Default Risk Detection[1]

Minyi Chen

Northeastern University

07/02/2018

[1] Final project of ALY 6040 Data Mining Applications, Instructor Marcus Ellis

I. Introduction

Being able to predict the reliability of the borrowers is very important to the lenders. If the borrowers cannot repay their debt on time, then it will cause issues to lenders and borrowers. Usually, lenders can check people's credit scores or credit histories to make the lending decisions. However, there are a group of people who are hard to get loan because they don't have sufficient credit score history. Home Credit Group (HCG) focuses on lending money to people with little or no credit history. HCG try to help those people to be successful with their loan. At the same time, HCG want to make sure their borrowers can repay the debts.

Home Credit Group is a leading financial service company, which was founded in 1997 at Czech Republic. They want people to unlock the full potential of their massive data and help them to predict the liability of their clients. They create an ongoing competition "Home Credit Default Risk" on Kaggle.com and provide their datasets. They ask people "Can you predict how capable each applicant is of repaying a loan?". I find the project interesting because I can use their dataset to solve a real-world problem, and to practice data mining and supervised learning algorithms.

I will use supervised learning methods, such as decision trees, random forest, logistic regression to find out the characters of two groups of people: people who pay their debt on time and people who have default risk. People who have default risk is labeled as 1 under target variable in their dataset. It means people who have late payments more than a tolerance period on their loan.

They provide 9 excel CSV data files: 1) HomeCredit columns description.csv, 2) application train data, 3) application test data, 4) bureau.csv, 5) bureau balance.csv, 6) POS

CASH balance.csv, 7) credit card balance.csv, 8) previous application.csv, 9) installments payments.csv. The training dataset includes more than 100 variables. I will use some of them to explore the data and do supervised learning algorithms analysis.

II. Data and EDA

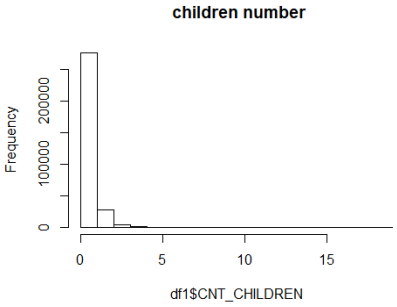
I first pick some variables out of the train dataset, include the ID variable and the dependent variable TARGET. The list of variables name, structure, and explanations are in the appendix Table 1. Then, read the updated csv file in R, use head(), str() to check the variables. I found that 1) target is not factor, 2) some variables I don't need anymore, 3) days_birth, and days_employed are negative, which means the days happen in the past. For example, if the days is -15750, that means the person was born for 15750 days before he submitted the application. Thus, I do three changes to the dataset, 1) change the TARGET variable into factor, 2) further delete more variables, 3) use mutate() to calculate two new variables age and working years, that is, turn the days into year. Let the new variables age = DAYS_BIRTH / (-365), and work_years= DAYS_EMPLOYED/ (-365). Finally, I save the changes into a new dataframe.

