깔삼한 3조 김나현 김준혁 이성희 허승민

깔쌈하게 대출받자!

AGENDA

1. 분석 목적 및 배경
2. 데이터 이해 및 시각화
3. 데이터 전처리 및 통합

분석 목적 및 배경

분석 배경

Home Credit의 상환 능력 예측 모델의 필요성

HOME CREDIT

OUR MISSION

We focus on responsible lending primarily to people with little or no credit history. Our services are simple, easy and fast.

홈 크레딧은 과거의 신용 정보가 불충분 하거나 존재하지 않아 은행에서 대출 서비스 이용이 어려운 고객들을 대상으로 서비스를 제공

예측목적

WHY? 홈 크레딧 고객에게 긍정적인 대출 경험을 제공

HOW? 신용이 낮은 고객들의 자사 데이터 신용평가 크레딧 뷰로의 데이터

WHAT? 고객들의 상환 능력을 예측

데이터 소개 – OWN DATA

Home Credit에 대출을 신청한 고객들의 APPLICATION 과 PREVIOUS APPLICATION

APPLICATION DATA

- TARGET 변수를 포함한 메인 테이블
- 대출 당시, 대출에 대한 정보와
  대출 신청자에 대한 개인적, 지역적인 정보
- 대출 신청자에 대한 외부 신용평가자료를
  포함

PREVIOUS_APPLICATION

- 메인테이블에 있는 샘플이 이전에 HOME CREDIT에 대출한 정보
- 하나의 행은 과거의 하나의 대출 상품을 의미
- 하나의 샘플이 여러 개의 과거 대출 기록이
  존재.

데이터 소개 – OWN DATA

Home Credit에 대출을 신청한 고객들의 APPLICATION 과 PREVIOUS APPLICATION

POS_CASH_INSTALLMENTS

- 샘플들의 과거 POS loan과 cash loan에 대한 월별 잔금
- 하나의 행은 샘플들과 관련된 과거 Credit에 대한 정보

CREDIT_CRAD_BALANCE

- 샘플들의 과거 credit cards loan에 대한 정보
- 하나의 행은 샘플들과 관련된 과거 Credit에 대한 정보

INSTALLMENTS_PAYMENTS

- 샘플들의 과거의 대출에 대한 상환 이력
- 하나의 행은 매달 상환 이력을 의미

데이터 소개 – Credit Bureau

Home Credit에 대출을 신청한 고객들의 APPLICATION 과 PREVIOUS APPLICATION

BUREAU

- HOME CREDIT이 아닌 외부 신용평가사가 sample을 평가한 자료
- 하나의 행은 외부 신용평가사의 고객의 하나의대출을 의미

BUREAU_BALANCE

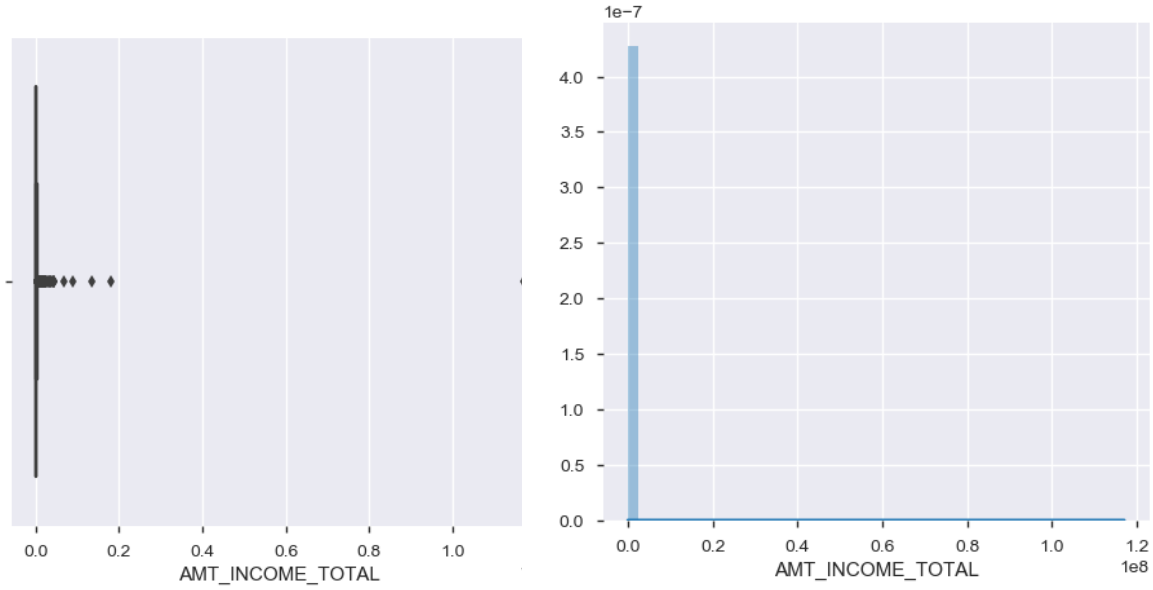
- 외부 신용평가사의 이전 credit에 대한 월별 잔액
- 하나의 행은 sample의 각각의 대출에 대해 Credit을 의미

데이터 이해 및 시각화

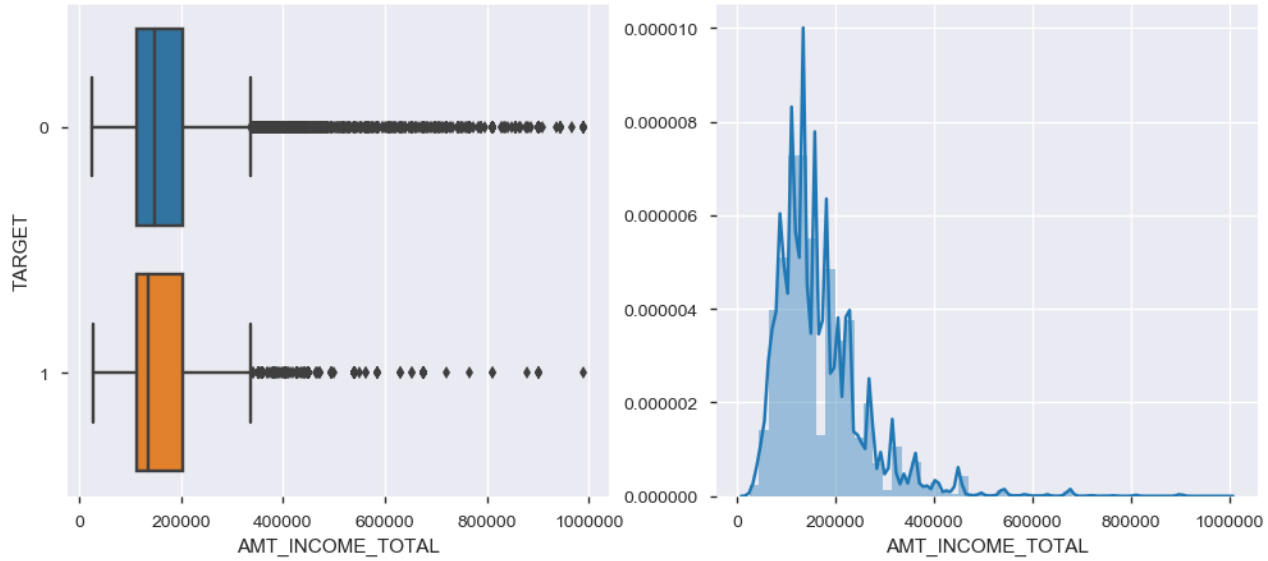
APPLICATION DATA

변수 판별을 위한 데이터 시각화

ANT_INCOME_TOTAL



RAW DATA

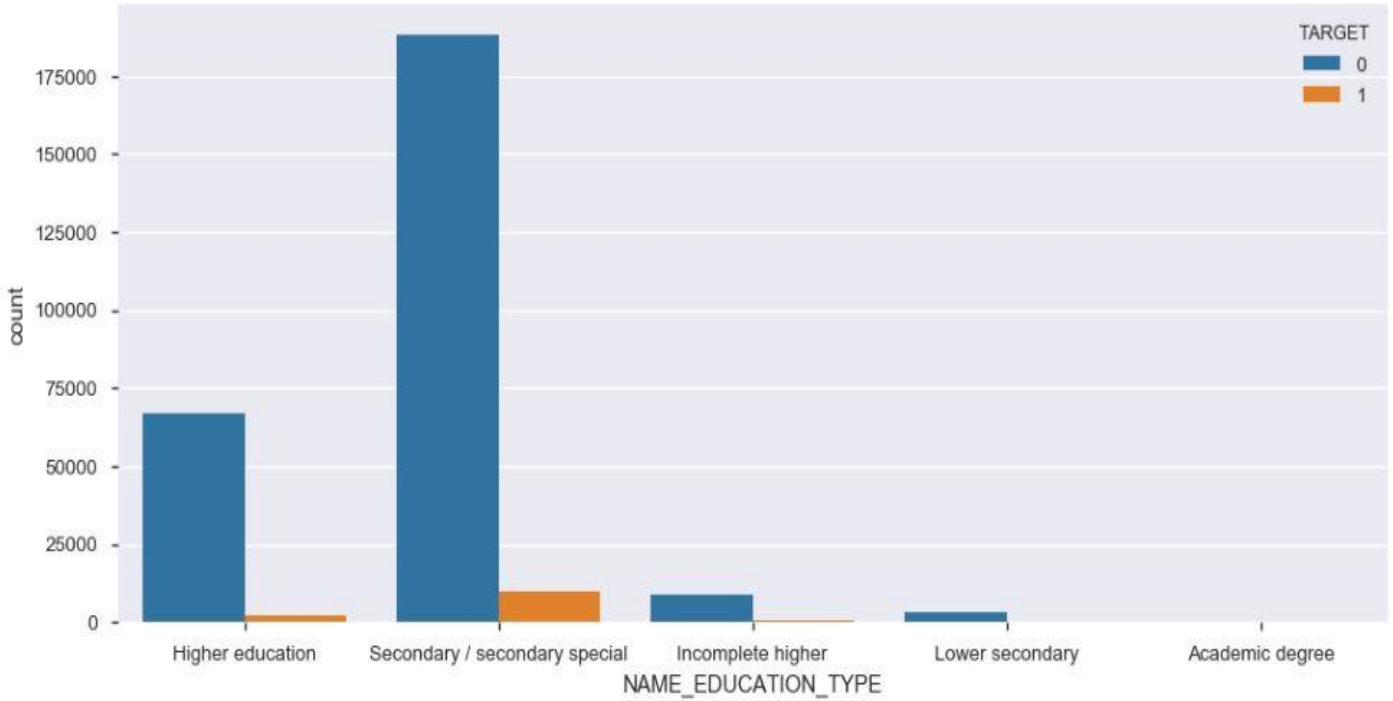


REMOVING OUTLIERS

APPLICATION DATA

변수 판별을 위한 데이터 시각화

NAME_EDUCATION_TYPE




```
NAME_EDUCATION_TYPE
Secondary / secondary special    5.164748
Higher education                 3.040878
Incomplete higher                4.719631
Lower secondary                  6.811820
Academic degree                  0.645161
Name: NAME_EDUCATION_TYPE, dtype: float64
```

고등 교육을 받은 고객들은 상환 지연을 하지 않는 경향성을 보였다.

