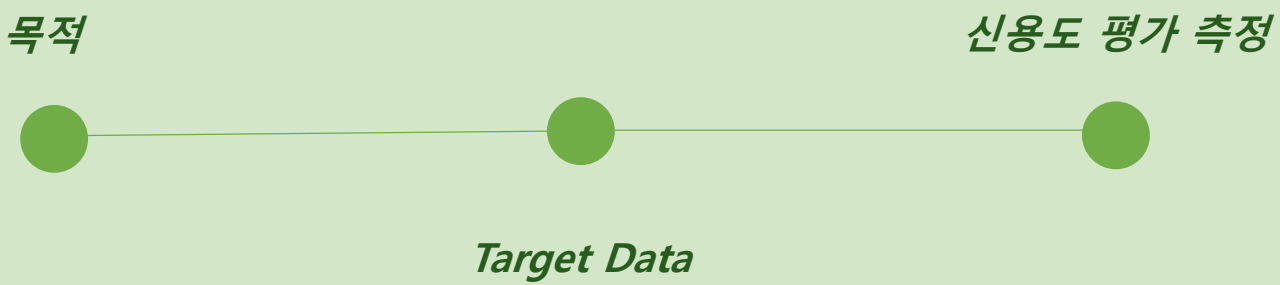
PSAT & YBIGTA

신용불량자

김수진, 안세훈, 유건욱, 이용하, 이정현

1 주제 설명



목적

공식 신용 기록이 없거나 부족한 사람들을 위해

*대체 신용 정보*를 이용함으로써

보다 합당한 대출 가능 여부 예측을 돕기 위함

Target Data

이번 대출 받았을 때 **연체**했니? ➡ *1*

밀리지 않고 꼬박꼬박 잘 **상환**했니? ➡ *0*

신용도 평가

1

상환이력정보

현재 연체 보유 여부 및
과거 채무 상환 이력

2

신용형태정보

신용거래종류, 신용거래형태
(상품별 건수, 활용비중)

3

현재부채수준

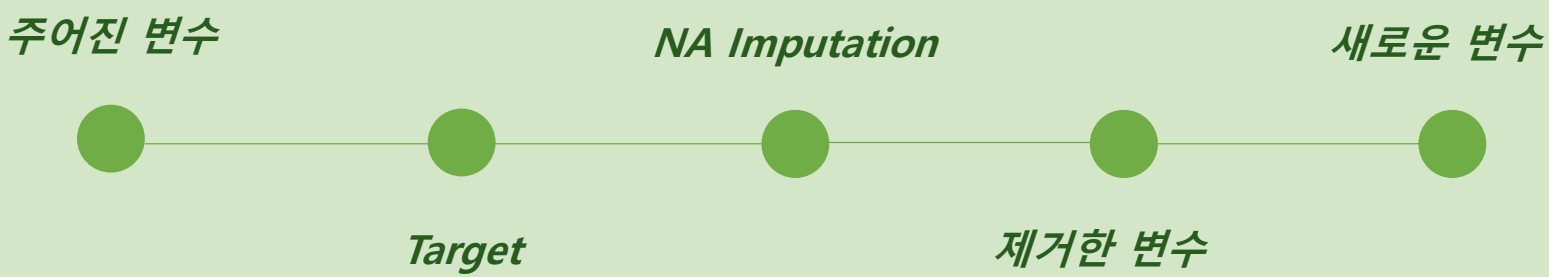
채무 부담 정보
(대출 및 보증 채무 등)

4

신용거래기간

신용거래 거래 기간
(최초/최근 개설로부터 기간)

2 데이터



데이터 개요

Data sets (총 7개의 세트)

Train

application_train
bureau
bureau_balance
credit_card_balance
installments_payments
POS_CASH_balance
previous_application

Test

application_test

적게는 28만개부터
많게는 1300만개의
다양한 OBS 수

&

총 220여개의 변수

0

Variable

총 변수 : 205

application_{train|test}.csv — SK_ID_CURR

	변수명	변수 설명
	TARGET	Target Variable
대출 관련 변수	SK_ID_CURR	대출 ID
	AMT_CREDIT	신용도
	AMT_GOODS_PRICE	담보
개인 정보 관련 변수	AMT_ANNUITY	연금
	NAME_INCOME_TYPE	소득
	NAME_FAMILY_STATUS	가족 형태
	DAYS_BIRTH	나이(생일)
	CNT_CHILDREN	자녀 수
	FLAG_MOBIL	휴대폰

0

Variable

총 변수 : 205

bureau / bureau_balance.csv — SK_ID_CURR / BUREAU_ID

	변수명	변수 설명
과거 신용도 관련 변수	CREDIT_ACTIVE	과거 신용 정보 상태
	AMT_CREDIT_SUM	현재 대출 상한선
	AMT_CREDIT_SUM_DEPT	현재 부채 상태
	CREDIT_TYPE	신용도 종류
과거 신용도 잔액 관련 변수	MONTHS_BALANCE	잔액 관련 개월 수
	STATUS	월별 잔액

POS_CASH_balance.csv — SK_ID_CURR / SK_ID_PREV

현금 대출 관련 변수	CNT_INSTALMENT_FUTURE	과거 신용 잔액
	SK_DPD	월별 연체 일수

0

Variable

총 변수 : 205

credit_card_balance.csv — SK_ID_CURR / SK_ID_PREV

	변수명	변수 설명
신용 카드 관련 변수	AMT_DRAWINGS_CURRENT	월별 출금 금액
	AMT_INST_MIN_REGULARITY	최소 할부금
	AMT_PAYMENT_CURRENT	월별 카드값
	MT_PAYMENT_TOTAL_CURRENT	누적 카드값

previous_application.csv — SK_ID_CURR / SK_ID_PREV

과거 대출 관련 변수	NAME_CONTRACT_TYPE	과거 대출 상품 종류
	AMT_APPLICATION	대출 신청액
	AMT_CREDIT	대출액
	NFLAG_MICRO_CASH	소액 금융 대출
	RATE_INTEREST_PRIVILEGED	이자율 비율

0

Variable

총 변수 : 205

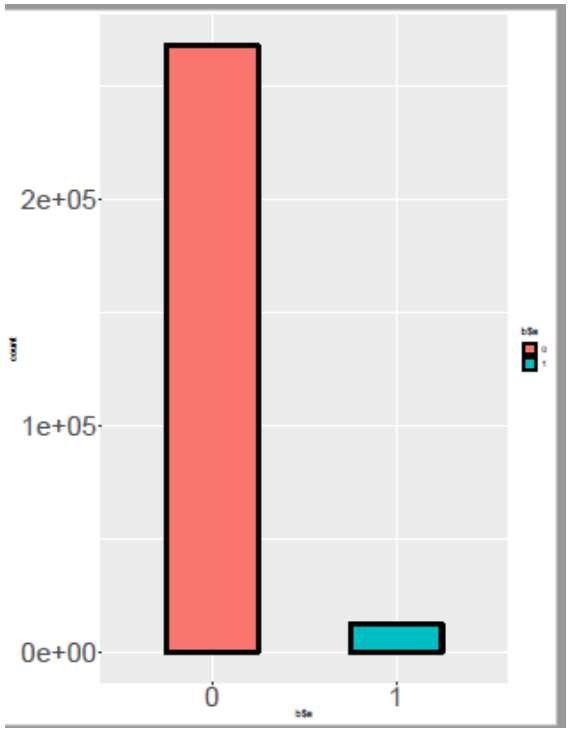
installments_payments.csv — SK_ID_CURR / SK_ID_PREV

