대출 상환 여부 예측

YBIGTA & P-SAT
연합세미나

팀 헛개수

황원영
곽지훈
이명진
백상현
나지윤

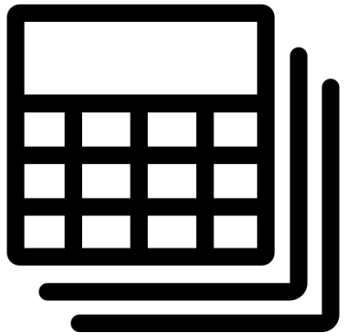
1주차 목차

- 데이터 셋 소개
- 데이터 전처리 및 EDA
  & Feature engineering
  - Application_Set
  - Bureau_Set
  - Previous_Application_Set
- 데이터 통합

데이터 셋 소개

주어진 data set은 각각의 loan 이 제때에 상환되었는지(target),
그리고 각각의 loan과 연관된 고객의 정보가 담겨있다

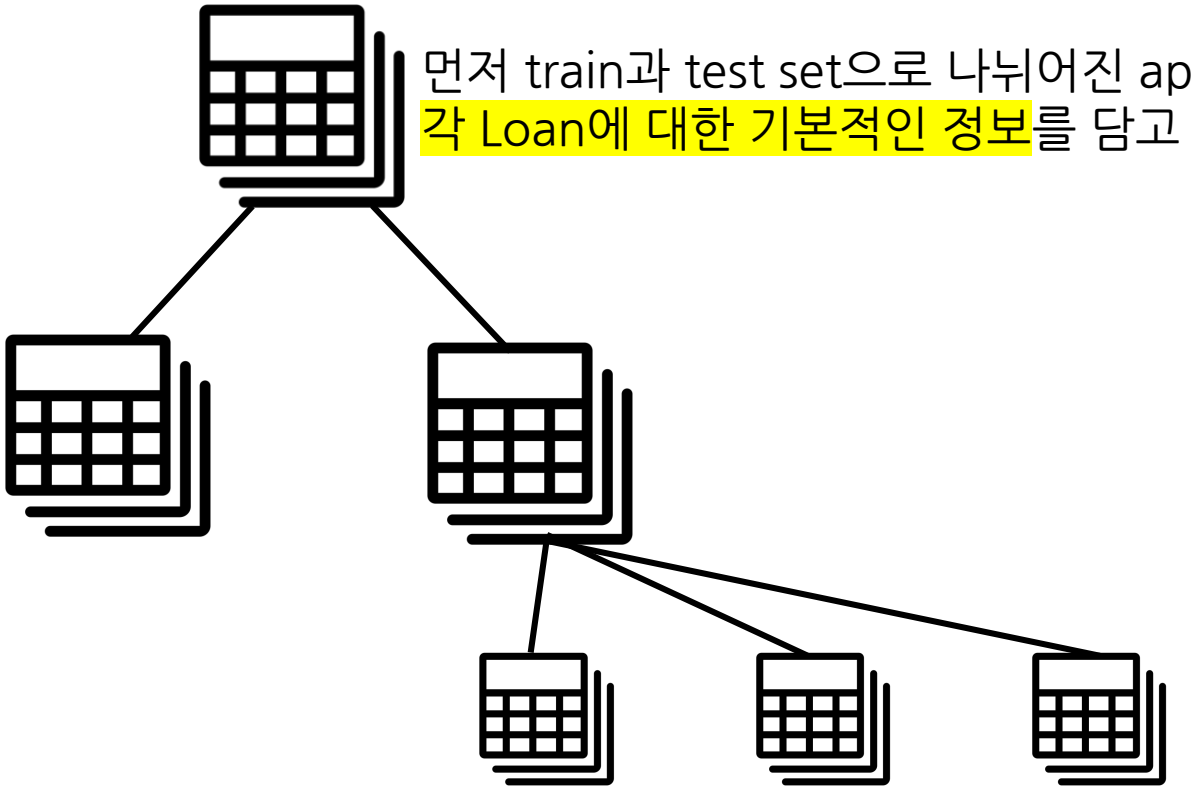
데이터 셋 소개



× 6

크게 나누면 6가지의 set으로 구성되어 있었는데,

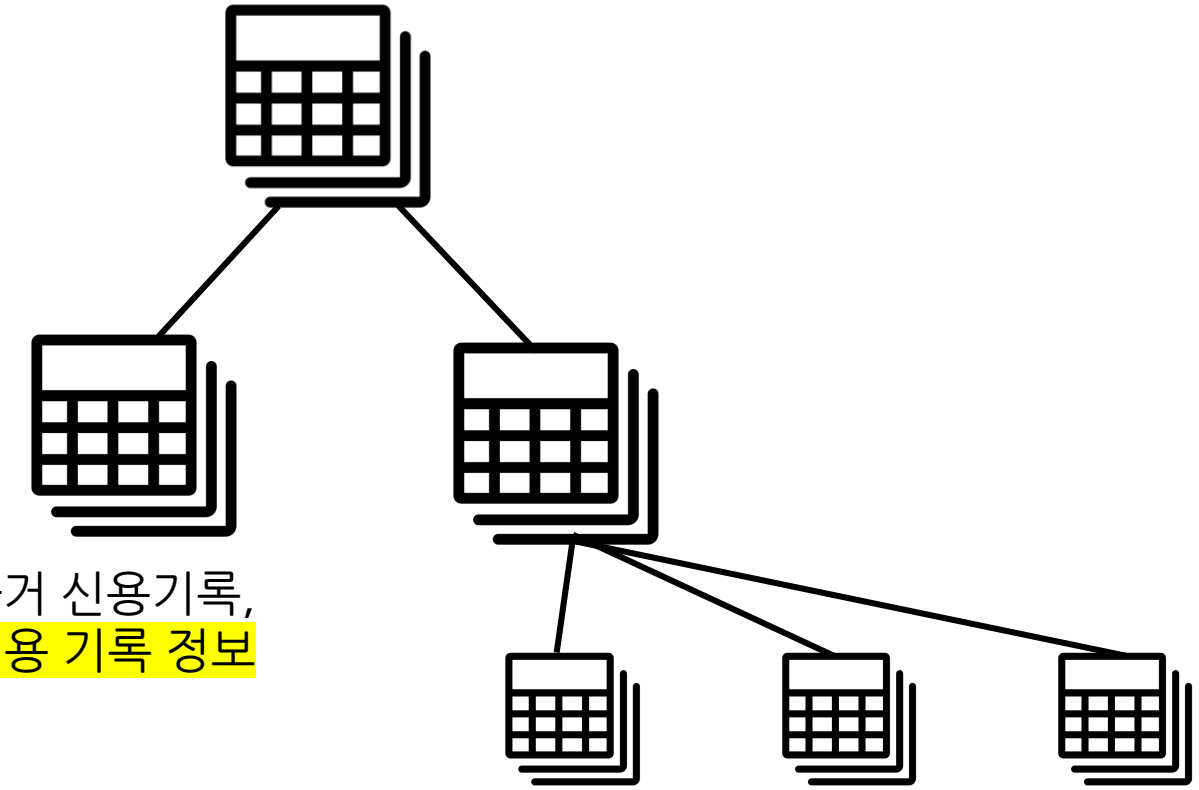
데이터 셋 소개



먼저 train과 test set으로 나뉘어진 application set,
각 Loan에 대한 기본적인 정보를 담고 있다

데이터 소개 | Application set | Bureau set | Previous application set | 데이터 통합

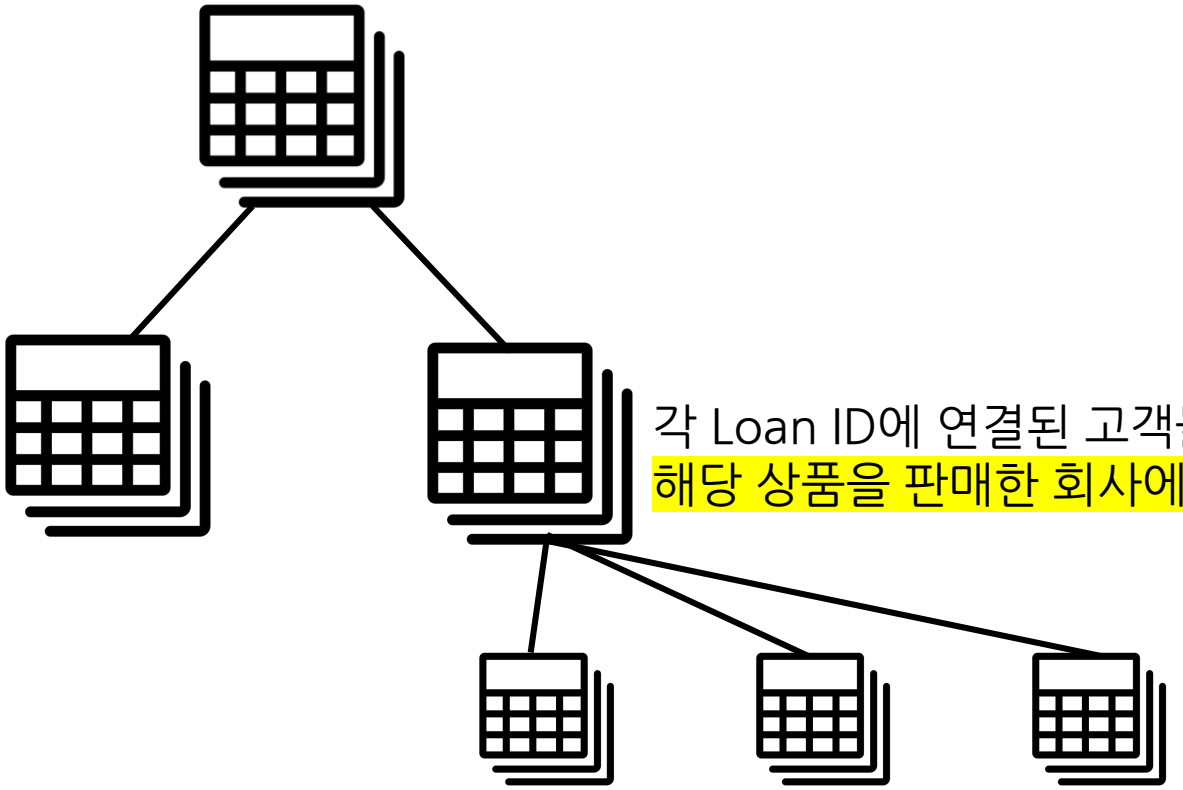
