for a data dictionary containing explanations of the variables used in the analysis). Our research questions are listed as follows:

- A. What static features and credit behaviors contribute to the repayment behavior?
- B. Can we conclude that there is a causal relationship between the LASSO selected variables and our response variable TARGET?
- C. Are there any patterns for credit histories devoted to features related to organization of clients?
- D. How can we accurately predict the default risk for a new customer given complete information?
- E. Are there distinct clusters of clients that have a higher probability of defaulting on their loans?

In the following part, we will present how we merged the data, how we dealt with the missing values as well as some exploratory analysis to have a clear insight into the data. Later we will conduct research on each question in order and eventually provide some conclusions of our research.

Data Preprocessing

As the main objective of our project is to analyze and predict whether applicants have repayment abilities, to ensure the quality of our analysis, comprehensive data cleaning and preprocessing are of utmost importance. Since these clients have insufficient or non-existent credit histories, there is some alternative data involved, which are provided by [Home Credit](#). They contain information about the applicants' past loans, repayment history, credit balance, and previous applications, and they supplement the main data file 'application_train.csv', which contains static data for all loan applications. A comprehensive data description that outlines the relationship between the different data points is delineated in Appendix I. First, we match these alternative data with each client's ID (SK_ID_CURR), and process and merge them into one dataset. Specifically, we convert categorical data into numerical representations using one-hot encoding and generate new features. For example, in the "previous_application.csv" dataset, we introduced "APP_CREDIT_PERC" which represented the ratio between the amount of loan applied for and the credit received. The date fields with placeholder values were also cleaned by replacing the placeholder with 'NaN' to avoid skewing the data. Then, we aggregated monthly balance snapshots for every loan in our sample. Key features like "SK_DPD" (Days Past Due), payment percentages, differences, and delay periods were also extracted, calculated, and integrated into our main dataset.

After merging the datasets together, we got 749 variables in total (including the response TARGET) which is quite big. However, when we browse through the data briefly, we could observe that there are many missing values in the data (some columns only few data are available to use which may due to the collection of data or other issues) and for the total 307,507 samples, there do not exist one single sample without any missing values, so we could not simply delete the samples with missing values. Hence in the ensuring section, we will present our attempts to deal with these missing values.

Initially, considering our sample size of 307,507, any column with more than 100,000 missing values poses a challenge for accurate imputation. Using mean values or other imputation methods could lead to significant issues in subsequent regression analyses; thus, we exclude variables with missing values exceeding this threshold. This reduction yields a total of 440 variables.

During our data integration process, we noted that a single ID number might contain multiple values for an object. Thus, we computed minimum, maximum, mean, and variance for such variables. For example, for the object BURO_DAYS_CREDIT, after the merging process, we got BURO_DAYS_CREDIT_MIN, BURO_DAYS_CREDIT_MAX, BURO_DAYS_CREDIT_MEAN and BURO_DAYS_CREDIT_VAR representing the BURO_DAYS_CREDIT's minimum, maximum mean and variance respectively. However, after

examining the data clearly, this looks quite redundant for us to use in prediction of the repayment behavior, so we will only keep the variables with MEAN as its ending and remove the other columns.

The subsequent phase involved several imputation procedures. We could observe some of the variables' name has mode in the end, some are with medi representing median and some are with AVG and MEAN which both state the mean. Thus, if the variable's name has mode, first we will remove all the missing values from that column, then compute the mode of that column and fill the missing values with the mode value we computed previously. Similarly, if the variable's name has medi, first we will remove all the missing values of that variable, then calculate the median value of the rest in the column and fill the missing values with the calculated median value. Meanwhile, for all the other columns with missing values including the columns whose name has AVG and MEAN inside, we will impute the mean values of each column to the corresponding missing values.

After all these steps, we got 368 variables in total. We give a further look into the data and observe that some columns contain most zero values and some values other than zero are extremely high, we will remove these columns as well since it may create problems in future regressions. Hence we will remove the columns whose zero value entries are more than 250,000(compared to the total sample size of 307,507). Also, SK_ID_CURR is just the ID number of each data which will be of no use for our further research so we would also remove it from our data. Finally we got our final cleaned data with 204 variables (including TARGET as our prediction).

Visualization

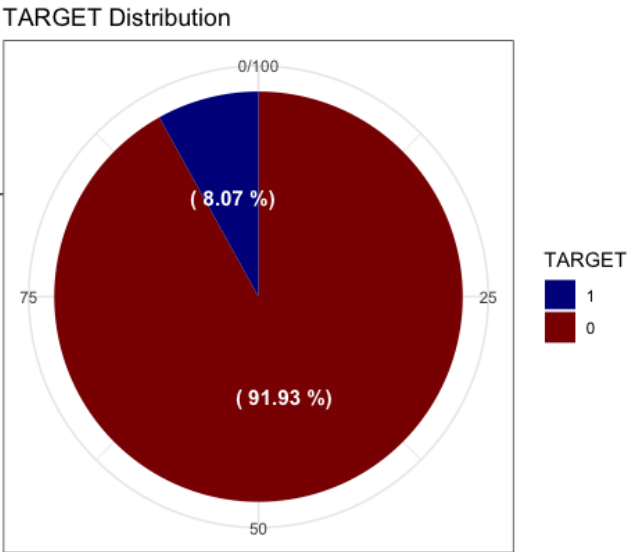
The dataset used in this analysis is quite extensive, and therefore, we have only selected a few key demographic variables for description. For variables that are identified as important in the model, they will be explained and visualized in the specific problem context. The analysis presented here is not an exhaustive exploration of the entire dataset. Depending on the specific research question, additional variables or different statistical techniques may be necessary for a comprehensive understanding of the data and to develop robust models.

Response Variable: Target

In the dataset, the 'Target' variable is the response that we would like to predict. It indicates whether a client has payment difficulties on their loan. A value of 1 is assigned if the client had late payment of more than X days on at least one of the first Y installments. Conversely, a value of 0 is assigned if the client made the payment on time. As plotted below, we could observe that 91.93% of the samples () had made the payment on time (i.e. Target=0) and the other 8.07% () had payment difficulties(i.e. Target=1). This suggests that a relatively small proportion of clients in our dataset experienced challenges with loan payments.

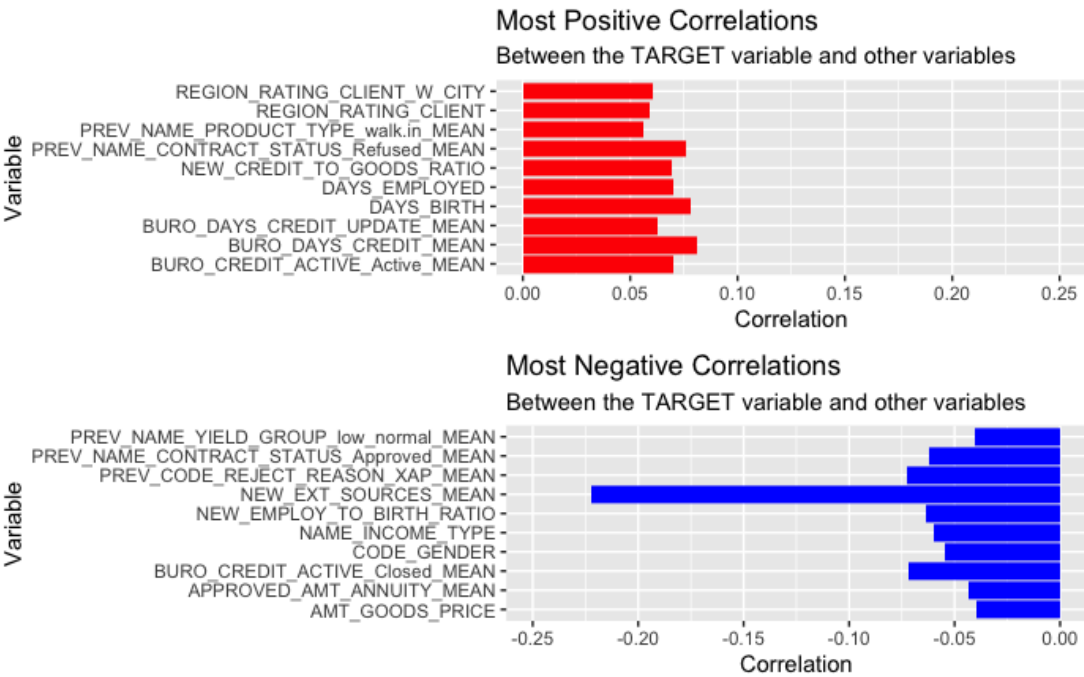
Additionally, we explored the correlation between the 'Target' and other variables. We identified the top positive and negative correlations, which provided insights into the relationship between variables and the likelihood of payment difficulties. Variables related to the client's region rating, credit history, employment duration, and credit-to-goods ratio show positive correlations with payment difficulties. This suggests that clients in certain regions, with more frequent credit updates, longer employment durations, and higher credit-to-goods ratios, may have a higher risk of payment difficulties. On the other hand, variables related to external data sources, rejection codes, closed credit accounts, employment-to-birth ratio, and approved loan contracts show negative correlations with payment difficulties. This suggests that clients with more favorable

data,



external
fewer
rejections,
closed
credit
accounts,
higher

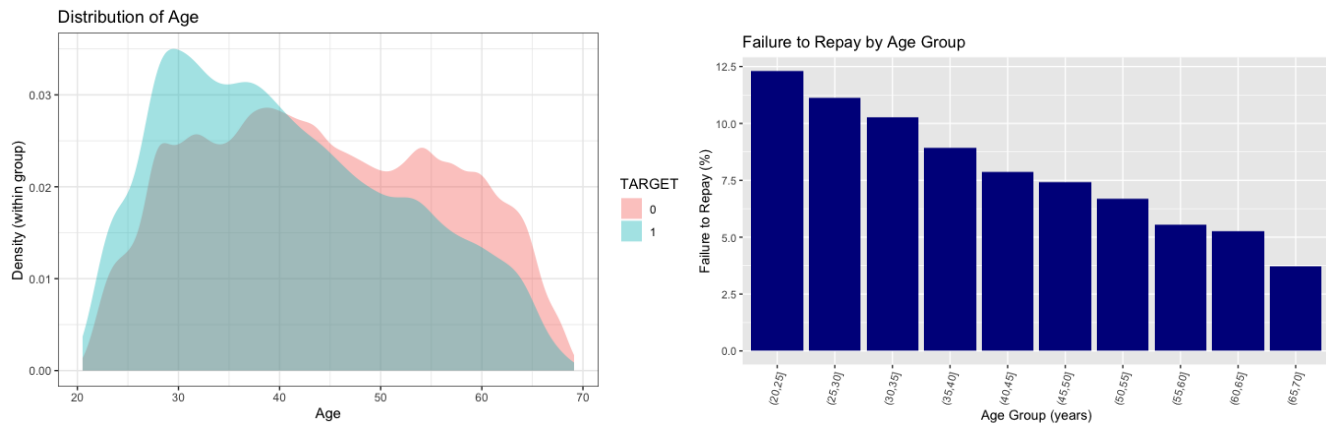
employment-to-birth ratios, and approved loan contracts have a lower risk of payment difficulties.



Clients' Demographics

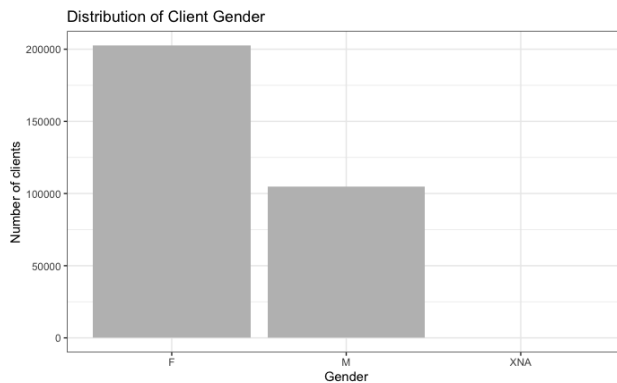
1. Age

It is noticed in the bivariate distribution plot, we observe that it skewed towards the younger end for target = 1, suggesting that younger applicants were more likely to default on their loans. To further inspect this relationship, we examined the average loan repayment failure rate across distinct age brackets. For this analysis, we segmented the age variable into bins of five-year increments. Subsequently, we calculated the average target value for each bin, thereby providing the proportion of unpaid loans in each age category. The results revealed a discernible pattern: younger applicants displayed a higher tendency towards loan default. The default rate exceeded 10% for the three youngest age groups and fell below 5% for the oldest age category. Such insights could be invaluable to financial institutions, which could provide younger clients with additional guidance or financial planning resources to enhance their likelihood of timely repayment.



2. Gender

Our data reveals that the number of female clients is almost twice as large as the number of male clients. This suggests that the necessity to borrow is more pronounced among women who possess inadequate or non-existent credit histories. However, when examining loan default rates, men exhibit a higher likelihood (10%) of not repaying their loans as compared to women (7%).

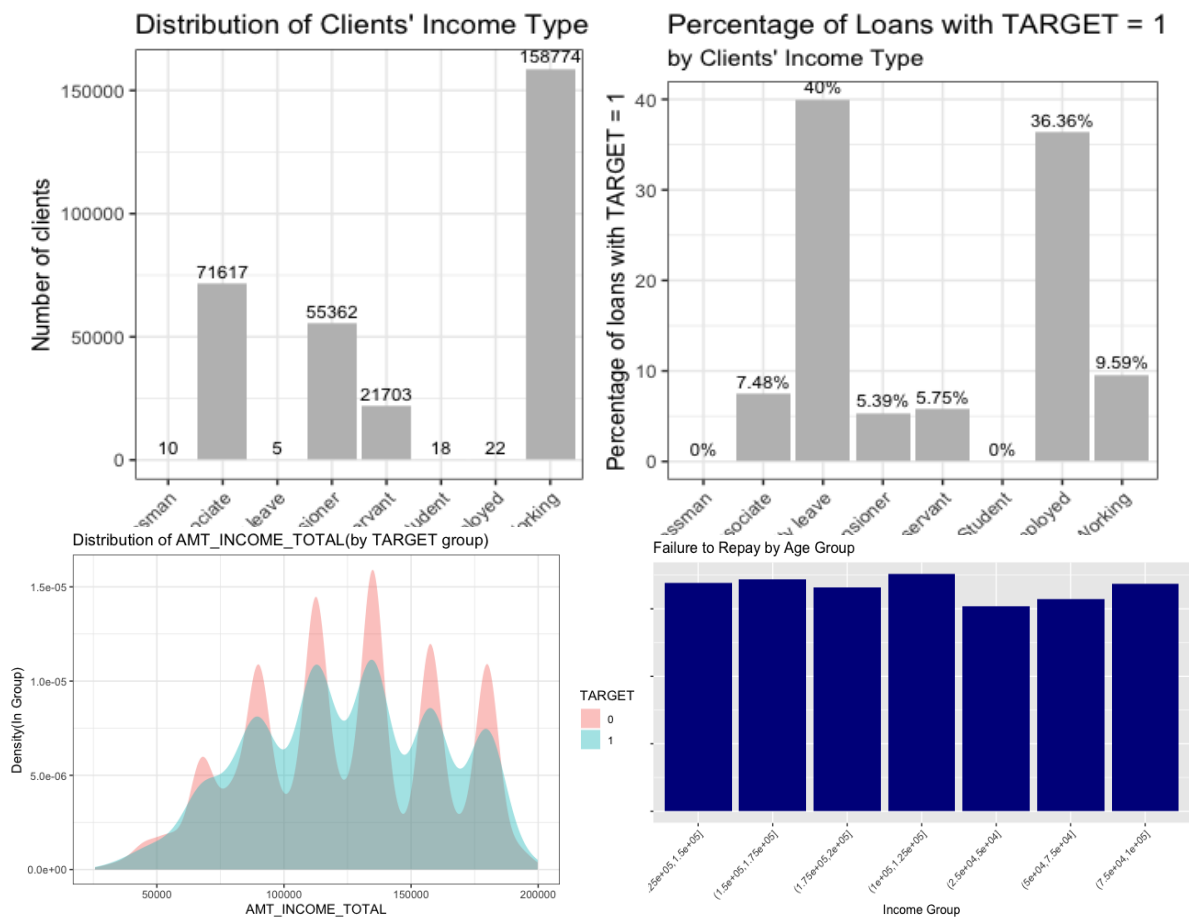


3. Income

Examining the income sources of loan applicants, we find that most applicants derive their income from employment, followed by commercial association, pension, and state service. Of note, applicants whose income is classified as 'Maternity leave' display an alarmingly high loan default rate of nearly 40%, followed closely by 'Unemployed' at 36.36%. All other income types exhibit default rates below the average of 10%. The number of "Maternity leave" and "Unemployed" in this dataset are very small. Therefore, this high percentage cannot accurately reflect the repayment ability of the population.