	변수명	변수 설명
할부 상환 관련 변수	NUM_INSTALMENT_NUMBER	할부 개월 표시
	DAYS_INSTALMENT	상환 예정 일자
	DAYS_ENTRY_PAYMENT	상환 일자
	AMT_INSTALMENT	할부금액
	AMT_PAYMENT	상환금액

Target

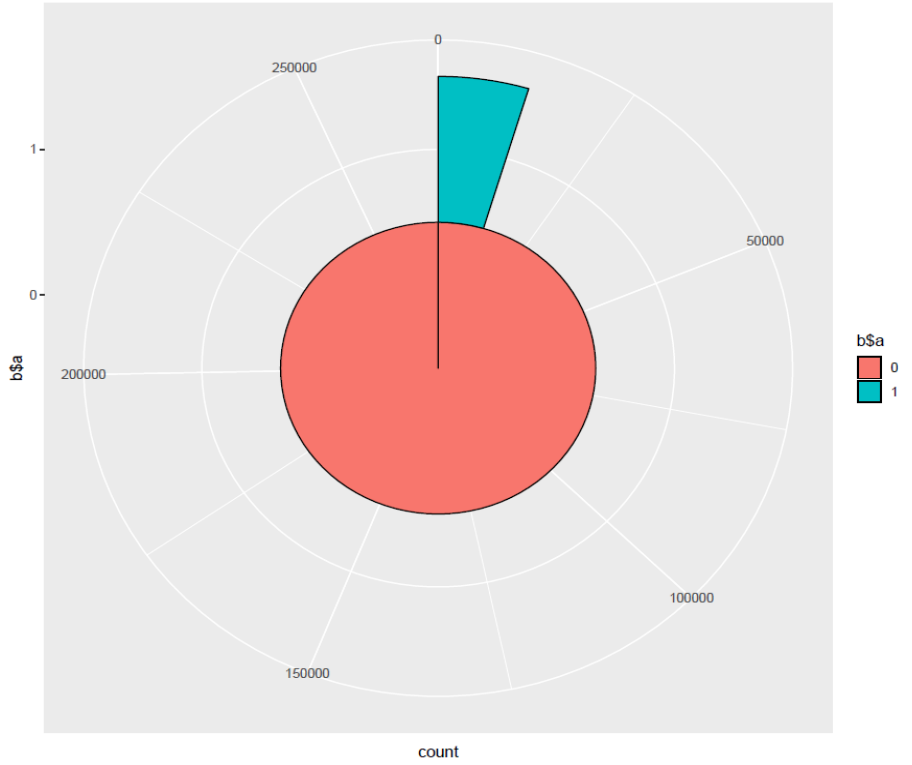
Binary data

: 0과 1을 factor처리 함.



Unbalanced Data

: 목표 값이 심하게 unbalanced 되어있다



Target

Unbalanced Data 처리

1 Sample 조정

- Oversampling
- Undersampling
- Synthetic Samples(SMOTE)

2 Model 조정

- Penalized-SVM
- Penalized-LDA

NA Imputation

기준 1

NA값이 극소수(1% 이하)인 경우

각각 변수의
median으로
대체

기준 2

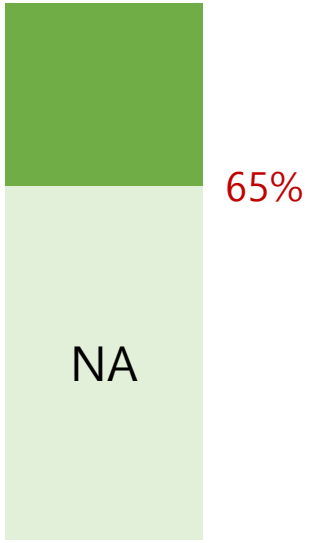
NA값이 50% 이상인 경우

변수 **제거**

1

application_train.csv

OWN_CAR_AGE

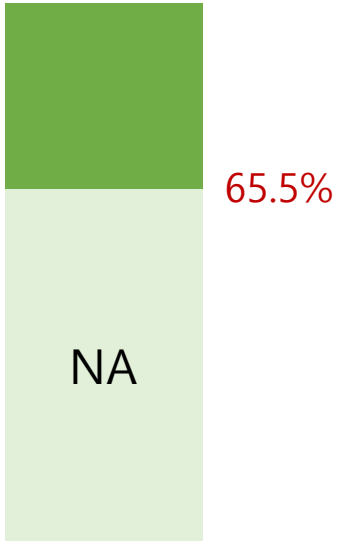


NA가 50% 이상 존재

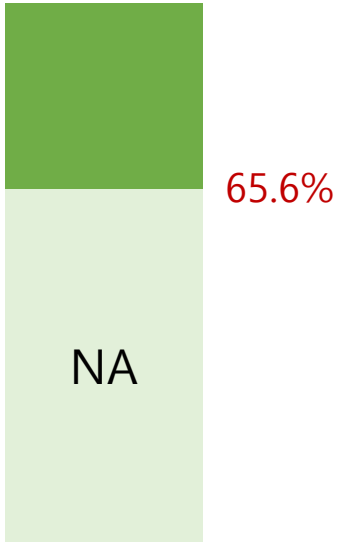
2

bureau.csv

AMT_CREDIT_MAX_OVERDUE



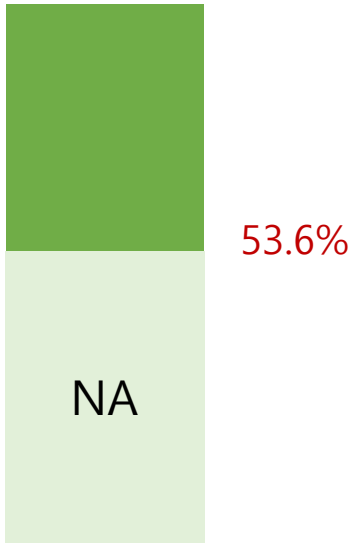
AMT_ANNUITY



NA가 50% 이상 존재

previous_application.csv

AMT_DOWN_PAYMENT
RATE_DOWN_PAYMENT



RATE_INTEREST_PRIMARY
RATE_INTEREST_PRIVILEGED

NA가 50% 이상 존재

4

installment_payments.csv

DAYS_INSTALMENT

DAYS_ENTRY_PAYMENT

AMT_INSTALMENT

AMT_PAYMENT

OVERED_TOTAL

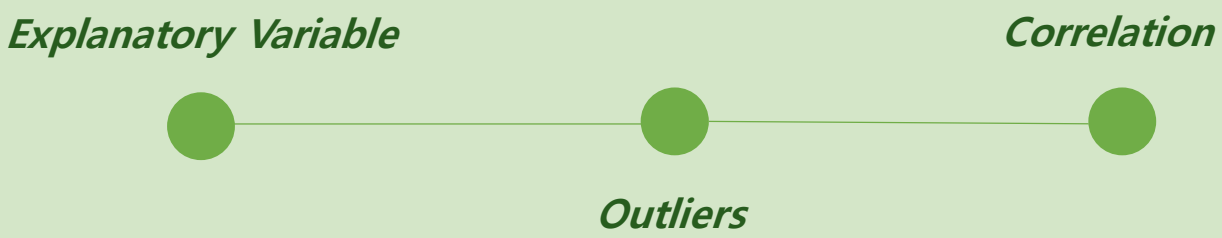
= AMT_INSTALMENT - AMT_PAYMENT

OVERED_COUNT

= (IF ["OVERED_TOTAL"] > 0] = 1)