데이터 셋 소개



각 Loan ID에 연결된 고객들의 과거 신용기록,
Credit Bureau에 보고된 신용 기록 정보

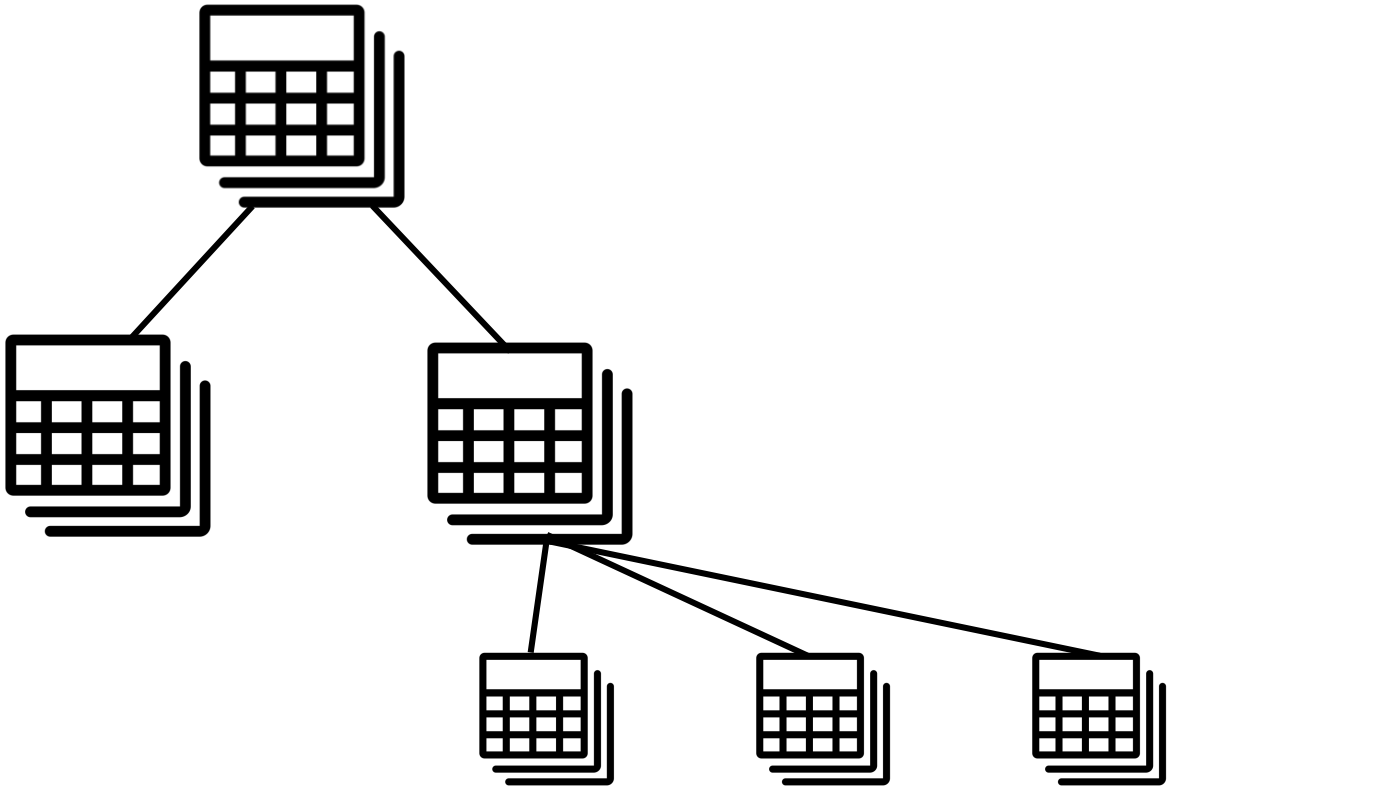
데이터 소개 | Application set | Bureau set | Previous application set | 데이터 통합

데이터 셋 소개

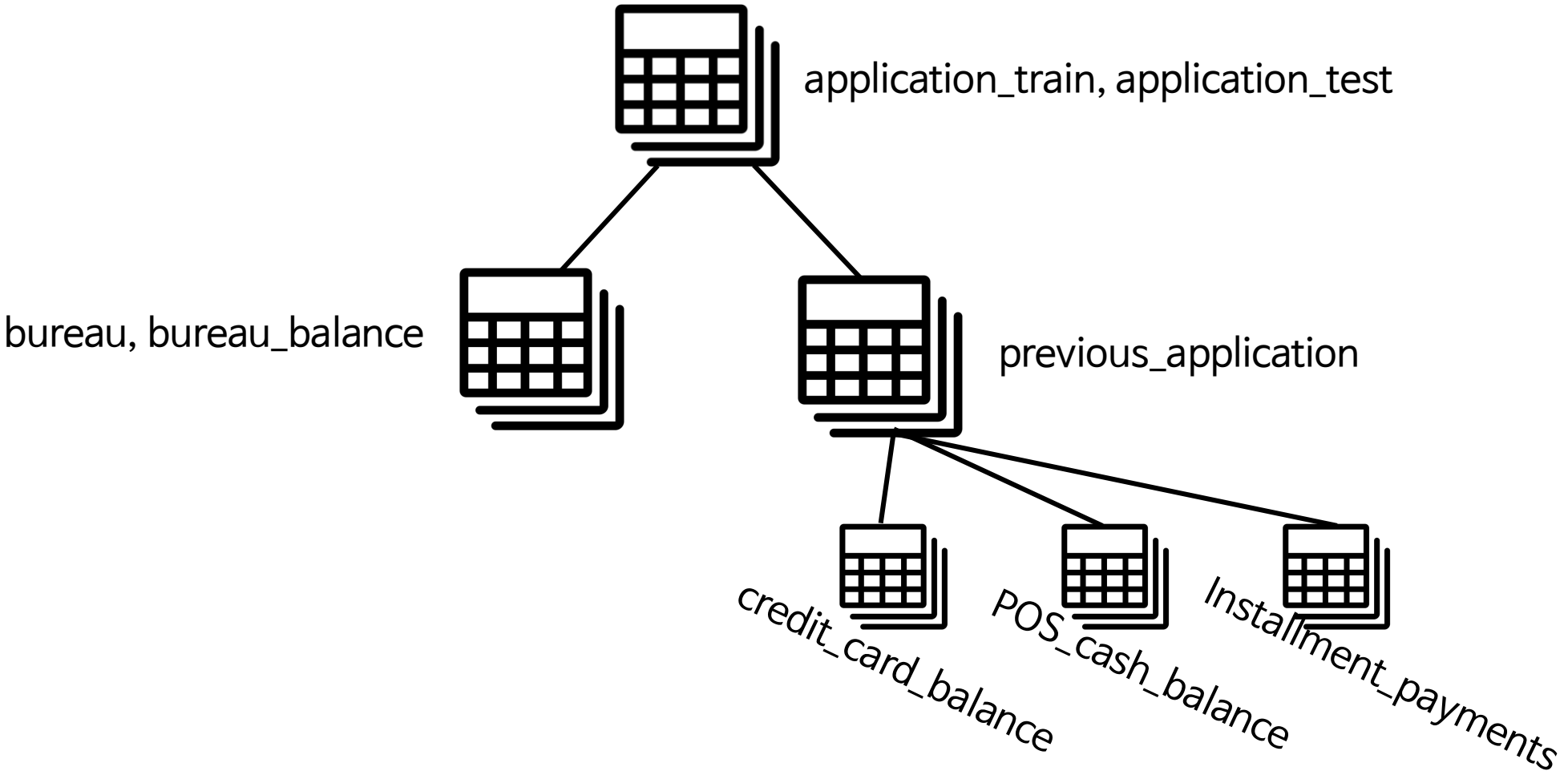


각 Loan ID에 연결된 고객들의 과거 신용기록,
해당 상품을 판매한 회사에 기록된 신용 기록 정보

데이터 셋 소개



데이터 셋 소개

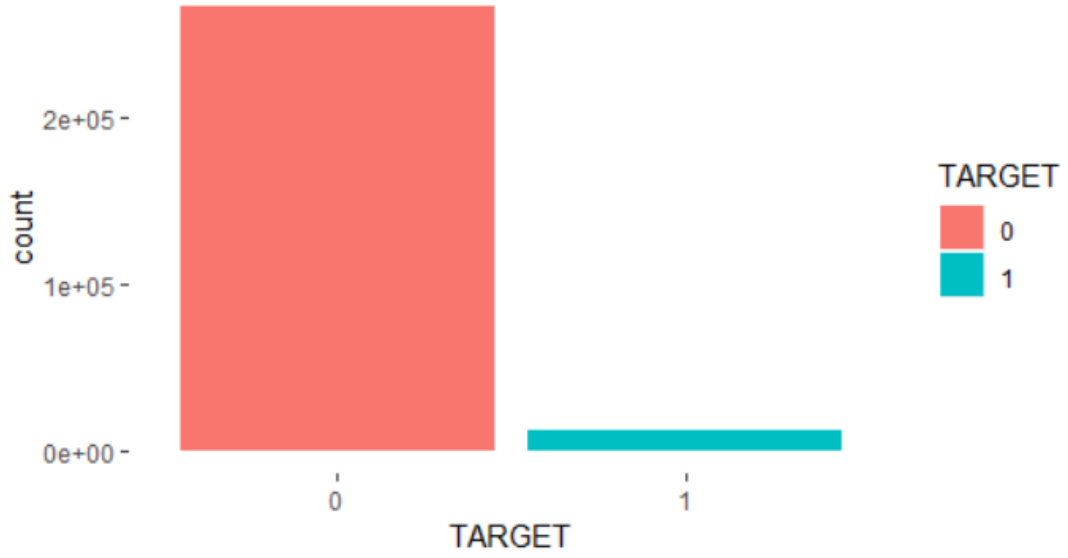

application_train, application_test
bureau, bureau_balance
previous_application
credit_card_balance
POS_cash_balance
Installment_payments

데이터 전처리 및 EDA

데이터 전처리 방향

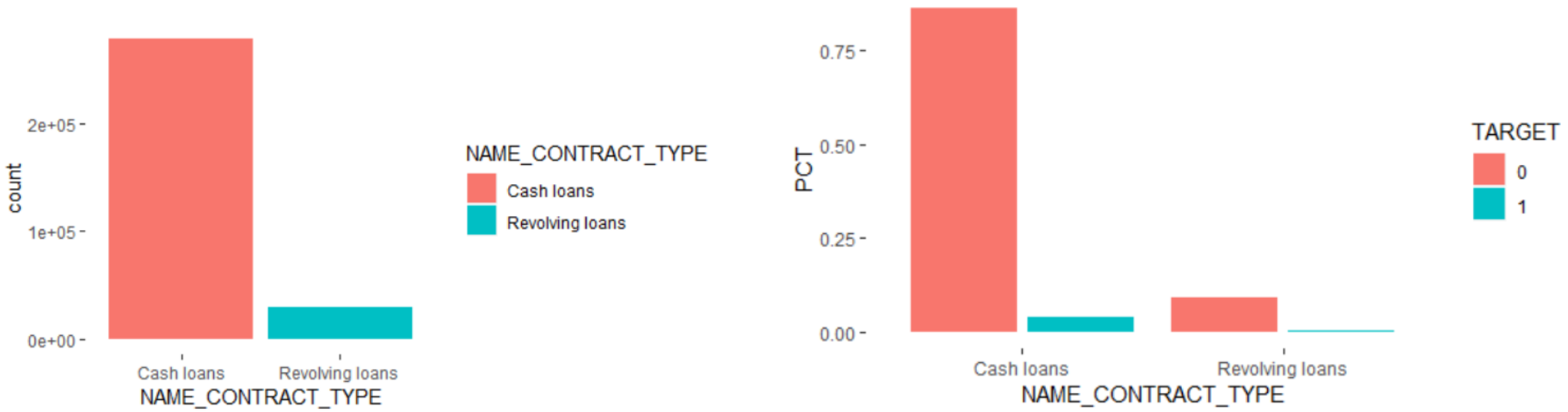
- ✓ 합리적인 결측치 처리
- ✓ 주어진 데이터 활용 최대화
- ✓ Redundant 한 변수 삭제