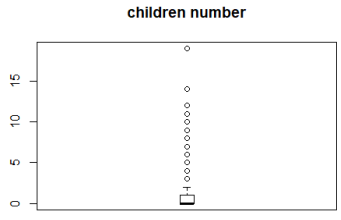
Once I get the final dataframe, I start doing Exploratory Data Analysis (EDA). The goal of doing EDA is to get more familiar with the data and each variables, and their relationships with default risk (variable TARGET), and one other important thing is to remove outliers. First, use summary() to get information of the dataset. The summary graph shows us min, 1st quartile, median, mean, 3rd quartile, and max of the continuous variables, and the count of categories variables.

```
> summary(df1)
   SK_ID_CURR      TARGET           NAME_CONTRACT_TYPE CODE_GENDER  FLAG_OWN_CAR FLAG_OWN_REALTY  CNT_CHILDREN     AMT_INCOME_TOTAL      AMT_CREDIT       AMT_ANNUITY
 Min.   :100002   0:282686   Cash loans     :278232   F  :202448   N:202924     N: 94199        Min.   : 0.0000   Min.   :    25650   Min.   :  45000   Min.   :  1616
 1st Qu.:189146   1: 24825   Revolving loans: 29279   M  :105059   Y:104587     Y:213312        1st Qu.: 0.0000   1st Qu.:   112500   1st Qu.: 270000   1st Qu.: 16524
 Median :278202                                       XNA:     4                                 Median : 0.0000   Median :   147150   Median : 513531   Median : 24903
 Mean   :278181                                                                                 Mean   : 0.4171   Mean   :   168798   Mean   : 599026   Mean   : 27109
 3rd Qu.:367143                                                                                 3rd Qu.: 1.0000   3rd Qu.:   202500   3rd Qu.: 808650   3rd Qu.: 34596
 Max.   :456255                                                                                 Max.   :19.0000   Max.   :117000000   Max.   :4050000   Max.   :258026
                                                                                                                                                    NA's   :12
 AMT_GOODS_PRICE                NAME_INCOME_TYPE                NAME_EDUCATION_TYPE            NAME_FAMILY_STATUS            NAME_HOUSING_TYPE
 Min.   :  40500   Working             :158774   Academic degree              :   164   Civil marriage      : 29775   Co-op apartment    :  1122
 1st Qu.: 238500   Commercial associate: 71617   Higher education             : 74863   Married             :196432   House / apartment  :272868
 Median : 450000   Pensioner           : 55362   Incomplete higher            : 10277   Separated           : 19770   Municipal apartment: 11183
 Mean   : 538396   State servant       : 21703   Lower secondary              :  3816   Single / not married: 45444   Office apartment   :  2617
 3rd Qu.: 679500   Unemployed          :    22   Secondary / secondary special:218391   Unknown             :     2   Rented apartment   :  4881
 Max.   :4050000   Student             :    18                                          Widow               : 16088   With parents       : 14840
 NA's   :278       (Other)             :    15
 REGION_POPULATION_RELATIVE    OCCUPATION_TYPE  CNT_FAM_MEMBERS               ORGANIZATION_TYPE  AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR      age
 Min.   :0.00029               :96391   Min.   : 1.000   Business Entity Type 3: 67992   Min.   :  0.00           Min.   : 0.0              Min.   :20.52
 1st Qu.:0.01001   Laborers    :55186   1st Qu.: 2.000   XNA                   : 55374   1st Qu.:  0.00           1st Qu.: 0.0              1st Qu.:34.01
 Median :0.01885   Sales staff:32102   Median : 2.000   Self-employed         : 38412   Median :  0.00           Median : 1.0              Median :43.15
 Mean   :0.02087   Core staff :27570   Mean   : 2.153   Other                 : 16683   Mean   :  0.27           Mean   : 1.9              Mean   :43.94
 3rd Qu.:0.02866   Managers   :21371   3rd Qu.: 3.000   Medicine              : 11193   3rd Qu.:  0.00           3rd Qu.: 3.0              3rd Qu.:53.92
 Max.   :0.07251   Drivers    :18603   Max.   :20.000   Business Entity Type 2: 10553   Max.   :261.00           Max.   :25.0              Max.   :69.12
                   (Other)    :56288   NA's   :2        (Other)               :107304   NA's   :41519            NA's   :41519
   work_years
 Min.   :-1000.6658
 1st Qu.:    0.7918
 Median :    3.3233
 Mean   : -174.8357
 3rd Qu.:    7.5616
 Max.   :   49.0740
```

Plot the graphs for each variable, starting with continuous variables.

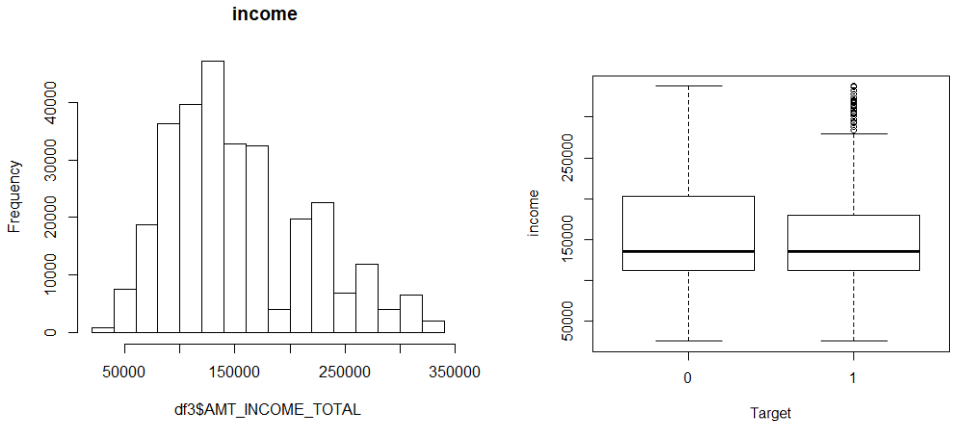
Children numbers (CNT_CHILDREN). We can see from the graph hist(), there are a lot of outliers in the children number. Use boxplot.stats() to check the outliers. The boxplot() shows that children number with more than 2 is outliers. It is common that we have more than 2 children. I delete the outliers that are greater than 4. In this way, I can get rid of the extreme outliers and keep the variety in the data. Also, the number of observation is very large, so it will be fine to remove some outliers. The last graph cro() shows the distribution of default risk across the children numbers.