: 연체 여부와 연체 금액으로 압축 가능하다고 판단

3 EDA

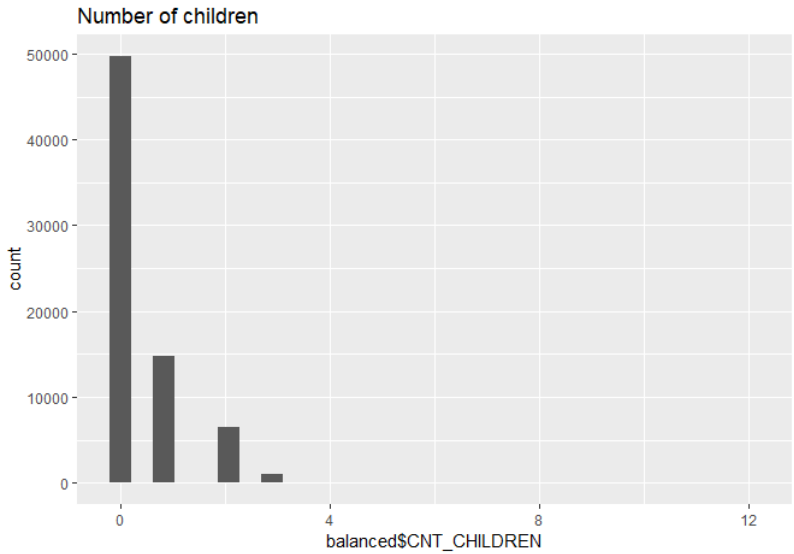


application_train.csv

CNT_CHILDREN

: 아이들의 수

```
> table(balanced2$CNT_CHILDREN)

    0     1     2     3     4     5     6     8     9    11    12
49728 14702  6460   958    96    12     3     1     4     1     1
```



application_train.csv

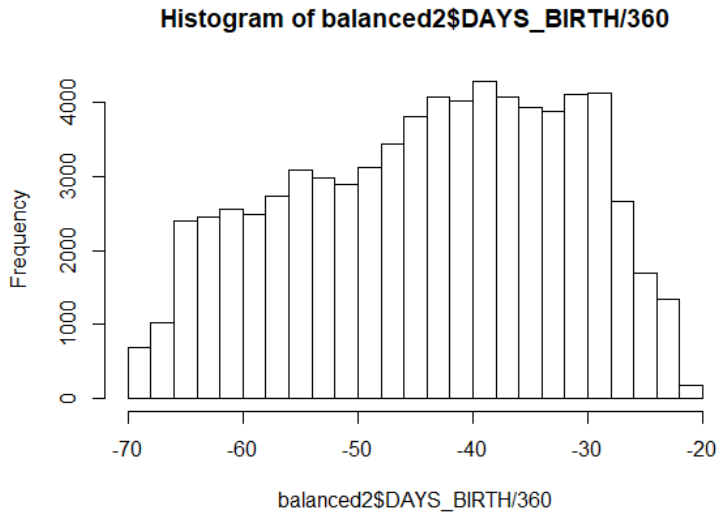
AMT_INCOME_TOTAL

: 소득

```
> summary(balanced2$AMT_INCOME_TOTAL)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    26550    112500    157500    171677    202500 117000000
```

DAYS_BIRTH

: 나이



application_train.csv

DAYS_EMPLOYED

: 현재 직장 근속연수

```
> summary(balanced2$DAYS_EMPLOYED)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -16429   -2687   -1233   58618    -335  365243
```

```
ind = (balanced2$DAYS_EMPLOYED==365243)
balanced2$DAYS_EMPLOYED[ind] <- NaN
##DAYS_EMPLOYED_FLAG 만듬
balanced2$DAYS_EMPLOYED_FLAG <- 0
balanced2$DAYS_EMPLOYED_FLAG[ind]<- 1
balanced2$DAYS_EMPLOYED_FLAG <- as.factor(balanced2$DAYS_EMPLOYED_FLAG )
```

application_train.csv

CNT_FAM_MEMBERS

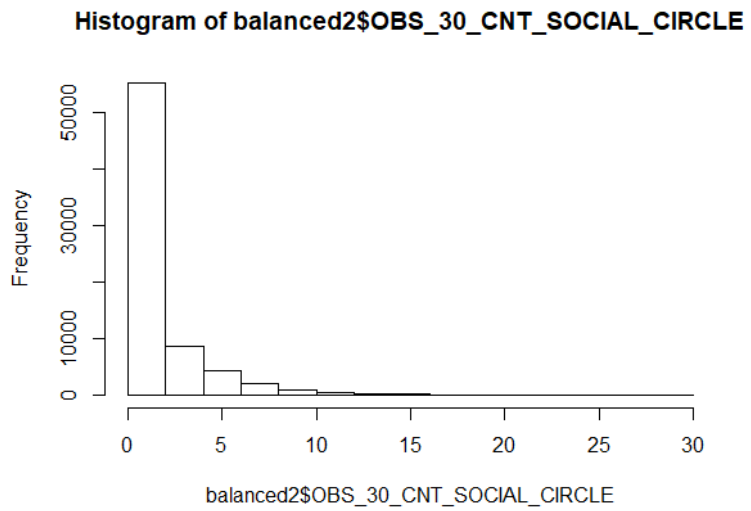
: 가족의 수

```
> table(balanced2$CNT_FAM_MEMBERS)

    1     2     3     4     5     6     7     8    10    11    13    14
15404 36956 12629  5970   891    94    12     3     2     3     1     1
```

OBS_30_CNT_SOCIAL_CIRCLE

: 친인척 중 연체한 사람 수

```
> summary(balanced2$OBS_30_CNT_SOCIAL_CIRCLE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00    1.57    2.00   30.00
```