데이터 전처리 및 EDA
Application_Set



TARGET 이 굉장히 unbalanced 임을 확인 할 수 있다
(1 : 제때에 상환 못함, 0 : 제때에 상환 함)

데이터 전처리 및 EDA
Application_Set



Contract type은 Cash loan이 압도적으로 높았다
(Revolving loan의 예시로는 credit card가 있다, 고정된 지불액을 가지고 있지 않은 credit의 종류)

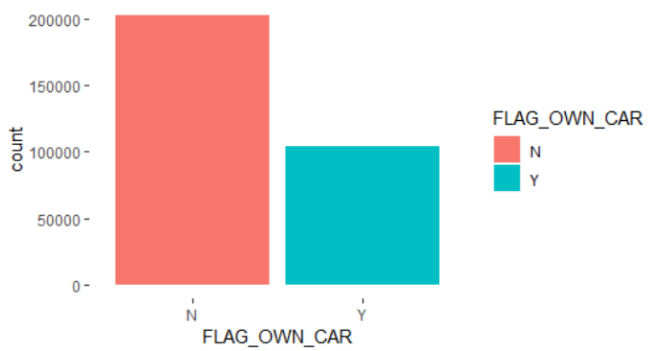
데이터 전처리 및 EDA

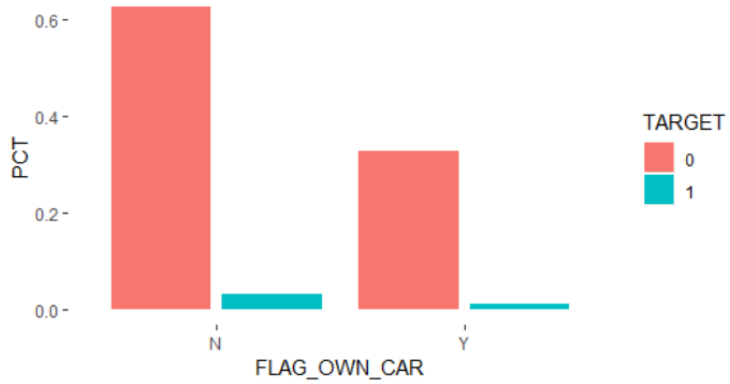
Application_Set

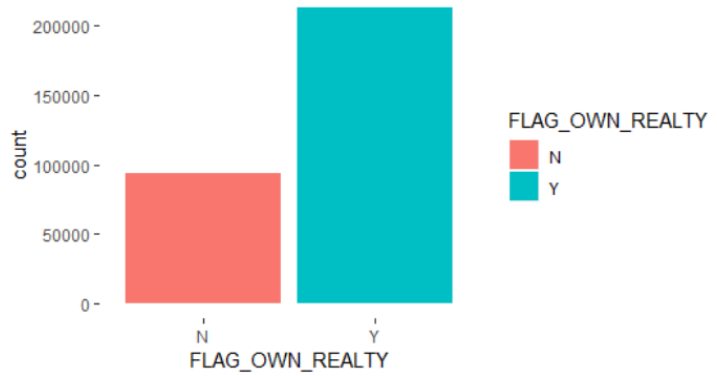
성별의 비율은 여성이 2배 정도로 높았으며,
성별이 기입되지 않은 obs가 2개 발견되었다

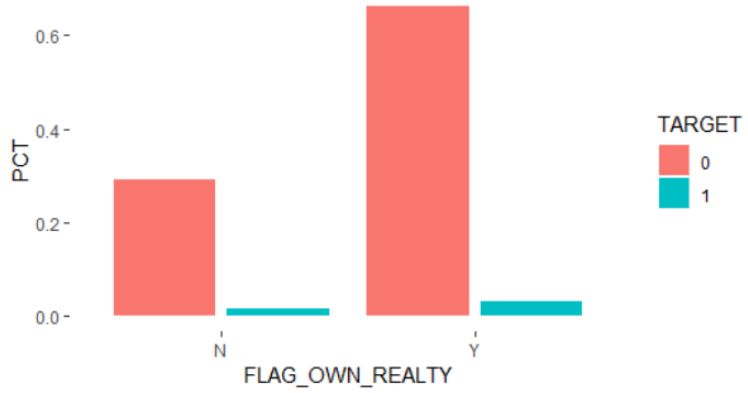
데이터 전처리 및 EDA

Application_Set



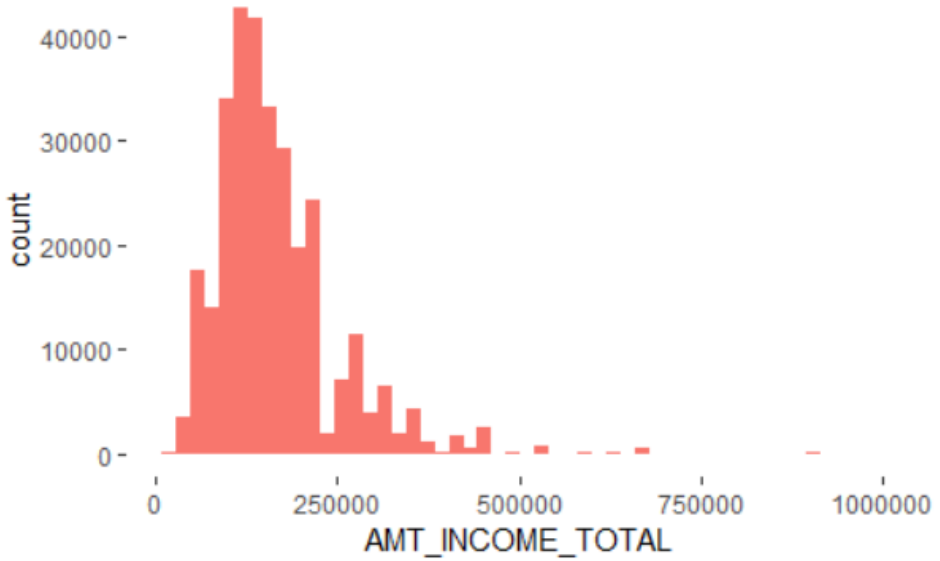






해당 고객이 차와 부동산을 소유했는지도 확인 할 수 있었다

데이터 전처리 및 EDA
Application_Set

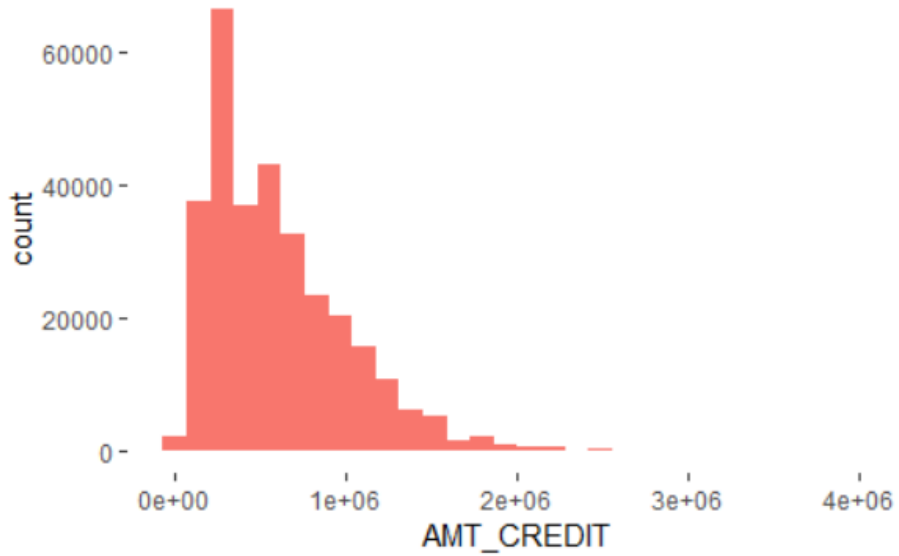


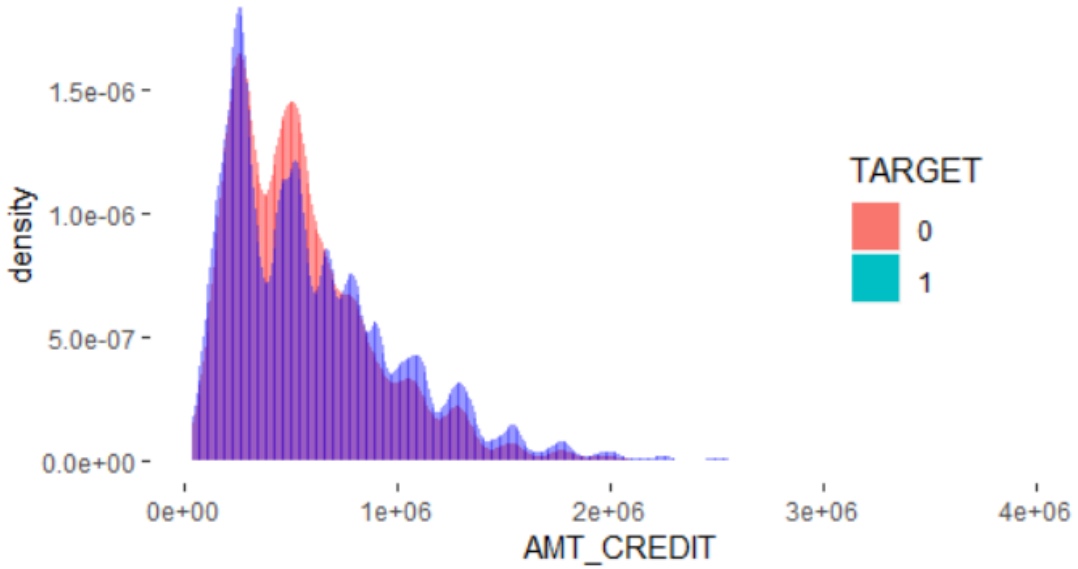
고객의 소득을 나타낸 plot인데 실제로는 더 right-skewed 하다,
(소득이 1000000$ 가 넘는 250개의 obs를 제외하고 plotting 한 결과)