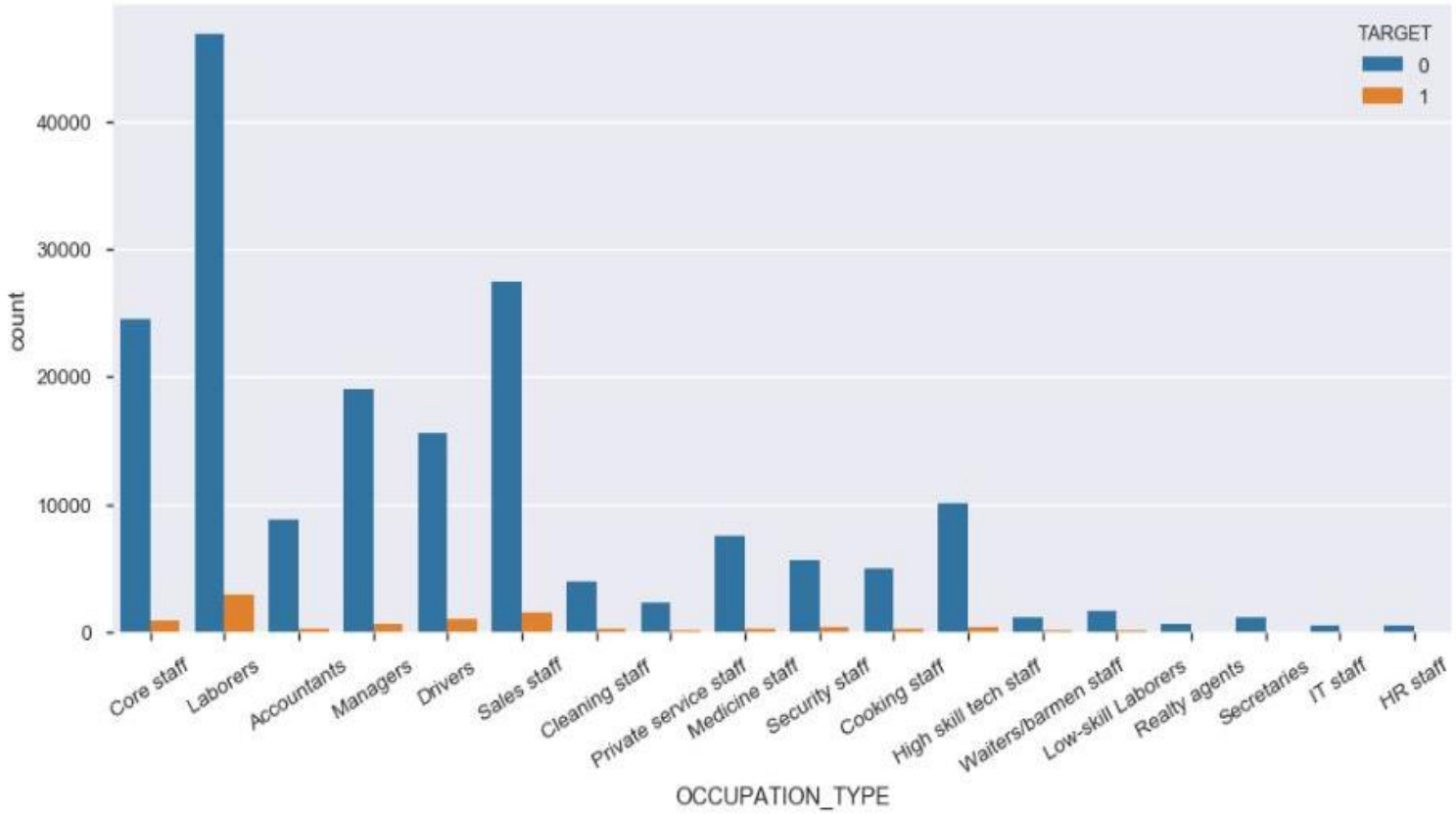
APPLICATION DATA

변수 판별을 위한 데이터 시각화

OCCUPATION_TYPE

OCCUPATION_TYPE	
Accountants	2.649757
Cleaning staff	5.392526
Cooking staff	6.536857
Core staff	3.646025
Drivers	6.620990
HR staff	3.807615
High skill tech staff	3.451004
IT staff	3.648069
Laborers	6.263485
Low-skill Laborers	10.292326
Managers	3.352777
Medicine staff	3.818206
Private service staff	3.594080
Realty agents	4.984424
Sales staff	5.802708
Secretaries	4.467354
Security staff	6.340261
Waiters/barmen staff	6.868867

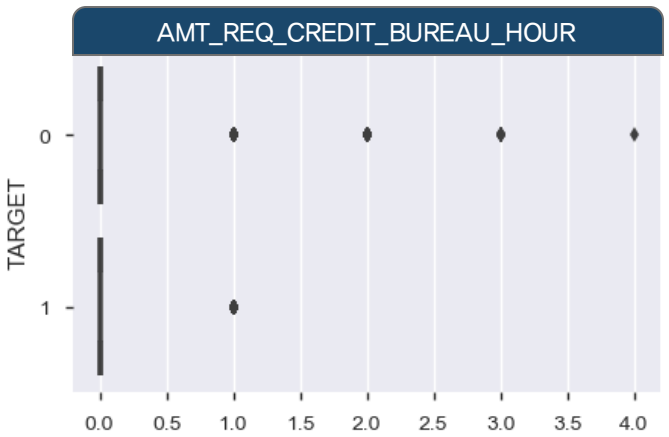
Name: OCCUPATION_TYPE, dtype: float64

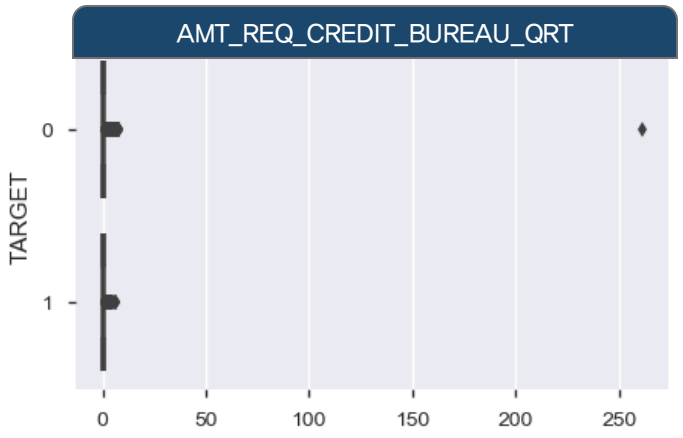


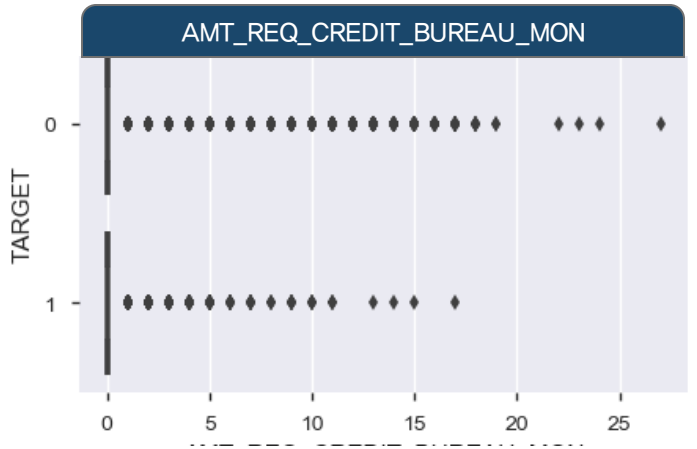
APPLICATION DATA

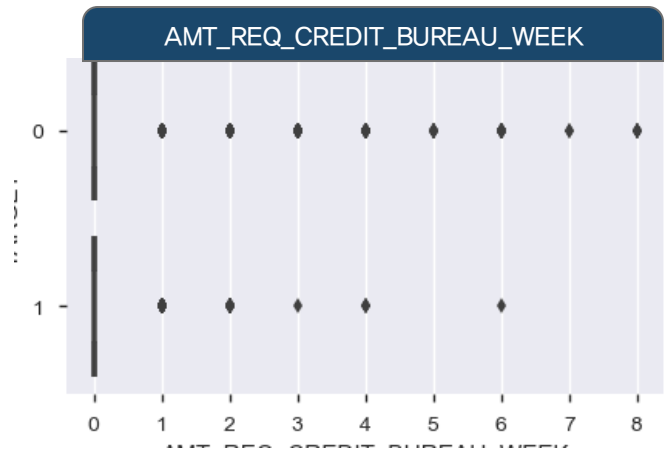
변수 판별을 위한 데이터 시각화

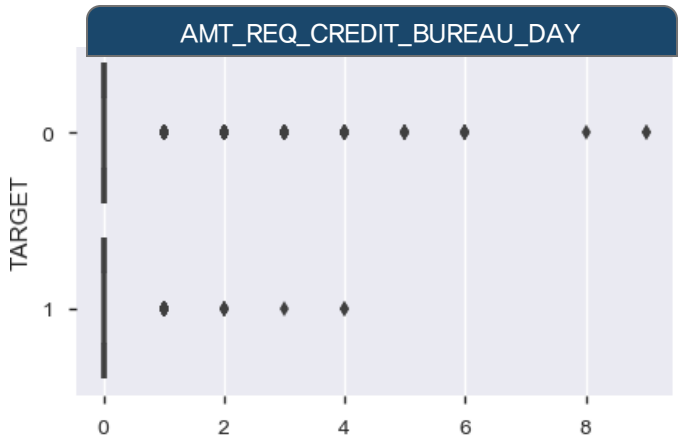
AMT_REQ_CREDIT_BUREAU

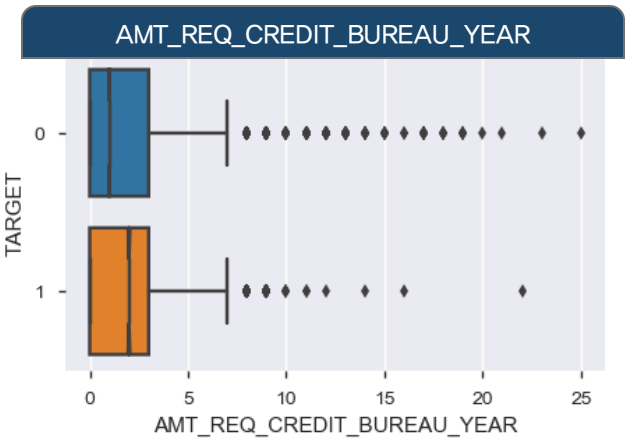










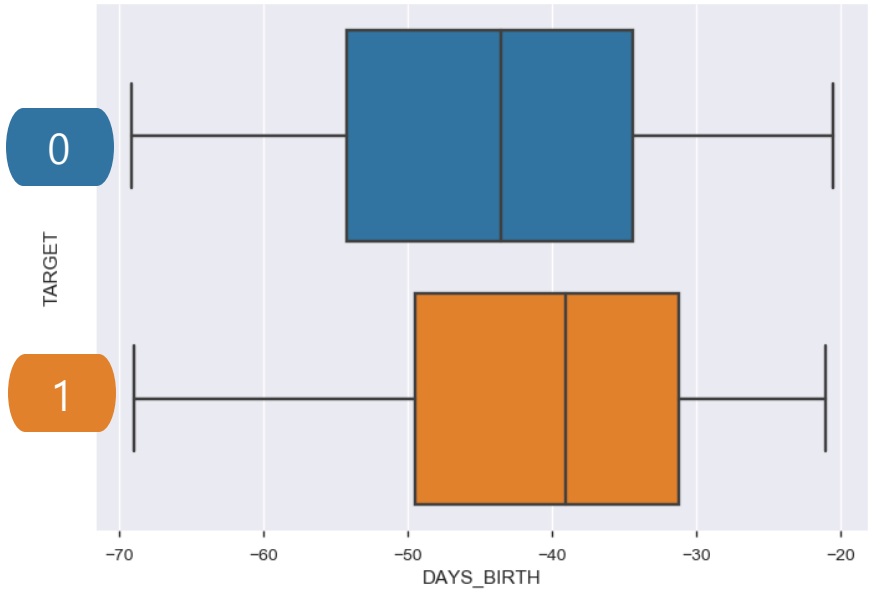


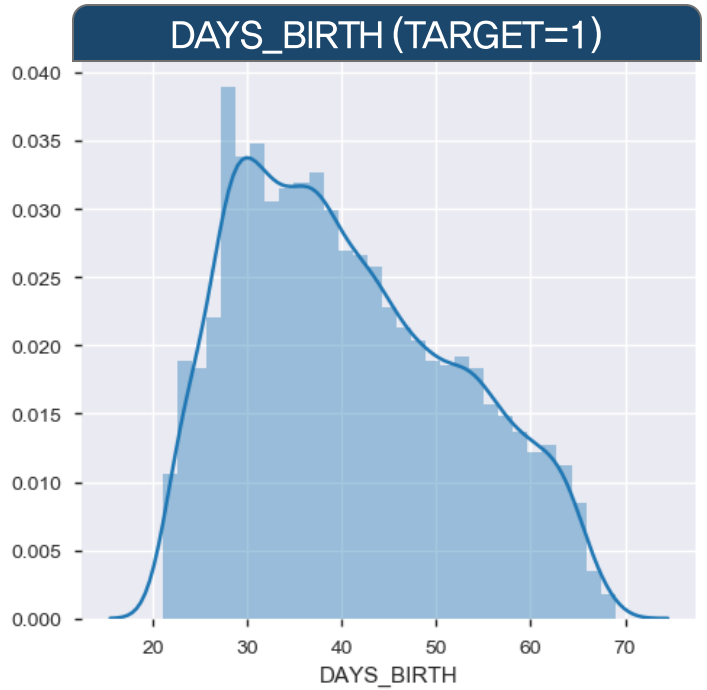
Credit Bureau에게 HOME CREDIT이 문의 횟수 데이터를 확인하였을 때 3달 이후부터 1년 이전까지 고객 정보만이 비교가 가능하여 주어진 기간 별 문의 횟수를 모두 합쳐 하나의 데이터로 만들었다.