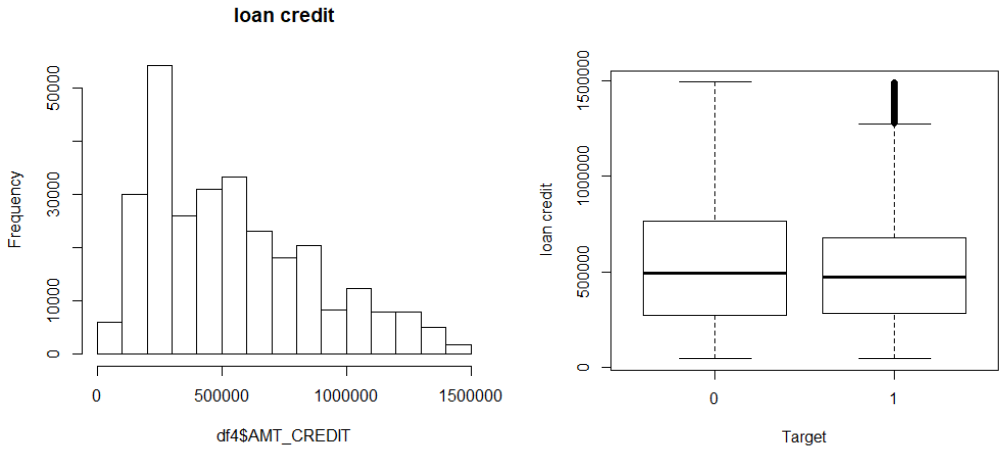


		default risk	
		0	1
children#	0	198762	16609
	1	55665	5454
	2	24416	2333
	3	3359	358
	4	374	55
	#Total cases	282576	24809

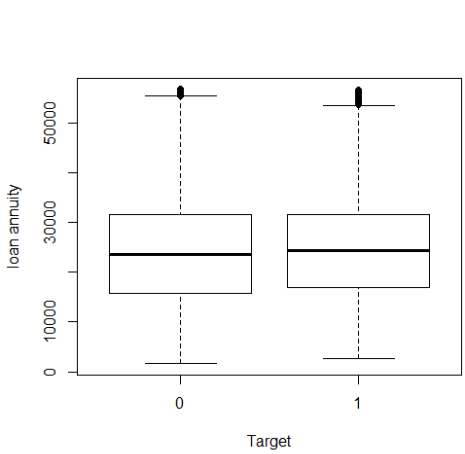
Next, we use the same method to check other variables. The outliers of income are income > 337500, so I remove those observations and get the new histogram of the income variable. The graph looks similar to normal distribution. Then, we get the boxplot for income cross TARGET, we can see some outliers for the default risks. People who have default risk tend to have lower income than people who don't have.

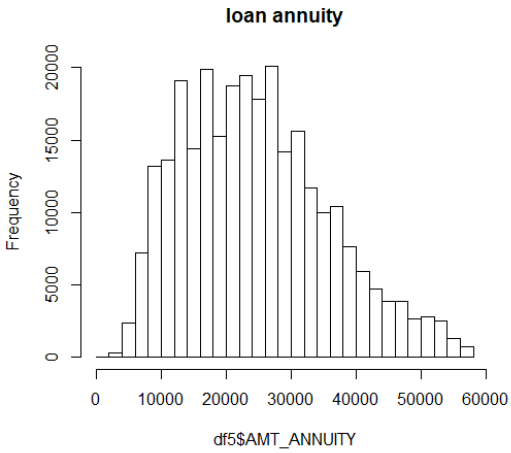


Credit amount of the loan has similar boxplot as income vs target. People who have default risk tend to borrow less money.

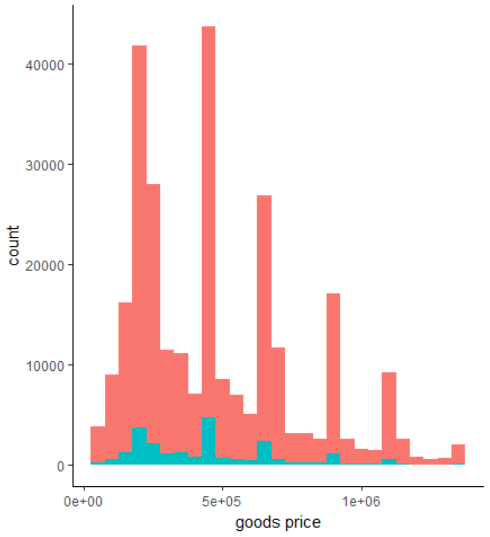


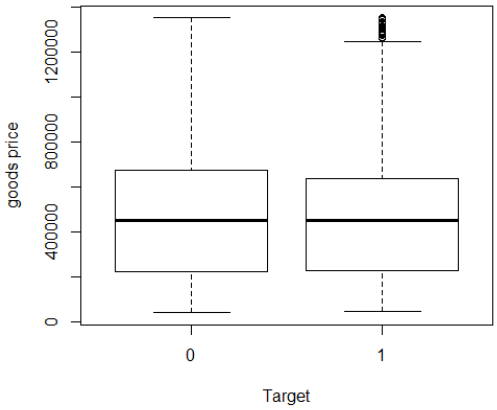
Loan annuity is a series of payments made at equal intervals, such as monthly payment. Below are the graphs after removing outliers. The histogram doesn't show significant outliers. From the boxplot, people who have default risk tend to have higher monthly payment.