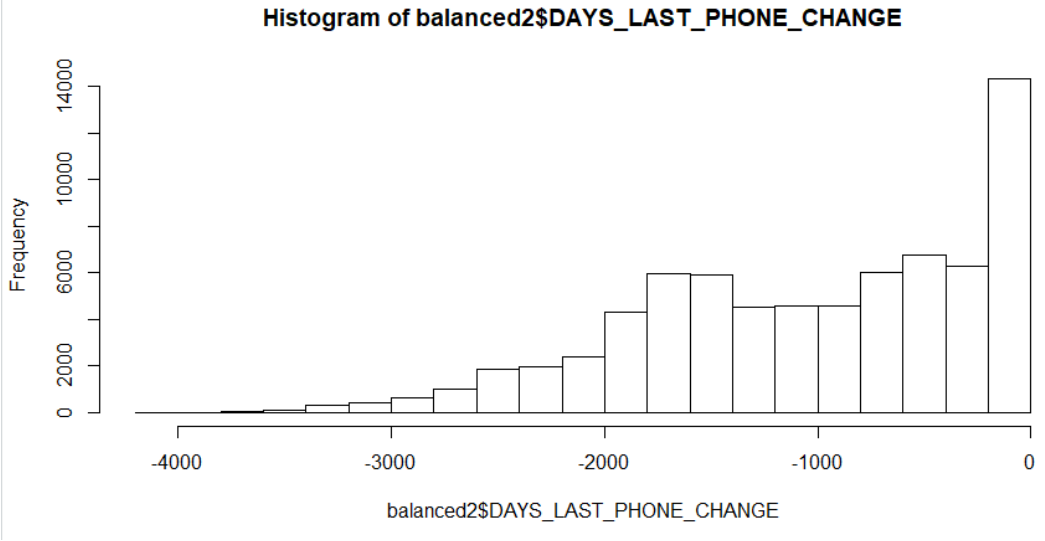


application_train.csv

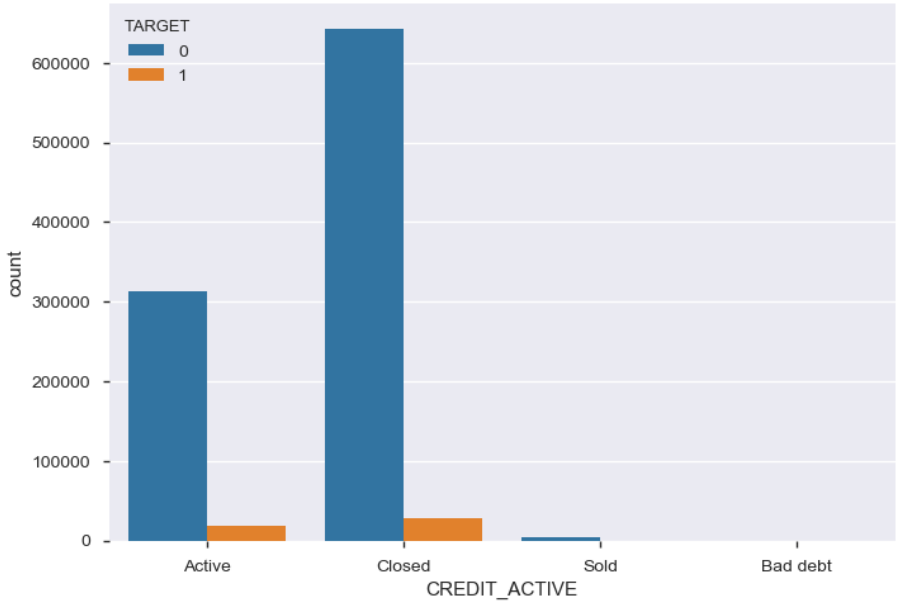
DAYS_LAST_PHONE_CHANGE

: 마지막으로 휴대폰 바꾼 날

```
> summary(balanced2$DAYS_LAST_PHONE_CHANGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -4121   -1634    -906   -1023    -317       0
```



bureau / bureau_balance.csv



CREDIT_ACTIVE
Categorical

Distinct count	4
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

Toggle details

Value	Count	Frequency (%)
Closed	669674	66.6%
Active	331291	33.0%
Sold	4370	0.4%
Bad debt	3	0.0%

bureau / bureau_balance.csv

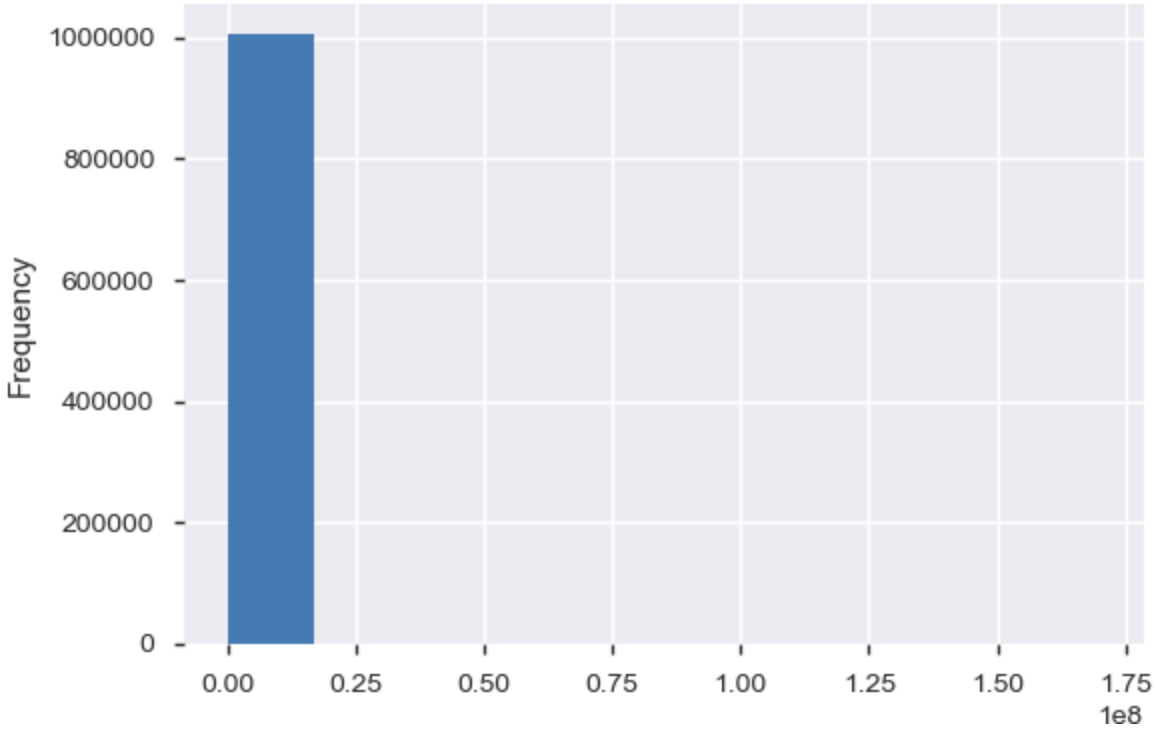
AMT_CREDIT_SUM

Numeric

Mean	369970
Minimum	0
Maximum	170100000
Zeros (%)	3.8%

Quantile statistics

Minimum	0
5-th percentile	13500
Q1	54000
Median	134420
Q3	333000
95-th percentile	1350000
Maximum	170100000
Range	170100000
Interquartile range	279000