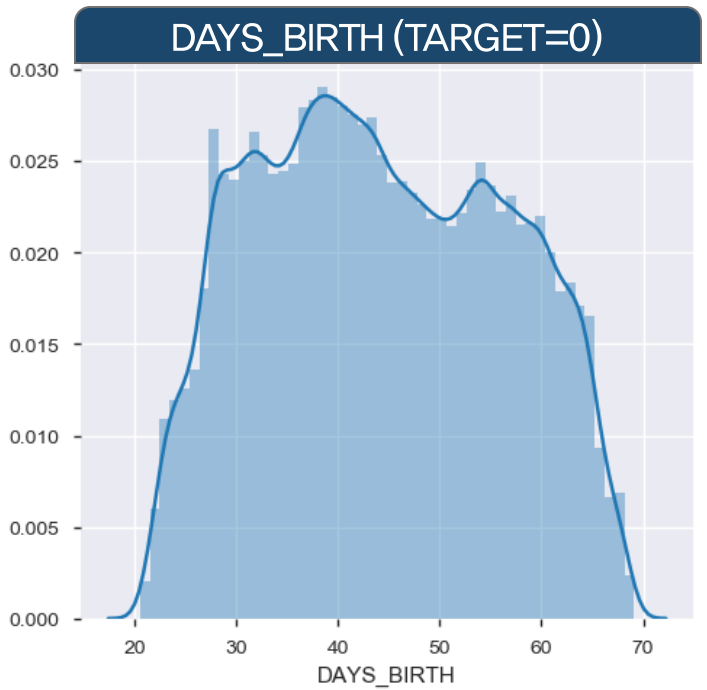
APPLICATION DATA

변수 판별을 위한 데이터 시각화

DAYS_BIRTH







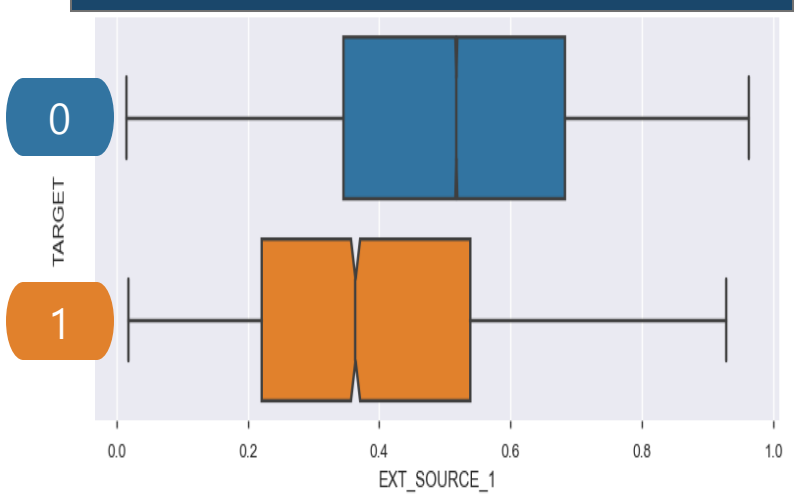
DAYS_BIRTH가 DAY 단위로 되어 있어 숫자가 크고 범위가 넓어 이를 연도로 변환한 뒤 다시 범주화 시켜 나누었다.
이때, 범주는 좌측 히스토그램을 참고하여 QUANTILE을 열 개 구간으로 나누었다. (DAYS_EMPLOYED 등도 유사 변환)

APPLICATION DATA

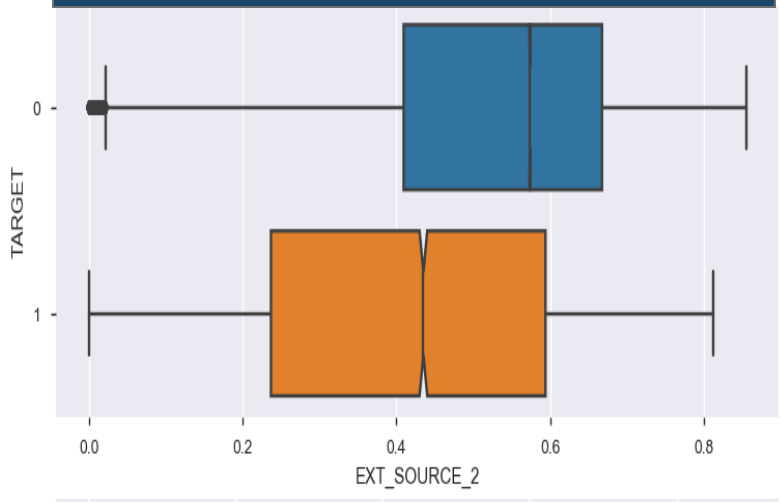
변수 판별을 위한 데이터 시각화

EXT_SOURCE

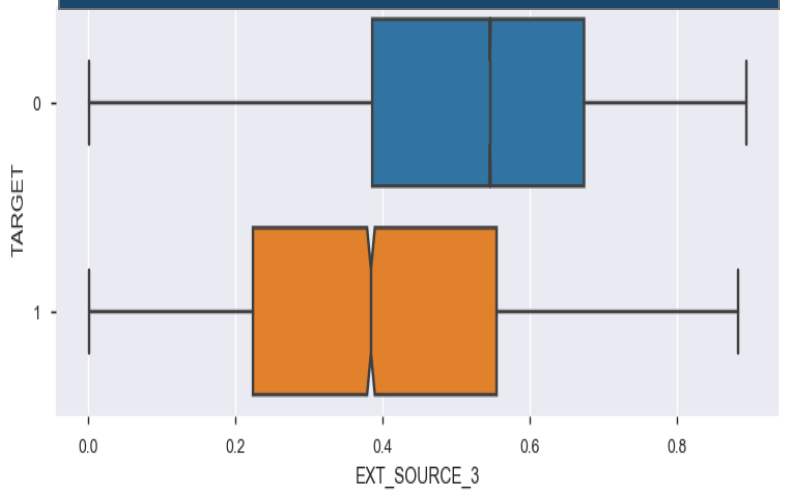
EXT_SOURCE_1



EXT_SOURCE_2



EXT_SOURCE_3

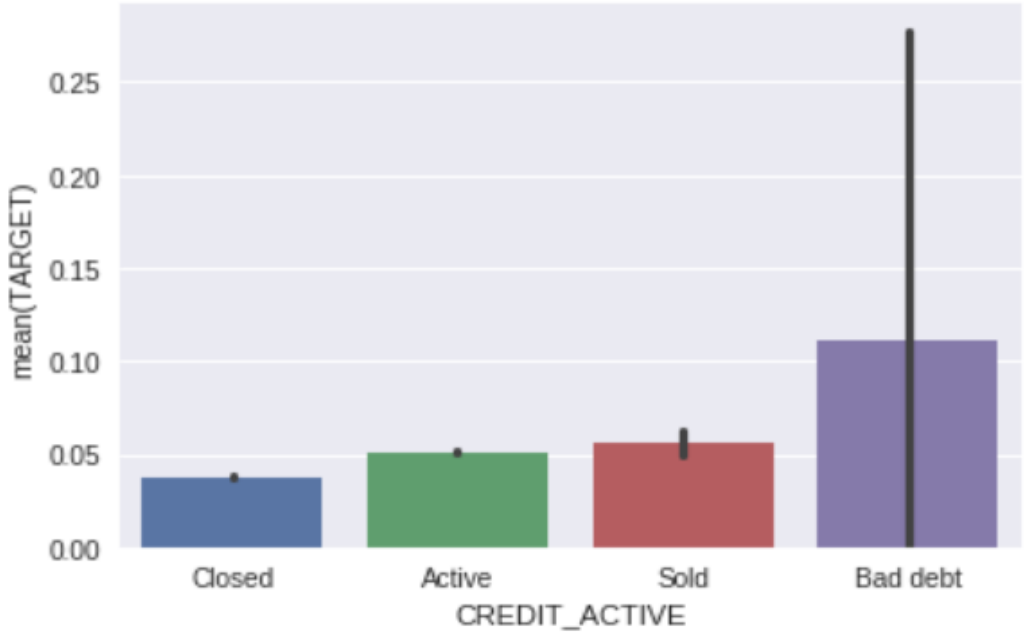


모든 외부 정보는 TARGET 데이터와 상관관계를 가진다.