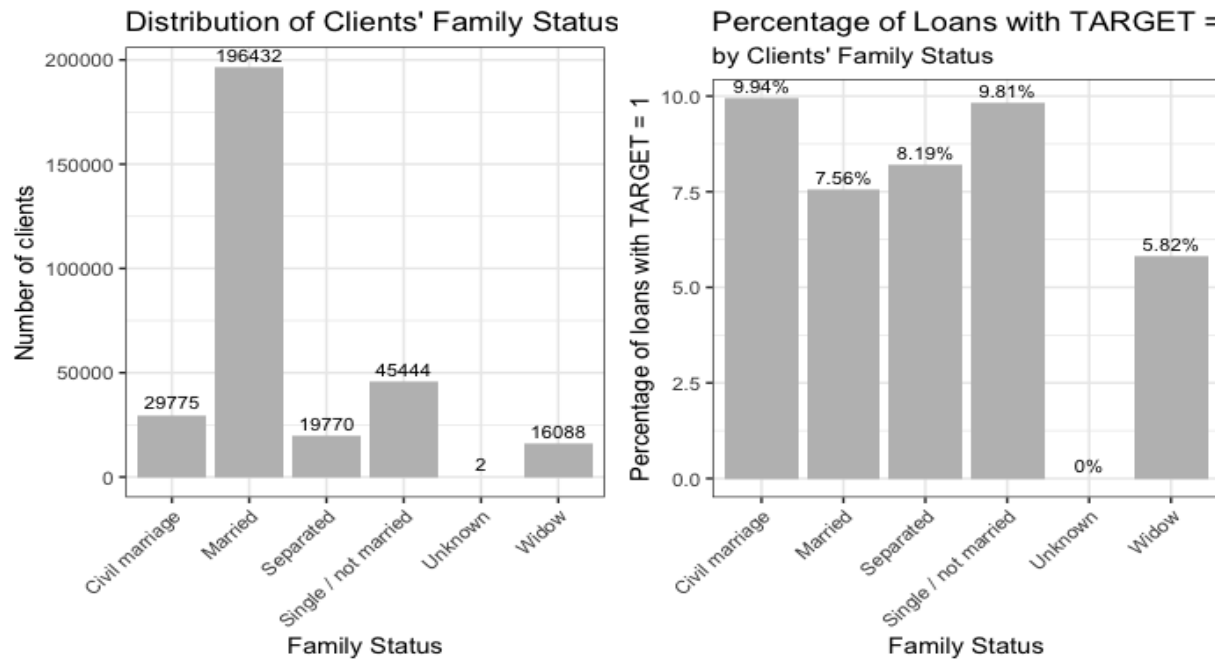
In order to assess the impact of annual income on repayment ability, we divided the annuity income into bins of \$25,000 increments. We then calculated the average target value (indicating loan non-repayment) for each income bin, which provided the proportion of unpaid loans in each income category. The results showed that the failure to repay ratio remained relatively consistent

across different income groups. This suggests that there is no clear trend between income level and repayment ability.



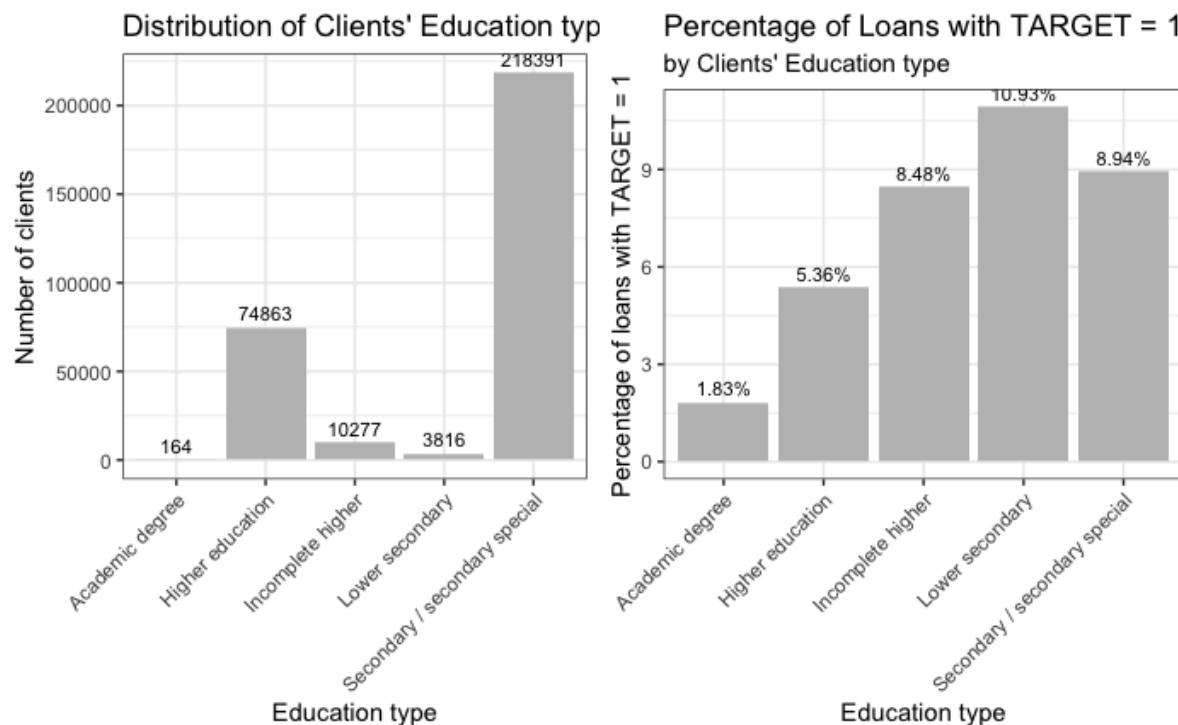
4. Family Status

The majority of clients in the dataset are classified as married, followed by clients who are single/not married and those in a civil marriage. However, when considering the percentage of not repaying the loan, clients in a civil marriage have the highest percentage of non-repayment at 10%, while clients classified as widows have the lowest percentage of non-repayment (excluding the category labeled as "Unknown"). This information provides insights into the relationship between marital status and loan repayment behavior. It suggests that individuals in a civil marriage may face more challenges or factors that contribute to a higher likelihood of not repaying their loans. On the other hand, widows seem to have a lower risk of non-repayment, indicating potentially more stable financial circumstances or a greater ability to meet their loan obligations.



5. Education

The majority of clients in the dataset have completed secondary/secondary special education, followed by clients with higher education. However, clients with an academic degree represent only a small portion of the dataset. It can be seen that the number of people with insufficient or non-existent credit histories who have a need to borrow is higher among the middle education groups. When analyzing the repayment behavior, it is interesting to note that clients with lower secondary education, although rare in the dataset, have the highest rate of not returning the loan at 11%. This suggests that individuals with lower secondary education may face more challenges or factors that contribute to a higher likelihood of non-repayment. On the other hand, clients with an academic degree have a significantly lower rate of non-repayment, at less than 2%. This indicates that individuals with an academic degree may exhibit more responsible financial behaviors or possess higher income levels, leading to a better ability to meet their loan obligations.

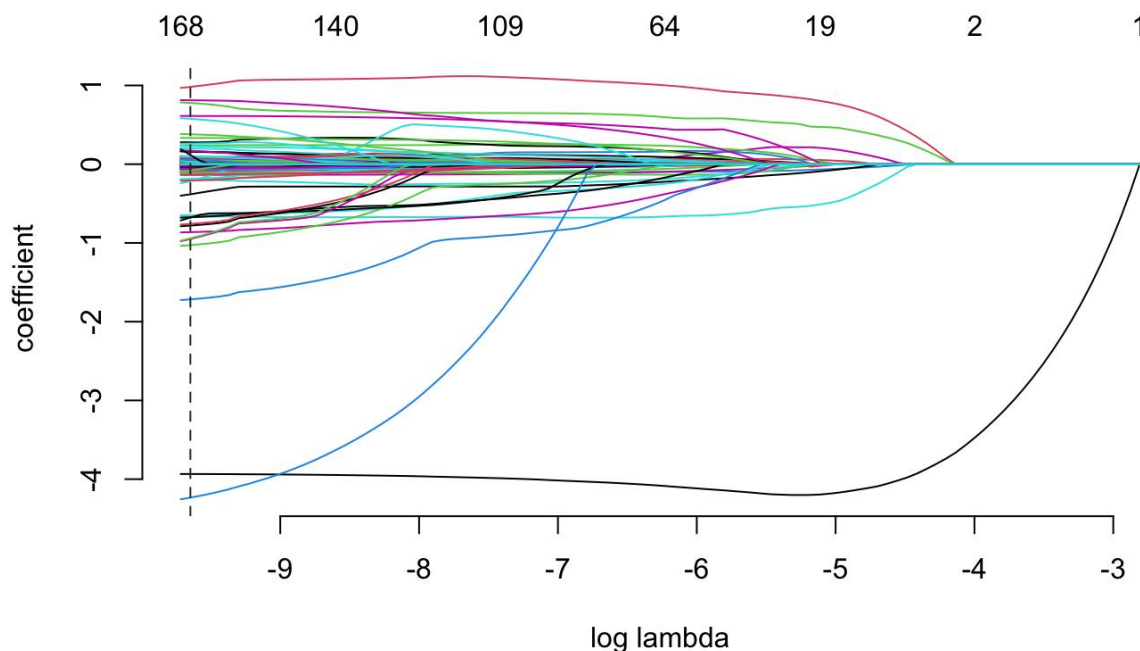


What static features of the applicants or credit histories contribute to the repayment behavior?

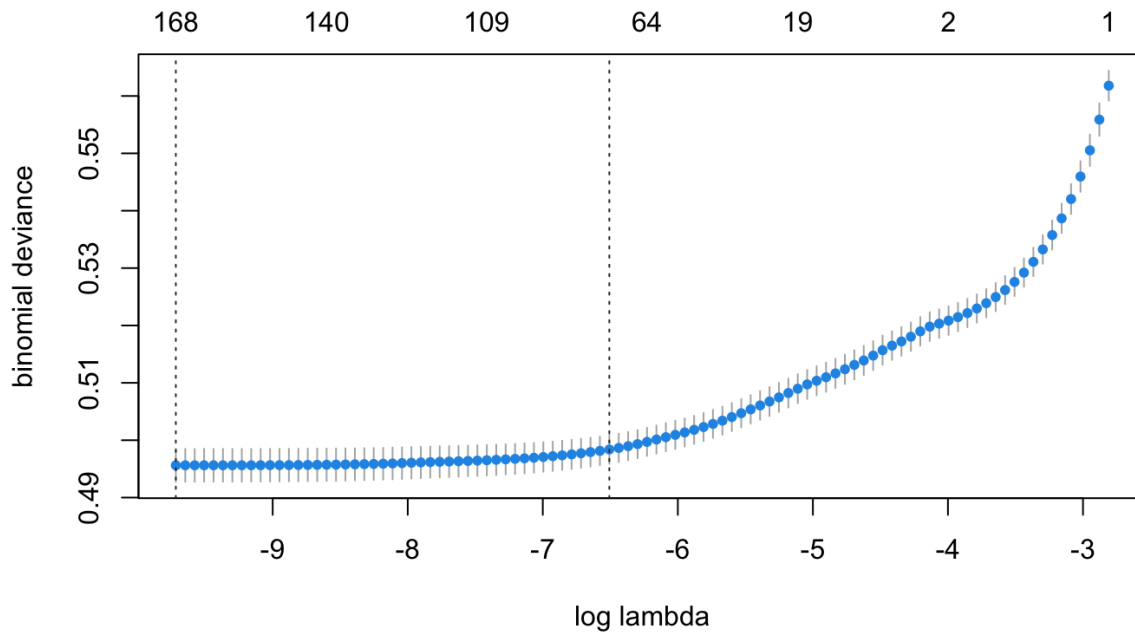
First we will split the data into training and testing with proportion 7:3 to see our prediction accuracies of different models. In this part, we were curious among all these total 203 variables in the data provided by Home Credit, which ones devote most to the repayment behavior and were interested in our ability to predict the repayment behavior by running lasso and random forest.

LASSO Model

First, we fit a Lasso model on our binary response TARGET using all the 203 predictors.



We will also utilize the cross validation method to run a `cv.lasso`, the plot is shown below. From the model, we have the best lambda value that minimizes OOS deviance is $6.024775e-05$, the number of nonzero coefficients for the best lambda is 168 and the number of nonzero coefficients for the 1se lambda is 74, the exact variables will be shown in the Appendix and here we also present the top 10 variables for both cases.



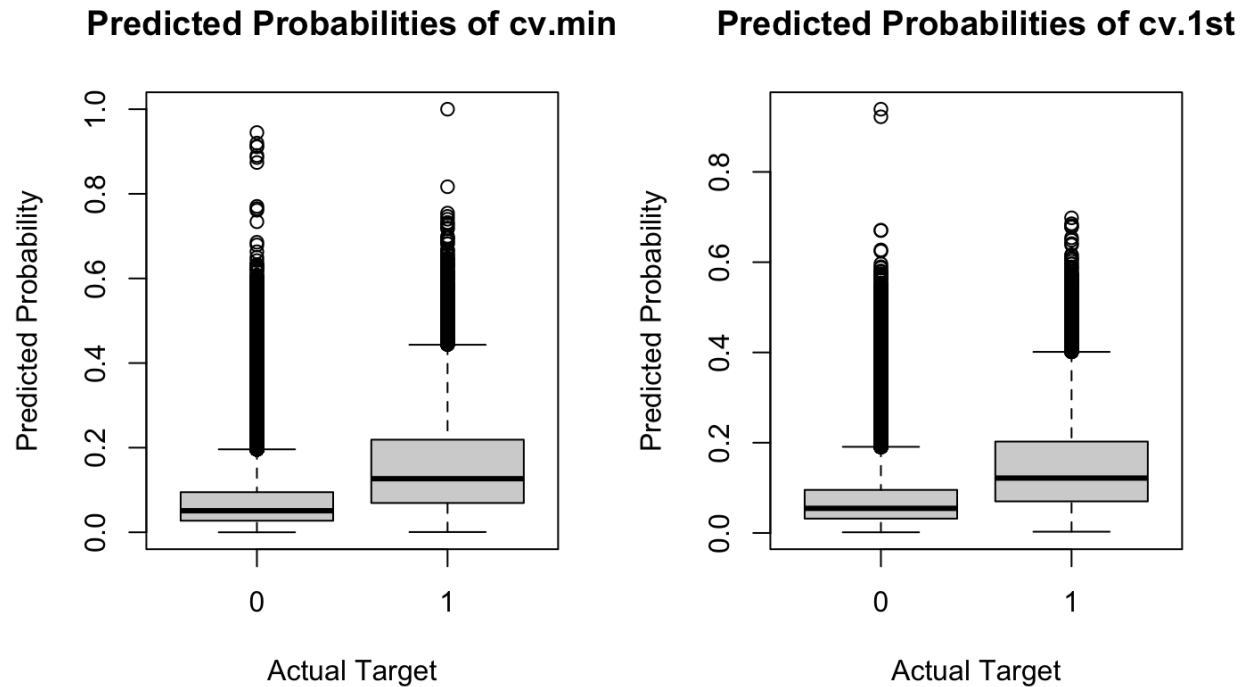
	x
NEW_EXT_SOURCES_MEAN	-4.0555732
NEW_CREDIT_TO_GOODS_RATIO	1.0325414
NEW_EMPLOY_TO_BIRTH_RATIO	-0.6782590
BURO_CREDIT_TYPE_Mortgage_MEAN	-0.6397333
PREV_NAME_CONTRACT_STATUS_Refused_MEAN	0.6271733
NEW_ANNUITY_TO_INCOME_RATIO	0.4822295
PREV_PRODUCT_COMBINATION_Cash.X.Sell.low_MEAN	-0.4813442
NEW_SCORES_STD	0.4123932
PREV_NAME_YIELD_GROUP_low_action_MEAN	-0.2900990
PREV_RATE_DOWN_PAYMENT_MEAN	-0.2613795

	x
NEW_DOC_IND_AVG	-4.2556258
NEW_EXT_SOURCES_MEAN	-3.9343831
BURO_CREDIT_TYPE_Mortgage_MEAN	-1.7256608
BURO_CREDIT_TYPE_Car.loan_MEAN	-1.0372734
POS_NAME_CONTRACT_STATUS_Active_MEAN	-0.9784014
POS_NAME_CONTRACT_STATUS_Completed_MEAN	-0.9709623
NEW_CREDIT_TO_GOODS_RATIO	0.9684843
PREV_PRODUCT_COMBINATION_Cash.X.Sell.low_MEAN	-0.8663109
NEW_ANNUITY_TO_INCOME_RATIO	0.8127775
BURO_CREDIT_TYPE_Consumer.credit_MEAN	-0.7881897

Table1. 10 variables have the most effect on the repayment behavior by using cv.lasso. The left table contains 10 most selected variables by cv.min and the right table contains 10 most selected variables by cv.1se.