소득 구간을 10분위로 나눈 뒤
각 구간에서 상환율을 비교해보았다

데이터 전처리 및 EDA

Application_Set

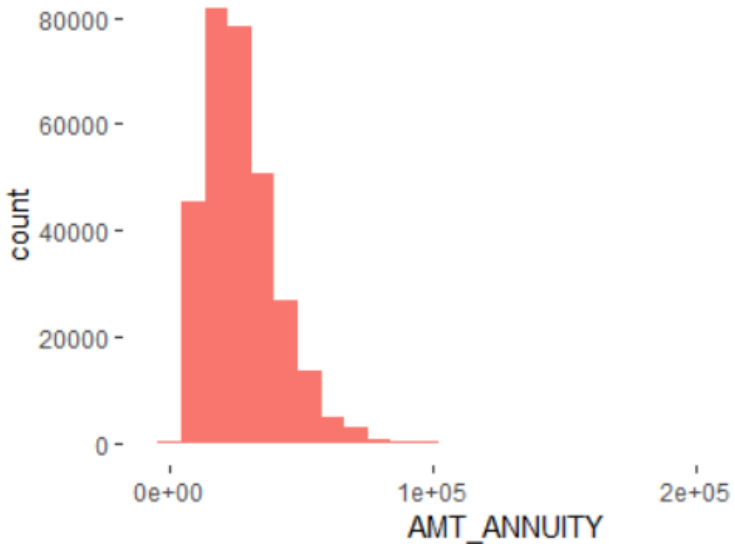


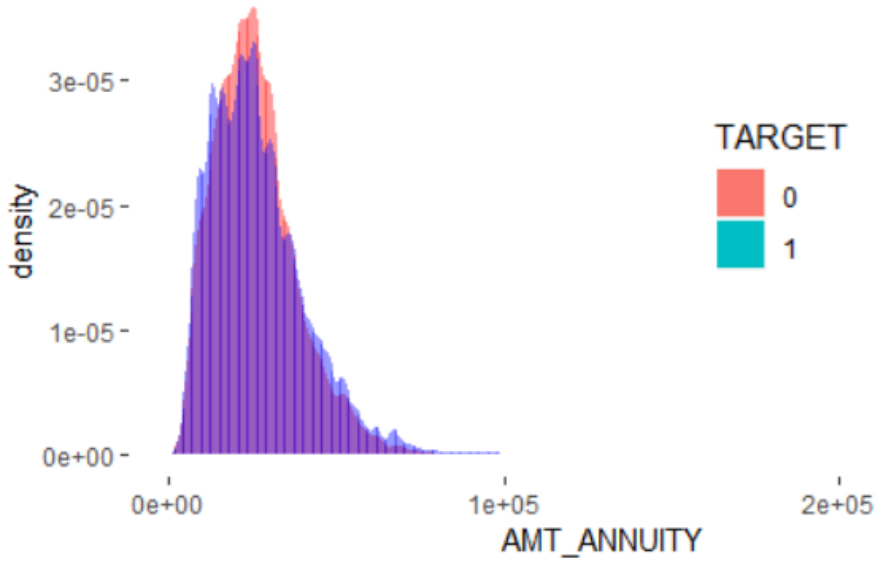


Credit amount도 right skewed함을 확인 가능하다,
오른쪽 plot은 target별 credit amount의 분포

데이터 전처리 및 EDA

Application_Set

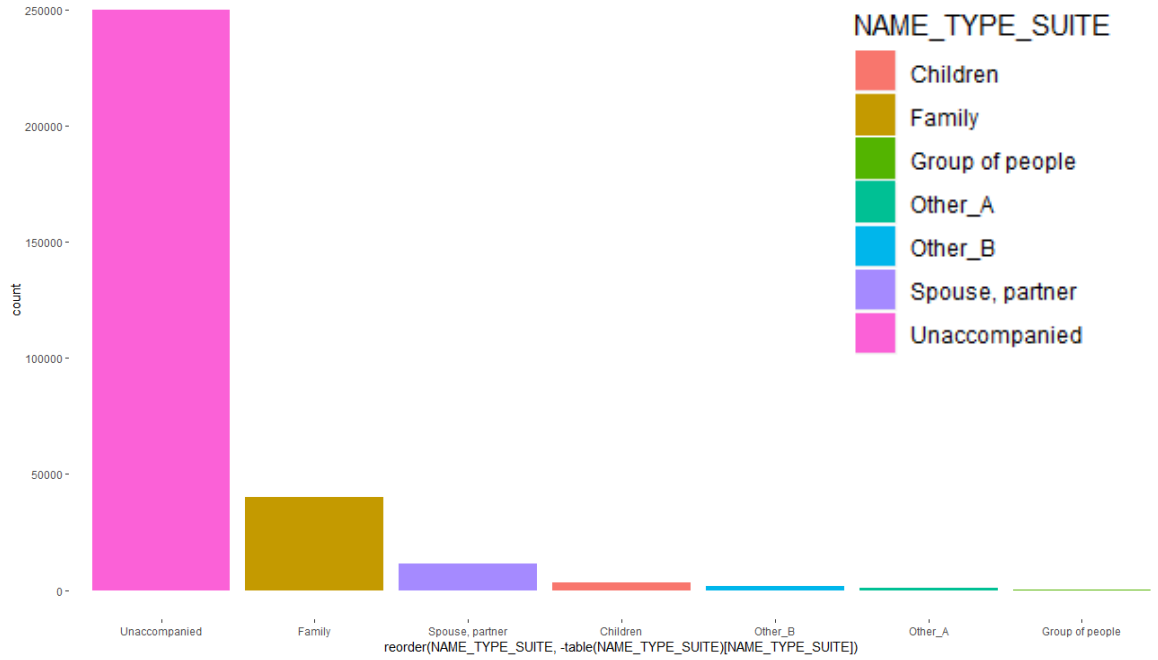




AMT_ANNUITY에서는 12개의 NA값이 발견 되었는데,
Loan annuity(정기적으로 갚아야 하는 돈)과 Amount credit의 correlation이 0.77임을 발견 regression Imputation 진행