BUREAU

변수 판별을 위한 데이터 시각화

CREDIT_ACTIVE

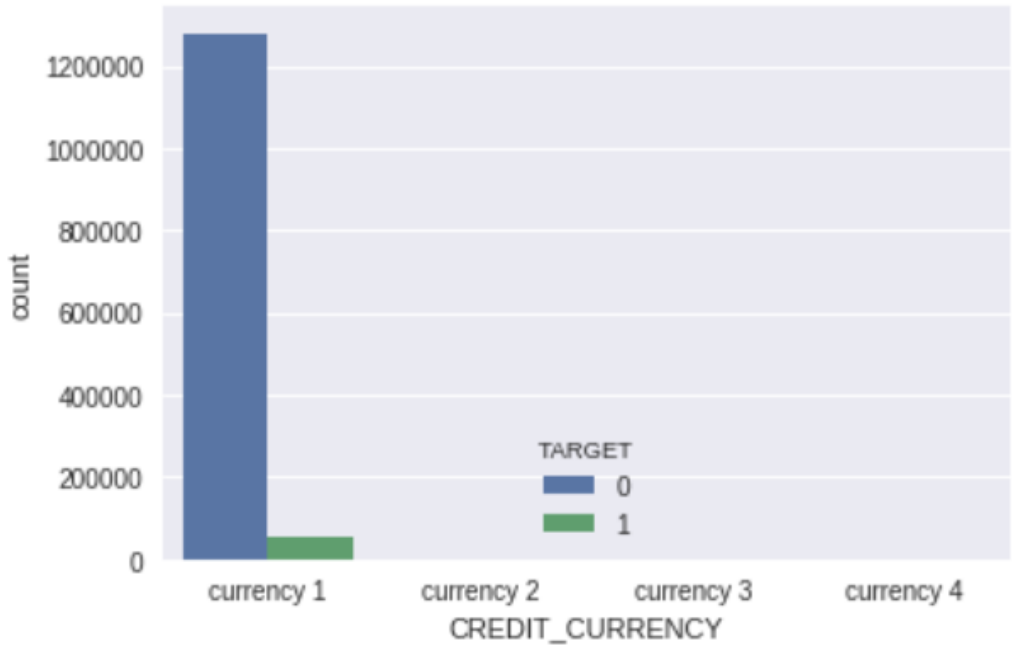


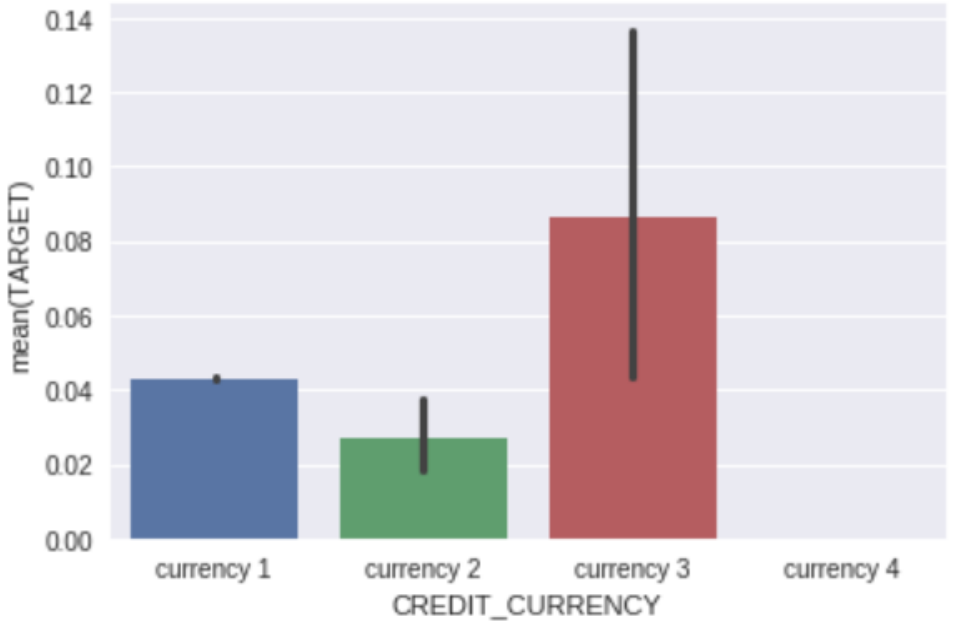
TARGET 데이터 별 CREDIT_ACTIVE에서는 유의미한 결과를 찾을 수 없었지만
CREDIT_ACTIVE 별로 TARGET = 1 의비율을 보았을 때 Bad Debt과 Sold에서 높은 값이 도출되었다.

BUREAU

변수 판별을 위한 데이터 시각화

CREDIT_CURRENCY



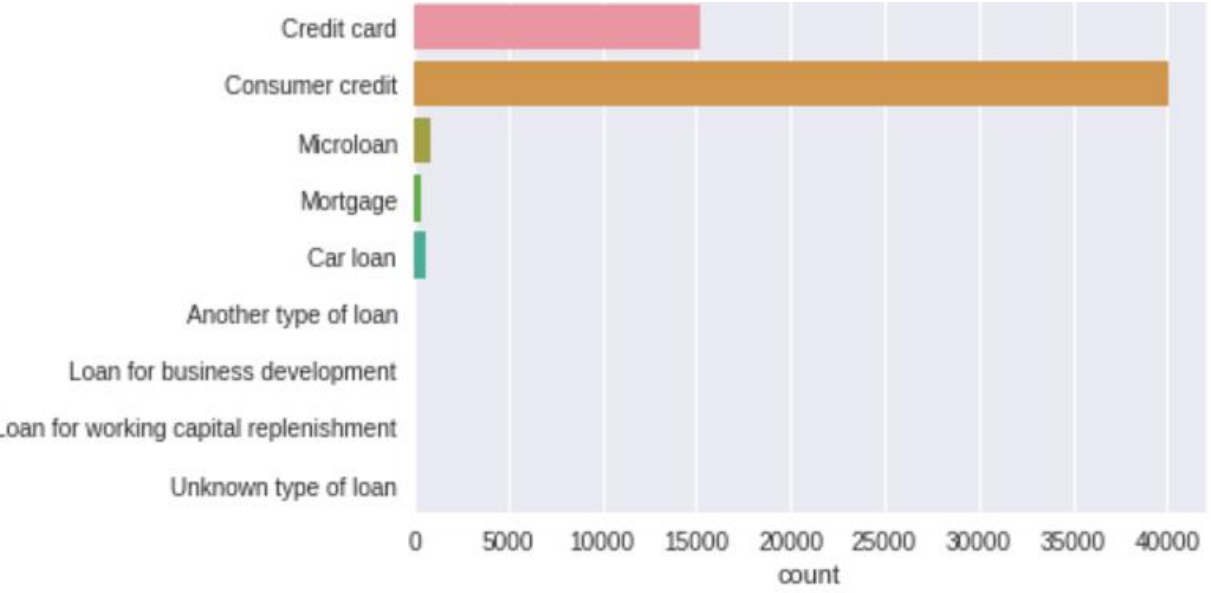


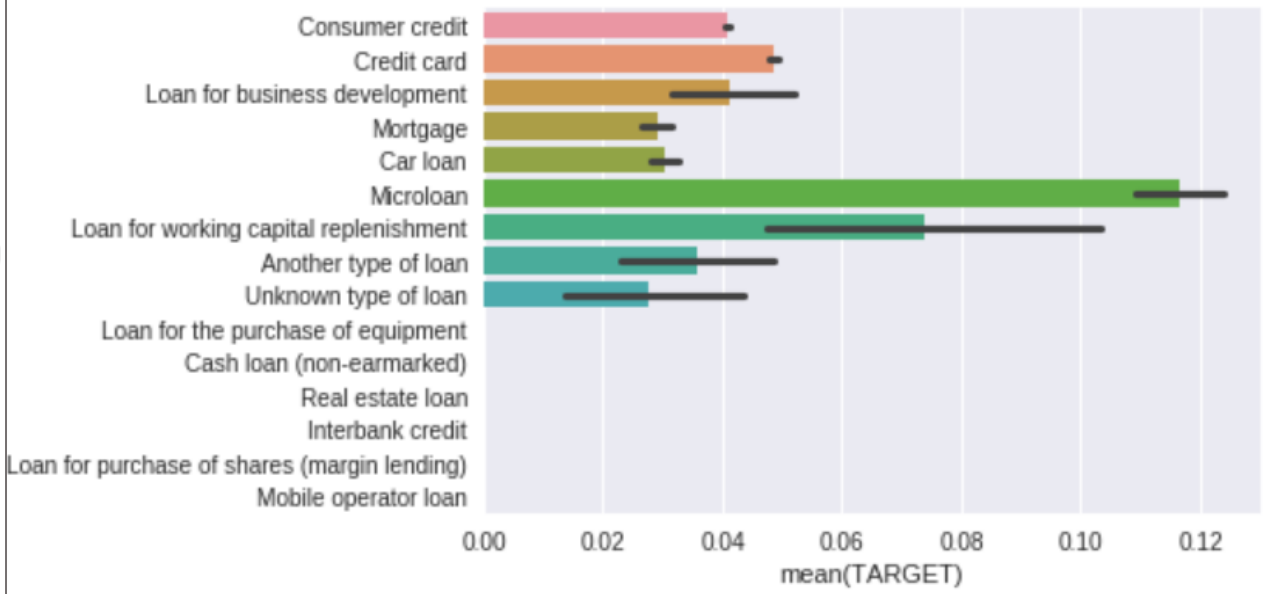
전체 데이터를 확인 하였을 때는 CURRENCY 1 이 TARGET 1, 0 모두에서 높은 값을 차지 하고 있었다.
CREDIT_CURRENCY 별로 TARGET = 1 의 비율을 보았을 때 CURRENCY 3 일 때 TARGET 1의 확률이 높은 것으로 확인 되었다.

BUREAU

변수 판별을 위한 데이터 시각화

CREDIT_TYPE



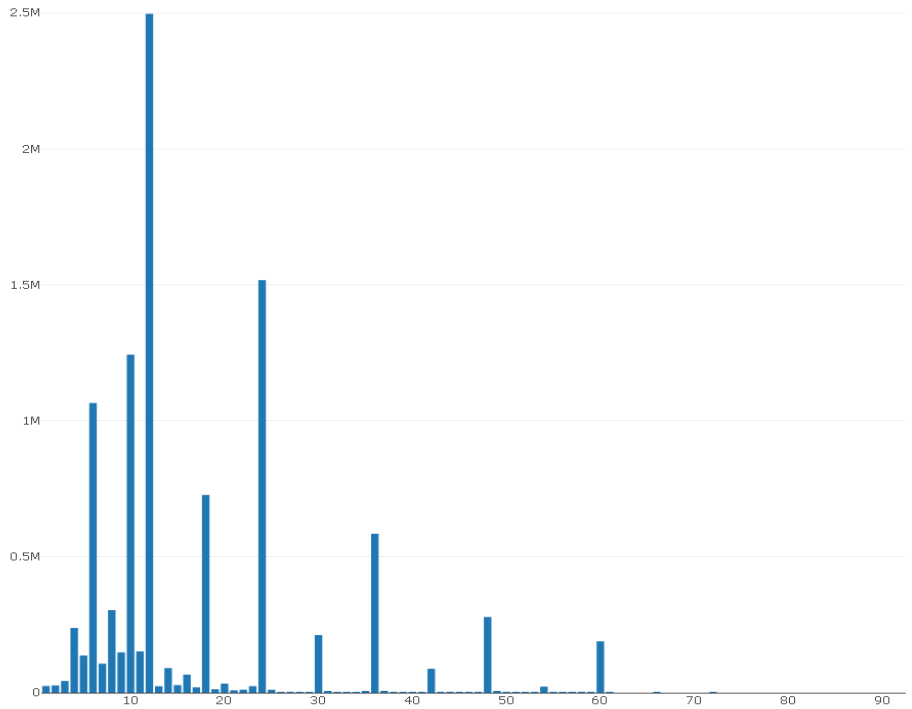


가장 높은 비율의 CREDIT CARD와 CONSUMER CREDIT 이나
CREDIT_TYPE 별로 TARGET = 1 의 비율을 보았을 때 MICROLOAN 의 비율이 가장 높았다.

POS_CASH

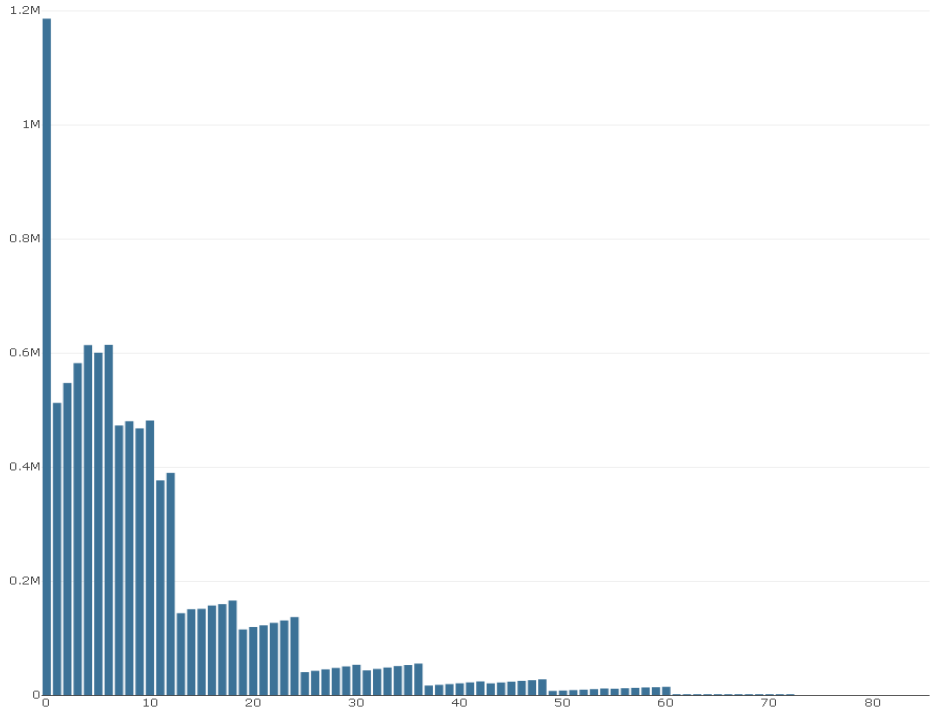
변수 판별을 위한 데이터 시각화

CNT_INSTALLMENT



계약서에 명시된 상환 기간을 의미한다. (단위 : 월)