We could observe that NEW_EXT_SOURCES_MEAN has contributed a lot to the prediction of TARGET in both models. Here NEW_EXT_SOURCES_MEAN is the mean of the normalized credit scores from different 3 external data sources.

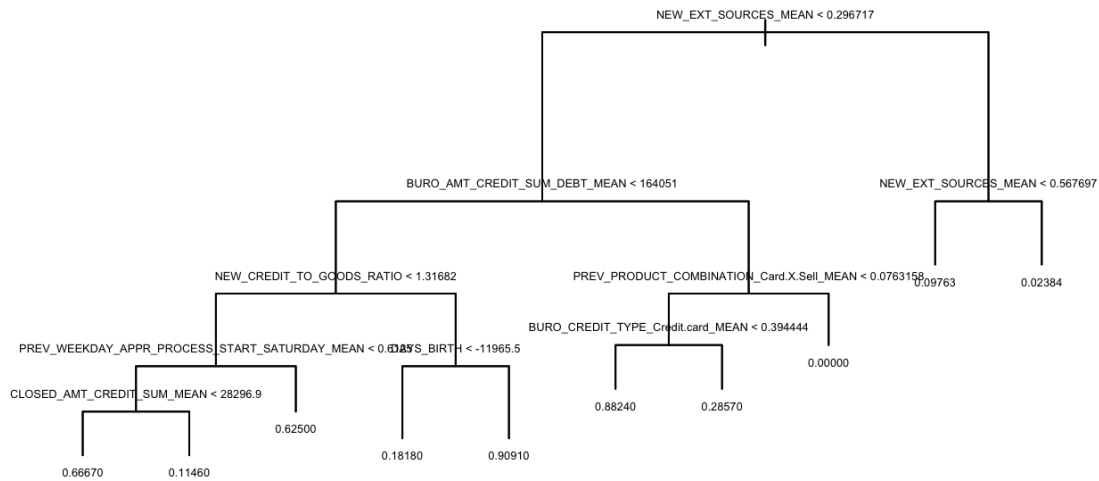
Next we also plot the TARGET with the predicted TARGET by cv.min and cv.1se respectively to have a brief look of their predictive ability.



Within our expectation, cv.min performs slightly better in prediction than cv.1se. Also, the in-sample R^2 of cv.min is 0.08544 compared to 0.07975 of cv.1se. Since the R^2 of both cases are not high, we could say that the lasso models do not perform well on this data. We also use the test data to see the prediction accuracy. On the test dataset, we have the prediction accuracy 91.98% for cv.min and 91.99% for cv.1se which looks pretty good. However, we need to notice that the data contains most TARGET=0 (the ratio TARGET=0 and TARGET=1 is greater than 9), so the prediction accuracy higher than 90% will be quite common. We also calculate the R^2 , obtaining 0.08409 and 0.08067 which are still quite low.

Since the lasso models did not perform well, we would like to utilize more models to see their ability to predict the repayment behavior.

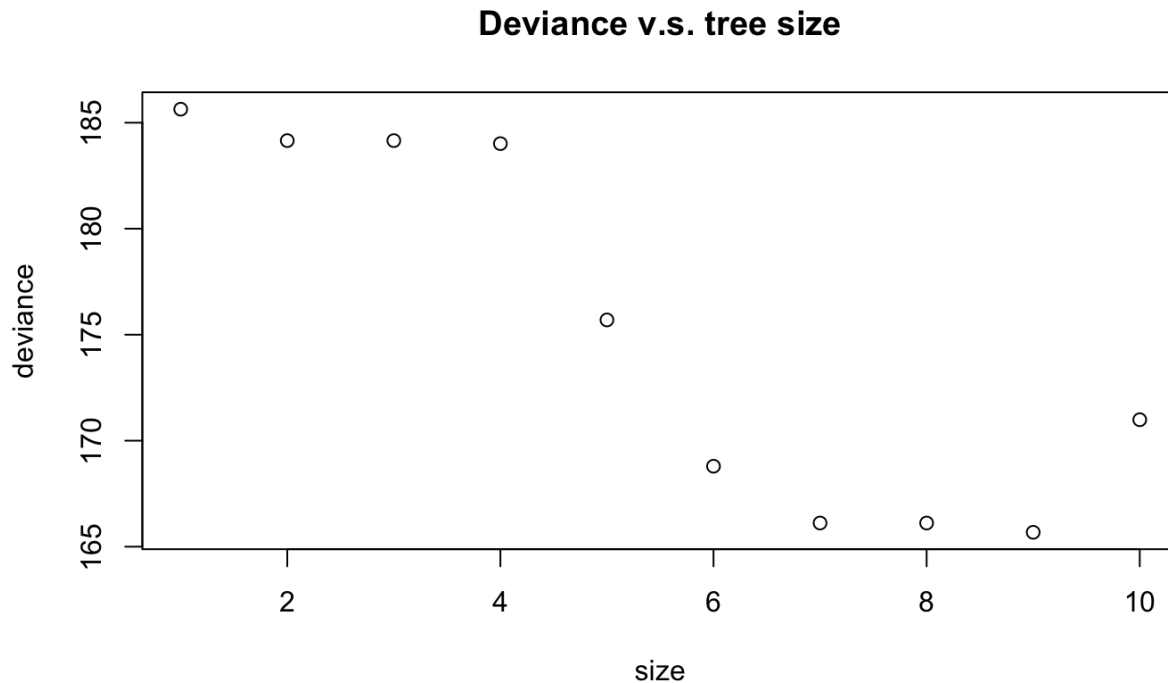
Next, we will use a cart tree model to make the prediction. Again, we present a non-parametric tree model by using the 203 variables using the training dataset to predict the TARGET. The tree is presented below.



The model selected 8 variables among the 203 variables in its construction and they are listed in the table below.

NEW_EXT_SOURCES_MEAN
BURO_AMT_CREDIT_SUM_DEBT_MEAN
NEW_CREDIT_TO_GOODS_RATIO
PREV_WEEKDAY_APPR_PROCESS_START_SATURDAY_MEAN
CLOSED_AMT_CREDIT_SUM_MEAN
DAYS_BIRTH
PREV_PRODUCT_COMBINATION_Card.X.Sell_MEAN
BURO_CREDIT_TYPE_Credit.card_MEAN

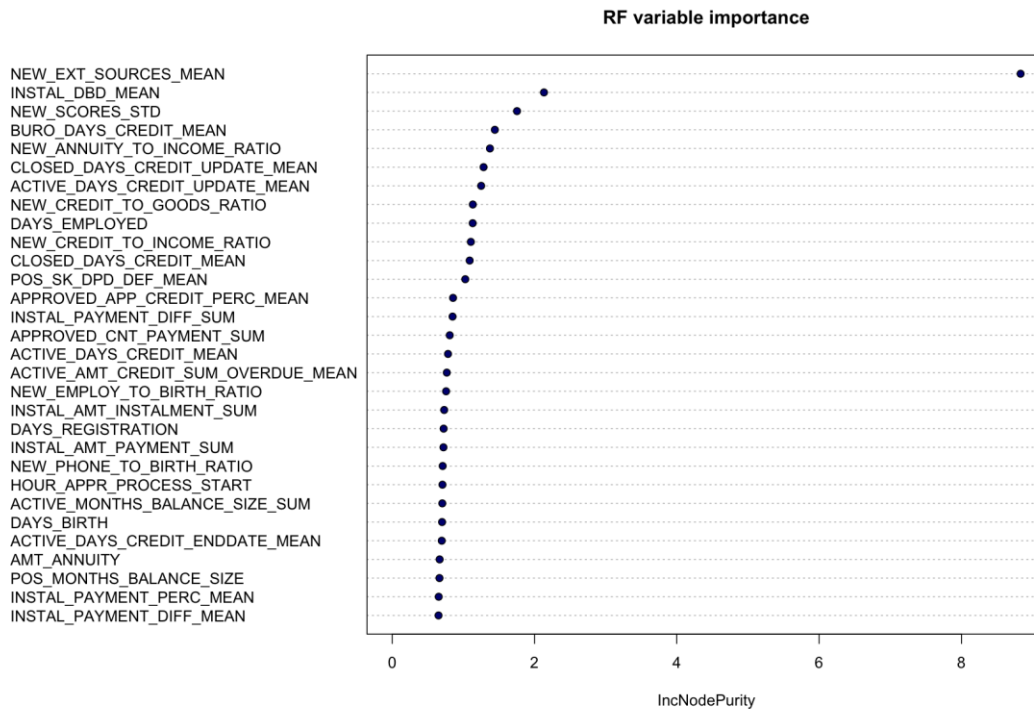
The R^2 of the tree model is 0.1505 which is much higher than the previous lasso model. We also plotted the cv.tree to see the deviance clearly to avoid the possibility of still selecting too many variables in the tree.



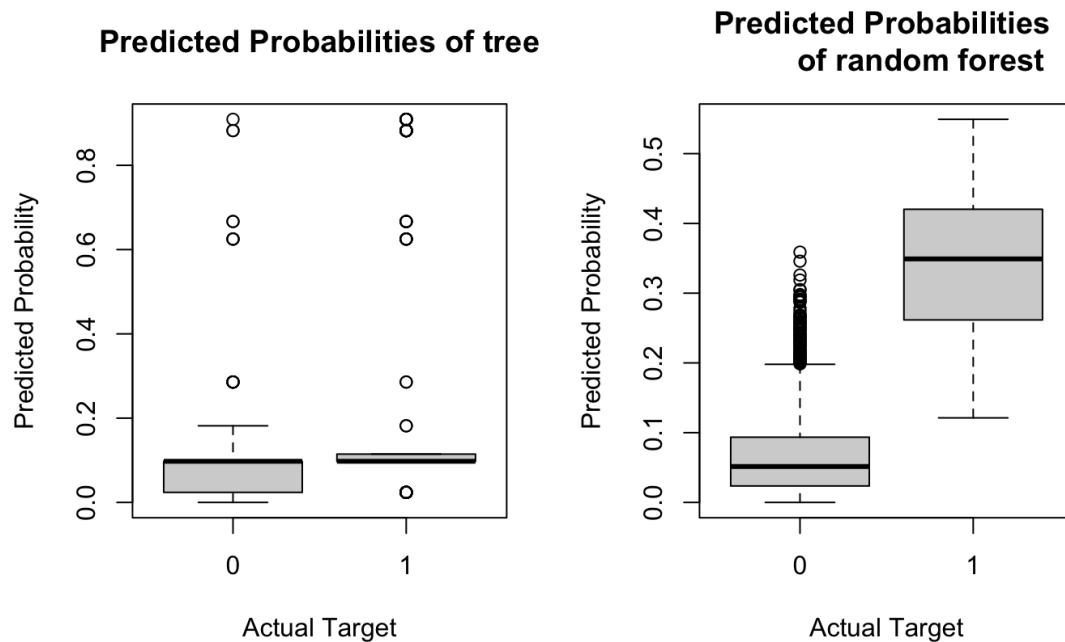
Here we could observe that 8 variables may be a good choice since it has the smallest deviance and also the size smaller than 8 have quite large deviance. Examine the tree model more carefully, we could observe that if the applicant has NEW_EXT_SOURCES_MEAN smaller than 0.297, BURO_AMT_CREDIT_SUM_DEBT_MEAN greater than \$164051 and PREV_PRODUCT_COMBINATION_Card.X.Sell_MEAN greater than 0.076, he/she will not has repayment difficulty. Similar to the lasso model, the variable NEW_EXT_SOURCES_MEAN plays an important role in the prediction of repayment behavior.

Random Forest

Finally, since the previous R^2 and prediction accuracies are still not that satisfying, we will fit a random forest for our data to see if the result of prediction has any improvement. First we will plot the most important variables in the random forest in order with their relative importance.



The random forest also selected `NEW_EXT_SOURCES_MEAN` as its most important variables in predicting whether there will be a repayment difficulty or not. The random forest has the best R^2 so far with a value of 0.4186 which is quite satisfying. Similarly to lasso, we also plotted the prediction of the tree model and the random forest model below.



Here we could observe that the prediction ability of tree is really bad but the prediction ability of random forests is pretty good and is better than all the other models shown above. We will then use the test data to see its prediction accuracy. Here we have the prediction accuracy of random forest is 92.52% which is highest among all the models presented above but the prediction accuracy for a single tree is pretty low so it's not a good way to simply use a tree to predict the repayment behavior.

Can we conclude that there is a causal relationship between the LASSO selected variables and our response variable TARGET?

Following the original full LASSO model, a natural question comes to our mind: can we conclude that there is a causal relationship between the LASSO selected variables and our response variable TARGET, and how would TARGET change as a specific treatment variable moves independently from all the other variables? To study the causal treatment effects, we conduct the Two Stage Treatment Effects Lasso procedure to the first ten significant variables excluding the intercept selected by the full LASSO model.

To get an insight of the approach, the Treatment Effect Model can be written as

$$\mathbb{E}(y | d, x) = \alpha + d\gamma + x'\beta,$$

where y denotes the response variable TARGET, d denotes the treatment variable we would like to study, and x' denotes all the other variables except d . The idea is to remove the effect of any other influences that are correlated with d , which are called ‘controls’ or ‘confounders’. They are variables whose effect can be confused with that of d . To control for confounding effects, we remove confounders from the predicted γ by including them in regression. Writed $= x'\gamma + noise$,

$$\begin{aligned}\mathbb{E}[y|x, d] &= d\gamma + x'\beta \\ &= (x'\tau + \nu)\gamma + x'\beta \\ &= \nu\gamma + x'(\gamma\tau + \beta) = \nu\gamma + x'\beta^*\end{aligned}$$

In this case γ is identified as the effect of ν , the independent part of d . Note this is an inference problem, not a prediction problem.

In our dataset, as described in the previous part, we have a total of 203 columns of variables that were used to predict for the response variable TARGET, which combined the information from over seven hundred variables. As before, since there are relatively many covariates, we want to analyze the treatment effect while controlling for confounders in a high dimension. To be more specific, we want to forecast y for new $x \in R^{m \times N}$, $m = 203$, $N = \text{No. of data points}$, while d changes independently from x . This naturally leads us to the Causal Lasso method introduced in class where we transfer the causal estimation to two prediction problems:

1. Estimate $\widehat{d(x)}$ with LASSO regression of d on x
2. Fit another LASSO model of y on $[d, \widehat{d(x)}, x]$ with $\widehat{d(x)}$ unpenalized

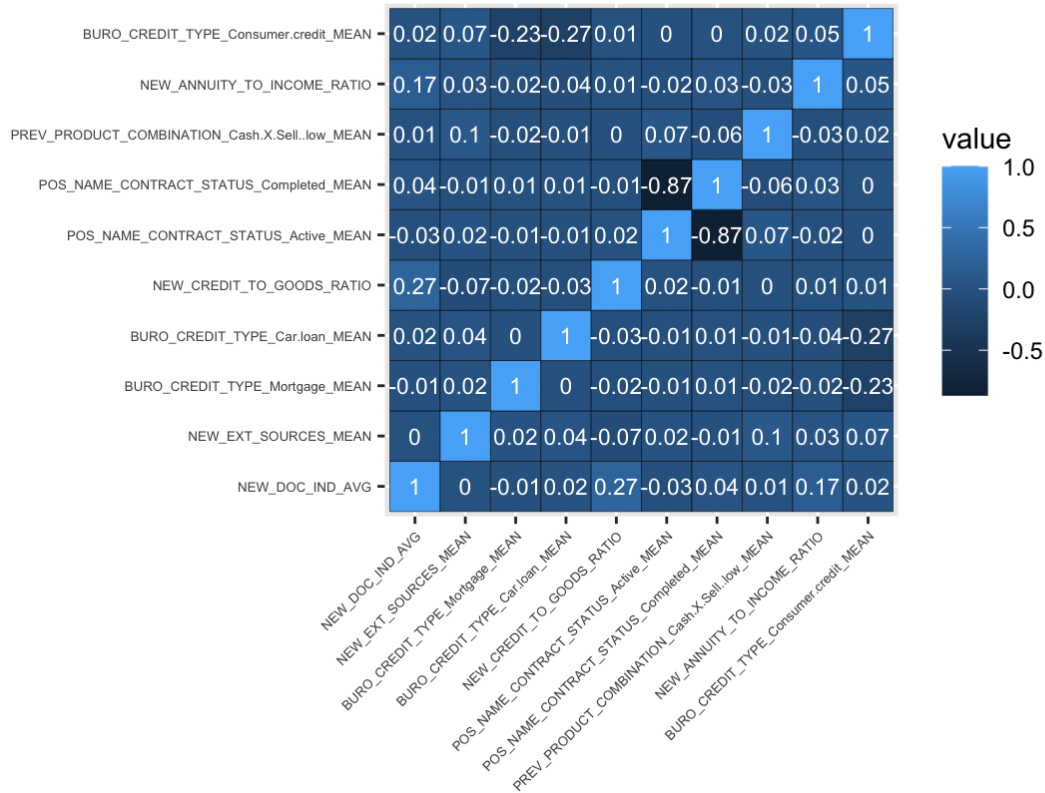
Here, by including $\widehat{d(x)}$ unpenalized in stage 2, we remove the confounder effects on d and thus $\hat{\gamma}$ measures the effect of v .