데이터 전처리 및 EDA

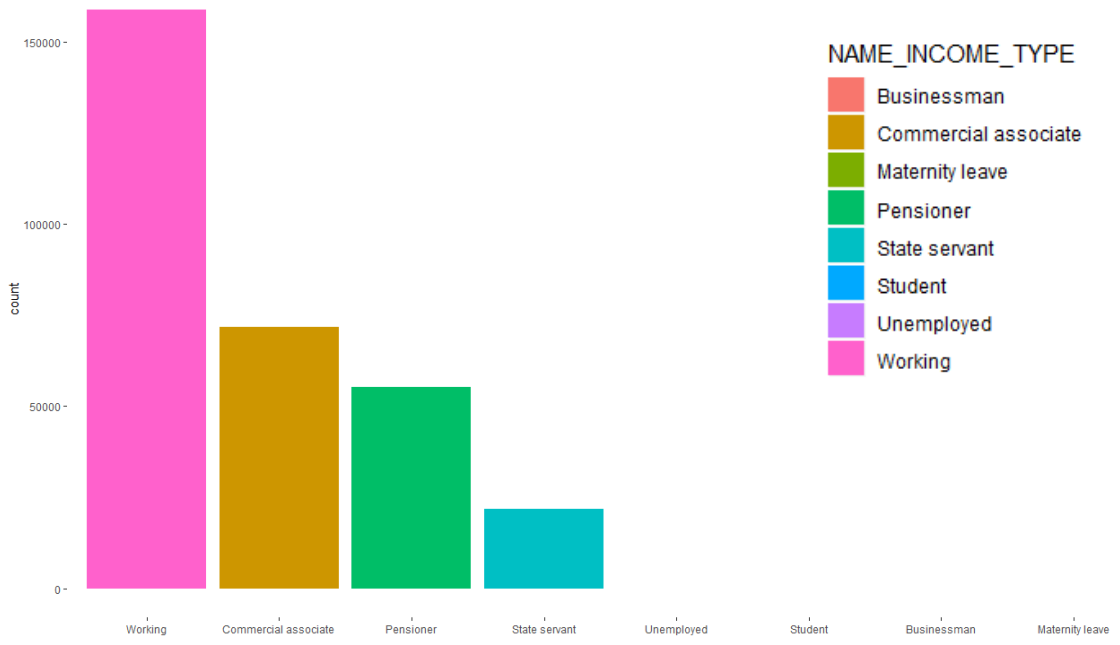
Application_Set

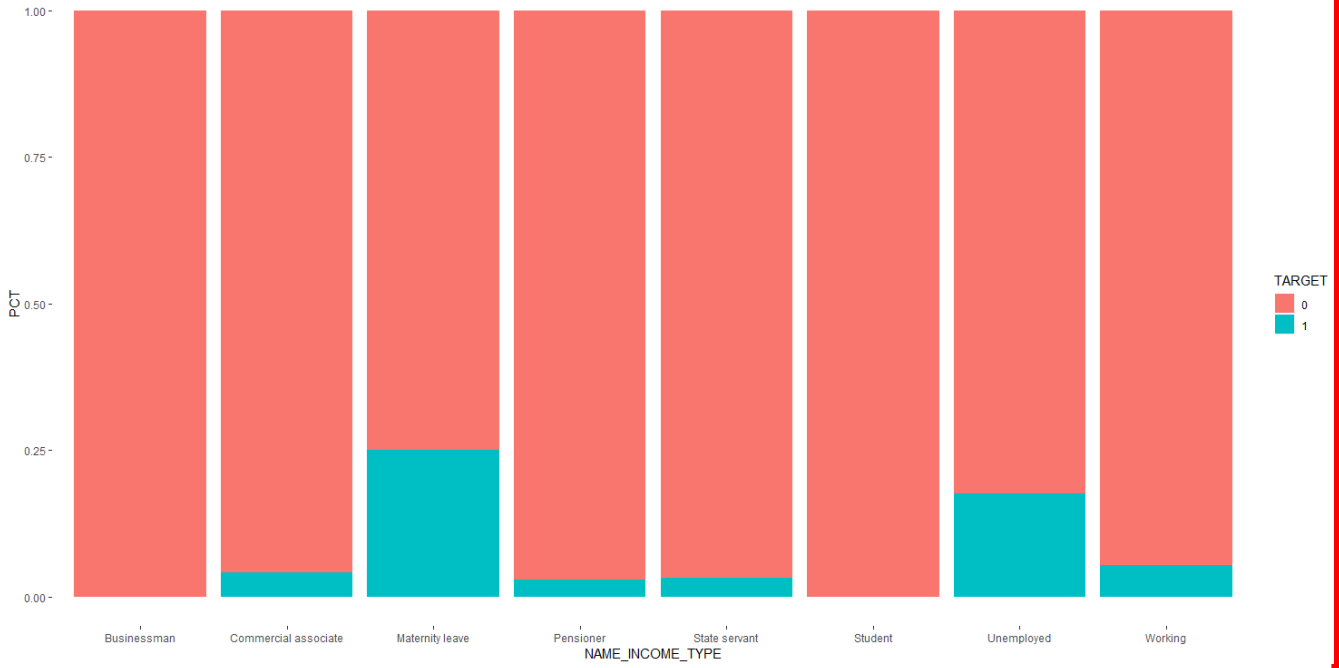


Loan 신청 당시 동반한 사람의 유형에서 발견된 1292개의 결측치는
동반자가 없었다는 것으로 판단 Unaccompanied로 교체했다

데이터 전처리 및 EDA

Application_Set

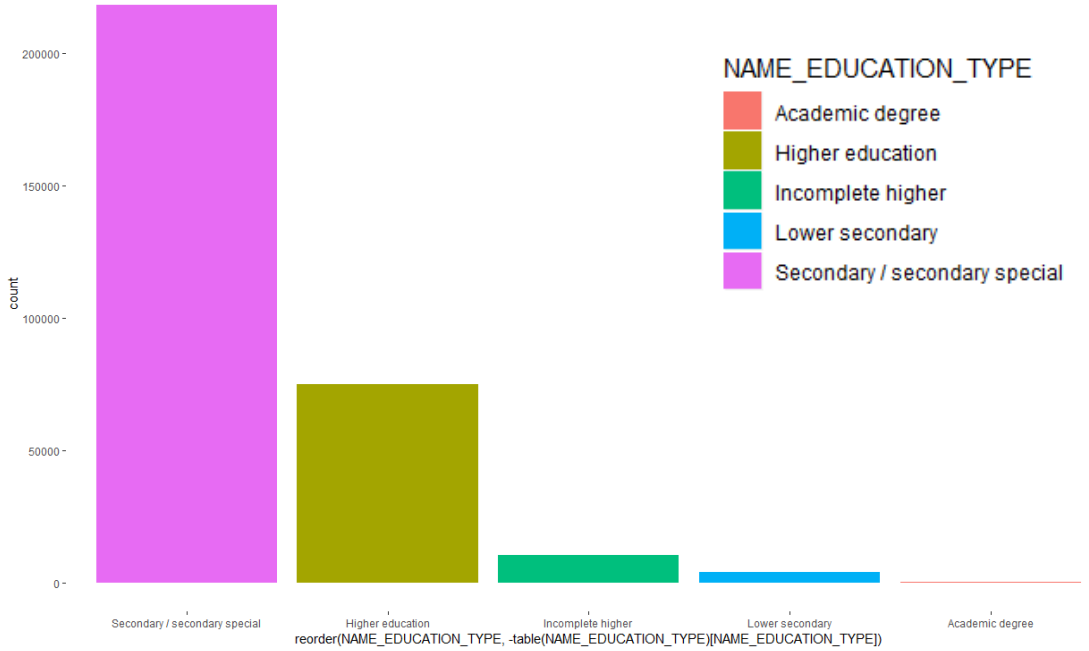


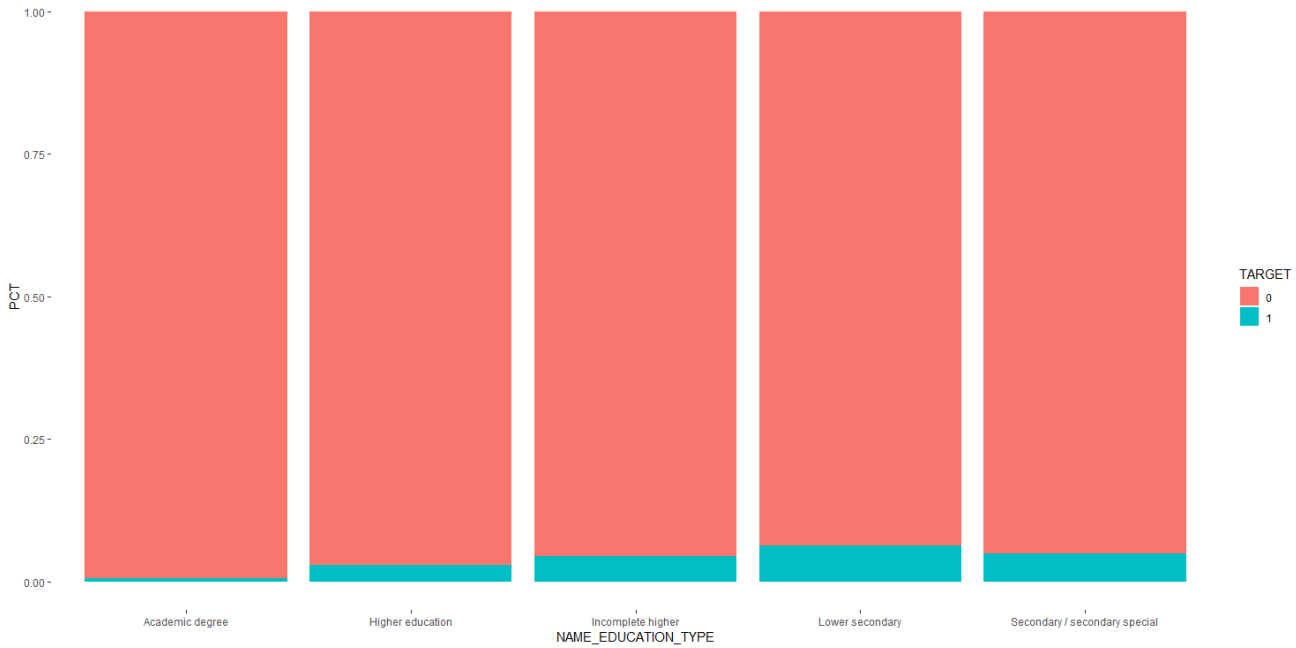


Income type에 대한 정보도 포함되어 있으며 각 Income type별 상환율을 비교한 결과
Maternity leave(출산휴가 수당)과 Unempolyed 의 경우 상환 실패 비율이 다소 높았다

데이터 전처리 및 EDA

Application_Set

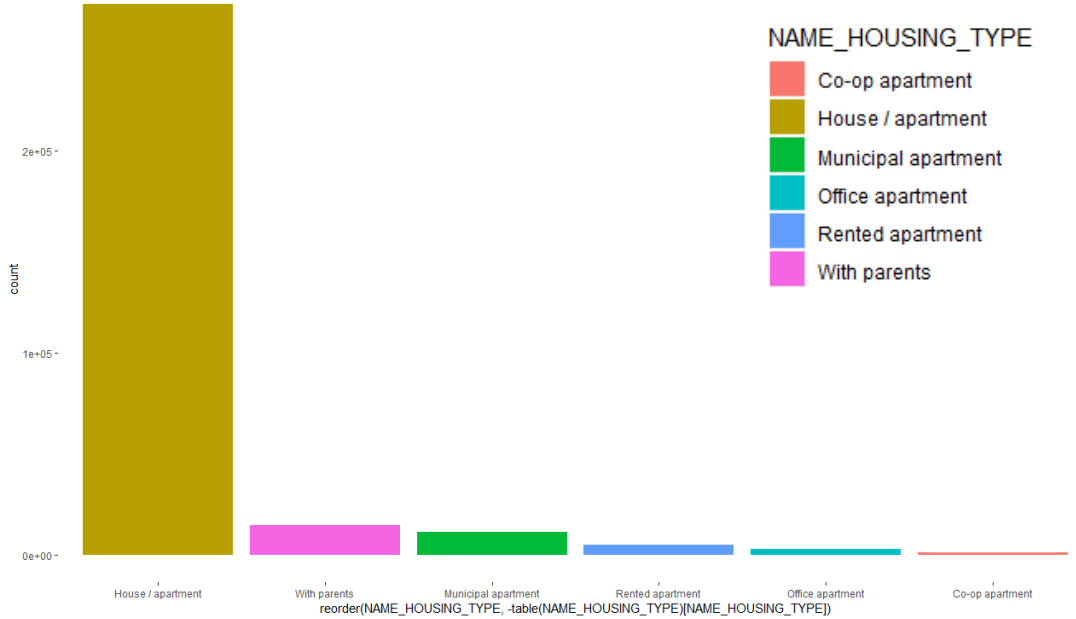


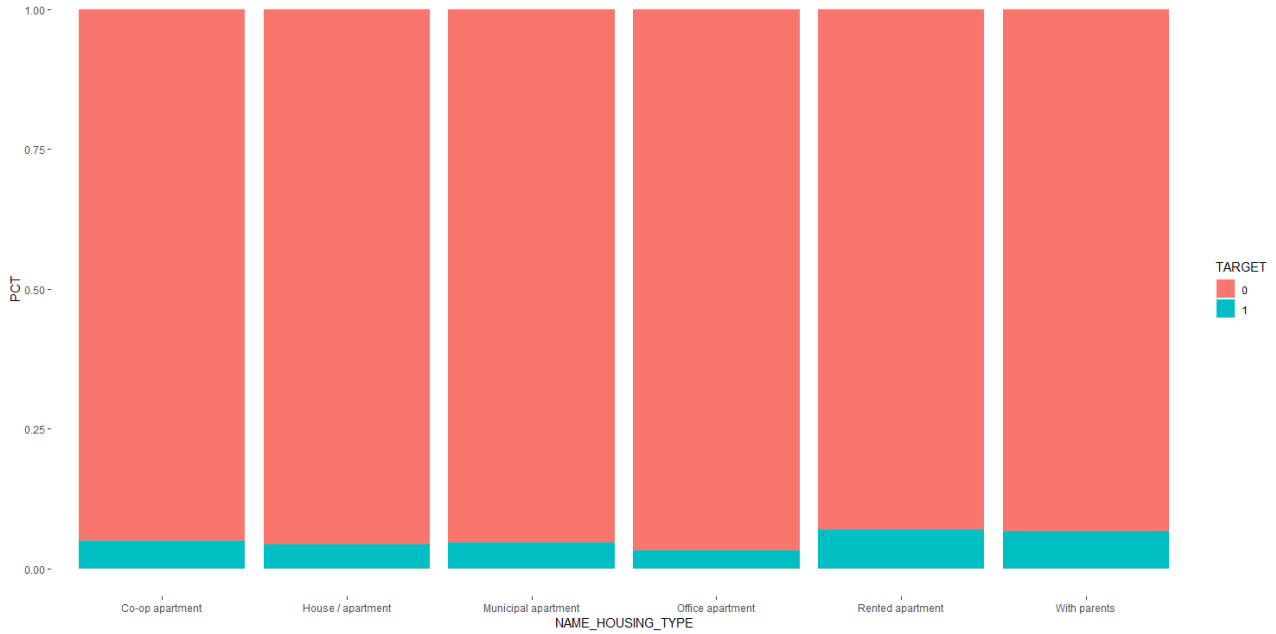


Education type에 대한 정보도 포함되어 있으며 각 Education type별 상환율을 비교한 결과
Lower secondary의 경우 상환 실패 비율이 다소 높았다