bureau / bureau_balance.csv

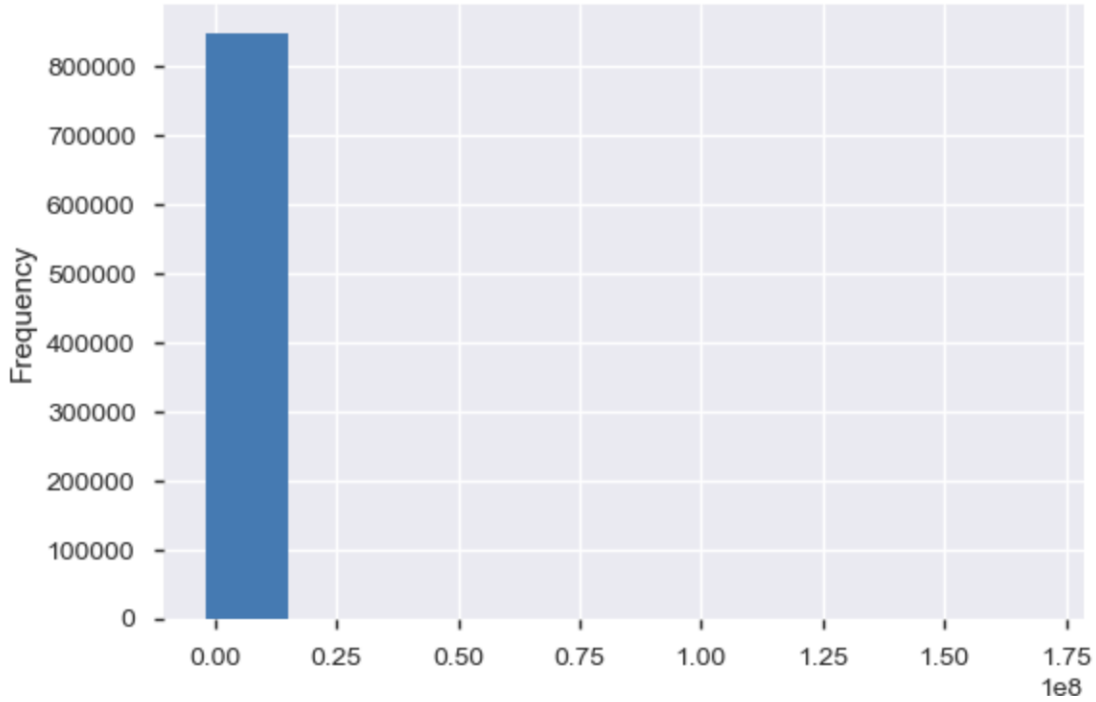
AMT_CREDIT_SUM_DEBT

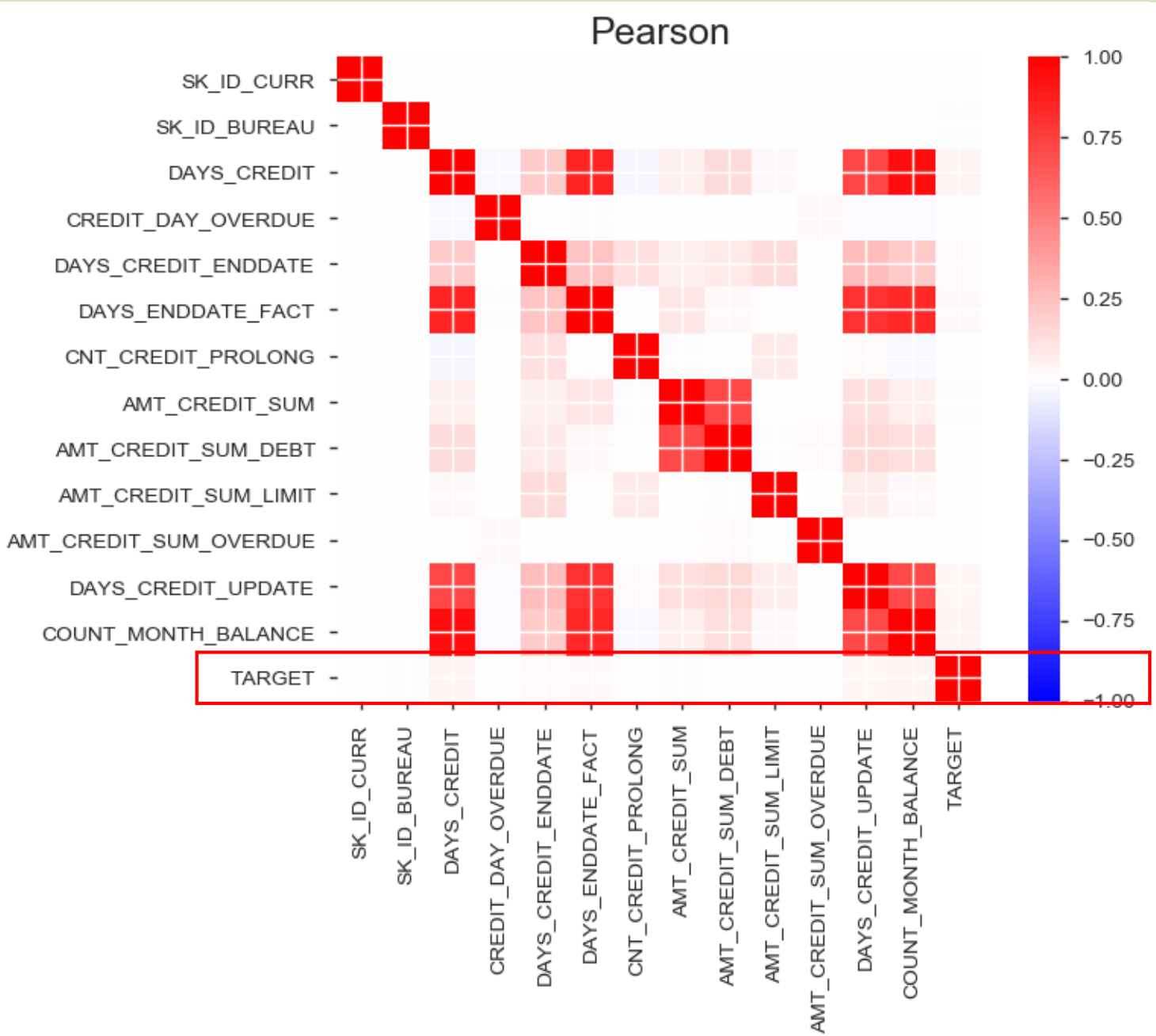
Numeric

Mean	133400
Minimum	-2014800
Maximum	170100000
Zeros (%)	61.1%

Quantile statistics

Minimum	-2014800
5-th percentile	0
Q1	0
Median	0
Q3	23638
95-th percentile	615930
Maximum	170100000
Range	172110000
Interquartile range	23638




Pearson
SK_ID_CURR
SK_ID_BUREAU
DAYS_CREDIT
CREDIT_DAY_OVERDUE
DAYS_CREDIT_ENDDATE
DAYS_ENDDATE_FACT
CNT_CREDIT_PROLONG
AMT_CREDIT_SUM
AMT_CREDIT_SUM_DEBT
AMT_CREDIT_SUM_LIMIT
AMT_CREDIT_SUM_OVERDUE
DAYS_CREDIT_UPDATE
COUNT_MONTH_BALANCE
TARGET
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