Understanding the method, we get back to our data set. To start up, we first look at the first ten significant covariates, their meanings and data summaries.

Ind.	Variable Names	Coefficients	Description	Min	Mean	Max
1	NEW_DOC_IND_AVG	-4.23	Avg. NO. of FLAG_DOCUMENTs provided by each client	0.00	0.05	0.20
2	NEW_EXT_SOURCES_MEAN	-3.93	Avg. normalized score from the three external data source	0.00	0.51	0.88
3	BURO_CREDIT_TYPE_Mortgage_MEAN	-1.72	Avg. NO. of clients' related previous credits in credit bureau having Credit_Type == Mortgage	0.00	0.01	1.00
4	BURO_CREDIT_TYPE_Car_loan_MEAN	-1.03	Avg. NO. of clients' related previous credits in credit bureau having Credit_Type == Car_loan	0.00	0.02	1.00
5	NEW_CREDIT_TO_GOODS_RATIO	0.98	Current credit amount for the Credit Bureau credit / Goods price of good that client asked for (if applicable) on the previous application	0.15	1.12	6.00
6	POS_NAME_CONTRACT_STATUS_Active_MEAN	-0.95	Avg. NO. of clients' contract status (approved, canceled, ...) of previous application == Active	0.0	0.9	1.0
7	POS_NAME_CONTRACT_STATUS_Completed_MEAN	-0.94	Avg. NO. of clients' contract status (approved, canceled, ...) of previous application == Completed	0.00	0.08	1.00
8	PREV_PRODUCT_COMBINATION_Cash.X.Sell.low_MEAN	-0.86	Avg. NO. of product combination of the previous application == Cash.X.Sell.low	0.00	0.05	1.00
9	NEW_ANNUIITY_TO_INCOME_RATIO	0.81	AMT_ANNUIITY / 1 + AMT_INCOME_TOTAL Annuity of previous application	0.00	0.18	1.88
10	BURO_CREDIT_TYPE_Consumer.credit_MEAN	-0.78	Avg. NO. of clients' related previous credits in credit bureau having Credit_Type == Consumer.credit	0.00	0.72	1.00

All variables are numerical and range mostly from 0 to 1. The variables align with the results from other models that NEW_EXT_SOURCES_MEAN, NEW_CREDIT_TO_GOODS_RATIO, NEW_ANNUIITY_TO_INCOME_RATIO, etc. are some of the most significant variables. We can see the most significant variable is NEW_DOC_IND_AVG with a corresponding coefficient of around -4. This indicates a negative relationship between the variables and the response variable TARGET. However, we cannot say that the independent movement of NEW_DOC_IND_AVG would affect TARGET at this time since the effect of NEW_DOC_IND_AVG on TARGET may be because of the effects of other variables. We calculate the correlation matrix between the 10 top significant covariates and demonstrate it using a heatmap.

Correlation Heatmap for Full LASSO Top 10 Significant Covariates



According to the heatmap, except for POS_NAME_CONTRACT_STATUS_Active_MEAN and POS_NAME_CONTRACT_STATUS_Completed_MEAN, the correlation between which two variables is -0.87, other variables seem to be quite independent from each other with most pairs of them bearing a correlation around 0. However, since we have another 193 variables, we can still not be optimistic to conclude causal relationships.

To actually see if the variables could independently affect the change of TARGET, we first fit a LASSO model for each of the top 10 covariates, the treatment variables d 's, on the x 's formed by the other 202 variables in the data set. We select for the best score model and compute for $\widehat{d(x)}$, which is hence the best predictor for d from x . Next we find the best predictor for y from d and x , after influence of $\widehat{d(x)}$ is removed. Store the coefficient on d , this is what we want as the causal effect $\hat{\gamma}$.

We repeat the above process ten times to get $\widehat{\gamma}_d$'s for each of the ten treatment variables d . We compute the in-sample R^2 for each first stage fit to see how much of the examining d can be explained by x and equivalently how much of d can be considered as independent to other variables. Then we report the second stage fitted coefficients $\widehat{\gamma}_d$. The result table is shown below.

Ind.	Treatment Variable Names	In-Sample R^2	Causal Effect
1	NEW_DOC_IND_AVG	0.9991	0
2	NEW_EXT_SOURCES_MEAN	0.3304	-0.3167
3	BURO_CREDIT_TYPE_Mortgage_MEAN	0.7256	-0.1253
4	BURO_CREDIT_TYPE_Car.loan_MEAN	0.7441	-0.0930
5	NEW_CREDIT_TO_GOODS_RATIO	0.6963	0.0722
6	POS_NAME_CONTRACT_STATUS_Active_MEAN	0.9540	0
7	POS_NAME_CONTRACT_STATUS_Completed_MEAN	0.9415	0
8	PREV_PRODUCT_COMBINATION_Cash.X.Sell..low_MEAN	0.6376	-0.0284
9	NEW_ANNUITY_TO_INCOME_RATIO	0.9161	0.0267
10	BURO_CREDIT_TYPE_Consumer.credit_MEAN	0.9594	-0.0396

We may easily find there are three different kinds of results by sorting the in-sample R^2 from high to low (color from deep to light):

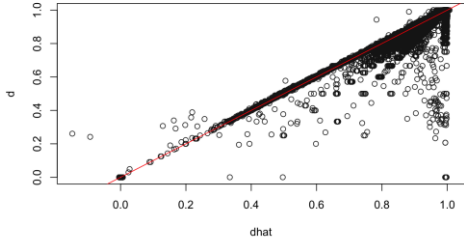
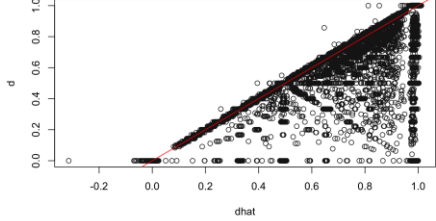
1. High in-sample R^2 , few causal effect
2. Relatively high in-sample R^2 , some causal effect
3. Relatively Low in-sample R^2 , relatively much causal effect

Here we separately analyze the three kinds of results by giving examples for each kind.

Start with the first group with the highest in-sample R^2 .

For Treatment Variables 1, 6, 7, 9, 10, we have a nearly 1 in-sample R^2 , suggesting over 90% variation of d could be explained by x . Note $v = d - d(x)$, we say there is almost no independent movement of the five Treatment Variables to measure as effecting TARGET. In this case, after including $\widehat{d(x)}$, we would be quite sure we would get a $\widehat{\gamma}_{d_i} \approx 0$ for $i = 1, 6, 7, 9, 10$, as there is nearly no independent movement for these five variables and the significance while fitting for the full LASSO model come from their interactions with x 's.

Note for Treatment Variables 9 and 10, although it seems to have an absolute causal effect around 0.03, we may need to consider carefully and probably introduce other methods to say there is actually a causal effect as the in-sample R^2 's are large and the desired coefficients are negligibly small.

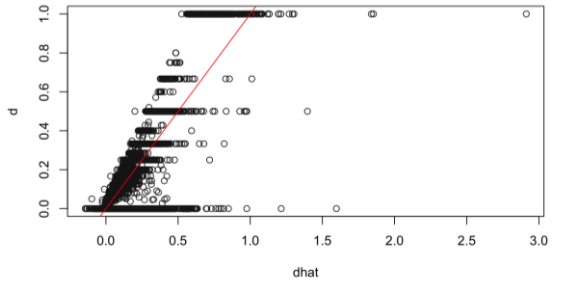
Group 1: Relatively Low in-sample R^2 , relatively much causal effect		
Treat Var. Index	6	10
Treat Variable Name	POS_NAME_CONTRACT_STATUS_Active_MEAN	BURO_CREDIT_TYPE_Consumer.credit_MEAN
In-sample R-squared	0.9540	0.9594
Treat True vs. Predicted Plot		
Causal Effect of Treat	0	-0.0396

We select POS_NAME_CONTRACT_STATUS_Active_MEAN as an example and plot for the $\widehat{d(x)}$ vs. $d(x)$. We may see the vast majority of the points fall onto the line $d(x)=\widehat{d(x)}$, aligning with what was shown by the R squared. Also by the plot, there is still some nonconstant variance, but looks better to me compared to raw. For BURO_CREDIT_TYPE_Consumer.credit_MEAN, the plot looks good.

Now we see to the second group where there are around 0.7 in-sample R^2 and around 10% causal effect.

We include Treatment Variables 3, 4, 5 to be in this group. The in-sample R^2 indicates an around 70% variation of d can be explained by x , and hence there is some information in d independent of x and upon which we can measure a treatment effect. Observe $\widehat{\gamma}_{d_i} \approx 10\%$ for $i = 3,4,5$, we conclude that at least amongst these controls, we have evidence showing the Treatment Variables 3, 4, 5 and related to TARGET.

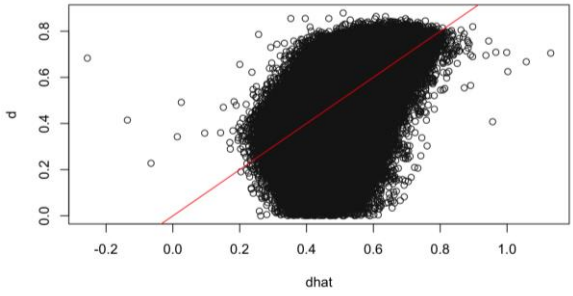
Group 2: Relatively high in-sample R^2 , some causal effect

Treat Var. Index	Treat Variable Name	In-sample R-squared	Treat True vs. Predicted Plot	Causal Effect of Treat
3	BURO_CREDIT_TYPE_Mortgage_MEAN	0.7256		-0.1253

Take the third significant variable `BURO_CREDIT_TYPE_Mortgage_MEAN` as an instance of Group 2. We plot for the Treat True vs. Predicted, the plot looks okay to me. Here $\widehat{\gamma}_{d_3} = 12.53\%$, we conclude there is a causal effect amongst the control. To moreover interpret the coefficient, we may say each extra unit of `BURO_CREDIT_TYPE_Mortgage_MEAN` would cause TARGET to decrease 0.1253 units.

Lastly, for the second significant variable `NEW_EXT_SOURCES_MEAN`, which individually form the third group as it bears the lowest in-sample $R^2 = 0.3304$, indicating that there are around 33.04% variation of d that could be explained by x . In this case, we say there is much information in d independent of x , and upon which we can measure a treatment effect. By the plot of $\widehat{d}(x)$ and d , we notice a misalignment which aligns with our conclusion that there strong signal for us to measure for effect of d after controlling for x . Here we have $\widehat{\gamma}_{d_3} = 31.67\%$, we may conclude there is a causal effect on TARGET.

Group 3: Relatively Low in-sample R^2 , relatively much causal effect				
Treat Var.	Treat Variable Name	In-sample R-squared	Treat True vs. Predicted Plot	Causal Effect of Treat

Index				
2	NEW_EXT_SOURCES_MEAN	0.3304		-0.3167

However, note many things would actually affect this certain variable

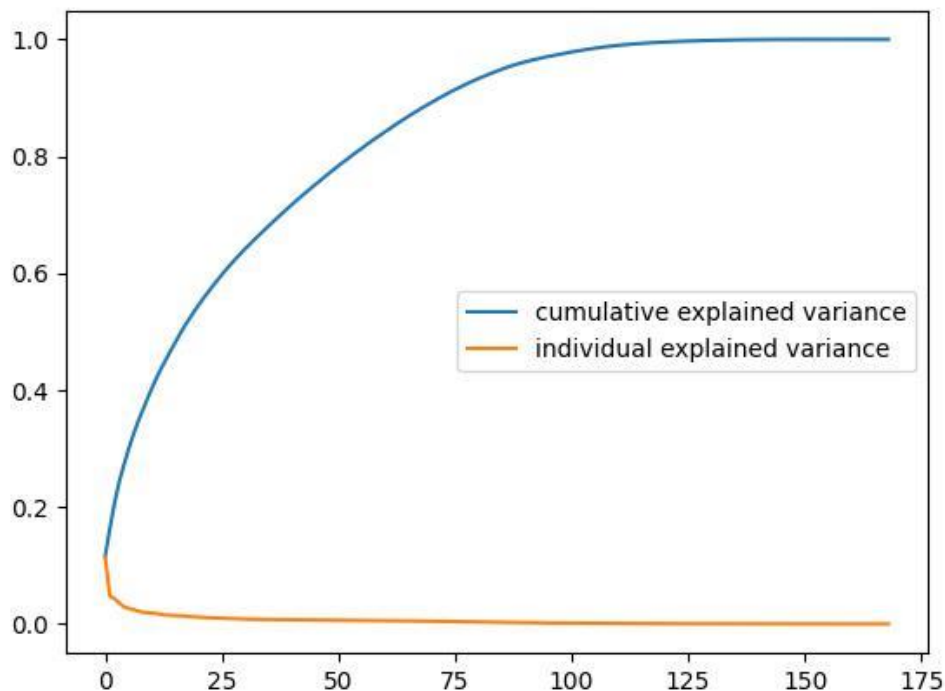
NEW_EXT_SOURCES_MEAN, as it is an external source data which we are not sure according to the data description. Thus, x here may not contain all relevant variables that could potentially influence the outcome. Thus although by fitting a lasso regression model $d \sim x$ and including the estimated $d(x)$ unpenalized as a control variable into our causal model, the potential for causal inference is improved and we could expect the confounder effect to be reduced, we cannot be confident that our result will be causal. Namely we should expect there to be a confounding effect although weakened, and less confident when making the final conclusion.

Are there any patterns for credit histories devoted to features related to organization of clients?

In this section, we delve deeper into the credit histories devote features of the data, rather than focusing solely on static features. Our goal is to investigate whether any interesting patterns exist in credit histories for different groups of clients.

Principal Component Analysis

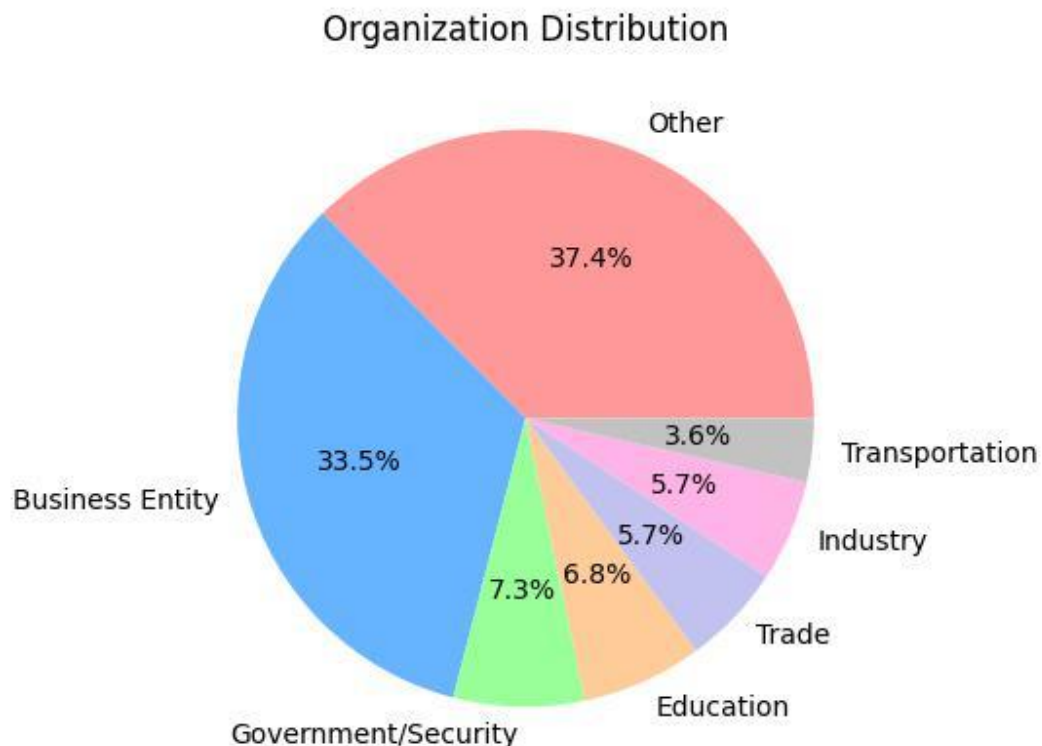
Given the large number of credit histories devote features per observation in the dataset, Principal Component Analysis (PCA) can be employed to gain a more insightful understanding of the data. PCA allows for dimensionality reduction while preserving as much information as possible. Prior to implementing the PCA algorithm, it is crucial to normalize each column (i.e., feature) of the dataset, as PCA is sensitive to scaling. Following the normalization step, PCA is applied, resulting in a plot that illustrates the cumulative information with respect to the number of components.