데이터 전처리 및 EDA

Application_Set

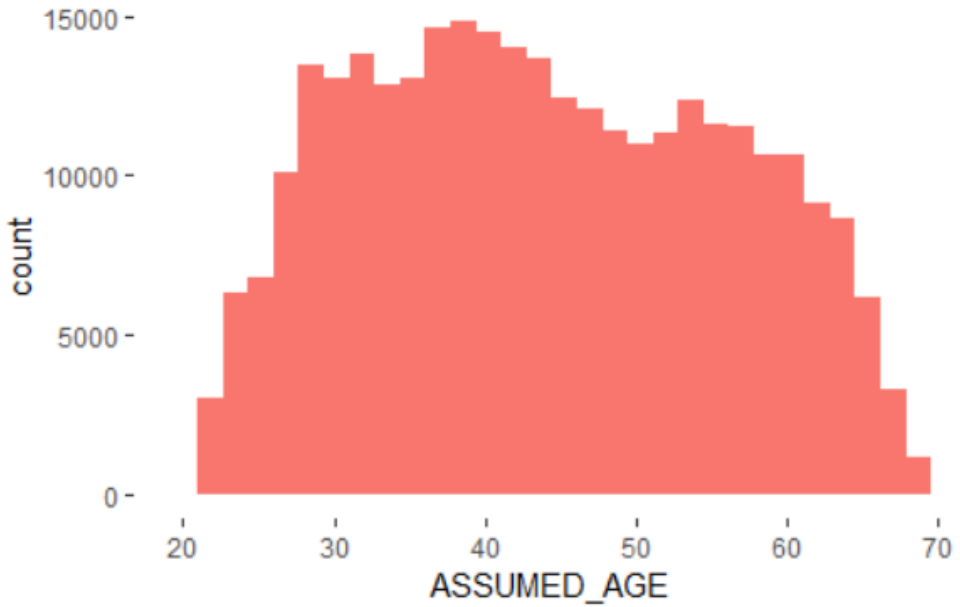


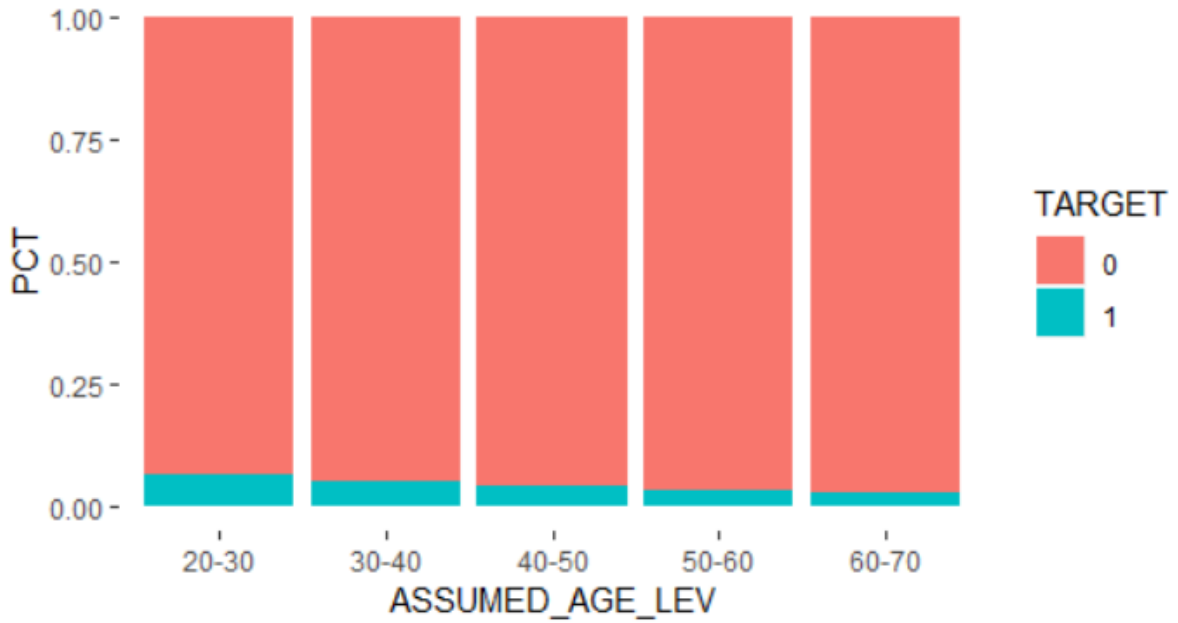


Housing type에 대한 정보도 포함되어 있으며 각 Housing type별 상환율을 비교한 결과
rented apartment의 경우 상환 실패 비율이 다소 높았다

데이터 전처리 및 EDA

Application_Set

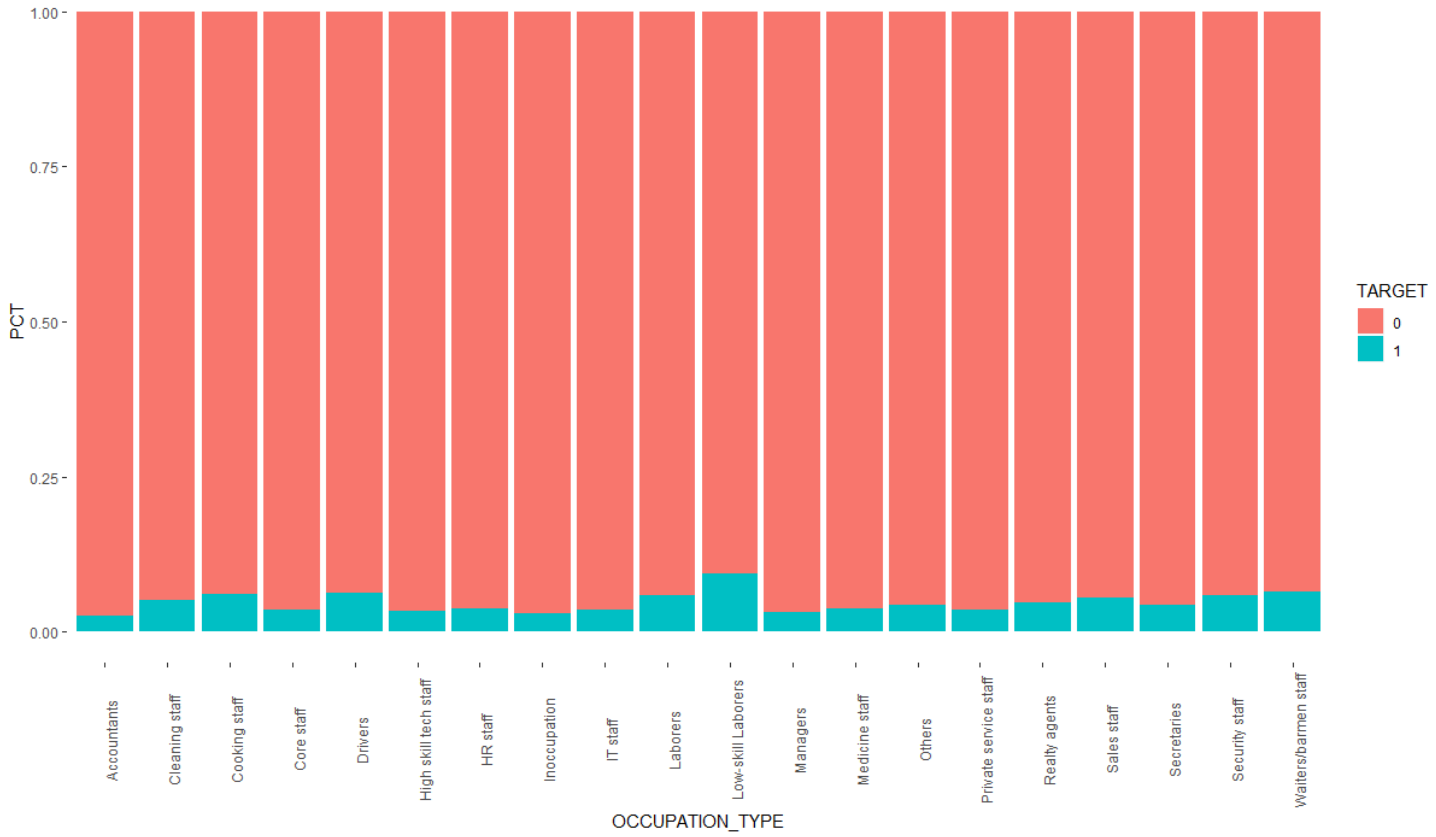




DAYS_BIRTH 변수는 해당 loan 상품 신청 당시 날과 고객이 태어난 날의 차이로 표현되어 있어
이를 사용해 고객의 나이(ASSUMED_AGE) 변수를 생성함,
오른쪽은 나이대별 상환율을 시각화한 plot, 20대의 상환 실패 비율이 다소 높다

데이터 전처리 및 EDA

Application_Set



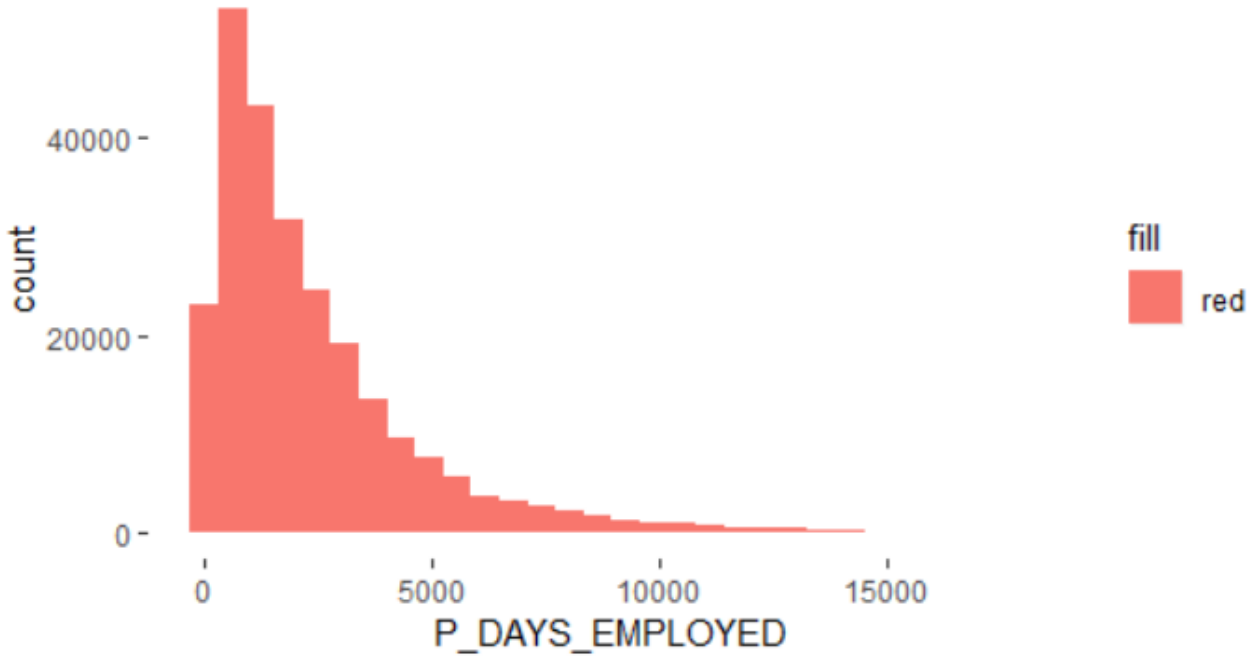
Occupation type 에서 NA로 기록 된 obs 중 Days employed도 NA인 obs는 무직이라고 판단 Inoccupation으로 교체