CNT_INSTALLMENT_FUTURE

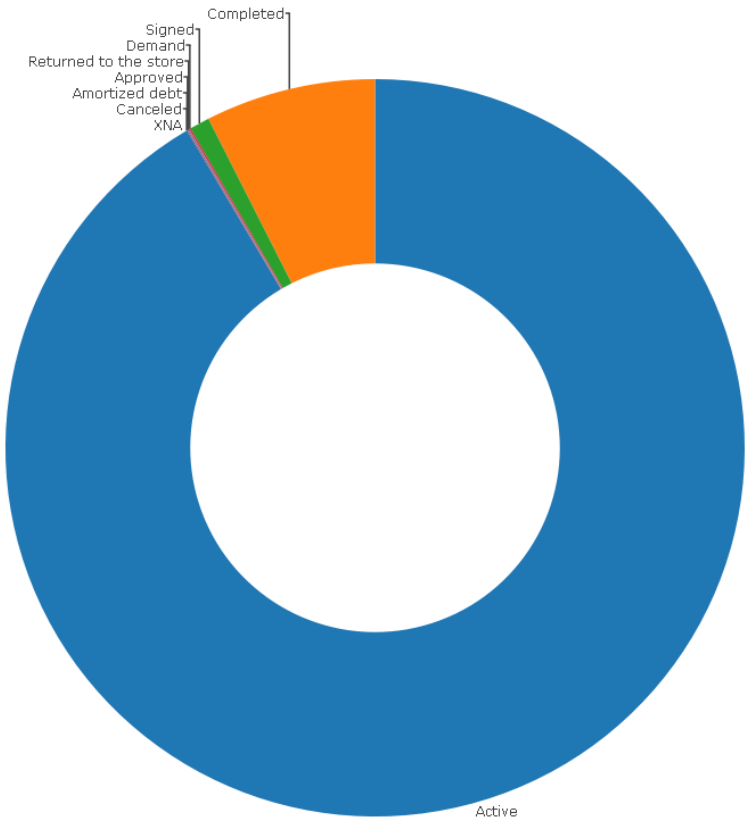


신청 당시를 기준으로 계약서를 기준으로
남은 상환 기간을 의미한다. (단위 : 월)

POS_CASH

변수 판별을 위한 데이터 시각화

NAME_CONTRACT_STATUS



- Active
- Completed
- Signed
- Demand
- Returned to the store
- Approved
- Amortized debt
- Canceled
- XNA

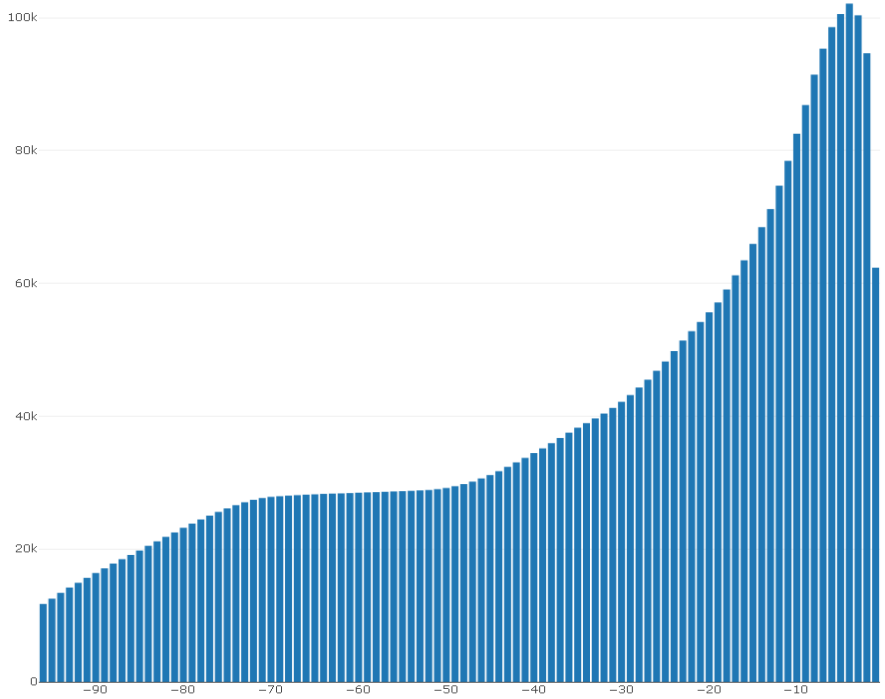
CREDIT_CARD

변수 판별을 위한 데이터 시각화

CNT_DRWAINGS_CURRENT

이전 크레딧에 대한 인출 횟수

MONTH_BALANCE

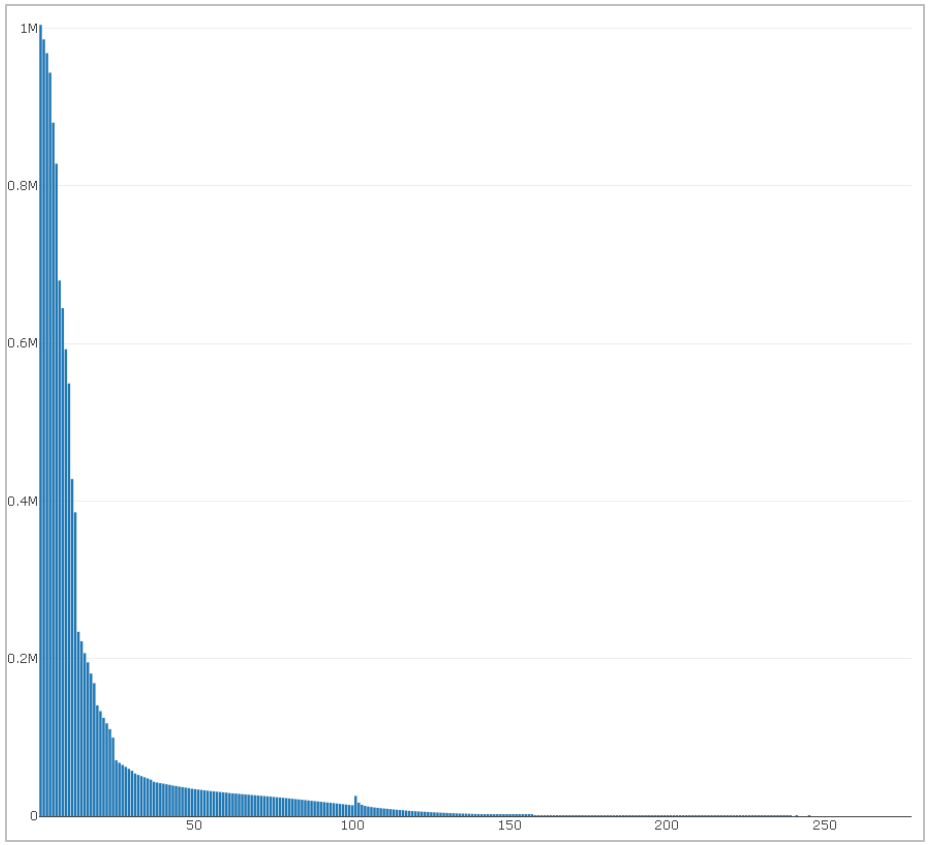


우측으로 갈 수록 , 즉 0에 가까워 질 수록 가장 최신의 CREDIT CARD 잔액 정보를 의미한다.