We set 0.85 as the threshold for percentage of explained variance. Choose `n_components = 75`. We have done dimensionality reduction for credit histories by focusing on essential features. The original dimensionality of 175 has been reduced to 75, while only a small amount of information has been lost in the process.

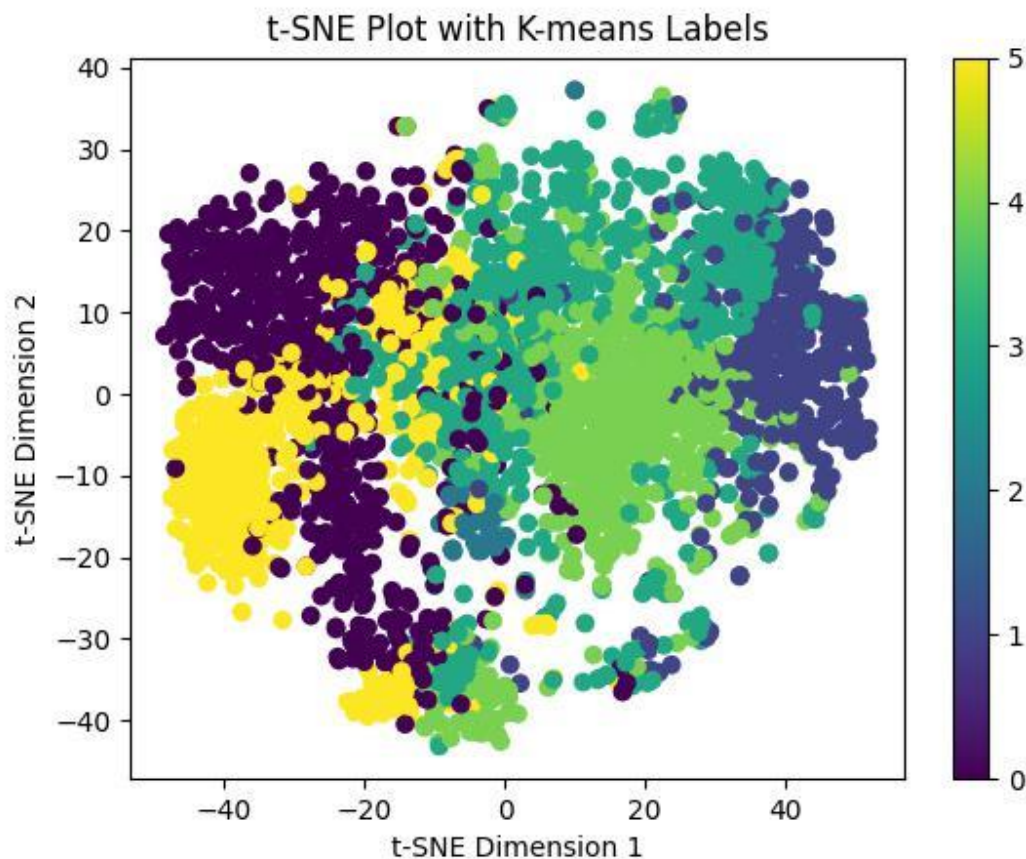
K-Means Clustering with t-SNE Visualization

We will utilize 75-dimensional data obtained through PCA to apply various clustering methods. Specifically, we will employ the K-Means technique for our clustering analysis. Our objective is to examine whether there are any discernible patterns between clients' organization types and different clusters based on their credit history features. The original dataset comprises 55 distinct organization types. To simplify the analysis, we have consolidated several organization types into seven final categories: 'Business Entity', 'Education', 'Government/Security', 'Transportation', 'Trade', 'Industry', and 'Other'.

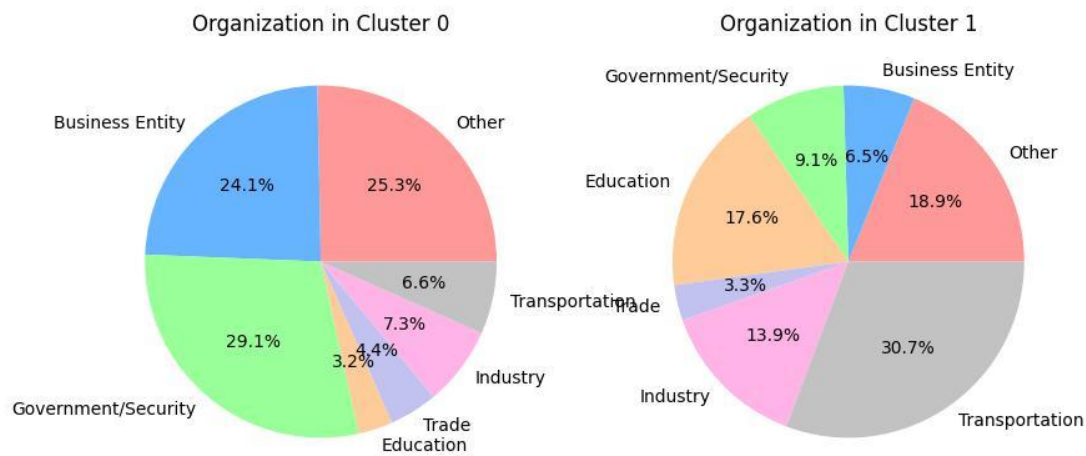


By merging similar organization types into broader categories, we aim to uncover potential associations between these categories and distinct credit history clusters. This analysis will help us gain insights into the relationship between organization type and credit history, contributing to a better understanding of our clients' credit profiles and potential risk factors.

After performing parameter tuning, we have determined that a value of $k = 6$ is optimal for our k-means clustering. To visualize the results of our clustering, we employed t-Distributed Stochastic Neighborhood Embedding (t-SNE). t-SNE is a dimensionality reduction technique that utilizes a heavy-tail distribution in a two-dimensional space, enabling clear differentiation between distant data points. The t-SNE plot presented below illustrates the outcomes of our K-means clustering.



We present the distribution of organization types for the first two clusters generated by the K-Means clustering algorithm in the plot below.



Clear patterns can be observed between the clusters and organization types in the plot. Cluster 0 predominantly comprises organization types such as 'Business Entity', 'Government/Security', and 'Other'. On the other hand, Cluster 1 is characterized by a higher representation of organization types such as 'Transportation', 'Education', 'Industry', and 'Other'. These findings highlight the distinct composition of organization types within each cluster, providing valuable insights into the relationships and distributions captured by the K-Means clustering algorithm.

How can we accurately predict the default risk for a new customer given complete information?

In this section, we will discover new models focusing on predictions. We would check the performance of Lasso, random forest, PCA and lightGBM trained in previous sections first. Then we will include SVM (Support Vector Machine) and Neural Network. Different from previous analysis on the factors, in this section we will not consider much on whether the covariate contributes to the model significantly or not. Instead, we focus on the prediction result of each model. It can be interpreted by comparing the predicted classes with the actual classes. If a predicted class does not match the actual class, that's a misclassification. Misclassifications can be organized into a confusion matrix, which breaks down the types of correct predictions (true positives and true negatives) and errors (false positives and false negatives). This matrix can be used to calculate various metrics such as precision, F1 score, and accuracy to measure the performance of the model.

LASSO

Among the previous models, the most typical model in analyzing the covariate contribution is LASSO. In this section, we use LASSO as an example to illustrate how these models perform in prediction.

Confusion matrix:

	Reference	
	0	1
Predictions		
0	91.67%	8.09%
1	0.12%	0.11%

Some statistics on lasso model prediction:

Accuracy : 0.9183

Sensitivity : 0.99875

Specificity : 0.01454

Pos Pred Value : 0.91924

Neg Pred Value : 0.50935

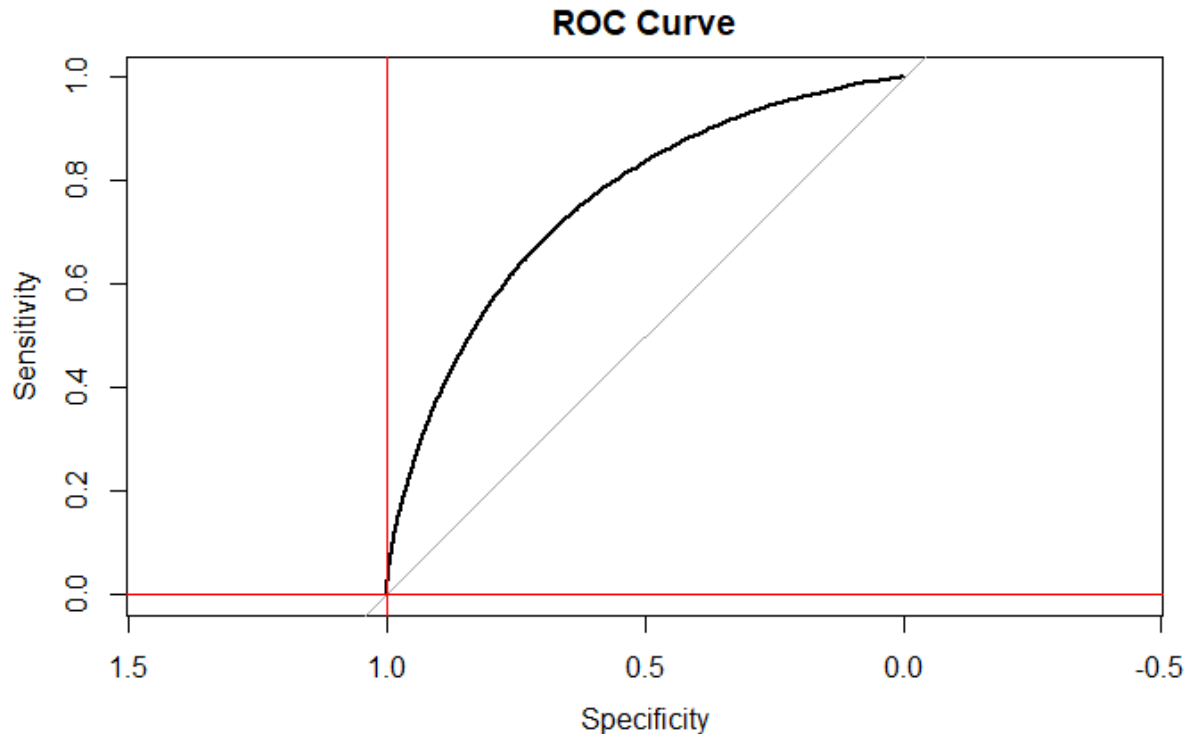
Prevalence : 0.91824

Detection Rate : 0.91709

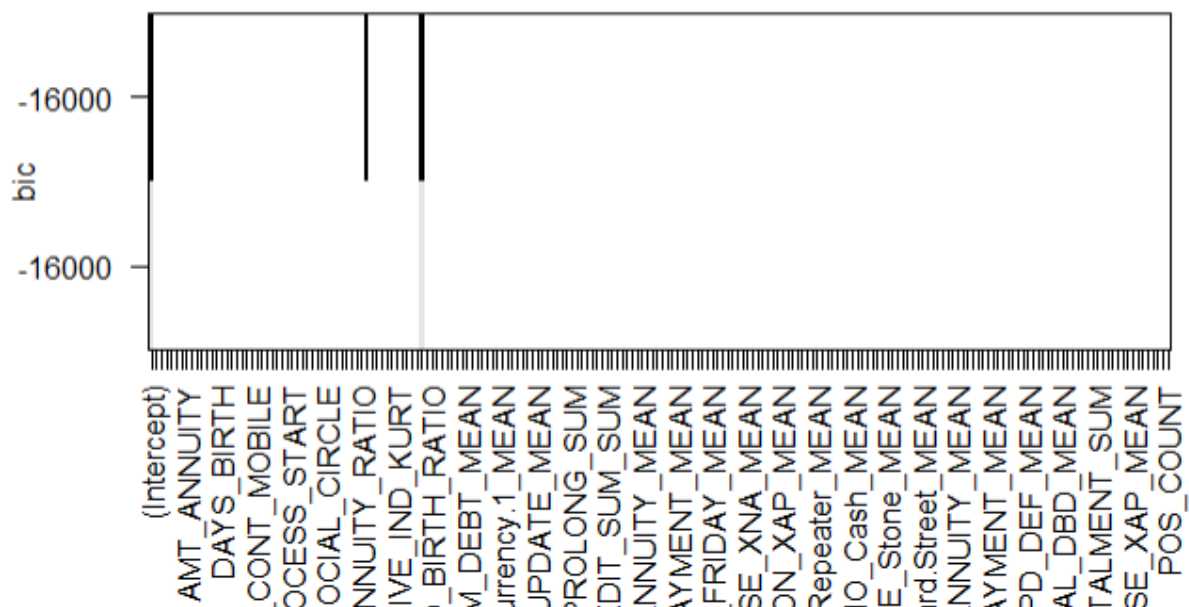
Detection Prevalence : 0.99767

Balanced Accuracy : 0.50665

The Receiver Operating Characteristic (ROC) curve is a useful tool for comparing different classifiers. It is a graphical plot that displays the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. If the curve of one classifier completely encloses that of another, the first classifier is always better than the second. In general, the worse the classifier is, the closer to the diagonal line its curve is.



The ROC curve shows that the LASSO model is not very good at performing the prediction. The result would lead to a huge false negative rate and a considerable false positive rate, which indicates that the lasso model would not suggest a lot for new clients to check if they would be accepted into the load based on the result. To take a deep understanding of the reason under the large misclassification rate, we run a similar step function based on the BIC criteria to see how the selection on covariates influences the prediction.



From the selection step graph, it is very obvious that most of the covariates are omitted at the early step, which leads to the tremendous drop of information entropy (measured as BIC). The model would not perform well with such a low contain of information, causing misclassifications.

Support Vector Machines (SVMs)

In order to avoid the drawback of systematic information lack, we introduce the SVMs. Support Vector Machines are a type of binary linear classification model. The goal of an SVM is to find the best separating hyperplane which classifies the data points into two classes. This hyperplane has the largest distance to the nearest training data points of any class.

SVMs use a kernel (Radial Basis Function in this case) to map the data to a higher-dimensional space where a hyperplane can be used to do the classification. The output of an SVM model in R is a list containing various information about the trained model. The key element is the "SV" element which represents the support vectors, i.e., the observations that lie closest to the hyperplane and thus determine the margin.

Confusion matrix:

		Reference	
Predictions		0	1
	0	91.57%	8.08%
	1	0.13%	0.11%

Compared to the Lasso model, the SVMs model performs better in the false negative part. It raises the accuracy to 92.38%, but the result is not significant enough. This is due to the many existing level categorical covariates (such as Superior, middle, lower) in the dataset, and they are treated to be continuous in the SVMs. This problem is not solvable because these categorical variables have order but cannot be carelessly turned into continuous hyperplanes.

In order to overcome this difficulty, we introduce neural networks to develop the model. In the NNs model, we can encode the categorical variables into continuous distributions by setting embedding layers.

Neural networks

Neural networks are a type of model inspired by the human brain. They consist of layers of nodes, with each node in a layer connected to all nodes in the previous and next layer. The nodes are where the computation happens: they take a weighted sum of their inputs, add a bias term, and then pass the result through a non-linear activation function.

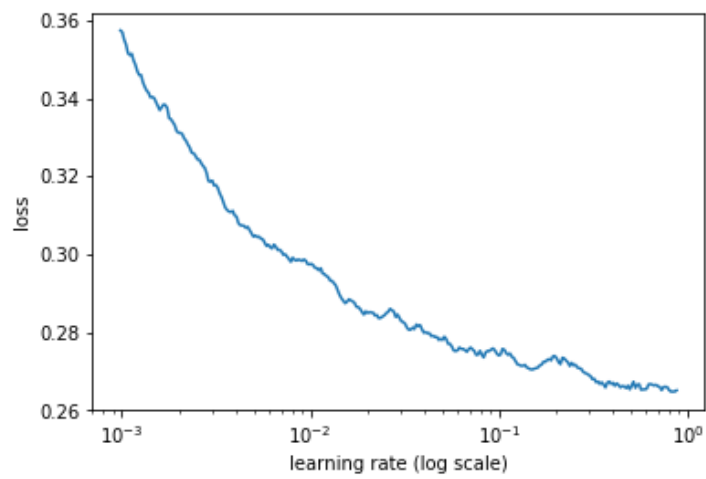
The weights and biases are the parameters of the model, and they're what get updated during training. The training process involves iteratively updating the weights and biases to minimize the difference between the model's predictions and the actual values. This difference is calculated using a loss function, and the updates are done using an optimization algorithm (gradient descent in this case).

We use an embedding layer for each categorical data, the embedding size is given as:

```
emb_szs = [(c, min(50, (c+1)//2)) for _, c in cat_sz]
```

Besides the embedding, we add 3 fully connected layers with size [500, 250, 250].