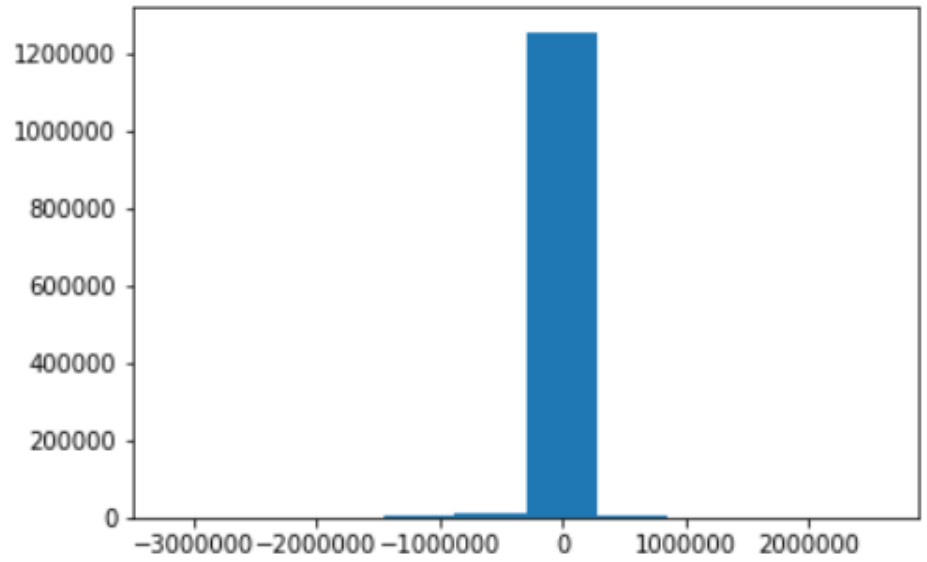
installment_payments.csv

```
In [24]: ins["OVERED_TOTAL"].describe().astype(int)

Out[24]: count    1270913
         mean       -1984
         std        84217
         min     -3195000
         25%            0
         50%            0
         75%            0
         max      2602348
         Name: OVERED_TOTAL, dtype: int32
```

```
In [38]: ins["OVERED_TOTAL"].quantile(0.025)

Out[38]: -63282.69899999999
```

```
In [39]: ins["OVERED_TOTAL"].quantile(0.975)

Out[39]: 74186.97299999947
```

```
In [40]: ins["OVERED_TOTAL"].hist(bins=10, grid=False)

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x19e241739b0>
```



installment_payments.csv

```
In [10]: ins["OVERED_COUNT"].describe().astype(int)
```

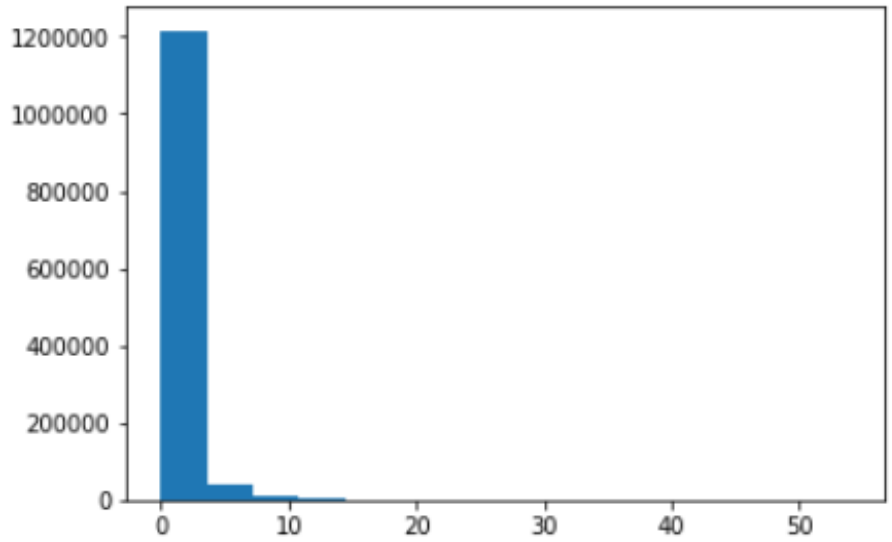
```
Out[10]: count    1270913
         mean           0
         std            1
         min            0
         25%            0
         50%            0
         75%            0
         max           54
         Name: OVERED_COUNT, dtype: int32
```

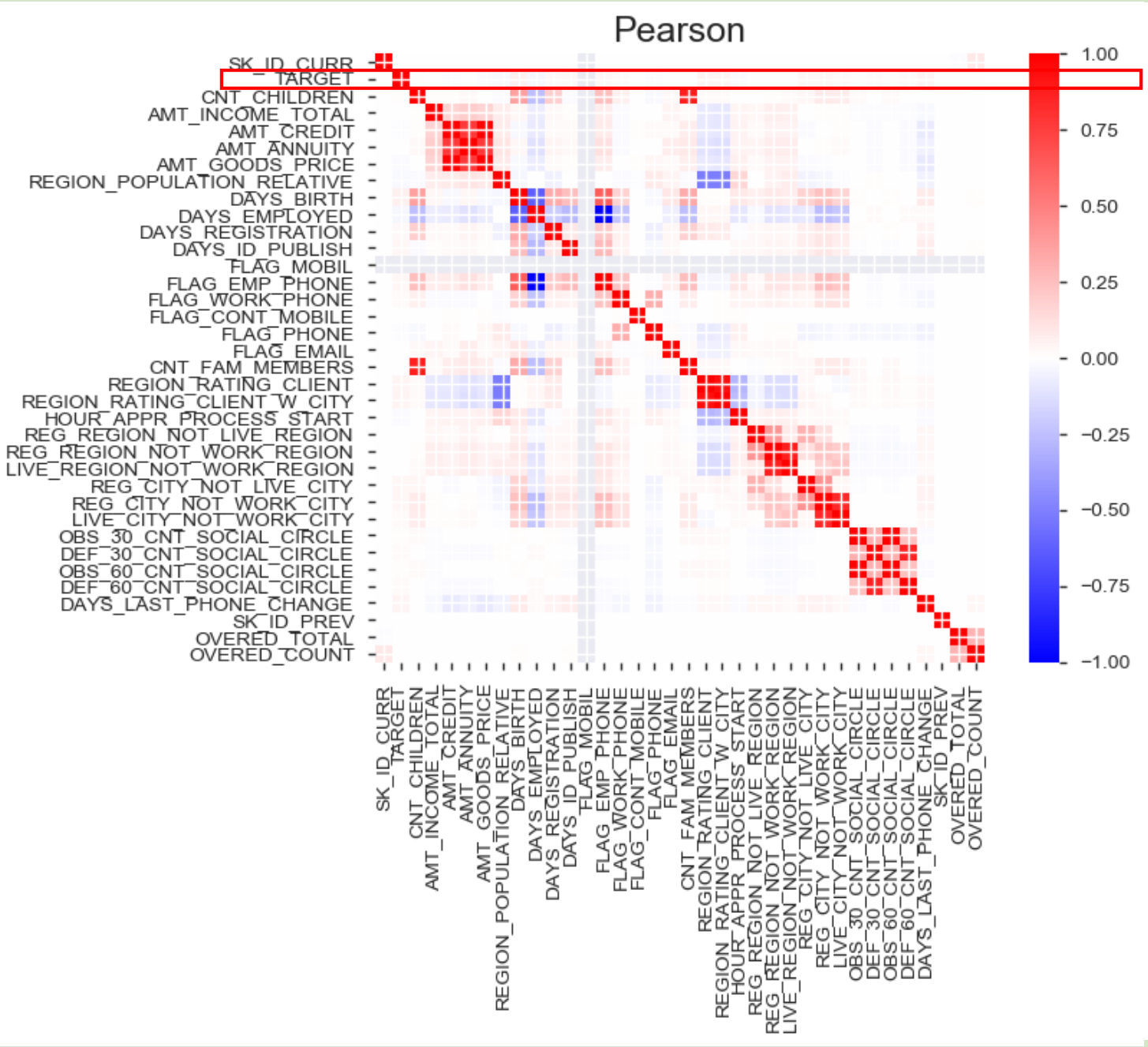
```
In [11]: ins["OVERED_COUNT"].quantile(0.95)
```

```
Out[11]: 3.0
```

```
In [12]: ins["OVERED_COUNT"].hist(bins=15, grid=False)
```