INSTALLMENTS

변수 판별을 위한 데이터 시각화

NUM_INSTALLMENT_NUMBER

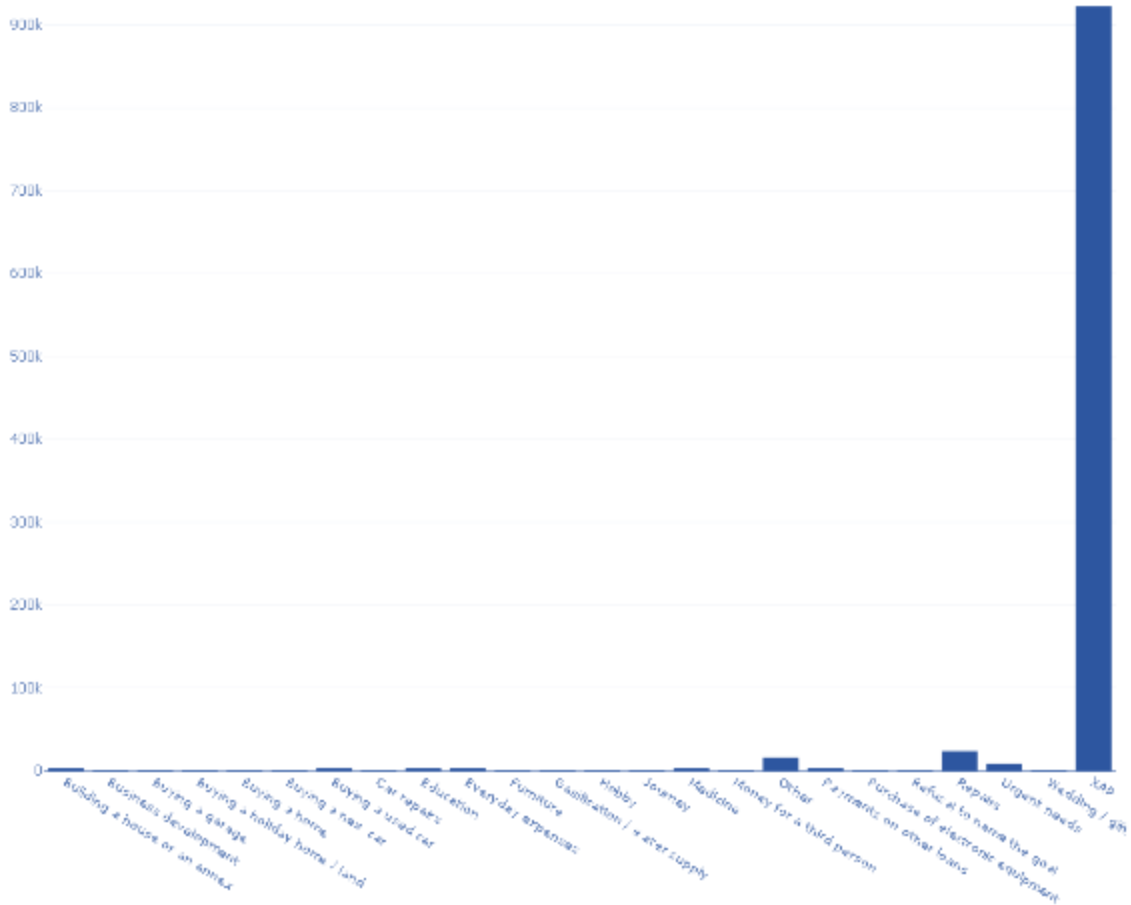


INSTALLMENT 상환 지불 횟수

PREVIOUS_APPLICATION

변수 판별을 위한 데이터 시각화

NAME_CASH_LOAN_PURPOSE

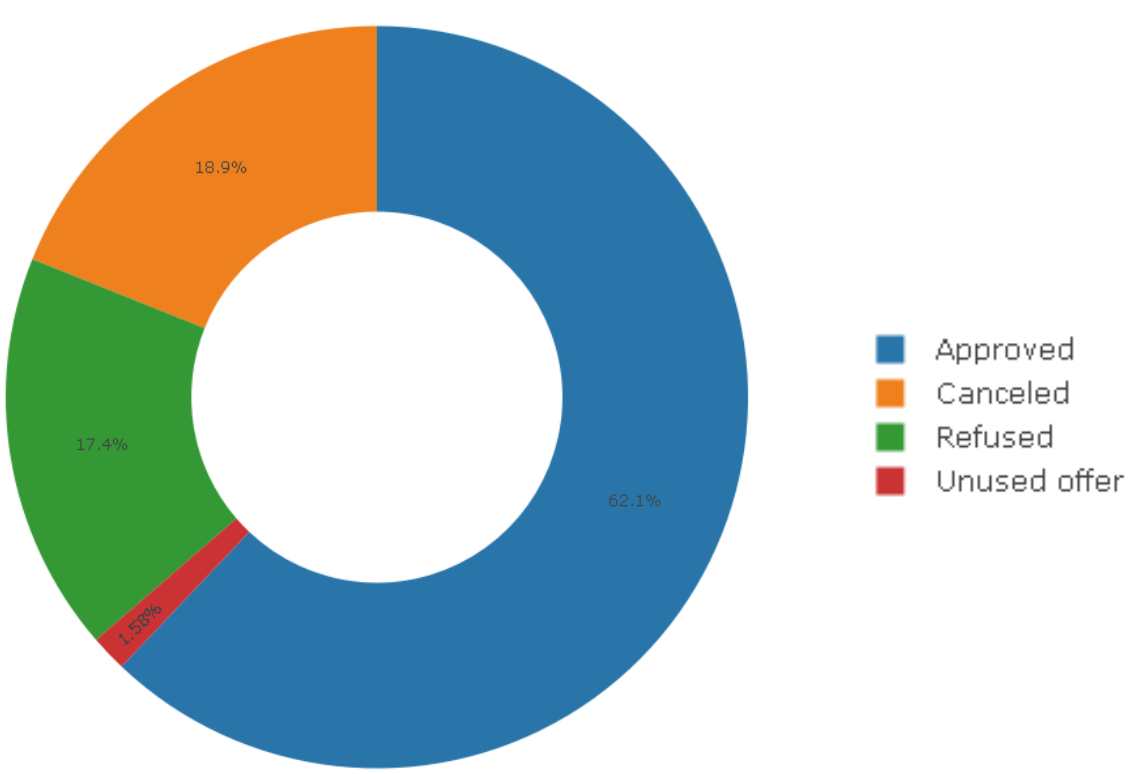


CASH LOAN 에 대한 이유 및 목적 정보
XAP 가 가장 높게 나왔다

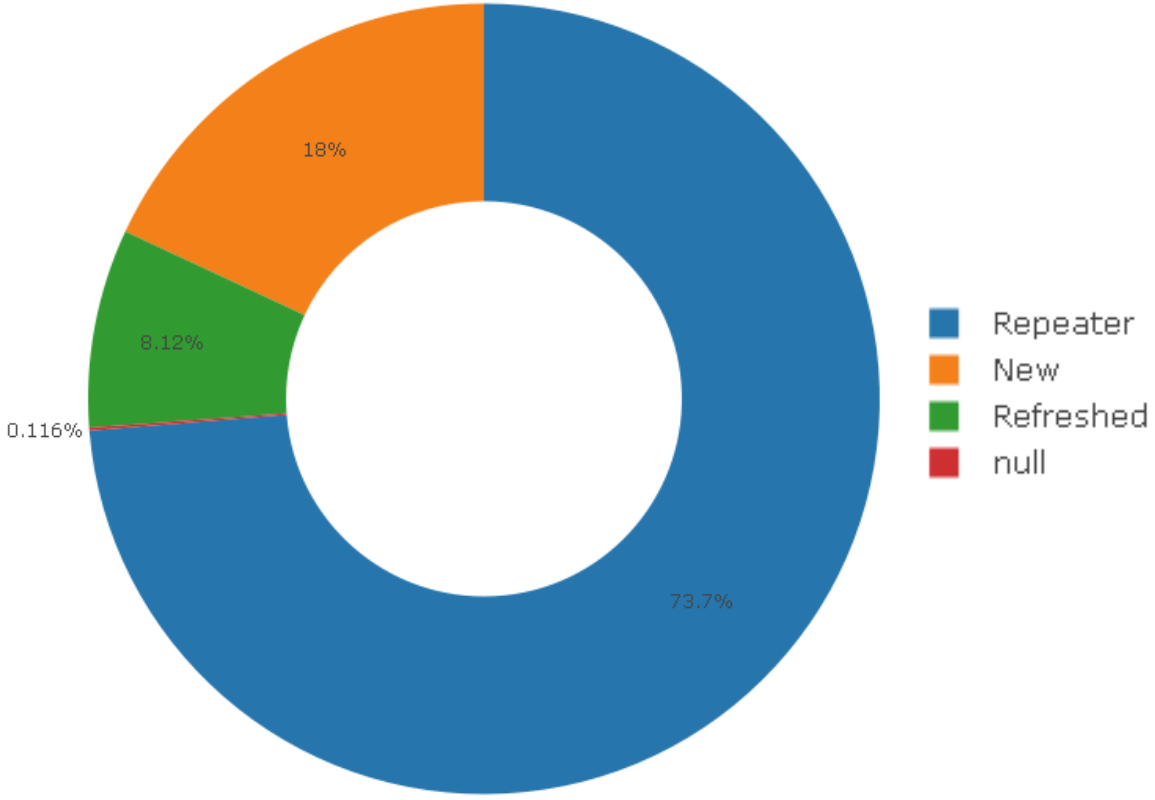
PREVIOUS_APPLICATION

변수 판별을 위한 데이터 시각화

NAME_CONTRACT_STATUS



NAME_CLIENT_TYPE

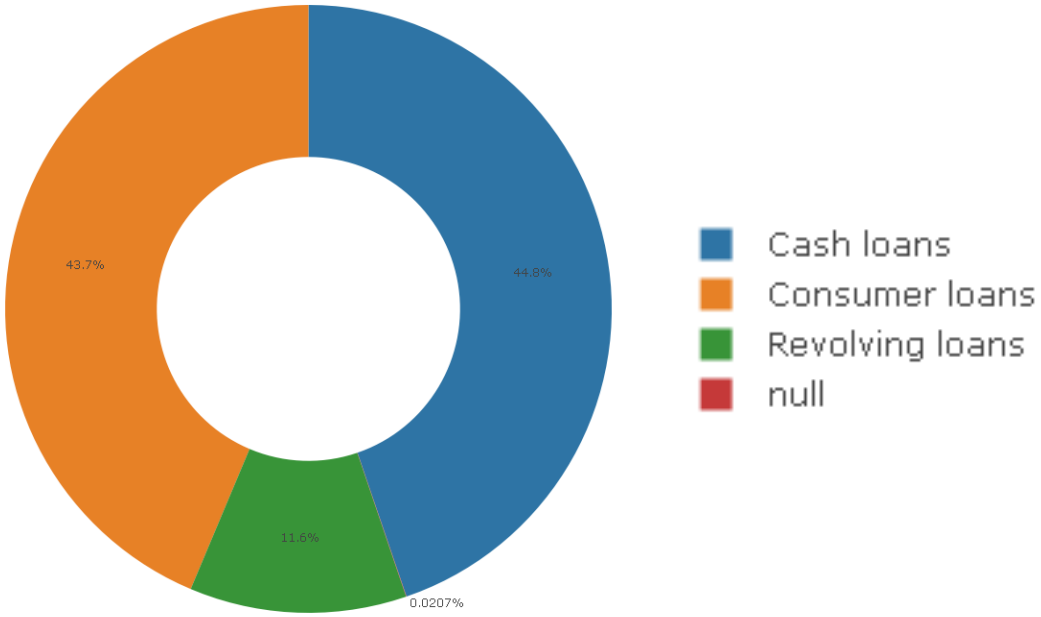


PREVIOUS_APPLICATION

변수 판별을 위한 데이터 시각화

NAME_PAYMENT_TYPE

NAME_CONTRACT_TYPE



Previous application에서 대출 계약 유형에 consumer Loans 이 포함되는데 application의 테스트, 트레인 데이터 모두 우연히도 Cash loans 와 Revolving loans으로만 구성되었다

파생 변수

결측 값이 존재하는 변수처리하기

데이터 파일	변수 명	설명
previous_application.csv	INTEREST_RATE_PRE	AMT_ANNUITY / AMT_CREDIT
previous_application.csv	APP_CREDIT_PERC	AMT_APPLICATION / AMT_CREDIT

INTEREST_RATE_PRE

$$\text{현재가치} = \frac{1-(1+\text{이자율})^{-\text{연금지급년도}}}{\text{이자율}} \times \text{동일금액}$$

고객이 기존에 HOME CREDIT 에서 신청했던 대출에 대한 이자를 의미하며,

이자율은 높을 수록 해당 상품이 위험하다고 간주한다.

APP_CREDIT_PERC

고객이 기존에 HOME CREDIT 에서 신청했던 대출에 대해 요청했던 금액과 실제로

승인된 금액 간의 비율을 의미한다

파생 변수

결측 값이 존재하는 변수처리하기

데이터 파일	변수 명	설명
installments_payments.csv	DAYS_RELATIVE_PAYMENT	DAYSINSTALMENT – DAYSENTRYPAYMENT
installments_payments.csv	PAYMENT_PERC	AMT_PAYMENT / AMT_INSTALMENT

DAYS_RELATIVE_PAYMENT

납부해야 하는 날짜보다 늦게 납부 했을 경우 negative

납부해야 하는 날짜에 납부하거나 더 일찍 납부했을 경우 positive

PAYMENT_PERC

실제로 낸 금액과 내야하는 금액의 비율

파생 변수

결측 값이 존재하는 변수처리하기

데이터 파일	변수 명	설명
bureau.csv	INTEREST_RATE	AMT_ANNUITY / AMT_CREDIT_SUM
bureau.csv	RATIO_DEBT_TO_CREDIT	AMT_CREDIT_SUM_DEBT / AMT_CREDIT_SUM

INTEREST_RATE_PRE

고객이 기존에 에서 신청했던 대출에 대한 이자를 의미하며,

이자율은 높을 수록 해당 상품이 위험하다고 간주한다.

RATIO_DEBT_TO_CREDIT

크레딧 뷰로에 신용도에 비해 빚을 얼마나 지고 있는 지에 대한 비율

데이터 전처리 및 통합

데이터 전처리 방향

결측 값이 존재하는 변수처리하기

행을 삭제하지 않는다.

행 삭제는 실시하지 않았다. TRAIN 데이터의 경우는 괜찮더라도 TEST 데이터의 경우는 없어지면 예측을 못하는 ID가 생기므로 행을 삭제 하지 않았다.

NA가 많지 않은 경우에는 평균과 최빈값을 이용하여 랜덤 샘플링 하였다.

ex) AMT_REQ_CREDIT_BUREAU 의 경우 변수를 통합하였고 NA가 13% 정도 되었지만, Hinge Spread 범위 내(0~7)에서 랜덤 샘플링하여 넣음.

XNA 범주에 대해서는 Kaggle Discussion에서 출제자의 가이드로 NA 처리하였다.