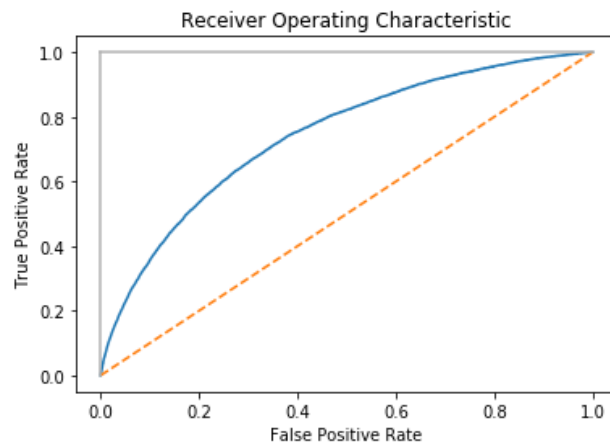
Under the NNs model, we dealt with all systematic problems that occurred previously. The model should give a more stable and reliable result.



The output Confusion matrix, with normalization:

		Reference	
		0	1
Predictions	0	91.92%	8.06%
	1	0.13%	0.12%

ROC curve:

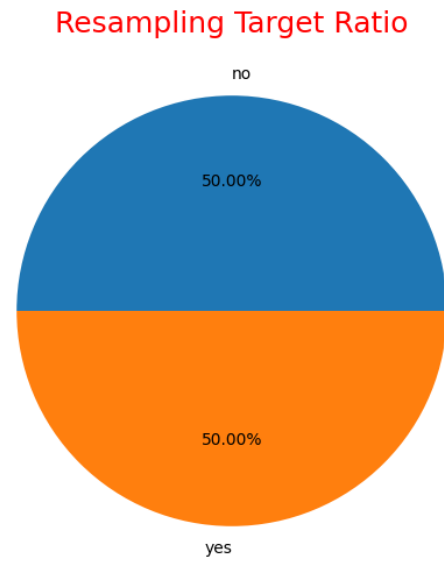
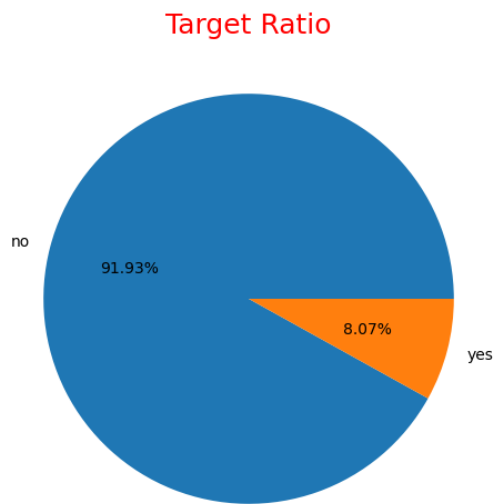


From the analysis result, we can see that the NNs still do not give a promising result, despite its slightly better performance than other models. We would not conclude any systematical error in the model. Hence the reason for poor performance of prediction might lay under the dataset itself. There would be several factors that might contribute to the high false positive rate as shown below.

Retrospect on the dataset

- **Imbalanced data:**

As we revisit the target column, the dataset is heavily imbalanced (i.e., one class is represented much more than the other). It caused the model to learn to always predict the majority class, causing a significantly high false positive rate. In this case, we could try techniques like oversampling the minority class, undersampling the majority class, or using a combination of both (like SMOTE or spline).

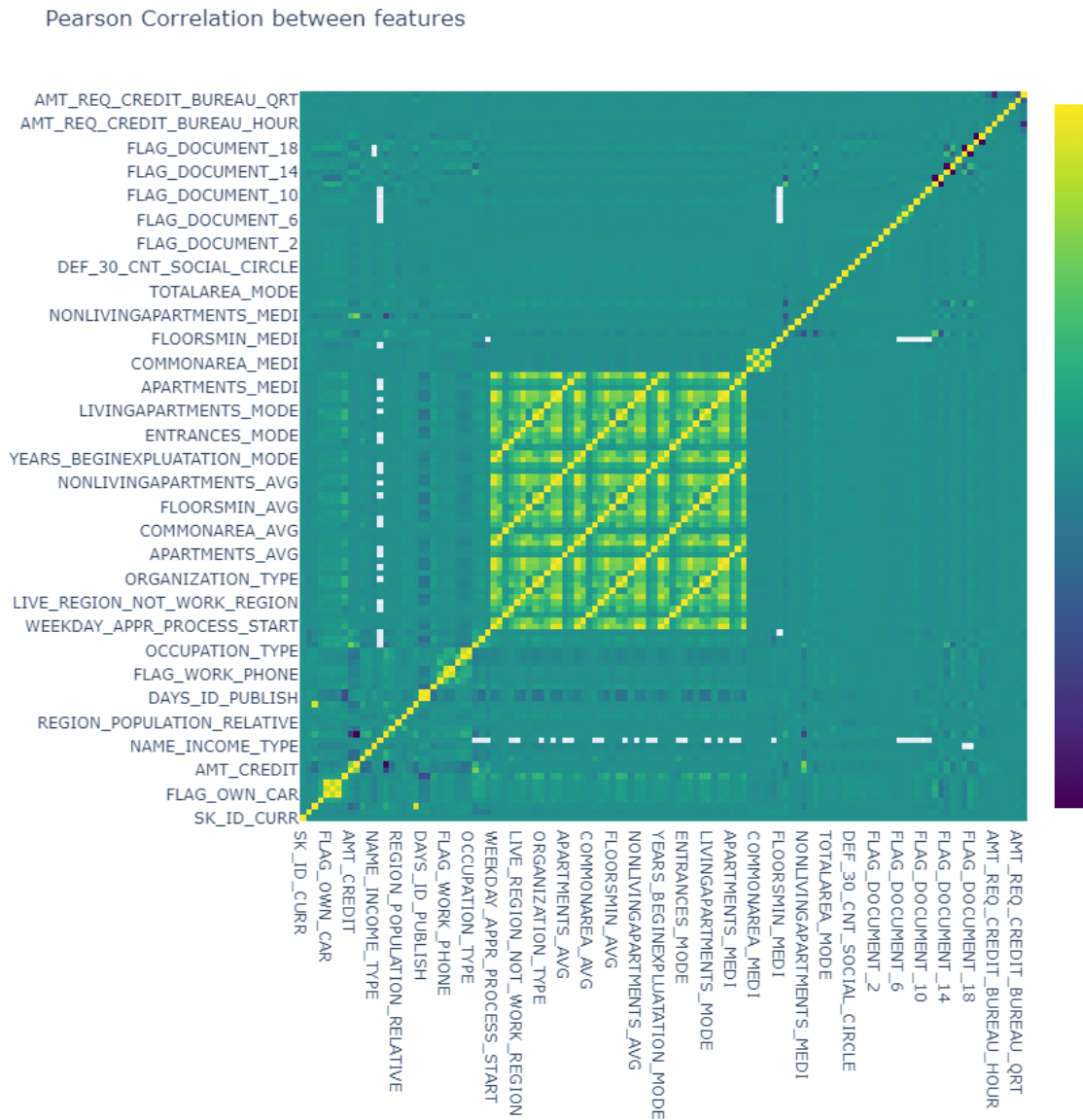


- **Low correlation between y and x:**

We can never make a successful model when y is completely independent from x. Similarly, it is of low expectation to train a model with low obviously observed patterns between x and y. In this case, even if there is no significant systematic error made in the model, the performance of prediction would still run low.

- **Low correlation between dummy variables:**

Most categorical variables in the dataset are not quite correlated with each other as shown in the below heatmap. This leads to an exponential expansion of interact terms in the model, while y has only two features. High covariance occurs between y and interaction terms, and causes the model to be unstable.



In conclusion, all the models we developed would not be good classification models for predictions.

Conclusion and Further Work

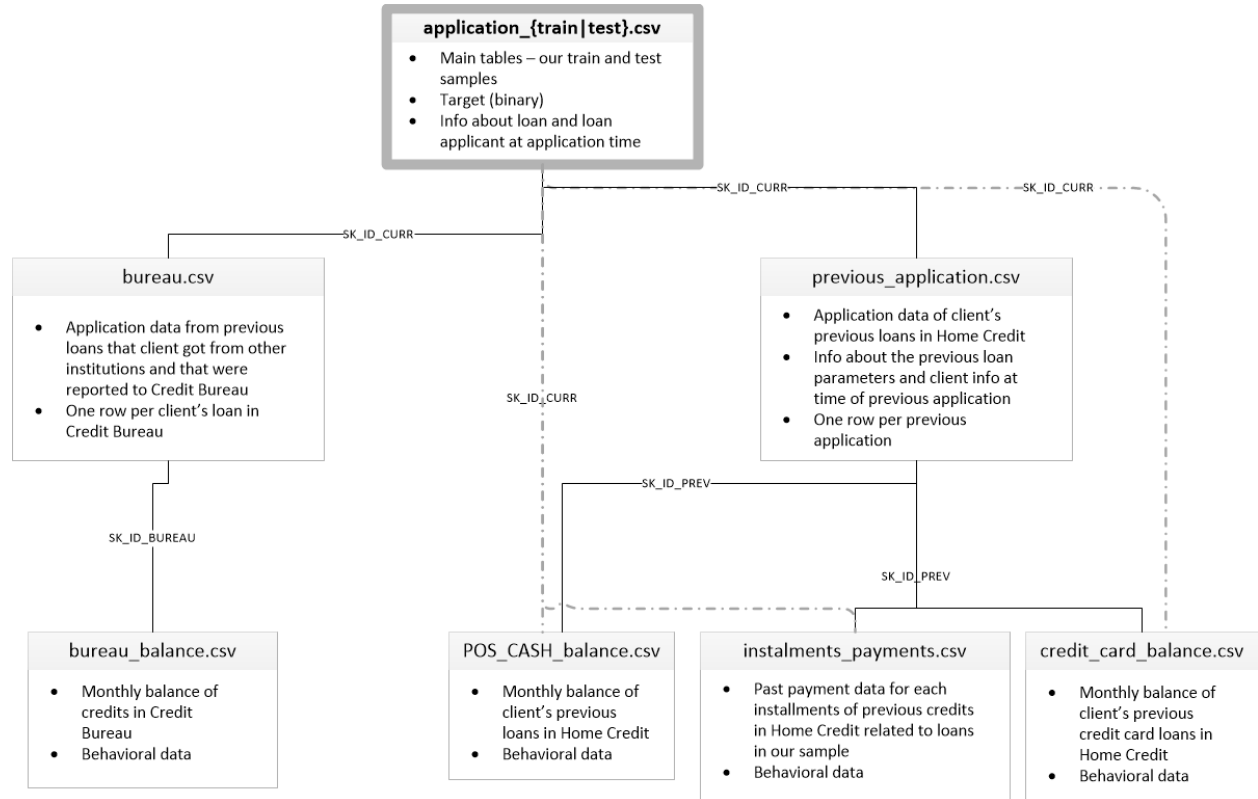
Our analysis led to several compelling insights into the predictive elements of loan repayment behaviors. Firstly, the external credit scores emerged as the most significant predictive variable across different regression models, accentuating the importance of these scores in the loan decision-making process. Secondly, causal inference investigations provided nuanced understanding about the influence of specific covariates on the outcome, highlighting the potential for both confounding and causal effects within our variables of interest. Lastly, the unsupervised learning techniques underscored patterns in credit histories corresponding to different organization types, indicating the presence of distinct financial behavior profiles within these groupings.

Despite these advancements, the overall prediction performance of our models remained suboptimal. Further introspection into the dataset revealed potential structural issues, which might have contributed to the poor performance. Future work should therefore focus on addressing these potential data issues and exploring other machine learning techniques or ensemble methods that may improve model performance. Moreover, investigating more complex and possibly non-linear relationships between the predictors and the outcome might yield more accurate models. Additionally, applying feature engineering to create new variables, especially from the non-static features, could potentially lead to the unearthing of more predictive power.

In sum, while our analysis succeeded in unearthing important factors contributing to loan repayment behaviors and potential avenues for improvements in prediction accuracy, further exploration and refinement are necessary. The need for accurate models in this context is crucial, not only from a business perspective but also for broadening financial inclusion for those struggling with credit access due to insufficient credit histories.

Appendix

Appendix I: Datasets description that outlines the relationship between the different data points



Appendix II: Variables' description after cleaning the data

	variables	Description
1	CODE_GENDER	Gender of the client
2	FLAG_OWN_CAR	Flag if the client owns a car
3	FLAG_OWN_REALTY	Flag if client owns a house or flat
4	CNT_CHILDREN	Number of children the client has
5	AMT_INCOME_TOTAL	Income of the client
6	AMT_CREDIT	Credit amount of the loan
7	AMT_ANNUITY	Loan annuity
8	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given

9	NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
10	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,)
11	NAME_EDUCATION_TYPE	Level of highest education the client achieved
12	NAME_FAMILY_STATUS	Family status of the client
13	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
14	DAYS_BIRTH	Client's age in days at the time of application
15	DAYS_EMPLOYED	How many days before the application the person started current employment
16	DAYS_REGISTRATION	How many days before the application did client change his registration
17	DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
18	FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
19	FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
20	FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)
21	FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
22	FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
23	OCCUPATION_TYPE	What kind of occupation does the client have
24	CNT_FAM_MEMBERS	How many family members does client have
25	REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
26	REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
27	WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
28	HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
29	REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
30	ORGANIZATION_TYPE	Type of organization where client works
31	FONDKAPREMONT_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
32	HOUSETYPE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

33	WALLSMATERIAL_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
34	EMERGENCYSTATE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
35	OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
36	OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
37	DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
38	FLAG_DOCUMENT_3	Did client provide document 3
39	AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
40	AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
41	AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)
42	NEW_CREDIT_TO_ANNUITY_RATIO	The ratio of new credit amount to annuity payment.
43	NEW_CREDIT_TO_GOODS_RATIO	The ratio of new credit amount to goods price.
44	NEW_DOC_IND_AVG	The average number of new documents submitted.
45	NEW_DOC_IND_STD	The standard deviation of new document indices.
46	NEW_DOC_IND_KURT	The kurtosis of new document indices.
47	NEW_LIVE_IND_SUM	The sum of new live indices.
48	NEW_LIVE_IND_STD	The standard deviation of new live indices.
49	NEW_LIVE_IND_KURT	The kurtosis of new live indices.
50	NEW_INC_PER_CHLD	The new income divided by the number of children.
51	NEW_INC_BY_ORG	The new income divided by the number of organizations.
52	NEW_EMPLOY_TO_BIRTH_RATIO	The ratio of employment period to birth period.
53	NEW_ANNUITY_TO_INCOME_RATIO	The ratio of annuity payment to income.
54	NEW_EXT_SOURCES_MEAN	The mean value of new external sources.
55	NEW_SCORES_STD	The standard deviation of new scores.
56	NEW_PHONE_TO_BIRTH_RATIO	The ratio of phone usage period to birth period.
57	NEW_CREDIT_TO_INCOME_RATIO	The ratio of new credit amount to income.