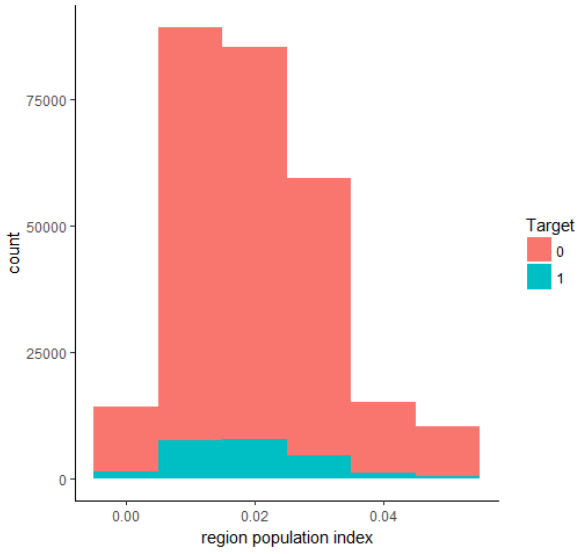


Goods price is the price of the goods for which the loan is given. The outliers in the boxplot suggests that most prices of the goods are lower for the people who have default risk. We can also see the distribution in green for default risk applications. Most of the green area are below 500000.





After removing the outliers of region population index. We can see that most of the observations in the sample live in a place with population index around 0.02. People who have default risk tend to live in a place with relatively smaller population.

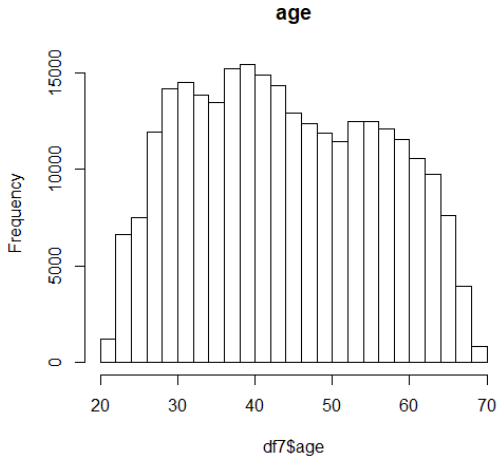


From the summary graph at the beginning, we know family member has max=20, so we need to remove the outliers. After checking the variable in our newest dataframe, I found that the max=6, which makes sense even though the outliers are member > 4. I keep the variable in the dataframe without any change. In the graph below, family with 6 members has 14.6% chance to be default risk in the sample.

		default risk	
		0	1
family member	1	91.4	8.6
	2	92.1	7.9
	3	90.8	9.2
	4	91.0	9.0
	5	90.2	9.8
	6	85.4	14.6
	#Total cases	250106.0	22970.0

The age variable looks good. People are from 20s to 70s. The working years variable has a lot of outliers over 20 years, and most people don't have much working experience. However, I

didn't remove the outliers because it makes sense that people have working experience from 0 to 50 years.