데이터 전처리 방향

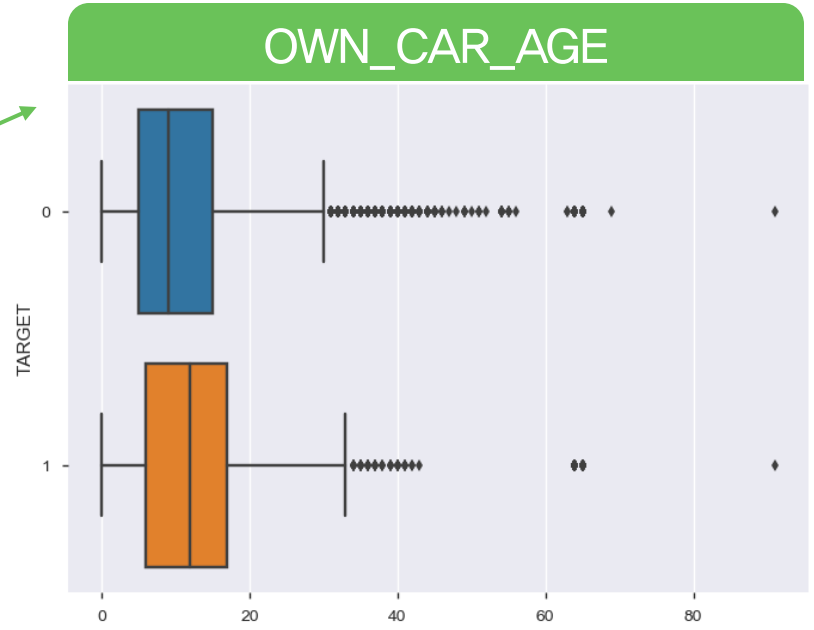
결측값이 존재하는 변수처리하기

Na가 50%를 넘으면 Column 삭제를 원칙으로 하나,

TARGET 데이터와 관련이 있을 경우 변수를 삭제하지 않는다

ex) OWN_CAR_AGE : NA가 65% 정도의 비율로 굉장히 높지만 TARGET과 관련이 있다고 판단하여
NA를 범주화시키고(XNA) 수치형 자료도 일정 구간으로 범주화 시켰다.

	count_NA	ratio_NA
SK_ID_CURR	0	0.000
TARGET	0	0.000
CNT_CHILDREN	0	0.000
AMT_INCOME_TOTAL	0	0.000
AMT_CREDIT	0	0.000
AMT_ANNUITY	12	0.004
AMT_GOODS_PRICE	254	0.091
REGION_POPULATION_RELATIVE	0	0.000
DAYS_BIRTH	0	0.000
DAYS_EMPLOYED	0	0.000
DAYS_REGISTRATION	0	0.000
DAYS_ID_PUBLISH	0	0.000
OWN_CAR_AGE	184561	65.866
FLAG_MOBIL	0	0.000
FLAG_EMP_PHONE	0	0.000
FLAG_WORK_PHONE	0	0.000
REG_REGION_NOT_WORK_REGION	0	0.000
LIVE_REGION_NOT_WORK_REGION	0	0.000
REG_CITY_NOT_LIVE_CITY	0	0.000
REG_CITY_NOT_WORK_CITY	0	0.000
LIVE_CITY_NOT_WORK_CITY	0	0.000
EXT_SOURCE_1	157655	56.264
EXT_SOURCE_2	599	0.214
EXT_SOURCE_3	55268	19.724
APARTMENTS_AVG	141388	50.459
BASEMENTAREA_AVG	163267	58.267
YEARS_BEGINEXPLUATATION_AVG	135911	48.504
YEARS_BUILD_AVG	185757	66.293
COMMONAREA_AVG	195289	69.695
ELEVATORS_AVG	148566	53.020
ENTRANCES_AVG	140266	50.058
FLOORSMAX_AVG	138624	49.472
FLOORSMIN_AVG	189582	67.658



데이터 전처리 방향

결측값이 존재하는 변수처리하기

Na가 50%를 넘으면 Column 삭제를 원칙으로 하나,

TARGET 데이터와 관련이 있을 경우 변수를 삭제하지 않는다

EXT_SOURCE_1	157655	56.264
EXT_SOURCE_2	599	0.214
EXT_SOURCE_3	55268	19.724

Ex) EXT_SOURCE : 타 변수들을 이용해

Regression 을 돌려 NA 값을 채워 넣었다.

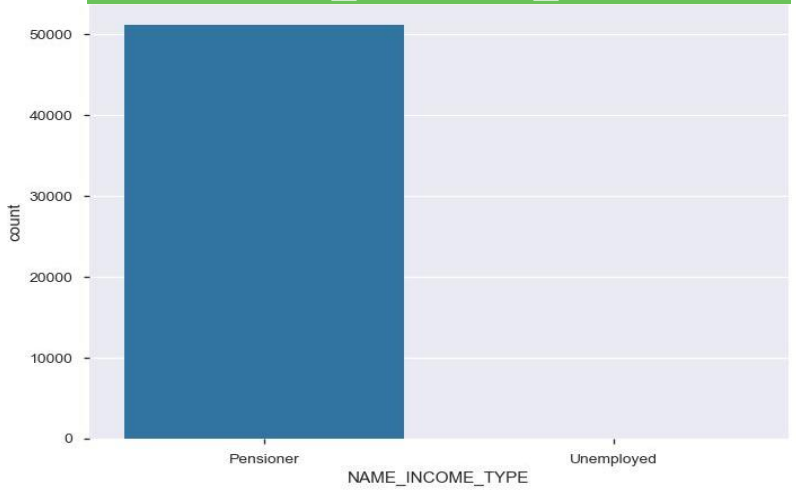
데이터 전처리 방향

결측 값이 존재하는 변수처리하기

TARGET과의 관계 외에도 NA자체가 MNAR 상태라면, 하나의 범주로 해석하였다

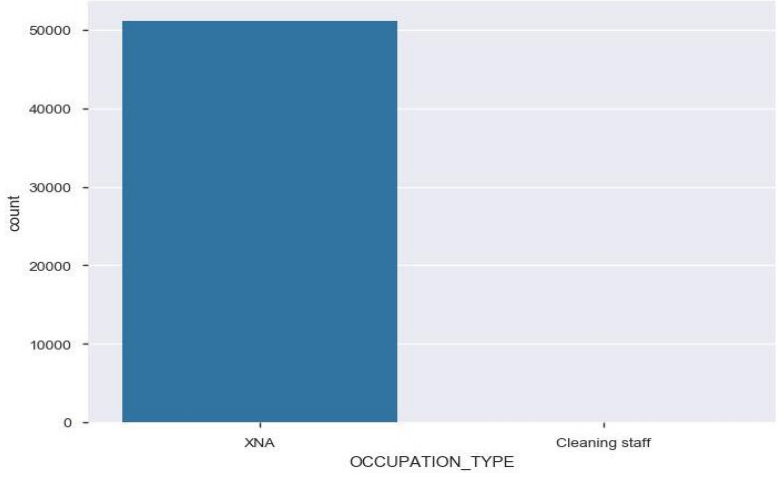
ex) DAY_EMPLOYED의 경우 고용 기간이 1000년에 해당하는 자료가 존재하는데 이는 NA라고 간주해야 한다.
하지만 이러한 값들은 실제로 NAME_INCOME_TYPE에서 PENSIONER로 확인되며, 따라 OCCUPATION_TYPE 에서도 NA라고 판단된다.
즉 이러한 NA들은 아무런 의미가 없는 것이 아니라 특정한 의미를 지니기 때문에 이를 NA로 범주화 시켰다

NAME_INCOME_TYPE



Pensioner	51084
Unemployed	17

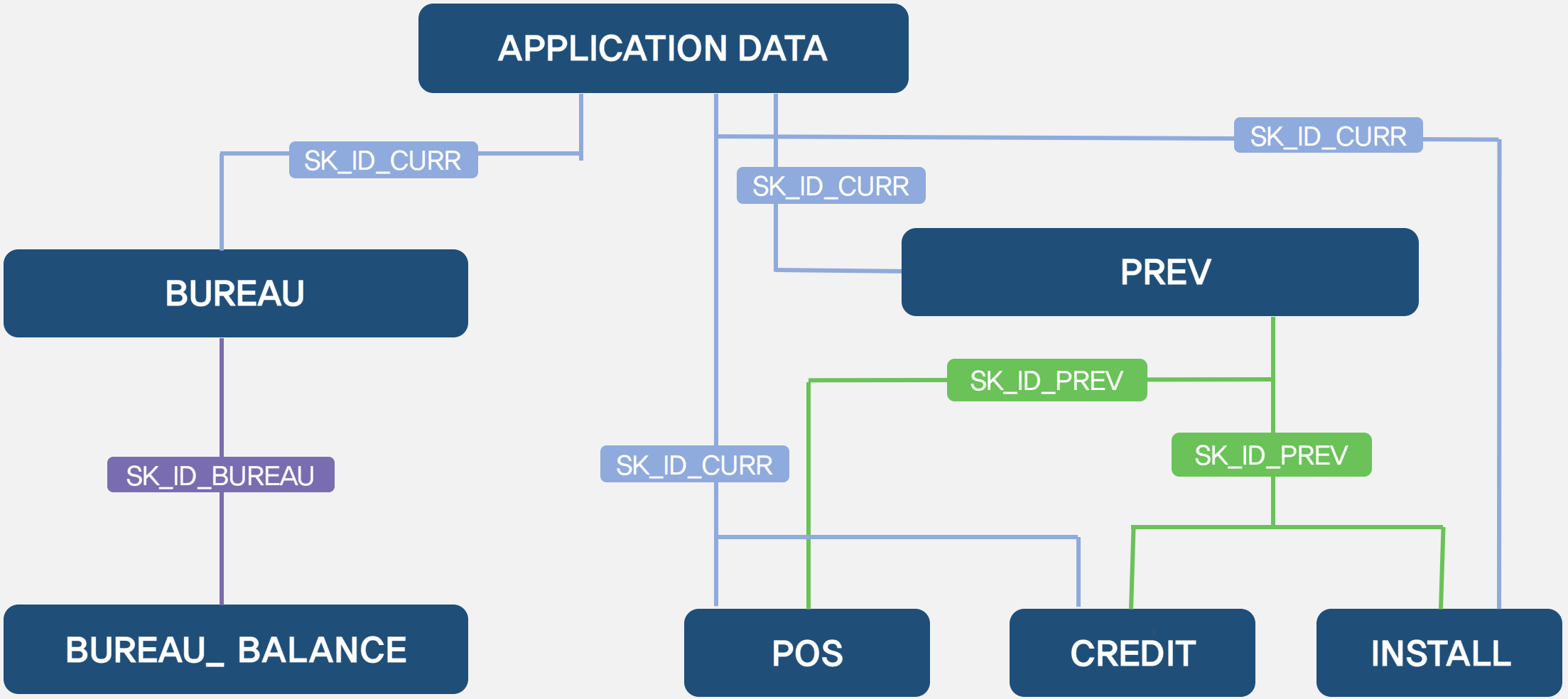
OCCUPATION_TYPE



XNA	51099
Cleaning staff	2

데이터 통합

로드맵


APPLICATION DATA
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
BUREAU
PREV
SK_ID_PREV
SK_ID_PREV
SK_ID_BUREAU
SK_ID_CURR
BUREAU_ BALANCE
POS
CREDIT
INSTALL

데이터 통합

bureau_balance 테이블을 bureau 에 통합

1 BUREAU_BALANCE

SK_ID_BUREAU	MONTHS_BALANCE	DAYS_CREDIT
5715448	0	C
5715448	–1	–1
5715448	–2	X

SK_ID_BUREAU	MONTHS_TOTAL	DAYS_CREDIT_MEAN
5715448	3	–0.33

MONTHS_TOTAL이라는 총 빈도수를 나타내는 변수로 생성
DAYS_CREDIT에서 C,X,0은 0으로 취급하여 평균을 냄.

데이터 통합

bureau_balance 테이블을 bureau 에 통합

2 BUREAU_BALANCE 와 BUREAU 조인

SK_ID_CURR	SK_ID_BUREAU	CREDIT_CURRENCY	DAYS_CREDIT	MONTHS_TOTAL
215354	5714462	Currency 1	-497	3
215354	5714463	Currency 2	-208	20
215354	5715448	Currency 1	-203	33

SK_ID_CURR	CREDIT_CURRENCY	DAYS_CREDIT	MONTHS_TOTAL
215354	Currency 1	-385	18

Bureau에 Bureau_Balance를 조인하였다.
이때 범주형은 최빈값, 수치형은 평균값으로 설정

데이터 통합

Application 테이블에 previous_application에 통합

3 PREVIOUS_APP

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_APPLICATION
113582	232145	Cash Loan	11000
113583	232145	Cash Loan	84000
113584	232145	Consumer Loan	50500

SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_APPLICATION
232145	Cash Loan	48500

범주형은 최빈값, 수치형은 평균값으로 설정

데이터 통합

변수 판별을 위한 탐색적 데이터 조사

4 COMBINED APP

SK_ID_CURR	GENDER	NAME_CONTRACT_TYPE	AMT_APPLICATION	CREDIT_CURRENCY
113582	F	Cash Loan	11000	Currency 1
113583	M	Cash Loan	84000	Currency 1
113584	M	Consumer Loan	50500	Currency 2

Application_train 데이터에 모두 Left_Join

데이터 통합

변수 판별을 위한 탐색적 데이터 조사

5 범주형 → One Hot Encoding

SK_ID_CURR	GENDER	NAME_CONTRACT_TYPE	AMT_APPLICATION	CREDIT_CURRENCY
113582	F	Cash Loan	11000	Currency 1
113583	M	Cash Loan	84000	Currency 1
113584	M	Consumer Loan	50500	Currency 2

SK_ID_CURR	GENDER_M	GENDER_W	NAME_CONTRACT_TYPE _Cash Loan	NAME_CONTRACT_TYPE _Consumer Loan	AMT_APPLICATION
113582	0	1	1	0	11000

범주형 변수들을 모두 One Hot Encoding 시킴!