58	BURO_DAYS_CREDIT_MEAN	The average days before current application did client apply for Credit Bureau credit
59	BURO_DAYS_CREDIT_ENDDATE_MEAN	The average remaining duration of CB credit (in days) at the time of application in Home Credit
60	BURO_DAYS_CREDIT_UPDATE_MEAN	The average days before loan application did last information about the Credit Bureau credit come
61	BURO_AMT_CREDIT_SUM_MEAN	The average of current credit amount for the Credit Bureau credit
62	BURO_AMT_CREDIT_SUM_SUM	The sum of current credit amounts for Credit Bureau credits related to the loan in the sample.
63	BURO_AMT_CREDIT_SUM_DEBT_MEAN	The average current debt on Credit Bureau credits related to the loan in the sample.
64	BURO_AMT_CREDIT_SUM_DEBT_SUM	The sum of current debts on Credit Bureau credits related to the loan in the sample.
65	BURO_AMT_CREDIT_SUM_LIMIT_MEAN	The average current credit limit of credit cards reported in the Credit Bureau related to the loan in the sample.
66	BURO_AMT_CREDIT_SUM_LIMIT_SUM	The sum of current credit limits of credit cards reported in the Credit Bureau related to the loan in the sample.
67	BURO_MONTHS_BALANCE_SIZE_SUM	The sum of months of balance relative to the application date for Credit Bureau credits related to the loan in the sample.
68	BURO_CREDIT_ACTIVE_Active_MEAN	The average proportion of active Credit Bureau credits related to the loan in the sample.
69	BURO_CREDIT_ACTIVE_Closed_MEAN	The average proportion of closed Credit Bureau credits related to the loan in the sample.
70	BURO_CREDIT_CURRENCY_currency.1_MEAN	The average proportion of Credit Bureau credits with currency "currency.1" related to the loan in the sample.
71	BURO_CREDIT_TYPE_Car.loan_MEAN	The average proportion of Credit Bureau credits with type "Car loan" related to the loan in the sample.
72	BURO_CREDIT_TYPE_Consumer.credit_MEAN	The average proportion of Credit Bureau credits with type "Consumer credit" related to the loan in the sample.
73	BURO_CREDIT_TYPE_Credit.card_MEAN	The average proportion of Credit Bureau credits with type "Credit card" related to the loan in the sample.
74	BURO_CREDIT_TYPE_Mortgage_MEAN	The average proportion of Credit Bureau credits with type "Mortgage" related to the loan in the sample.
75	ACTIVE_DAYS_CREDIT_MEAN	The average number of days before the current application that the client applied for active Credit Bureau credits related to the loan in the sample.
76	ACTIVE_DAYS_CREDIT_ENDDATE_MEAN	The average remaining duration of active Credit Bureau credits related to the loan in the sample.
77	ACTIVE_DAYS_CREDIT_UPDATE_MEAN	The average number of days before the loan application that the last information about active Credit Bureau credits came.
78	ACTIVE_CREDIT_DAY_OVERDUE_MEAN	The average number of days past due on active Credit Bureau credits at the time of application for the related loan in the sample.
79	ACTIVE_AMT_CREDIT_SUM_MEAN	The average current credit amount for active Credit Bureau credits related to the loan in the sample.
80	ACTIVE_AMT_CREDIT_SUM_SUM	The sum of current credit amounts for active Credit Bureau credits related to the loan in the sample.
81	ACTIVE_AMT_CREDIT_SUM_DEBT_SUM	The sum of current debts on active Credit Bureau credits related to the loan in the sample.
82	ACTIVE_AMT_CREDIT_SUM_OVERDUE_MEAN	The average current amount overdue on active Credit Bureau credits related to the loan in the sample.

83	ACTIVE_AMT_CREDIT_SUM_LIMIT_SUM	The sum of current credit limits of credit cards reported in the Credit Bureau for active credits related to the loan in the sample.
84	ACTIVE_CNT_CREDIT_PROLONG_SUM	The sum of how many times active Credit Bureau credits were prolonged related to the loan in the sample.
85	ACTIVE_MONTHS_BALANCE_SIZE_SUM	The sum of months of balance relative to the application date for active Credit Bureau credits related to the loan in the sample.
86	CLOSED_DAYS_CREDIT_MEAN	The average number of days before the current application that the client applied for closed Credit Bureau credits related to the loan in the sample.
87	CLOSED_DAYS_CREDIT_ENDDATE_MEAN	The average remaining duration of closed Credit Bureau credits related to the loan in the sample.
88	CLOSED_DAYS_CREDIT_UPDATE_MEAN	The average number of days before the loan application that the last information about closed Credit Bureau credits came.
89	CLOSED_CREDIT_DAY_OVERDUE_MEAN	The average number of days past due on closed Credit Bureau credits at the time of application for the related loan in the sample.
90	CLOSED_AMT_CREDIT_SUM_MEAN	The average current credit amount for closed Credit Bureau credits related to the loan in the sample.
91	CLOSED_AMT_CREDIT_SUM_SUM	The sum of current credit amounts for closed Credit Bureau credits related to the loan in the sample.
92	CLOSED_AMT_CREDIT_SUM_DEBT_MEAN	The average current debt on closed Credit Bureau credits related to the loan in the sample.
93	CLOSED_AMT_CREDIT_SUM_DEBT_SUM	The sum of current debts on closed Credit Bureau credits related to the loan in the sample.
94	CLOSED_AMT_CREDIT_SUM_OVERDUE_MEAN	The average current amount overdue on closed Credit Bureau credits related to the loan in the sample.
95	CLOSED_AMT_CREDIT_SUM_LIMIT_SUM	The sum of current credit limits of credit cards reported in the Credit Bureau for closed credits related to the loan in the sample.
96	CLOSED_CNT_CREDIT_PROLONG_SUM	The sum of how many times closed Credit Bureau credits were prolonged related to the loan in the sample.
97	CLOSED_MONTHS_BALANCE_SIZE_SUM	The sum of months of balance relative to the application date for closed Credit Bureau credits related to the loan in the sample.
98	PREV_AMT_ANNUITY_MEAN	The average annuity of previous applications related to the loan in the sample.
99	PREV_AMT_APPLICATION_MEAN	The average amount applied for on previous applications related to the loan in the sample.
100	PREV_AMT_CREDIT_MEAN	The average final credit amount on previous applications related to the loan in the sample.
101	PREV_APP_CREDIT_PERC_MEAN	The average ratio of the final credit amount to the amount applied for on previous applications related to the loan in the sample.
102	PREV_AMT_DOWN_PAYMENT_MEAN	The average down payment amount on previous applications related to the loan in the sample.
103	PREV_AMT_GOODS_PRICE_MEAN	The average goods price of goods that clients asked for on previous applications related to the loan in the sample.
104	PREV_HOUR_APPR_PROCESS_START_MEAN	The average hour of the day at which the client applied for previous applications related to the loan in the sample.
105	PREV_RATE_DOWN_PAYMENT_MEAN	The average down payment rate normalized on previous credits related to the loan in the sample.
106	PREV_DAYS_DECISION_MEAN	The average number of days relative to the current application when the decision about previous applications was made.

107	PREV_CNT_PAYMENT_MEAN	The average term of previous credits at the application of previous applications related to the loan in the sample.
108	PREV_CNT_PAYMENT_SUM	The sum of terms of previous credits at the application of previous applications related to the loan in the sample.
109	PREV_NAME_CONTRACT_TYPE_Cash.loans_MEAN	The average proportion of previous applications with a contract type of "Cash loans" related to the loan in the sample.
110	PREV_NAME_CONTRACT_TYPE_Consumer.loans_MEAN	The average proportion of previous applications with a contract type of "Consumer loans" related to the loan in the sample.
111	PREV_NAME_CONTRACT_TYPE_Revolving.loans_MEAN	The average proportion of previous applications with a contract type of "Revolving loans" related to the loan in the sample.
112	PREV_WEEKDAY_APPR_PROCESS_START_FRIDAY_MEAN	The average proportion of previous applications that were applied on a Friday related to the loan in the sample.
113	PREV_WEEKDAY_APPR_PROCESS_START_MONDAY_MEAN	The average proportion of previous applications that were applied on a Monday related to the loan in the sample.
114	PREV_WEEKDAY_APPR_PROCESS_START_SATURDAY_MEAN	The average proportion of previous applications that were applied on a Saturday related to the loan in the sample.
115	PREV_WEEKDAY_APPR_PROCESS_START_SUNDAY_MEAN	The average proportion of previous applications that were applied on a Sunday related to the loan in the sample.
116	PREV_WEEKDAY_APPR_PROCESS_START_THURSDAY_MEAN	The average proportion of previous applications that were applied on a Thursday related to the loan in the sample.
117	PREV_WEEKDAY_APPR_PROCESS_START_TUESDAY_MEAN	The average proportion of previous applications that were applied on a Tuesday related to the loan in the sample.
118	PREV_WEEKDAY_APPR_PROCESS_START_WEDNESDAY_MEAN	The average proportion of previous applications that were applied on a Wednesday related to the loan in the sample.
119	PREV_FLAG_LAST_APPL_PER_CONTRACT_Y_MEAN	The average proportion of previous applications that were the last application per contract related to the loan in the sample.
120	PREV_NAME_CASH_LOAN_PURPOSE_XAP_MEAN	The average proportion of previous applications with a cash loan purpose of "XAP" related to the loan in the sample.
121	PREV_NAME_CASH_LOAN_PURPOSE_XNA_MEAN	The average proportion of previous applications with a cash loan purpose of "XNA" related to the loan in the sample.
122	PREV_NAME_CONTRACT_STATUS_Approved_MEAN	The average proportion of previous applications with a contract status of "Approved" related to the loan in the sample.
123	PREV_NAME_CONTRACT_STATUS_Canceled_MEAN	The average proportion of previous applications with a contract status of "Canceled" related to the loan in the sample.
124	PREV_NAME_CONTRACT_STATUS_Refused_MEAN	The average proportion of previous applications with a contract status of "Refused" related to the loan in the sample.
125	PREV_NAME_PAYMENT_TYPE_Cash.through.the.bank_MEAN	The average proportion of previous applications with a payment type of "Cash through the bank" related to the loan in the sample.
126	PREV_NAME_PAYMENT_TYPE_XNA_MEAN	The average proportion of previous applications with a payment type of "XNA" related to the loan in the sample.
127	PREV_CODE_REJECT_REASON_HC_MEAN	The average proportion of previous applications with a rejection reason of "HC" related to the loan in the sample.
128	PREV_CODE_REJECT_REASON_XAP_MEAN	The average proportion of previous applications with a rejection reason of "XAP" related to the loan in the sample.

129	PREV_NAME_TYPE_SUITE_Family_MEAN	The average proportion of previous applications with a type suite of "Family" related to the loan in the sample.
130	PREV_NAME_TYPE_SUITE_Spouse.partner_MEAN	The average proportion of previous applications with a type suite of "Spouse/partner" related to the loan in the sample.
131	PREV_NAME_TYPE_SUITE_Unaccompanied_MEAN	The average proportion of previous applications with a type suite of "Unaccompanied" related to the loan in the sample.
132	PREV_NAME_TYPE_SUITE_nan_MEAN	The average proportion of previous applications with a type suite of "NaN" (missing value) related to the loan in the sample.
133	PREV_NAME_CLIENT_TYPE_New_MEAN	The average proportion of previous applications from new clients related to the loan in the sample.
134	PREV_NAME_CLIENT_TYPE_Refreshed_MEAN	The average proportion of previous applications from refreshed clients related to the loan in the sample.
135	PREV_NAME_CLIENT_TYPE_Repeater_MEAN	The average proportion of previous applications from repeater clients related to the loan in the sample.
136	PREV_NAME_GOODS_CATEGORY_Audio.Video_MEAN	The average proportion of previous applications with a goods category of "Audio/Video" related to the loan in the sample.
137	PREV_NAME_GOODS_CATEGORY_Computers_MEAN	The average proportion of previous applications with a goods category of "Computers" related to the loan in the sample.
138	PREV_NAME_GOODS_CATEGORY_Consumer.Electronics_MEAN	The average proportion of previous applications with a goods category of "Consumer Electronics" related to the loan in the sample.
139	PREV_NAME_GOODS_CATEGORY_Mobile_MEAN	The average proportion of previous applications with a goods category of "Mobile" related to the loan in the sample.
140	PREV_NAME_GOODS_CATEGORY_XNA_MEAN	The average proportion of previous applications with a goods category of "XNA" related to the loan in the sample.
141	PREV_NAME_PORTFOLIO_Cards_MEAN	The average proportion of previous applications with a portfolio of "Cards" related to the loan in the sample.
142	PREV_NAME_PORTFOLIO_Cash_MEAN	The average proportion of previous applications with a portfolio of "Cash" related to the loan in the sample.
143	PREV_NAME_PORTFOLIO_POS_MEAN	The average proportion of previous applications with a portfolio of "POS" related to the loan in the sample.
144	PREV_NAME_PORTFOLIO_XNA_MEAN	The average proportion of previous applications with a portfolio of "XNA" related to the loan in the sample.
145	PREV_NAME_PRODUCT_TYPE_XNA_MEAN	The average proportion of previous applications with a product type of "XNA" related to the loan in the sample.
146	PREV_NAME_PRODUCT_TYPE_walk.in_MEAN	The average proportion of previous applications with a product type of "Walk-in" related to the loan in the sample.
147	PREV_NAME_PRODUCT_TYPE_x.sell_MEAN	The average proportion of previous applications with a product type of "X-sell" related to the loan in the sample.
148	PREV_CHANNEL_TYPE_Country.wide_MEAN	The average proportion of previous applications acquired through a country-wide channel related to the loan in the sample.
149	PREV_CHANNEL_TYPE_Credit.and.cash.offices_MEAN	The average proportion of previous applications acquired through credit and cash offices related to the loan in the sample.
150	PREV_CHANNEL_TYPE_Regional...Local_MEAN	The average proportion of previous applications acquired through regional or local channels related to the loan in the sample.
151	PREV_CHANNEL_TYPE_Stone_MEAN	The average proportion of previous applications acquired through stone channels related to the loan in the sample.

152	PREV_NAME_SELLER_INDUSTRY_Connectivity_MEAN	The average proportion of previous applications with a seller industry of "Connectivity" related to the loan in the sample.
153	PREV_NAME_SELLER_INDUSTRY_Consumer.electronics_MEAN	The average proportion of previous applications with a seller industry of "Consumer electronics" related to the loan in the sample.
154	PREV_NAME_SELLER_INDUSTRY_XNA_MEAN	The average proportion of previous applications with a seller industry of "XNA" related to the loan in the sample.
155	PREV_NAME_YIELD_GROUP_XNA_MEAN	The average proportion of previous applications with a yield group of "XNA" related to the loan in the sample.
156	PREV_NAME_YIELD_GROUP_high_MEAN	The average proportion of previous applications with a yield group of "High" related to the loan in the sample.
157	PREV_NAME_YIELD_GROUP_low_action_MEAN	The average proportion of previous applications with a yield group of "Low action" related to the loan in the sample.
158	PREV_NAME_YIELD_GROUP_low_normal_MEAN	The average proportion of previous applications with a yield group of "Low normal" related to the loan in the sample.
159	PREV_NAME_YIELD_GROUP_middle_MEAN	The average proportion of previous applications with a yield group of "Middle" related to the loan in the sample.
160	PREV_PRODUCT_COMBINATION_Card.Street_MEAN	The average proportion of previous applications with a product combination of "Card Street" related to the loan in the sample.
161	PREV_PRODUCT_COMBINATION_Card.X.Sell_MEAN	The average proportion of previous applications with a product combination of "Card X-Sell" related to the loan in the sample.
162	PREV_PRODUCT_COMBINATION_Cash_MEAN	The average proportion of previous applications with a product combination of "Cash" related to the loan in the sample.
163	PREV_PRODUCT_COMBINATION_Cash.X.Sell..low_MEAN	The average proportion of previous applications with a product combination of "Cash X-Sell (low)" related to the loan in the sample.
164	PREV_PRODUCT_COMBINATION_Cash.X.Sell..middle_MEAN	The average proportion of previous applications with a product combination of "Cash X-Sell (middle)" related to the loan in the sample.
165	PREV_PRODUCT_COMBINATION_POS.household.with.interested_MEAN	The average proportion of previous applications with a product combination of "POS household with interest" related to the loan in the sample.
166	PREV_PRODUCT_COMBINATION_POS.household.without.interested_MEAN	The average proportion of previous applications with a product combination of "POS household without interest" related to the loan in the sample.
167	PREV_PRODUCT_COMBINATION_POS.industry.with.interested_MEAN	The average proportion of previous applications with a product combination of "POS industry with interest" related to the loan in the sample.
168	PREV_PRODUCT_COMBINATION_POS.mobile.with.interested_MEAN	The average proportion of previous applications with a product combination of "POS mobile with interest" related to the loan in the sample.
169	APPROVED_AMT_ANNUITY_MEAN	The average annuity of approved previous applications related to the loan in the sample.
170	APPROVED_AMT_APPLICATION_MEAN	The average amount applied for on approved previous applications related to the loan in the sample.
171	APPROVED_AMT_CREDIT_MEAN	The average final credit amount on approved previous applications related to the loan in the sample.
172	APPROVED_APP_CREDIT_PERC_MEAN	The average ratio of the final credit amount to the amount applied for on approved previous applications related to the loan in the sample.
173	APPROVED_AMT_DOWN_PAYMENT_MEAN	The average down payment amount on approved previous applications related to the loan in the sample.
174	APPROVED_AMT_GOODS_PRICE_MEAN	The average goods price of goods that clients asked for on approved previous applications related to the loan in the sample.