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x1c08ed570b8>
```




Pearson
SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
SK_ID_PREV
OVERED_TOTAL
OVERED_COUNT
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

4 Modeling

CART tree

Classification and Regression Tree

- Decision Tree의 일종으로 분류나 회귀 예측모형
- Gini Index를 통해 분류됨
- 이진 트리
- Full model 형성 후 pruning

CART tree

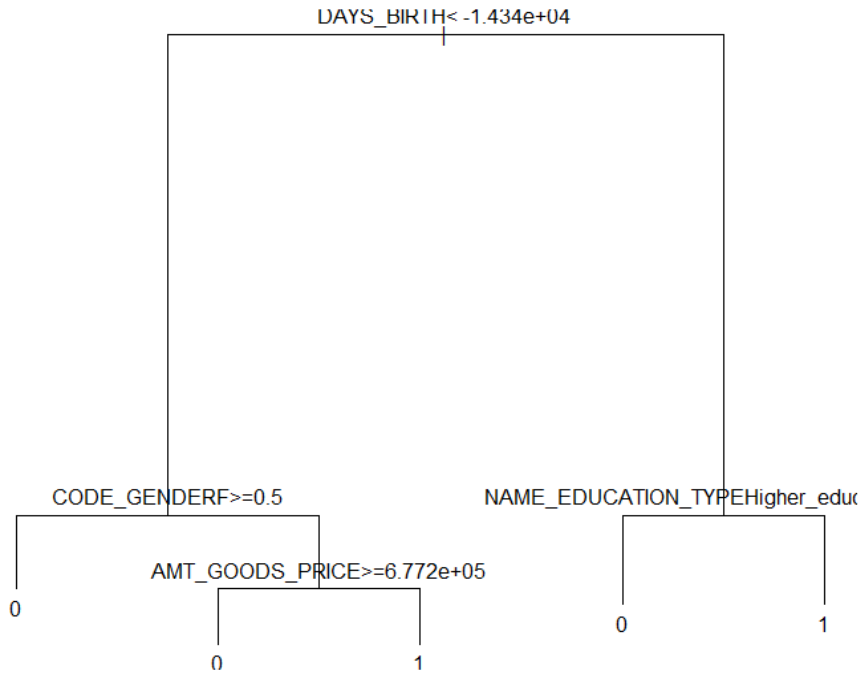
기준 :

DAYS_BIRTH

CODE_GENDER

AMT_GOODS_PRICE

NAME_EDUCATION_TYPEHigher_education



F1-score: 0.5662452

Random Forest

- Decision Tree를 여러 번 만들어서 결과 값의 최빈값 사용
- Train set에 과적합 방지
- OOB error으로 모델 평가

Random Forest

전처리

CODE_GENDER	NAME_TYPE_SUITE	NAME_INCOME_TYPE
NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE
OCCUPATION_TYPE	WEEKDAY_APPR_PROCESS_START	ORGANIZATION_TYPE

범주형 변수는 one-hot encoding 진행

Random Forest

전처리

CONTRACT_TYPE	OWN_CAR	OWN_REALITY

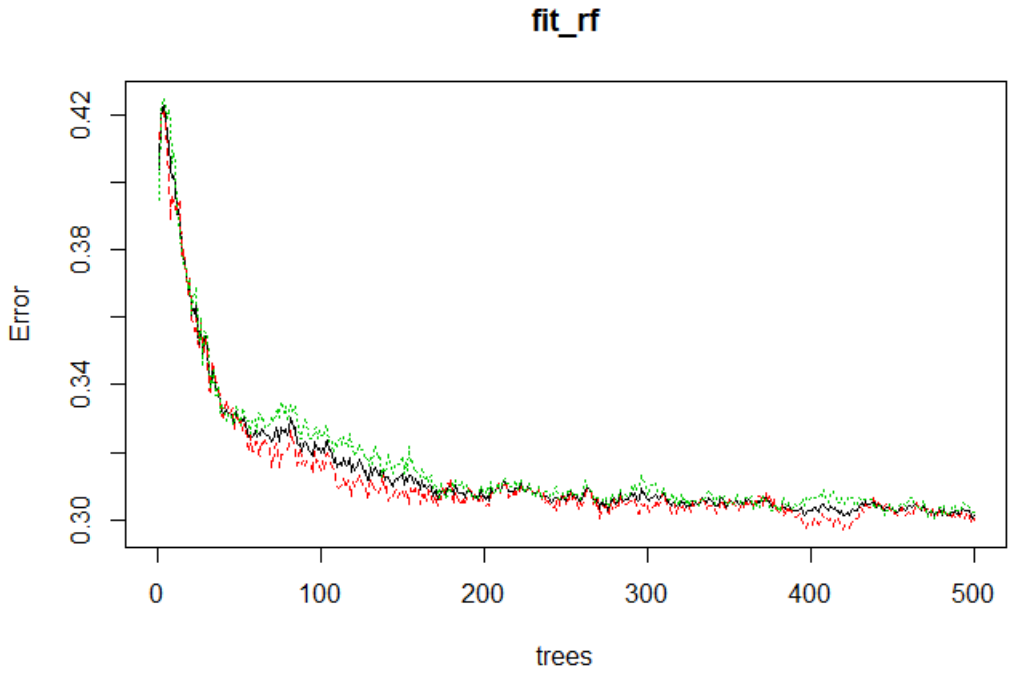
Factor가 두개인 범주형 변수는 0과 1로 표기

Random Forest

Number of trees(ntree): 500
No. of variables at each split(mtry): 12

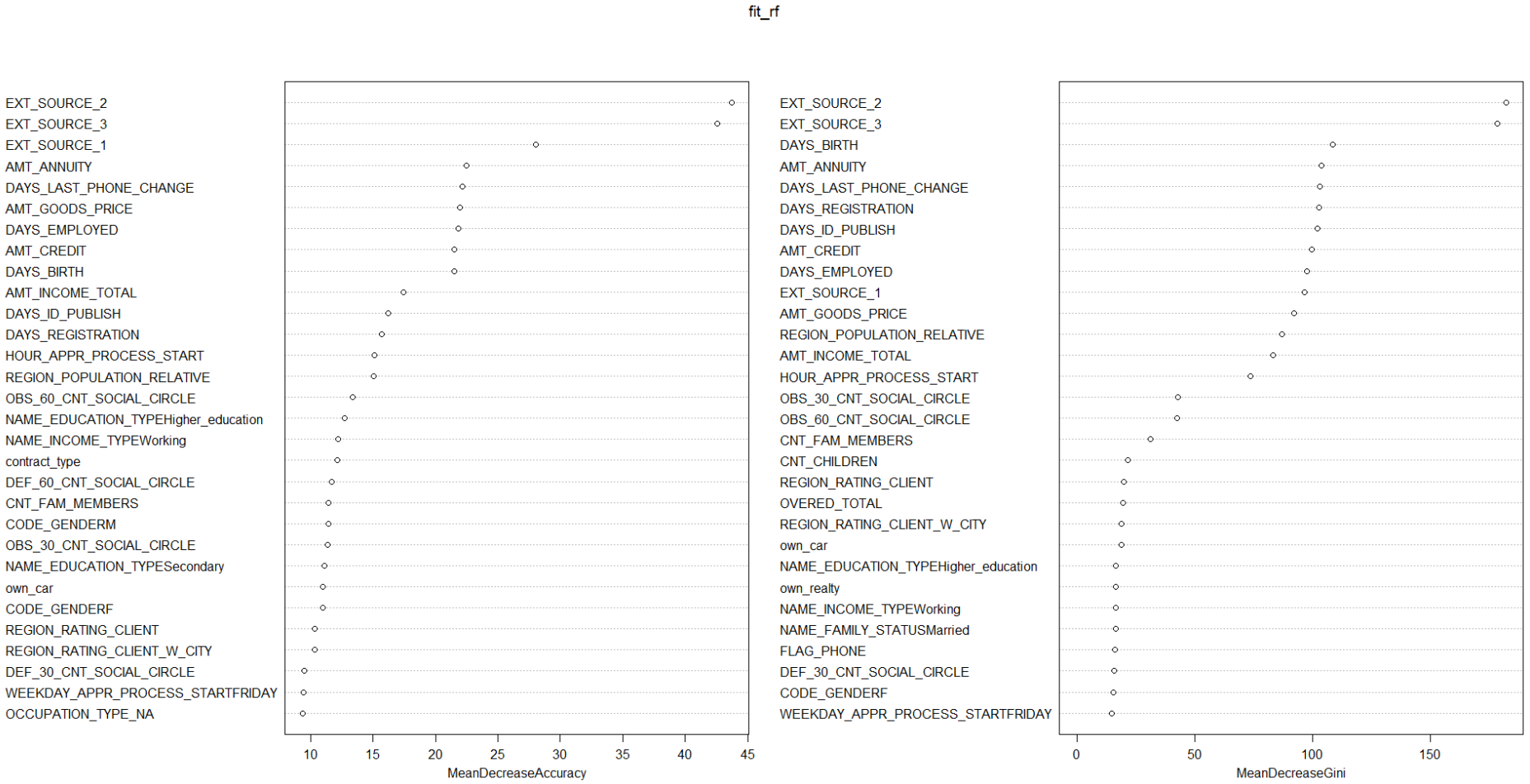
Obs: 5000개 중 NA가 없는 1727개 사용됨

predictions_rf	0	1
0	701	262
1	203	561



***F1-score:* 0.7069943**

Random Forest

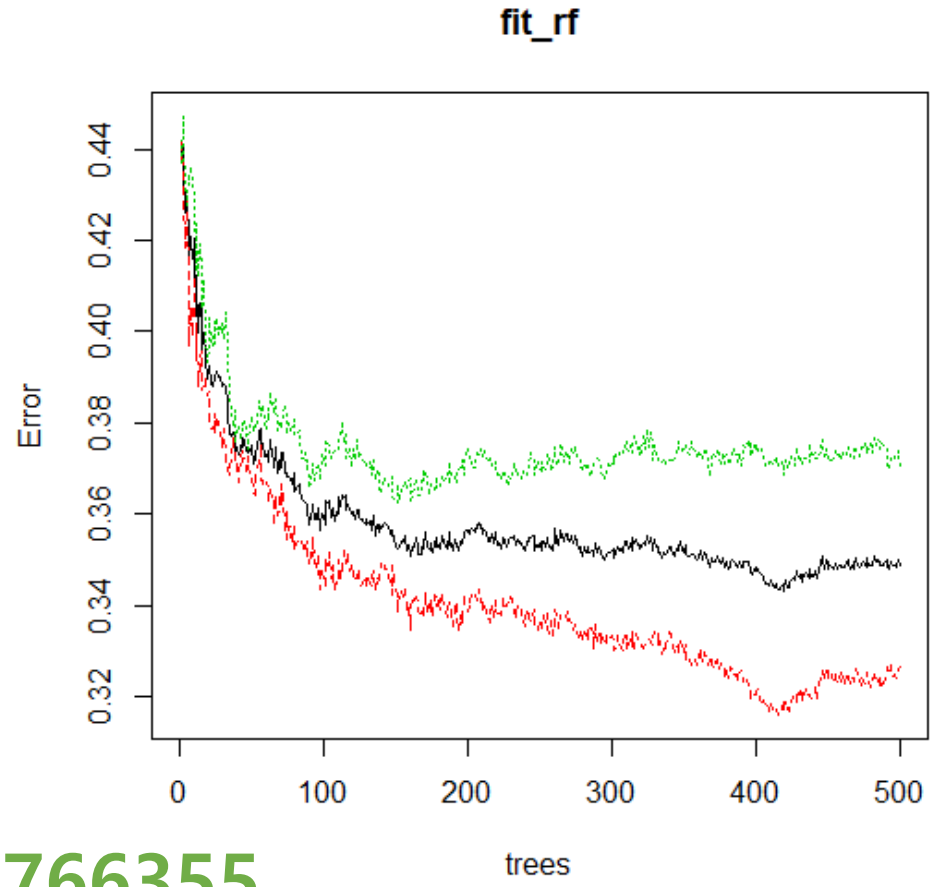


Random Forest

Number of trees(ntree): 500
No. of variables at each split(mtry): 12

Obs: 5000개 중 NA가 없는 4926개 사용됨

predictions_rf	0	1
0	1740	816
1	741	1629



***F1-score:* 0.6766355**

다음주 예고

Anomaly detection / Change detection

Penalized models

Penalized SVM

Penalized LDA

THANK YOU

 # 신용불량자