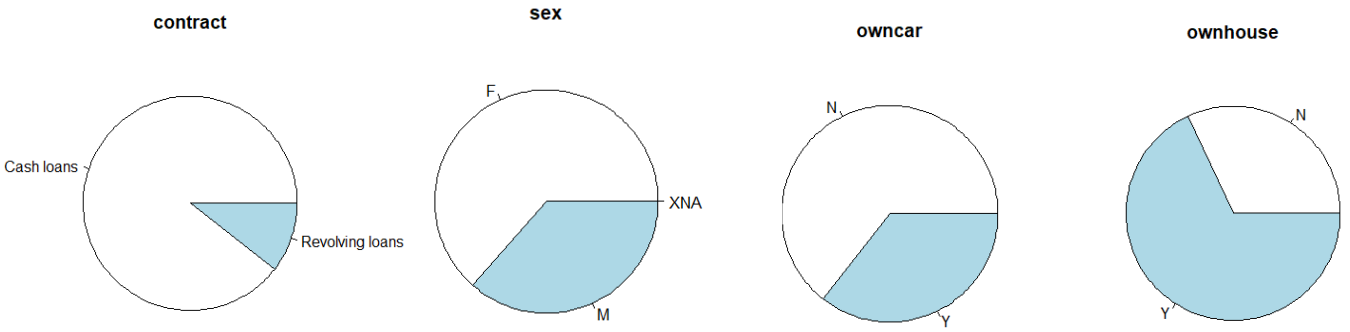


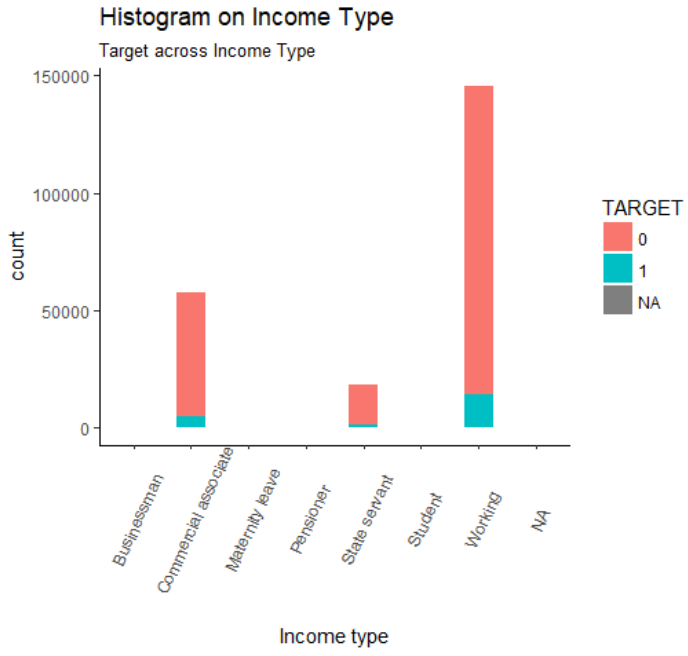
Enquiries/qrt is the number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application), Enquiries/year is the number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application). From the tables below, the person in our sample how has 19 times enquiries in the last 3 month (exclude one month before application) has default risk. The person has 22 times enquires in the last year (exclude last 3 months before application) also has default risk. This suggests people who have very significant enquiries are most likely have very high default risk.

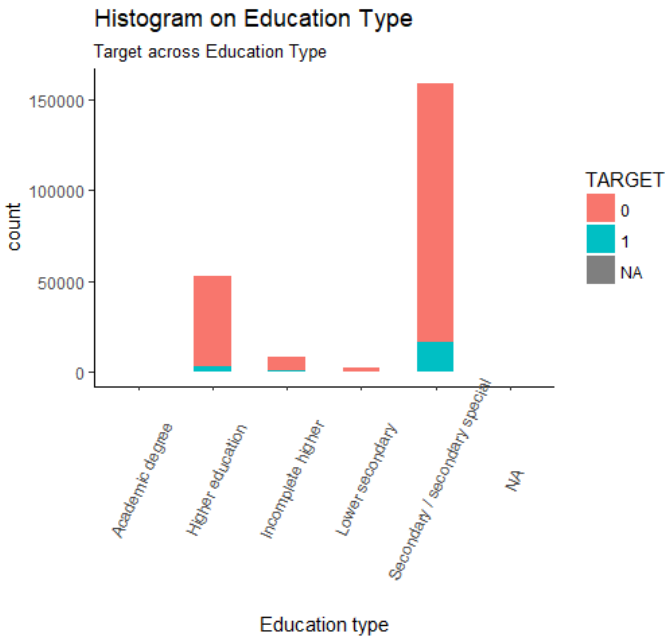
		default risk	
		0	1
enquiries/qrt	0	142484	13772
	1	21965	1790
	2	9168	954
	3	1106	100
	4	289	44
	5	41	5
	6	13	7
	7	4	1
	8	5	
	19		1
	#Total cases	175075	16674

		default risk	
		0	1
enquiries/yr	0	47074	4120
	1	43108	3849
	2	33400	3356
	3	22076	2139
	4	13362	1403
	5	7535	790
	6	4265	498
	7	2314	274
	8	1271	140
	9	603	94
	10	12	3
	11	17	2
	12	16	1
	13	7	
	14	5	3
	15	2	
	16	1	1
	17	5	
	19	1	
	21	1	
	22		1
	#Total cases	175075	16674

From the graphs below, we can roughly gain more information about our sample data from the categorial variables. Most contract types are cash loans, most people are female, more people don't own car, most people own house. Most people's income type is working and commercial associate. Most people went to secondary school. 10.8% of civil marriage has default risk and 10.9% of unmarried has default risk. People who live in rented apartment or with parents have 13% and 12% of default risk respectively. Low skill laborers have relatively high default risk (17.3%). From the organization type, we can see that some organization type tends to have more default risk people such as transportation: type 3. Also, some organization type tends to have lower default risk such as industry type 12.