175	APPROVED_HOUR_APPR_PROCESS_START_MEAN	The average hour of the day at which the client applied for approved previous applications related to the loan in the sample.
176	APPROVED_RATE_DOWN_PAYMENT_MEAN	The average down payment rate normalized on approved previous credits related to the loan in the sample.
177	APPROVED_DAYS_DECISION_MEAN	The average number of days relative to the current application when the decision about approved previous applications was made.
178	APPROVED_CNT_PAYMENT_MEAN	The average term of approved previous credits at the application of previous applications related to the loan in the sample.
179	APPROVED_CNT_PAYMENT_SUM	The sum of terms of approved previous credits at the application of previous applications related to the loan in the sample.
180	POS_MONTHS_BALANCE_MEAN	The average month of balance relative to the application date for POS-related data.
181	POS_MONTHS_BALANCE_SIZE	The number of months of POS-related data.
182	POS_SK_DPD_MEAN	The average DPD (days past due) for POS-related data.
183	POS_SK_DPD_DEF_MEAN	The average DPD with tolerance (debts with low loan amounts are ignored) for POS-related data.
184	POS_NAME_CONTRACT_STATUS_Active_MEAN	The average proportion of POS contracts with a status of "Active" related to the loan in the sample.
185	POS_NAME_CONTRACT_STATUS_Completed_MEAN	The average proportion of POS contracts with a status of "Completed" related to the loan in the sample.
186	POS_NAME_CONTRACT_STATUS_Signed_MEAN	The average proportion of POS contracts with a status of "Signed" related to the loan in the sample.
187	POS_COUNT	The count of POS-related data.
188	INSTAL_NUM_INSTALLMENT_VERSION_UNIQUE	The number of unique installment versions for previous credits related to the loan in the sample.
189	INSTAL_DPD_MEAN	The average DPD (days past due) for previous installments related to the loan in the sample.
190	INSTAL_DPD_SUM	The sum of DPD (days past due) for previous installments related to the loan in the sample.
191	INSTAL_DBD_MEAN	The average DBD (days before due) for previous installments related to the loan in the sample.
192	INSTAL_DBD_SUM	The sum of DBD (days before due) for previous installments related to the loan in the sample.
193	INSTAL_PAYMENT_PERC_MEAN	The average payment amount as a percentage of the prescribed installment amount for previous installments related to the loan in the sample.
194	INSTAL_PAYMENT_PERC_SUM	The sum of payment amounts as a percentage of the prescribed installment amount for previous installments related to the loan in the sample.
195	INSTAL_PAYMENT_DIFF_MEAN	The average difference between the prescribed installment amount and the actual payment amount for previous installments related to the loan in the sample.
196	INSTAL_PAYMENT_DIFF_SUM	The sum of differences between the prescribed installment amount and the actual payment amount for previous installments related to the loan in the sample.
197	INSTAL_AMT_INSTALLMENT_MEAN	The average prescribed installment amount for previous installments related to the loan in the sample.
198	INSTAL_AMT_INSTALLMENT_SUM	The sum of prescribed installment amounts for previous installments related to the loan in the sample.

199	INSTAL_AMT_PAYMENT_MEAN	The average actual payment amount for previous installments related to the loan in the sample.
200	INSTAL_AMT_PAYMENT_SUM	The sum of actual payment amounts for previous installments related to the loan in the sample.
201	INSTAL_DAYS_ENTRY_PAYMENT_MEAN	The average number of days between the installment due date and the actual payment date for previous installments related to the loan in the sample.
202	INSTAL_DAYS_ENTRY_PAYMENT_SUM	The sum of the number of days between the installment due date and the actual payment date for previous installments related to the loan in the sample.
203	INSTAL_COUNT	The count of previous installments related to the loan in the sample.
204	TARGET	

Appendix III: CV.LASSO selected variables

CV.min	CV.lse
intercept	intercept
FLAG_OWN_CAR	FLAG_OWN_CAR
FLAG_OWN_REALTY	AMT_INCOME_TOTAL
CNT_CHILDREN	AMT_ANNUITY
AMT_INCOME_TOTAL	NAME_INCOME_TYPE
AMT_CREDIT	NAME_EDUCATION_TYPE
AMT_ANNUITY	NAME_FAMILY_STATUS
AMT_GOODS_PRICE	DAYS_BIRTH
NAME_TYPE_SUITE	DAYS_EMPLOYED
NAME_INCOME_TYPE	DAYS_REGISTRATION
NAME_EDUCATION_TYPE	DAYS_ID_PUBLISH
NAME_FAMILY_STATUS	FLAG_EMP_PHONE
REGION_POPULATION_RELATIVE	FLAG_WORK_PHONE
DAYS_BIRTH	FLAG_PHONE
DAYS_EMPLOYED	REGION_RATING_CLIENT_W_CITY
DAYS_REGISTRATION	REG_CITY_NOT_WORK_CITY
DAYS_ID_PUBLISH	FONDKAPREMONT_MODE
FLAG_MOBIL	EMERGENCYSTATE_MODE

FLAG_EMP_PHONE	OBS_30_CNT_SOCIAL_CIRCLE
FLAG_WORK_PHONE	DAYS_LAST_PHONE_CHANGE
FLAG_CONT_MOBILE	FLAG_DOCUMENT_3
FLAG_PHONE	AMT_REQ_CREDIT_BUREAU_QRT
OCCUPATION_TYPE	NEW_CREDIT_TO_ANNUITY_RATIO
CNT_FAM_MEMBERS	NEW_CREDIT_TO_GOODS_RATIO
REGION_RATING_CLIENT	NEW_DOC_IND_KURT
REGION_RATING_CLIENT_W_CITY	NEW_INC_BY_ORG
WEEKDAY_APPR_PROCESS_START	NEW_EMPLOY_TO_BIRTH_RATIO
HOURLY_APPR_PROCESS_START	NEW_ANNUITY_TO_INCOME_RATIO
REG_CITY_NOT_WORK_CITY	NEW_EXT_SOURCES_MEAN
ORGANIZATION_TYPE	NEW_SCORES_STD
FONDKAPREMONT_MODE	BURO_DAYS_CREDIT_MEAN
HOUSETYPE_MODE	BURO_AMT_CREDIT_SUM_LIMIT_MEAN
WALLSMATERIAL_MODE	BURO_CREDIT_ACTIVE_Closed_MEAN
EMERGENCYSTATE_MODE	BURO_CREDIT_TYPE_Car.loan_MEAN
OBS_30_CNT_SOCIAL_CIRCLE	BURO_CREDIT_TYPE_Mortgage_MEAN
DAYS_LAST_PHONE_CHANGE	ACTIVE_DAYS_CREDIT_MEAN
FLAG_DOCUMENT_3	ACTIVE_AMT_CREDIT_SUM_SUM
AMT_REQ_CREDIT_BUREAU_MON	ACTIVE_AMT_CREDIT_SUM_DEBT_SUM
AMT_REQ_CREDIT_BUREAU_QRT	CLOSED_MONTHS_BALANCE_SIZE_SUM
AMT_REQ_CREDIT_BUREAU_YEAR	PREV_APP_CREDIT_PERC_MEAN
NEW_CREDIT_TO_ANNUITY_RATIO	PREV_RATE_DOWN_PAYMENT_MEAN
NEW_CREDIT_TO_GOODS_RATIO	PREV_CNT_PAYMENT_MEAN
NEW_DOC_IND_AVG	PREV_CNT_PAYMENT_SUM
NEW_DOC_IND_KURT	PREV_NAME_CONTRACT_TYPE_Consumer.loans_MEAN
NEW_LIVE_IND_SUM	PREV_NAME_CONTRACT_TYPE_Revolving.loans_MEAN
NEW_LIVE_IND_STD	PREV_NAME_CONTRACT_STATUS_Refused_MEAN

NEW_LIVE_IND_KURT	PREV_NAME_PAYMENT_TYPE_XNA_MEAN
NEW_INC_BY_ORG	PREV_CODE_REJECT_REASON_HC_MEAN
NEW_EMPLOY_TO_BIRTH_RATIO	PREV_NAME_TYPE_SUITE_nan_MEAN
NEW_ANNUITY_TO_INCOME_RATIO	PREV_NAME_CLIENT_TYPE_New_MEAN
NEW_EXT_SOURCES_MEAN	PREV_NAME_PRODUCT_TYPE_walk.in_MEAN
NEW_SCORES_STD	PREV_NAME_SELLER_INDUSTRY_Connectivity_MEAN
NEW_PHONE_TO_BIRTH_RATIO	PREV_NAME_SELLER_INDUSTRY_XNA_MEAN
BURO_DAYS_CREDIT_MEAN	PREV_NAME_YIELD_GROUP_XNA_MEAN
BURO_DAYS_CREDIT_ENDDATE_MEAN	PREV_NAME_YIELD_GROUP_high_MEAN
BURO_DAYS_CREDIT_UPDATE_MEAN	PREV_NAME_YIELD_GROUP_low_action_MEAN
BURO_AMT_CREDIT_SUM_MEAN	PREV_NAME_YIELD_GROUP_low_normal_MEAN
BURO_AMT_CREDIT_SUM_DEBT_MEAN	PREV_PRODUCT_COMBINATION_Cash.X.Sell..low_MEAN
BURO_AMT_CREDIT_SUM_DEBT_SUM	PREV_PRODUCT_COMBINATION_POS.industry.with.interest_MEAN
BURO_AMT_CREDIT_SUM_LIMIT_MEAN	APPROVED_AMT_ANNUITY_MEAN
BURO_AMT_CREDIT_SUM_LIMIT_SUM	APPROVED_HOUR_APPR_PROCESS_START_MEAN
BURO_MONTHS_BALANCE_SIZE_SUM	APPROVED_CNT_PAYMENT_MEAN
BURO_CREDIT_ACTIVE_Active_MEAN	APPROVED_CNT_PAYMENT_SUM
BURO_CREDIT_ACTIVE_Closed_MEAN	POS_MONTHS_BALANCE_SIZE
BURO_CREDIT_CURRENCY_currency.1_MEAN	POS_SK_DPD_MEAN
BURO_CREDIT_TYPE_Car.loan_MEAN	POS_SK_DPD_DEF_MEAN
BURO_CREDIT_TYPE_Consumer.credit_MEAN	POS_NAME_CONTRACT_STATUS_Signed_MEAN
BURO_CREDIT_TYPE_Credit.card_MEAN	POS_COUNT
BURO_CREDIT_TYPE_Mortgage_MEAN	INSTAL_DPD_MEAN
ACTIVE_DAYS_CREDIT_MEAN	INSTAL_DBD_MEAN
ACTIVE_DAYS_CREDIT_ENDDATE_MEAN	INSTAL_PAYMENT_DIFF_MEAN
ACTIVE_DAYS_CREDIT_UPDATE_MEAN	INSTAL_PAYMENT_DIFF_SUM
ACTIVE_CREDIT_DAY_OVERDUE_MEAN	INSTAL_AMT_INSTALMENT_SUM
ACTIVE_AMT_CREDIT_SUM_MEAN	INSTAL_DAYS_ENTRY_PAYMENT_SUM

ACTIVE_AMT_CREDIT_SUM_SUM	NA
ACTIVE_AMT_CREDIT_SUM_DEBT_SUM	NA
ACTIVE_AMT_CREDIT_SUM_OVERDUE_MEAN	NA
ACTIVE_CNT_CREDIT_PROLONG_SUM	NA
ACTIVE_MONTHS_BALANCE_SIZE_SUM	NA
CLOSED_DAYS_CREDIT_MEAN	NA
CLOSED_DAYS_CREDIT_ENDDATE_MEAN	NA
CLOSED_DAYS_CREDIT_UPDATE_MEAN	NA
CLOSED_CREDIT_DAY_OVERDUE_MEAN	NA
CLOSED_AMT_CREDIT_SUM_MEAN	NA
CLOSED_AMT_CREDIT_SUM_SUM	NA
CLOSED_AMT_CREDIT_SUM_DEBT_SUM	NA
CLOSED_AMT_CREDIT_SUM_OVERDUE_MEAN	NA
CLOSED_AMT_CREDIT_SUM_LIMIT_SUM	NA
CLOSED_CNT_CREDIT_PROLONG_SUM	NA
CLOSED_MONTHS_BALANCE_SIZE_SUM	NA
PREV_AMT_ANNUITY_MEAN	NA
PREV_APP_CREDIT_PERC_MEAN	NA
PREV_AMT_DOWN_PAYMENT_MEAN	NA
PREV_AMT_GOODS_PRICE_MEAN	NA
PREV_HOUR_APPR_PROCESS_START_MEAN	NA
PREV_RATE_DOWN_PAYMENT_MEAN	NA
PREV_DAYS_DECISION_MEAN	NA
PREV_CNT_PAYMENT_MEAN	NA
PREV_CNT_PAYMENT_SUM	NA
PREV_NAME_CONTRACT_TYPE_Consumer.loans_MEAN	NA
PREV_NAME_CONTRACT_TYPE_Revolving.loans_MEAN	NA
PREV_WEEKDAY_APPR_PROCESS_START_MONDAY_MEAN	NA

PREV_WEEKDAY_APPR_PROCESS_START_SATURDAY_MEAN	NA
PREV_WEEKDAY_APPR_PROCESS_START_SUNDAY_MEAN	NA
PREV_WEEKDAY_APPR_PROCESS_START_THURSDAY_MEAN	NA
PREV_WEEKDAY_APPR_PROCESS_START_TUESDAY_MEAN	NA
PREV_WEEKDAY_APPR_PROCESS_START_WEDNESDAY_MEAN	NA
PREV_FLAG_LAST_APPL_PER_CONTRACT_Y_MEAN	NA
PREV_NAME_CASH_LOAN_PURPOSE_XNA_MEAN	NA
PREV_NAME_CONTRACT_STATUS_Approved_MEAN	NA
PREV_NAME_CONTRACT_STATUS_Refused_MEAN	NA
PREV_NAME_PAYMENT_TYPE_XNA_MEAN	NA
PREV_CODE_REJECT_REASON_HC_MEAN	NA
PREV_NAME_TYPE_SUITE_Family_MEAN	NA
PREV_NAME_TYPE_SUITE_Spouse..partner_MEAN	NA
PREV_NAME_TYPE_SUITE_nan_MEAN	NA
PREV_NAME_CLIENT_TYPE_New_MEAN	NA
PREV_NAME_CLIENT_TYPE_Refreshed_MEAN	NA
PREV_NAME_GOODS_CATEGORY_Audio.Video_MEAN	NA
PREV_NAME_GOODS_CATEGORY_Computers_MEAN	NA
PREV_NAME_GOODS_CATEGORY_Mobile_MEAN	NA
PREV_NAME_PORTFOLIO_Cards_MEAN	NA
PREV_NAME_PORTFOLIO_XNA_MEAN	NA
PREV_NAME_PRODUCT_TYPE_walk.in_MEAN	NA
PREV_CHANNEL_TYPE_Country.wide_MEAN	NA
PREV_CHANNEL_TYPE_Credit.and.cash.offices_MEAN	NA
PREV_CHANNEL_TYPE_Stone_MEAN	NA
PREV_NAME_SELLER_INDUSTRY_Connectivity_MEAN	NA
PREV_NAME_SELLER_INDUSTRY_XNA_MEAN	NA
PREV_NAME_YIELD_GROUP_high_MEAN	NA

PREV_NAME_YIELD_GROUP_low_action_MEAN	NA
PREV_NAME_YIELD_GROUP_low_normal_MEAN	NA
PREV_PRODUCT_COMBINATION_Card.Street_MEAN	NA
PREV_PRODUCT_COMBINATION_Cash_MEAN	NA
PREV_PRODUCT_COMBINATION_Cash.X.Sell..low_MEAN	NA
PREV_PRODUCT_COMBINATION_Cash.X.Sell..middle_MEAN	NA
PREV_PRODUCT_COMBINATION_POS.household.with.interest_MEAN	NA
PREV_PRODUCT_COMBINATION_POS.household.without.interest_MEAN	NA
PREV_PRODUCT_COMBINATION_POS.industry.with.interest_MEAN	NA
PREV_PRODUCT_COMBINATION_POS.mobile.with.interest_MEAN	NA
APPROVED_AMT_ANNUIITY_MEAN	NA
APPROVED_AMT_CREDIT_MEAN	NA
APPROVED_APP_CREDIT_PERC_MEAN	NA
APPROVED_AMT_DOWN_PAYMENT_MEAN	NA
APPROVED_HOUR_APPR_PROCESS_START_MEAN	NA
APPROVED_DAYS_DECISION_MEAN	NA
APPROVED_CNT_PAYMENT_MEAN	NA
APPROVED_CNT_PAYMENT_SUM	NA
POS_MONTHS_BALANCE_MEAN	NA
POS_MONTHS_BALANCE_SIZE	NA
POS_SK_DPD_MEAN	NA
POS_SK_DPD_DEF_MEAN	NA
POS_NAME_CONTRACT_STATUS_Active_MEAN	NA
POS_NAME_CONTRACT_STATUS_Completed_MEAN	NA
POS_NAME_CONTRACT_STATUS_Signed_MEAN	NA
POS_COUNT	NA
INSTAL_NUM_INSTALMENT_VERSION_NUNIQUE	NA
INSTAL_DPD_MEAN	NA

INSTAL_DPD_SUM	NA
INSTAL_DBD_MEAN	NA
INSTAL_DBD_SUM	NA
INSTAL_PAYMENT_PERC_MEAN	NA
INSTAL_PAYMENT_DIFF_MEAN	NA
INSTAL_PAYMENT_DIFF_SUM	NA
INSTAL_AMT_INSTALMENT_SUM	NA
INSTAL_AMT_PAYMENT_MEAN	NA
INSTAL_DAYS_ENTRY_PAYMENT_SUM	NA
INSTAL_COUNT	NA