그렇지 않은 경우는 주어진 category에 해당하지 않는 직업이라고 판단 Others로 교체

왼쪽 plot은 직업별 상환율 비교, Low skilled Laborers 의 상환 실패 비율이 다소 높음

데이터 전처리 및 EDA

Application_Set



DAYS_EMPLOYED는 해당 loan 상품 신청 당시 날과 직업을 구한 날과의 차이(-)로 표현되어 있었으며
+365243라는 이상치를 지닌 55374개의 obs 발견,
해당 변수는 Occupation type에서 무직임이 발견되었고 Income type에서도 모두 Pensioner(연금 수급자)
로 나타났기에 직업이 없는 obs라 판단함

데이터 전처리 및 EDA

Application_Set

고객의 거주지 관련 정보(거실 크기, 건물 연식, 창고 크기 등등…)은 결측치의 비율이 매우 높아 사용할 수 없다고 판단함

대신 거주지 관련 정보를 하나라도 입력하면 1, 아니면 0이 들어가는 BUILDING_FILL변수 생성

데이터 전처리 및 EDA

Application_Set

AMT_REQ_CREDIT_BUREAU_HOUR

AMT_REQ_CREDIT_BUREAU_DAY

AMT_REQ_CREDIT_BUREAU_WEEK

AMT_REQ_CREDIT_BUREAU_MON

AMT_REQ_CREDIT_BUREAU_QRT

AMT_REQ_CREDIT_BUREAU_YEAR

Loan 신청 전 credit bureau 에 문의한 횟수,
모든 변수에 걸쳐 NA값을 갖는 obs가 발견되었는데
문의사항이 없었다고 판단 0으로 교체

OBS_30_CNT_SOCIAL_CIRCLE

DEF_30_CNT_SOCIAL_CIRCLE

OBS_60_CNT_SOCIAL_CIRCLE

DEF_60_CNT_SOCIAL_CIRCLE

Number of observations of client's social surroundings with observable DPD (Days Past Due)
모든 변수에 걸쳐 NA값을 갖는 obs가 발견되었는데
DPD observation이 없었다고 판단 0으로 교체

데이터 전처리 및 EDA
Application_Set

그 외...

DAYS_LAST_PHONE_CHANGE : 결측치 1개 존재, mean Imputation

AMT_GOODS_PRICE : AMT_CREDIT과의 correlation 이 0.98로 높아 열 삭제

OWN_CAR_AGE : missing value의 비율이 매우 높아 열 삭제

DOCUMENT_NUM : 총 제출한 문서의 수 변수 생성(FLAG_DOCUMENT 변수 활용)

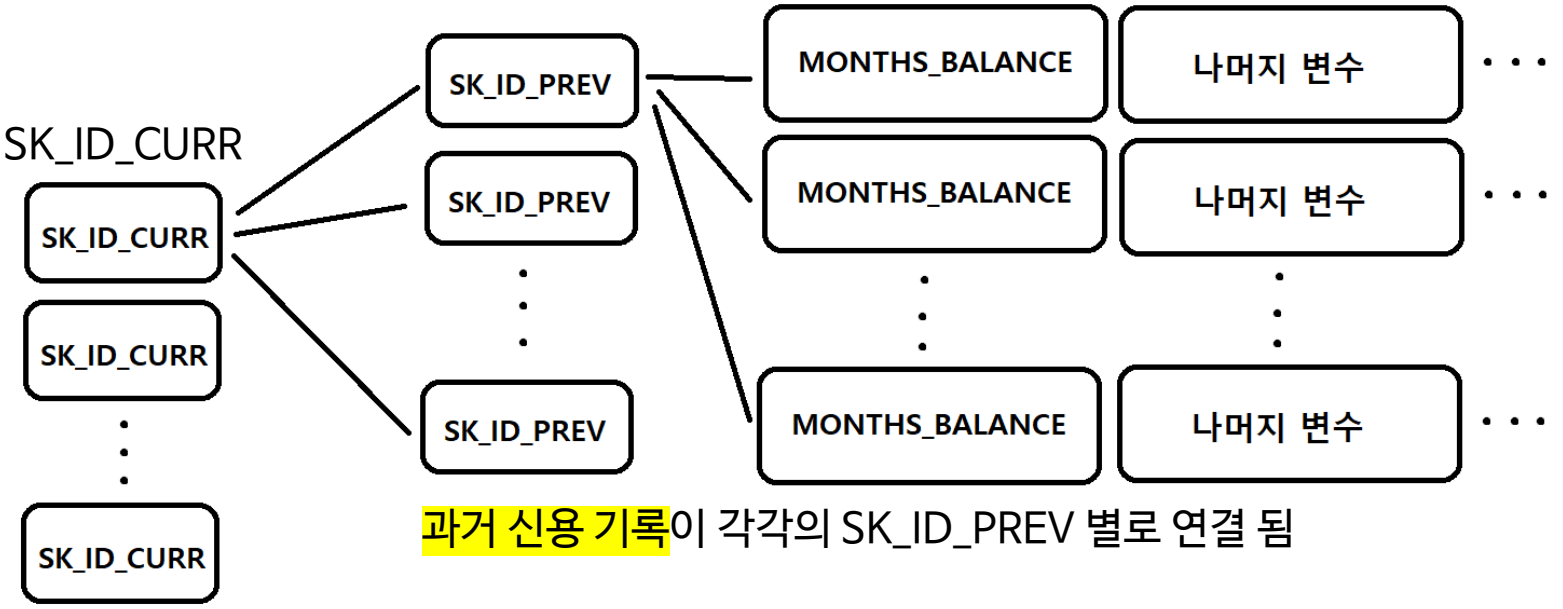
Feature Engineering

Application set 외의 Data set은 과거의 신용기록을 담은 Data set 이었기에,
각각의 Loan(SK_ID_CURR)별로 여러가지 정보를 담고 있었다

Feature Engineering

각 SK_ID_PREV 별로 여러가지 정보가 담겨 있음

Data 통합 시에 key가 되는 SK_ID_CURR

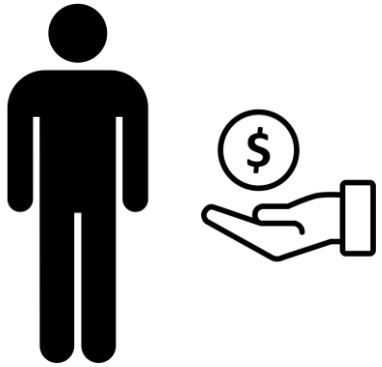


과거 신용 기록이 각각의 SK_ID_PREV 별로 연결 됨

Feature Engineering

SK_ID_CURR 당 row가 두개 이상 생기는 것을 방지하기 위해,
과거 신용기록의 통계량 또는 가장 의미 있어 보이는 값을 활용해 새로운 feature 생성

Feature Engineering
Bureau_Set



Bureau data set은 SK_ID_CURR별로 Credit Bureau에 보고된
이전 신용상품에 대한 정보가 담겨있다

Feature Engineering
Bureau_Set

주요 변수 설명

"ACT_NUM" : Credit bureau 에 보고된 대출 상품의 수 중 Status 가 Active 한 상품의 수

"ACT_PCT" : Credit bureau 에 보고된 대출 상품의 수 중 Status 가 Active 한 상품의 비율

"MEAN_DAYS_CREDIT_ACT" : application set에 있는 대출 신청한 시점과 CB 에 보고된 대출 상품 신청 시점 기간의 평균

Loan 신청 당시 이미 다른 신용 상품을 사용하고 있다면 그 신용상품을 얼마큼 전에 신청했는지 알 수 있다.

(Active하지 않은 상품은 의미 없다고 판단해서 Active한 신용 상품만 고려함)

"MEAN_AMT_CREDIT_SUM" : application set 에 대출 신청한 시점에 CB에 보고된 대출 상품 금액의 평균

Loan 신청 당시 이미 다른 신용 상품을 사용하고 있다면 그 신용상품의 크기가 얼마인지 알 수 있다.

(Active하지 않은 상품은 의미 없다고 판단해서 Active한 신용 상품만 고려함)

Feature Engineering
Bureau_Set

주요 변수 설명

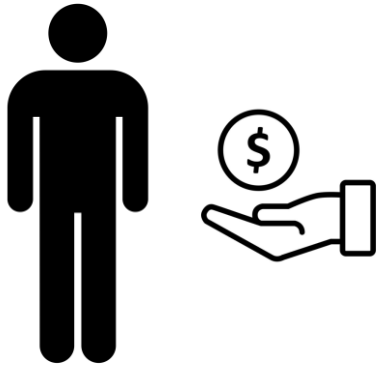
"SUM_CNT_CREDIT_PROLONG" : Credit bureau에 보고된 대출 상품 중 기한 연장 된 상품 수의 총합

"MEAN_AMT_CREDIT_SUM_OVERDUE" : Credit bureau 에 보고된 대출 상품 중 기한을 넘긴 상품 액수의 평균

"MOST_FREQ_CREDIT_TYPE" : 가장 빈번하게 쓰이는 CREDIT_TYPE

"HIGHEST_DPD" : 가장 높았던 days past due level (1인 경우는 기한을 넘긴 적이 없는 경우,
5인 경우는 기한을 120일 넘었거나 부채를 탕감한 경우)

Feature Engineering
Previous_Application_Set



Previous application data set은 SK_ID_CURR별로
과거에 같은 회사에서 신청한 신용상품에 대한 정보가 담겨있다

Feature Engineering
Previous_Application_Set

주요 변수 설명

"CNT_CASH" : cash 대출을 신청한 수의 총합

"CNT_POS" : POS 대출을 신청한 수의 총합

"CNT_CARD" : credit card 대출을 신청한 수의 총합

"MEAN_MONTH" : 평균적으로 할부를 건 개월의 수

"DIFF_APPL_FINAL" : 대출 신청 금액과 실제 수령 금액의 차이,

대출 신청 금액과 실제 수령 금액의 차이가 있을 수 있는 이유는 회사 승인 과정에서 실제 대출 가능한 금액이 줄어들 수 있기 때문

Feature Engineering
Previous_Application_Set

주요 변수 설명

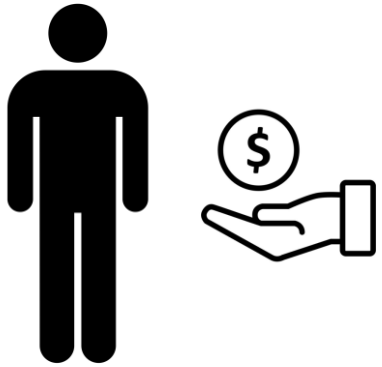
"RATE_APPROVED" : 전체 신청 건수 대비 승인 된 건수의 비율

"RATE_REFUSED" : 전체 신청 건수 대비 승인 되지 않은 건수의 비율

"RATE_NO_PAYMENT" : 전체 신청 건수 대비 상환 방법을 명시하지 않은 건수의 비율

"RATE_WALK_IN" : 전체 신청 건 중 walk-in (신용 평가를 받지 않은 상태)의 비율

Feature Engineering
Previous_Application_Set



Previous application set은 이전 대출 상품의 종류(Credit card, POS cash, Installment)에 따라 더 세분화된 data를 갖고 있었다.