		default risk	
		0	1
family status	Civil marriage	89.2	10.8
	Married	91.5	8.5
	Separated	91.0	9.0
	Single / not married	89.1	10.9
	Unknown		
	Widow	94.0	6.0
	#Total cases	201917.0	20172.0

		default risk	
		0	1
housing	Co-op apartment	91.5	8.5
	House / apartment	91.2	8.8
	Municipal apartment	90.3	9.7
	Office apartment	92.9	7.1
	Rented apartment	87.0	13.0
	With parents	88.0	12.0
	#Total cases	201917.0	20172.0

		default risk	
		0	1
occupation		91.5	8.5
	Accountants	94.8	5.2
	Cleaning staff	90.1	9.9
	Cooking staff	89.3	10.7
	Core staff	93.4	6.6
	Drivers	88.2	11.8
	High skill tech staff	93.5	6.5
	HR staff	93.6	6.4
	IT staff	92.7	7.3
	Laborers	89.1	10.9
	Low-skill Laborers	82.7	17.3
	Managers	93.5	6.5
	Medicine staff	93.2	6.8
	Private service staff	93.1	6.9
	Realty agents	91.9	8.1
	Sales staff	90.1	9.9
	Secretaries	93.1	6.9
	Security staff	88.9	11.1
	Waiters/barmen staff	88.7	11.3
	#Total cases	201917.0	20172.0

		default risk	
		0	1
organization	Advertising	92.1	7.9
	Agriculture	89.3	10.7
	Bank	94.3	5.7
	Business Entity Type 1	91.2	8.8
	Business Entity Type 2	91.0	9.0
	Business Entity Type 3	90.2	9.8
	Cleaning	88.1	11.9
	Construction	87.5	12.5
	Culture	93.8	6.2
	Electricity	92.9	7.1
	Emergency	92.8	7.2
	Government	92.7	7.3
	Hotel	93.2	6.8
	Housing	91.8	8.2
	Industry: type 1	88.2	11.8
	Industry: type 10	92.6	7.4
	Industry: type 11	91.1	8.9
	Industry: type 12	96.2	3.8
	Industry: type 13	85.2	14.8
	Industry: type 2	92.2	7.8
	Industry: type 3	89.0	11.0
	Industry: type 4	89.7	10.3
	Industry: type 5	93.1	6.9
	Industry: type 6	92.1	7.9

		Industry: type 7	91.7	8.3
		Industry: type 8	86.4	13.6
		Industry: type 9	92.8	7.2
		Insurance	94.3	5.7
		Kindergarten	92.9	7.1
		Legal Services	89.4	10.6
		Medicine	93.2	6.8
		Military	94.3	5.7
		Mobile	90.7	9.3
		Other	91.9	8.1
		Police	94.7	5.3
		Postal	91.3	8.7
		Realtor	88.2	11.8
		Religion	93.5	6.5
		Restaurant	88.0	12.0
		School	93.7	6.3
		Security	89.6	10.4
		Security Ministries	94.9	5.1
		Self-employed	89.5	10.5
		Services	93.2	6.8
		Telecom	91.7	8.3
		Trade: type 1	90.6	9.4
		Trade: type 2	92.3	7.7
		Trade: type 3	89.3	10.7
		Trade: type 4	96.6	3.4
		Trade: type 5	93.0	7.0
		Trade: type 6	95.1	4.9
		Trade: type 7	90.3	9.7
		Transport: type 1	95.5	4.5
		Transport: type 2	91.9	8.1
		Transport: type 3	83.7	16.3
		Transport: type 4	90.3	9.7
		University	94.6	5.4
		XNA		
		#Total cases	201917.0	20172.0

After exploring the data set, I found that some variables are not that important in the model and some variables are similar, so I decide not to put some variables in the model. Remove enquiries per quarter and per year because it seems that abnormally high enquiries means high default risk. Remove occupation type and organization type since they have a lot of categories. Remove whether own house because it is similar to variable housing type, also remove family status and family members since it is similar to the variable number of children. Remove income types since most people get income from working. Then I use the rest variables

to do logistic regression and select a few variables of them to do decision tree model in the next section.

III. Methods

In this section, I plan to use logistic regression and tree-based learning algorithms decision tree and random forest to find patterns of applicants who have default risk based on the chosen variables. Logistic Regression is used when the dependent variable is binary, either 0 or 1. In our dataset, the dependent variable is TARGET which equals to 1 if the application has payment difficulties and equals to 0 if otherwise. We use logistic regression to see how variables affect the probability of a client with payment difficulties. I will expect that people with car, house, kid, higher income from working, higher education, married, higher credit borrowed, lower annuity, lower goods price, will lower the probability of a client with payment difficulties.

Tree-based algorithms are supervised learning method that are used for classification problems. That is, we have the dataset and we know the data for an applicant with default risk. We want to use the information we have to predict the applicants who will have default risk. Unsupervised learning method such as clustering is another data mining method. If we don't know whether the applicant has default risk, and we only know those independent variables. Then we will use clustering to find the patterns based on those variables. Since we have the information of whether the applicant has default risk, we will use classification to identify the occupied rooms.

