Default Risk Detection

ALY6040 Data Mining Application

Minyi Chen

07/03/2018

Introduction

- Being able to predict the reliability of the borrowers is very important to the lenders.
- Home Credit Default Risk Competition by Home Credit Group:

"Can you predict how capable each applicant is of repaying a loan?"

Data

Data is from Kaggle.com

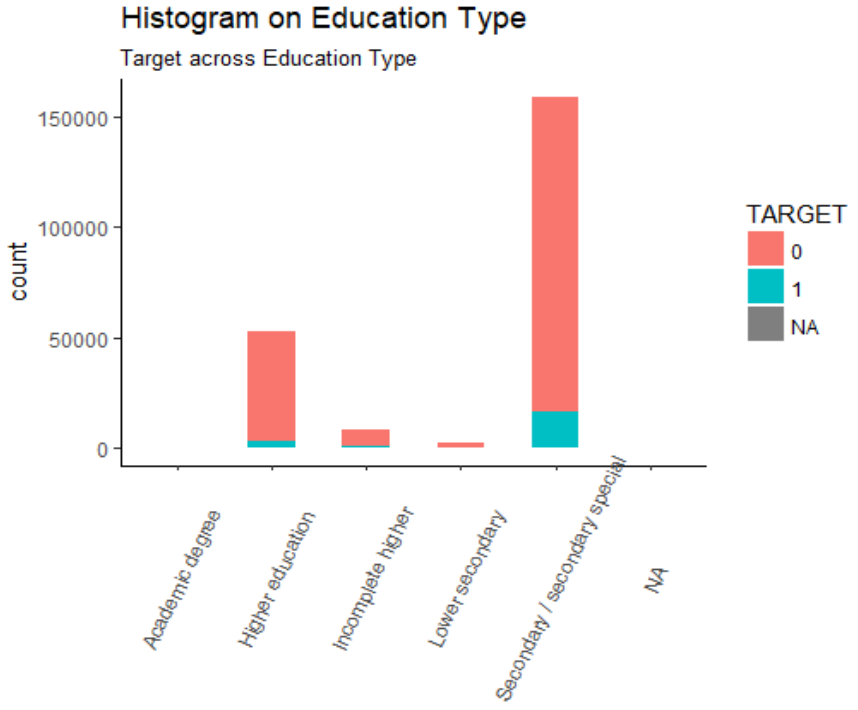
new variables:

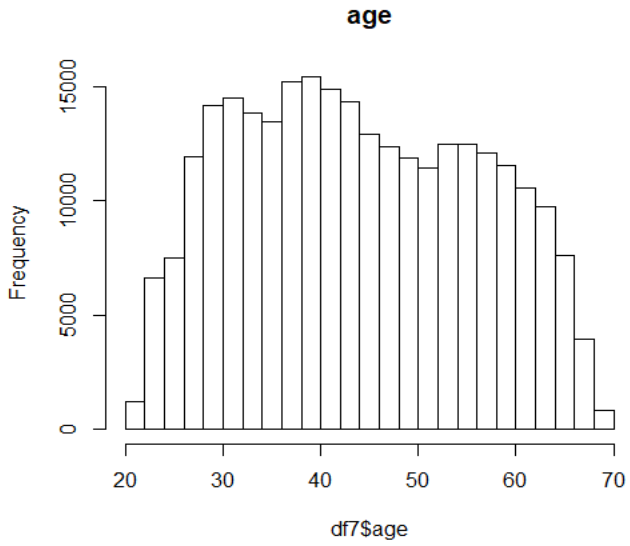
age = DAYS_BIRTH / (-365),

work_years= DAYS_EMPLOYED/ (-365)

	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)

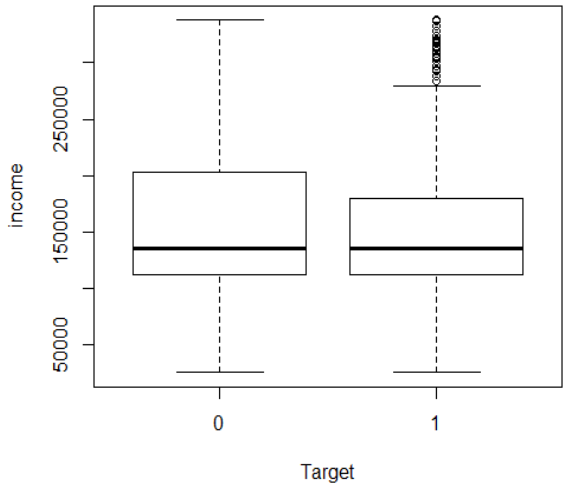
Exploratory Data Analysis (EDA)