데이터 통합

로드맵

범주형 변수들을 ONE HOT ENCODING 하는 이유?

XG Boost는 팩터 변수를 받지 않기 때문!

데이터 통합

변수 판별을 위한 탐색적 데이터 조사

EX) **범주형 → One Hot Encoding**

SK_ID_CURR	GENDER	NAME_CONTRACT_TYPE	AMT_APPLICATION	CREDIT_CURRENCY
113582	F	Cash Loan	11000	Currency 1
113583	M	Cash Loan	84000	Currency 1
113584	M	Consumer Loan	50500	Currency 2

SK_ID_CURR	GENDER_M	GENDER_W	NAME_CONTRACT_TYPE _Cash Loan	NAME_CONTRACT_TYPE _Consumer Loan	AMT_APPLICATION
113582	0	1	1	0	11000

범주형 변수들을 모두 One Hot Encoding 시킴!

변수 선택

Feature Selection

STEP1 도메인 지식 이해를 통한 주관적 선택

173	credit_card_balance.csv	NAME_CONTRACT_STATUS	Contract status (active signed,...) on the previous credit	
174	credit_card_balance.csv	SK_DPD	DPD (Days past due) during the month on the previous credit	
175	credit_card_balance.csv	SK_DPD_DEF	DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored)	of the pre
176	previous_application.csv	SK_ID_PREV	ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can h	hashed
177	previous_application.csv	SK_ID_CURR	ID of loan in our sample	hashed
178	previous_application.csv	NAME_CONTRACT_TYPE	Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application	
179	previous_application.csv	AMT_ANNUITY	Annuity of previous application	
180	previous_application.csv	AMT_APPLICATION	For how much credit did client ask on the previous application	
181	previous_application.csv	AMT_CREDIT	Final credit amount on the previous application. This differs from AMT_APPLICATION in a way that the AMT	
182	previous_application.csv	AMT_DOWN_PAYMENT	Down payment on the previous application	
183	previous_application.csv	AMT_GOODS_PRICE	price of goods that client asked for (if applicable) on the previous application	
184	previous_application.csv	WEEKDAY_APPR_PROCESS_STAR	On which day of the week did the client apply for previous application	
185	previous_application.csv	HOUR_APPR_PROCESS_START	Approximately at what day hour did the client apply for the previous application	rounded
186	previous_application.csv	FLAG_LAST_APPL_PER_CONTRAC	Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there cou	
187	previous_application.csv	NFLAG_LAST_APPL_IN_DAY	Flag if the application was the last application per day of the client. Sometimes clients apply for more appli	
188	previous_application.csv	NFLAG_MICRO_CASH	Flag Micro finance loan	

도메인 지식을 활용하여 제거할 데이터를 선택

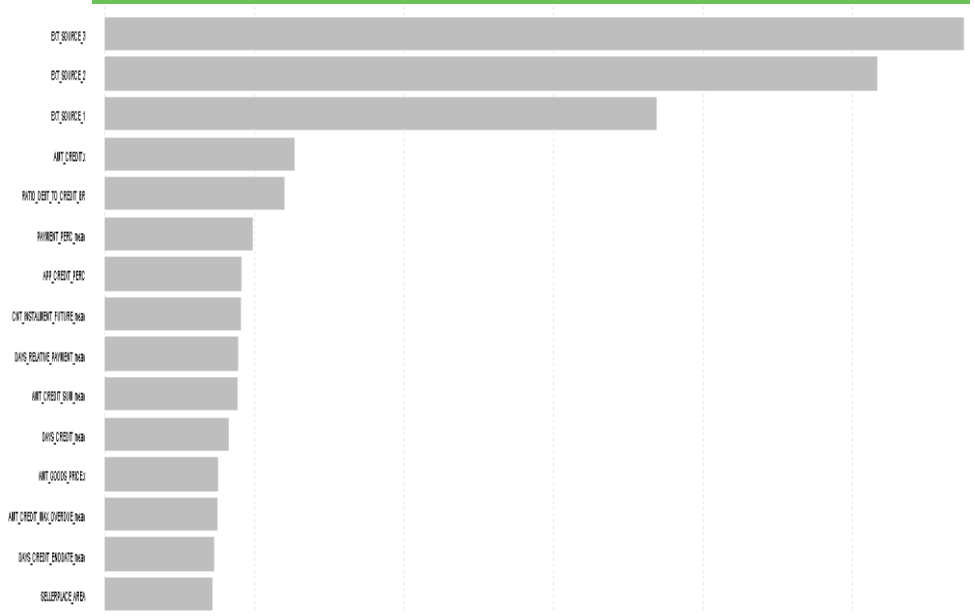
변수 선택

Feature Selection

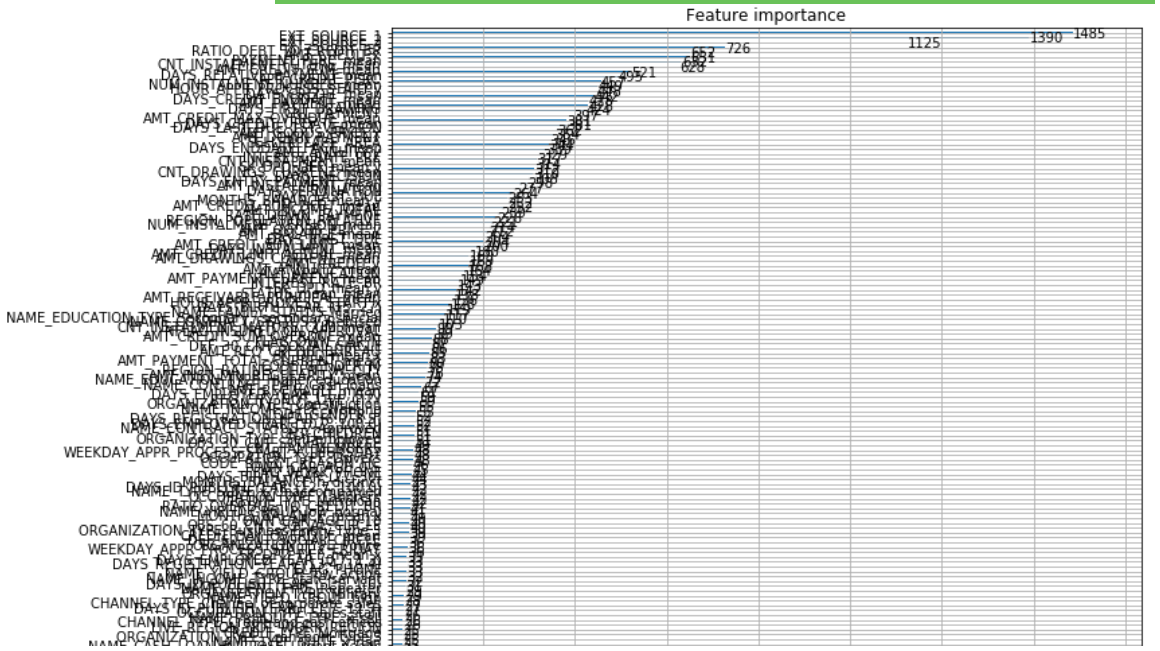
STEP2 XGBoost 와 LGBM 을 통한 importance 확인

앞선 도메인 지식 기반의 주관적 선택과 비교하며 변수를 선택하여 모델링 계획

XGBoost



LGBM



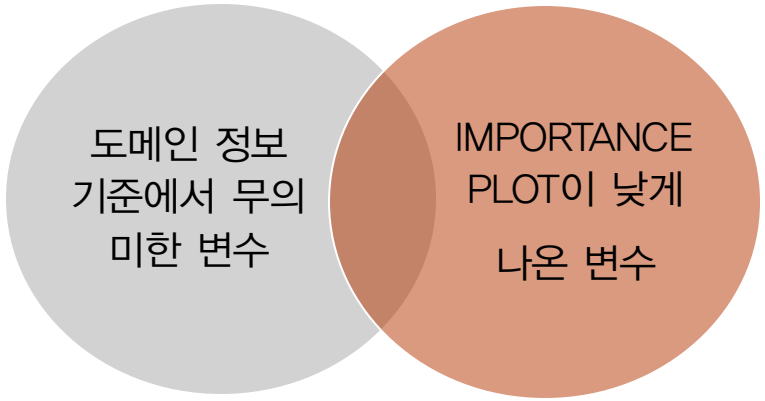
변수 선택

Feature Selection

STEP3 최종 데이터 셋

STEP1 STEP2 조건에
모두 해당되는 변수 제거

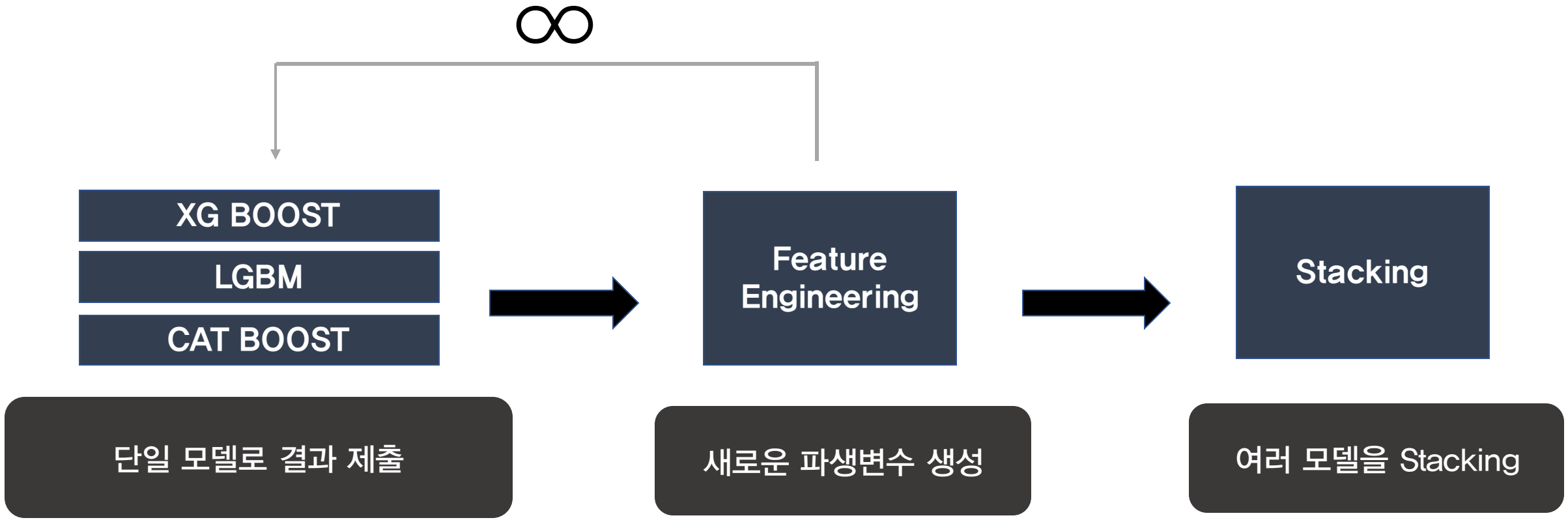
IMPORTANCE PLOT이
낮게 나온 변수들은
파생 변수로 고려



총 423개의 변수와 80만개의 OBS

다음 주차 흐름

모델링 기법 및 방향


∞
XG BOOST
LGBM
CAT BOOST
Feature Engineering
Stacking
단일 모델로 결과 제출
새로운 파생변수 생성
여러 모델을 Stacking