III. Analysis

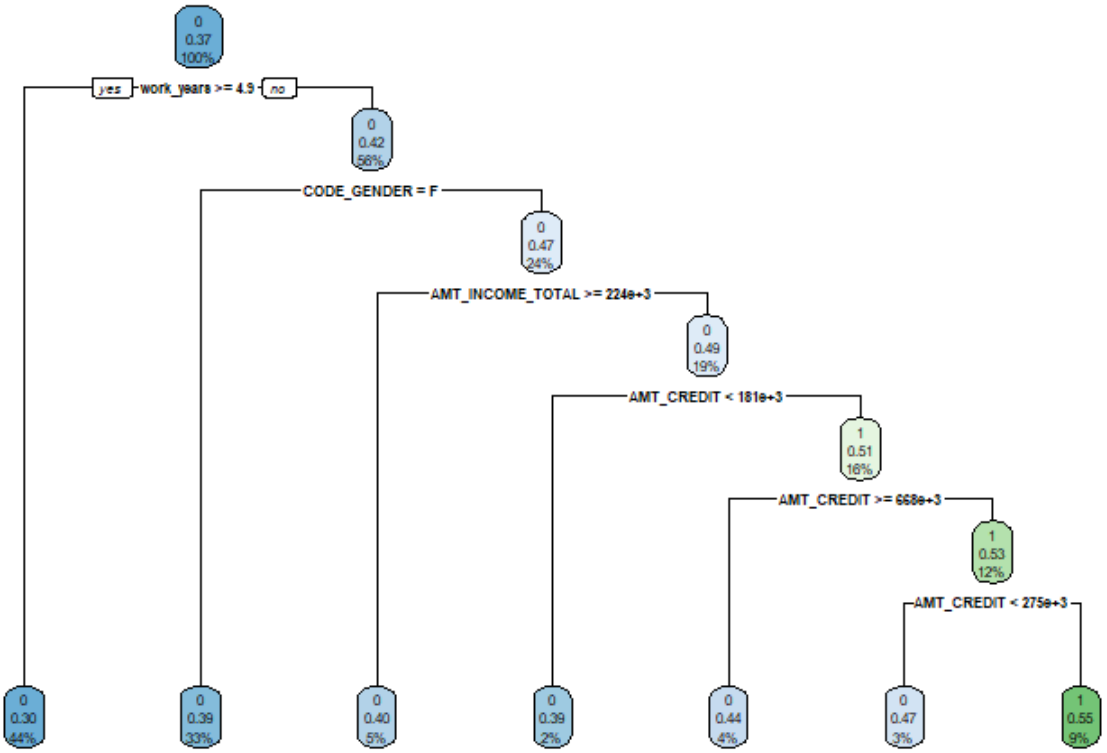
The results of logistic model are shown below. Variables loan type, sex, owncar, children number, income, credit, annuity, goods price, age, working years, and region population are

significantly affect the probability of a client with payment difficulties. More specifically, revolving loans other than cash loan, own a car, more children, higher income, higher good price, higher population region, older, work longer will lower the probability of a client with payment difficulties. Male other than female, loan credit, loan annuity will increase the probability of default risk. The results meet the expectation, expect for the education. In the EDA, applicants' education with academic degree has lower default risk.

```
Call:
glm(formula = TARGET ~ NAME_CONTRACT_TYPE + CODE_GENDER + FLAG_OWN_CAR +
    CNT_CHILDREN + AMT_INCOME_TOTAL + AMT_CREDIT + AMT_ANNUITY +
    AMT_GOODS_PRICE + NAME_EDUCATION_TYPE + NAME_HOUSING_TYPE +
    REGION_POPULATION_RELATIVE + age + work_years, family = "binomial",
    data = df11.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6621  -0.9770  -0.7464   1.2312   2.5642

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -1.025e+01  5.989e+01  -0.171 0.864170
NAME_CONTRACT_TYPERevolving loans           -3.805e-01  4.760e-02  -7.994 1.31e-15 ***
CODE_GENDERM                                 3.730e-01  2.672e-02  13.959  < 2e-16 ***
CODE_GENDERXNA                              -8.785e+00  1.970e+02  -0.045 0.964425
FLAG_OWN_CARY                               -3.144e-01  2.731e-02 -11.512  < 2e-16 ***
CNT_CHILDREN                                -4.079e-02  1.616e-02  -2.524 0.011586 *
AMT_INCOME_TOTAL                            -7.476e-07  2.241e-07  -3.335 0.000852 ***
AMT_CREDIT                                   2.484e-06  1.980e-07  12.544  < 2e-16 ***
AMT_ANNUITY                                  1.635e-05  1.773e-06   9.225  < 2e-16 ***
AMT_GOODS_PRICE                             -3.387e-06  2.228e-07 -15.201  < 2e-16 ***
NAME_EDUCATION_TYPEHigher education          1.028e+01  5.989e+01   0.172 0.863760
NAME_EDUCATION_TYPEIncomplete higher         1.047e+01  5.989e+01   0.175 0.861287
NAME_EDUCATION_TYPELower secondary           1.107e+01  5.989e+01   0.185 0.853379
NAME_EDUCATION_TYPESecondary / secondary special  1.080e+01  5.989e+01   0.180 0.856899
NAME_HOUSING_TYPEHouse / apartment           1.171e-01  2.009e-01   0.583 0.560032
NAME_HOUSING_TYPEMunicipal apartment         1.183e-01  2.106e-01   0.562 0.574256
NAME_HOUSING_TYPEOffice apartment           -1.837e-01  2.412e-01  -0.762 0.446168
NAME_HOUSING_TYPERented apartment            2.802e-01  2.164e-01   1.295 0.195463
NAME_HOUSING_TYPEWith parents                1.914e-01  2.056e-01   0.931 0.351760
REGION_POPULATION_RELATIVE                  -6.472e+00  1.147e+00  -5.642 1.68e-08 ***
age                                         -1.916e-02  1.391e-03 -13.777  < 2e-16 ***
work_years                                  -3.722e-02  2.337e-03 -15.929  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41771  on 31670  degrees of freedom
Residual deviance: 39682  on 31649  degrees of freedom
AIC: 39726

Number of Fisher Scoring iterations: 10
```