Feature Engineering
Credit_Card_Set

주요 변수 설명

"CONTRACT_STATUS" : 해당 고객의 신용카드 상태(더미화 된 7개의 column 존재)

"credit_card_balance_credit_limit_ratio" : 고객의 신용카드 당월 채무 금액 / 고객의 당월 신용카드 한도

"credit_card_ATM/ALL_Drawings_AMT_Ratio" : ATM 인출 금액 / 전체 신용카드 한도 사용 금액

"credit_card_POS/ALL_Drawings_AMT_Ratio" : 상품 구매 금액 / 전체 신용카드 한도 사용 금액

"credit_card_Other/ALL_Drawings_AMT_Ratio" : 그 외의 금액 / 전체 신용카드 한도 사용 금액

금액 뿐만 아니라 횟수로도 위와 같이 3개의 변수를 생성

Feature Engineering
Credit_Card_Set

주요 변수 설명

“First_DPD_After_n_months” : 첫 신용카드 연체까지 걸린 개월 수(여러 개 카드라면 평균 사용)

“AVG_DPD_TOTAL” : 평균 연체 일

Feature Engineering
Credit_Card_Set

CNT_DRAWINGS_ATM_CURRENT	CNT_DRAWINGS_CURRENT	CNT_DRAWINGS_OTHER_CURRENT	CNT_DRAWINGS_POS_CURRENT
NaN	0	NaN	NaN
NaN	0	NaN	NaN
NaN	0	NaN	NaN
NaN	0	NaN	NaN
NaN	0	NaN	NaN
NaN	0	NaN	NaN
AMT_DRAWINGS_ATM_CURRENT	**AMT_DRAWINGS_CURRENT**	**AMT_DRAWINGS_OTHER_CURRENT**	**AMT_DRAWINGS_POS_CURRENT**
NaN	0.00	NaN	NaN
NaN	0.00	NaN	NaN
NaN	0.00	NaN	NaN
NaN	0.00	NaN	NaN
NaN	0.00	NaN	NaN
NaN	0.00	NaN	NaN

⋮

신용카드를 사용하지 않은 시점의
기록에서는 결측치 발생

↓

소비가 없었다는 의미로 보고
0으로 교체

Feature Engineering
Installment_Set

주요 변수 설명

“GAP_DAYS” : 상환 예정일과 실제 상환일의 차이

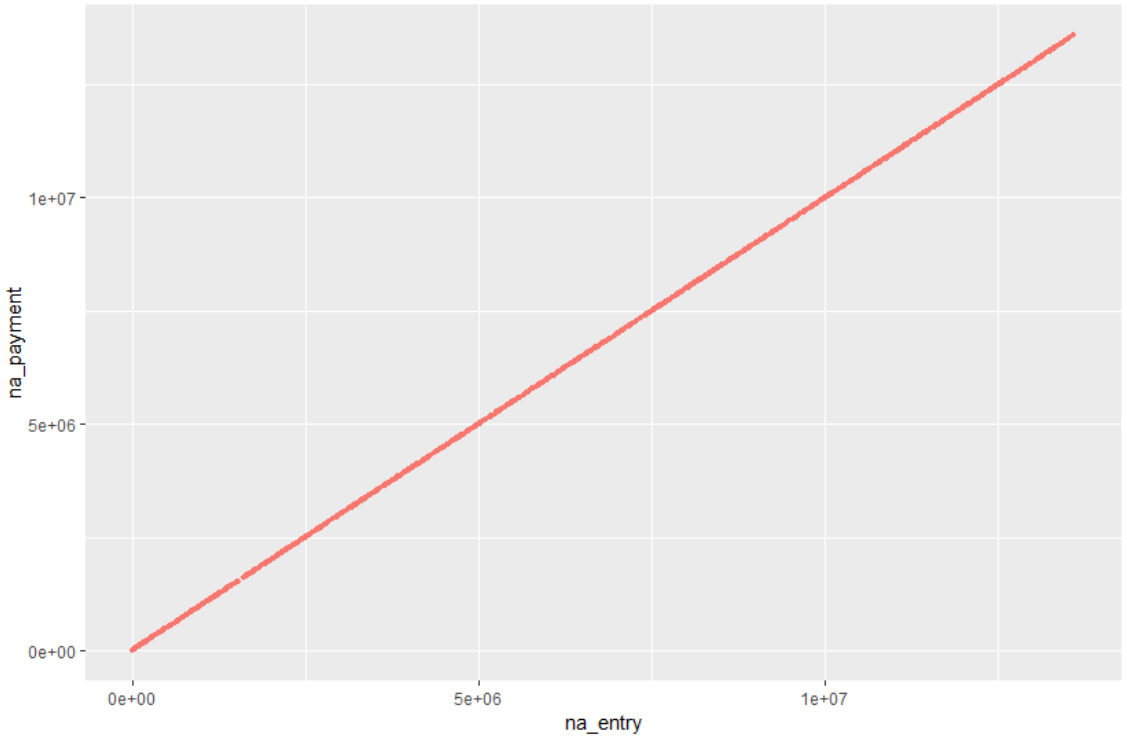
“GAP_AMOUNT” : 상환 예정 금액과 실제 상환 금액의 차이

“AVG_INTALLMENT” : 상환 예정 금액 평균

“AVG_PAYMENT” : 실제 상환 금액 평균

위의 4가지 값들과 함께 sum, average 등의 방법을 사용해 새로운 feature 생성해 냄
새로 생성된 변수 명 : SUM_GAP_DAYS, MEAN_GAP_DAYS, SUM_GAP_AMOUNT, MEAN_GAP_AMOUNT

Feature Engineering
Installment_Set



DAYS_ENTRY_PAYMENT(상환일)가 NA값을 가지는 index와
AMT_PAYMENT(상환액)가 NA값을 가지는 index가 정확히 일치

↓

해당 obs들은 해당 날짜(상환해야 하는 날짜)에 상환하지 못한 것으로 판단하여 0으로 교체

Feature Engineering
POS_Cash_Set

주요 변수 설명

"SK_DPD_SUM" : 상환 예정일과 실제 상환일의 차이

"SK_DPD_MEAN" : 상환 예정 금액과 실제 상환 금액의 차이

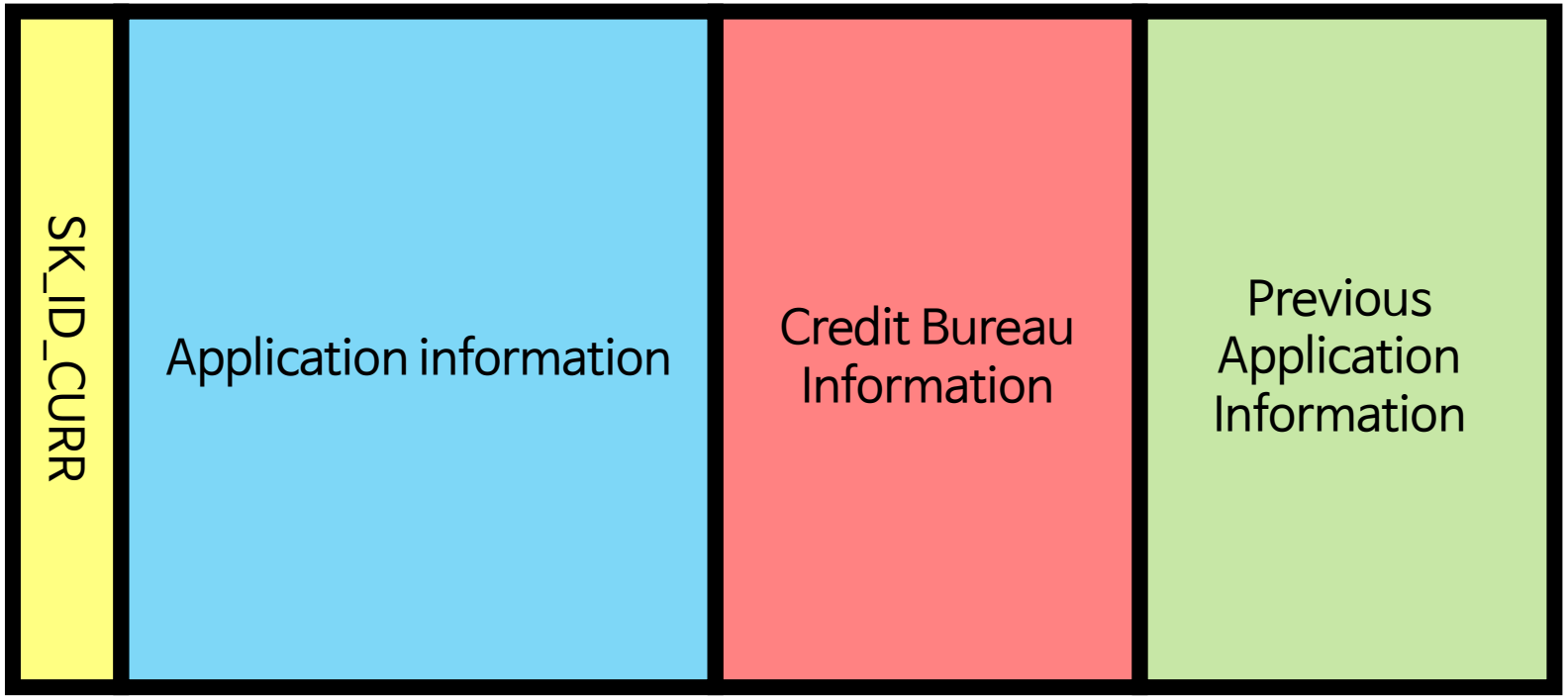
"SK_DPD_DEF_SUM" : 상환 예정일과 실제 상환일의 차이(낮은 부채는 무시)

"SK_DPD_DEF_MEAN" : 상환 예정 금액과 실제 상환 금액의 차이(낮은 부채는 무시)

데이터 통합

307511개의 obs와 237개의 variable을 가진
통합 데이터 셋 완성

데이터 통합



각각의 ID(Loan)가 해당 Loan과 연관된 기본적인 정보와 과거의 신용상품 기록 정보를 가지게 되었다

감사합니다