Education type

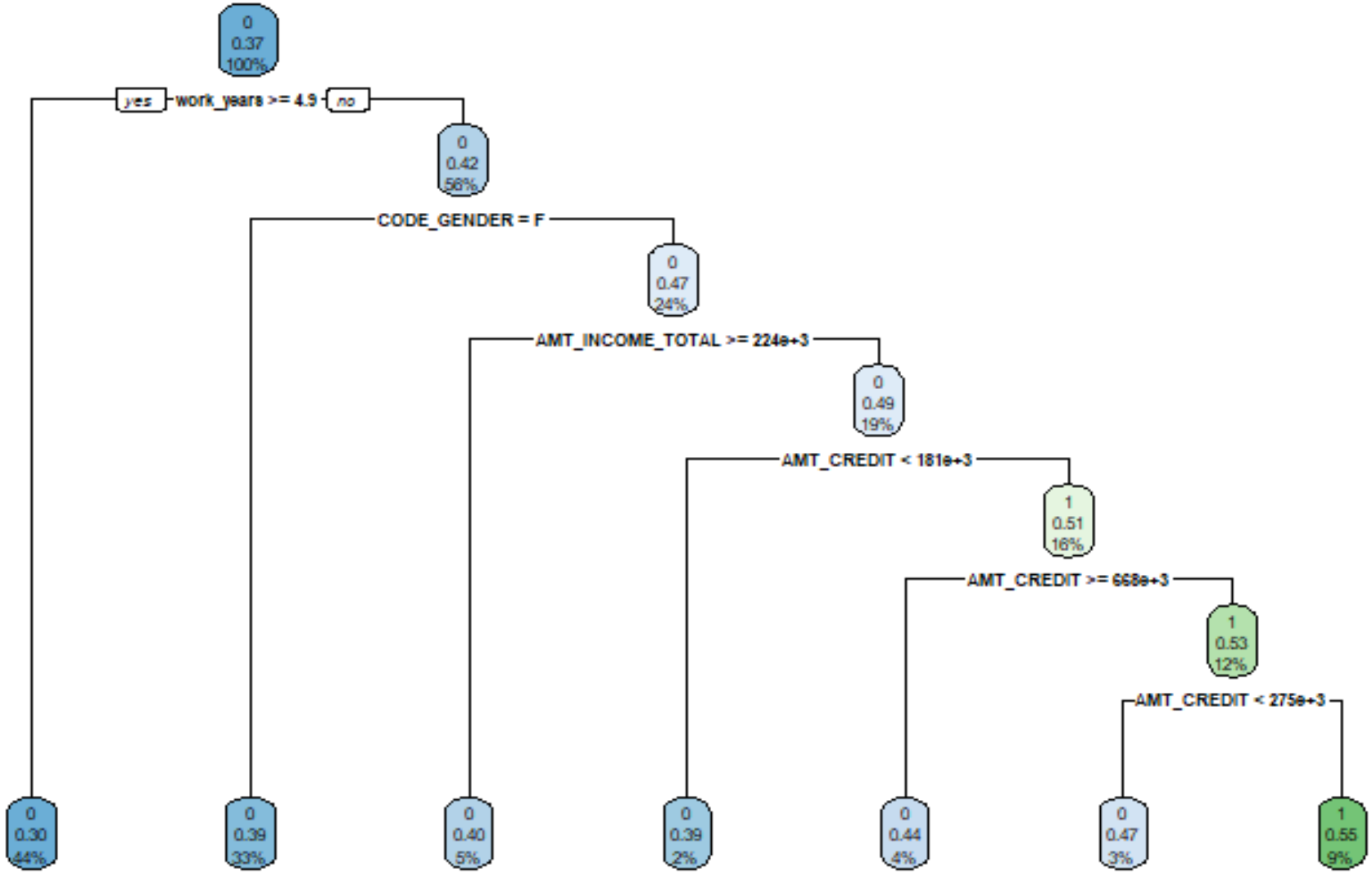
		0	1
occupation		91.5	8.5
	Accountants	94.8	5.2
	Cleaning staff	90.1	9.9
	Cooking staff	89.3	10.7
	Core staff	93.4	6.6
	Drivers	88.2	11.8
	High skill tech staff	93.5	6.5
	HR staff	93.6	6.4
	IT staff	92.7	7.3
	Laborers	89.1	10.9
	Low-skill Laborers	82.7	17.3
	Managers	93.5	6.5
	Medicine staff	93.2	6.8
	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



Logistic Regression

```
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



Conclusion

- 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.

Thank you!