Decision Tree method is performed using rpart function in R. I put a lot of variables in the model first, then check the feature importance. Based on my experience, I then pick the model with variables gender, income, loan credit, age and working years. I prune this model and get the resulted graph below. When the dependent variable is binary, in the decision tree, each node shows the predicted class (default risk or not), the predicted probability of default risk, and

the percentage of observations in the node. In graph below, if working years is greater than 4.9 years, then the applicant is not with default risk with 44% of the observations. When applicant is female, even if the number of working years is shorter than 4.9, the applicant is not with default risk. If applicant is male, and income level is greater than 224000, then the applicant is not with default risk.



The predicted result using the testing data (see table below) shows that there are 333 errors for applicants without default risk, where it should be no default risk, but the model predicts the applicant with default risk. Also, there are 4425 errors in predicting default risk. I also try other models and check the feature importance, there is no significant sign that the model will be improved. Other models that I tried still have around 4000 errors is predicting the default risk.

```
  predictions3
     0    1
0 5926  333
1 4425  501
```

```
tree1$variable.importance
          work_years     NAME_EDUCATION_TYPE                     age      AMT_ANNUITY    AMT_GOODS_PRICE
          227.1247561          125.5098662           90.0683276           50.3076216           49.0232447
          CODE_GENDER           AMT_CREDIT   NAME_CONTRACT_TYPE      FLAG_OWN_CAR   AMT_INCOME_TOTAL
           40.8043954           39.6576410           33.6595415           25.6974760            3.8221696
     CNT_CHILDREN REGION_POPULATION_RELATIVE     NAME_HOUSING_TYPE
            2.6764717            0.5074309            0.3781231
```

IV. Conclusion

This paper explores the train dataset from home credit default risk competition. Then, the paper applies logistics model and decision tree model to predict applicants' default risk. The logistic model shows that default risk might be resulted from cash loan, not own a car, less children, lower income, lower good price, lower population region, younger, work shorter, male, higher credit, and higher annuity will increase the probability of default risk. The classification model also suggests that work more years, female, higher income, lower credit will lower default risk, even though the model has errors in predicting the default risk. The reason for these could be that the variables spread out the class and they don't have a high degree of orders as shown in the EDA.

References

Home Credit Group (2018). Retrieved from: http://www.homecredit.net/.

Home Credit Default Risk (2018). Retrieved from: https://www.kaggle.com/c/home-credit-default-risk.

Michy Alice (2015). How to perform a Logistic Regression in R. Retrieved from: https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

Stephen Milborrow (2018). Plotting rpart trees with the rpart.plot package. Retrieved from: http://www.milbo.org/rpart-plot/prp.pdf

ANALYTICS VIDHYA CONTENT TEAM (2016). A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python). Retrieved from: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

Appendix

Table 1: Variables list

	Variable name	Description
1	SK_ID_CURR	ID of loan in our sample
2	TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
3	NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
4	CODE_GENDER	Gender of the client
5	FLAG_OWN_CAR	Flag if the client owns a car
6	FLAG_OWN_REALTY	Flag if client owns a house or flat
7	CNT_CHILDREN	Number of children the client has
8	AMT_INCOME_TOTAL	Income of the client
9	AMT_CREDIT	Credit amount of the loan
10	AMT_ANNUITY	Loan annuity
11	AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
12	NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,…)
13	NAME_EDUCATION_TYPE	Level of highest education the client achieved
14	NAME_FAMILY_STATUS	Family status of the client
15	NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
16	REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
17	DAYS_BIRTH	Client's age in days at the time of application
18	DAYS_EMPLOYED	How many days before the application the person started current employment
19	OCCUPATION_TYPE	What kind of occupation does the client have
20	CNT_FAM_MEMBERS	How many family members does client have
21	ORGANIZATION_TYPE	Type of organization where client works
22	AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
23	AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)