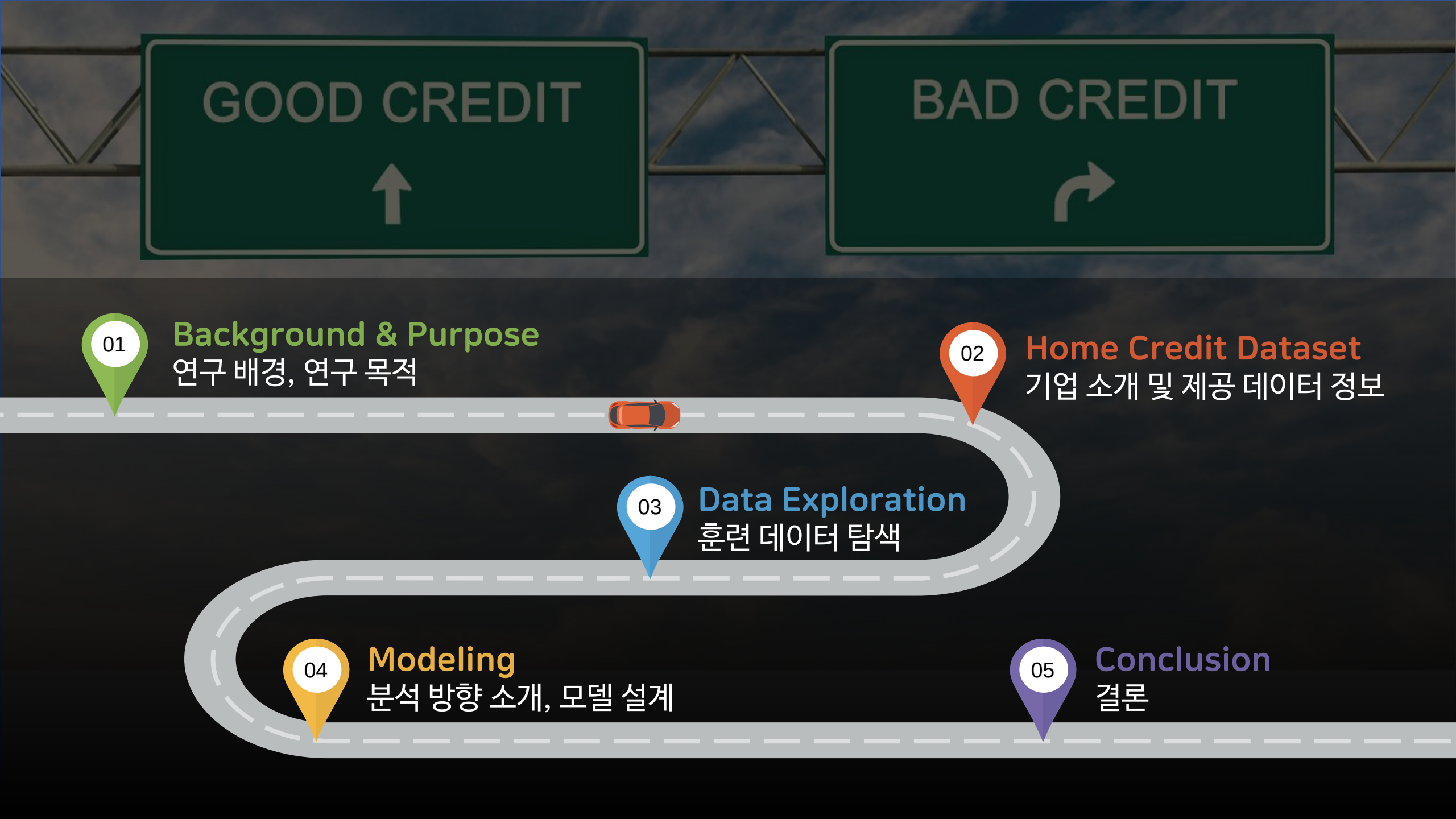
개인 신용도 예측 변수 분석

서울대학교 빅데이터 아카데미
2018-3 고급 빅데이터 분석 기법
BA 노은선 이현호 최의관



01 **Background & Purpose**
연구 배경, 연구 목적

02 **Home Credit Dataset**
기업 소개 및 제공 데이터 정보

03 **Data Exploration**
훈련 데이터 탐색

04 **Modeling**
분석 방향 소개, 모델 설계

05 **Conclusion**
결론

02

Home Credit Dataset
기업 소개 및 제공 데이터 정보

Home Credit Group 제공 데이터 사용

독립변수 X 소개

1. 일반 개인 정보
2. Credit Bereau 기반 정보
3. Home Credit 기반 정보

*변수 종류 : 총 221개

종속변수 Y 정의

1	0
실제 연체 고객 (상환 능력 부족)	그 외 고객

CB [Credit Bureau]

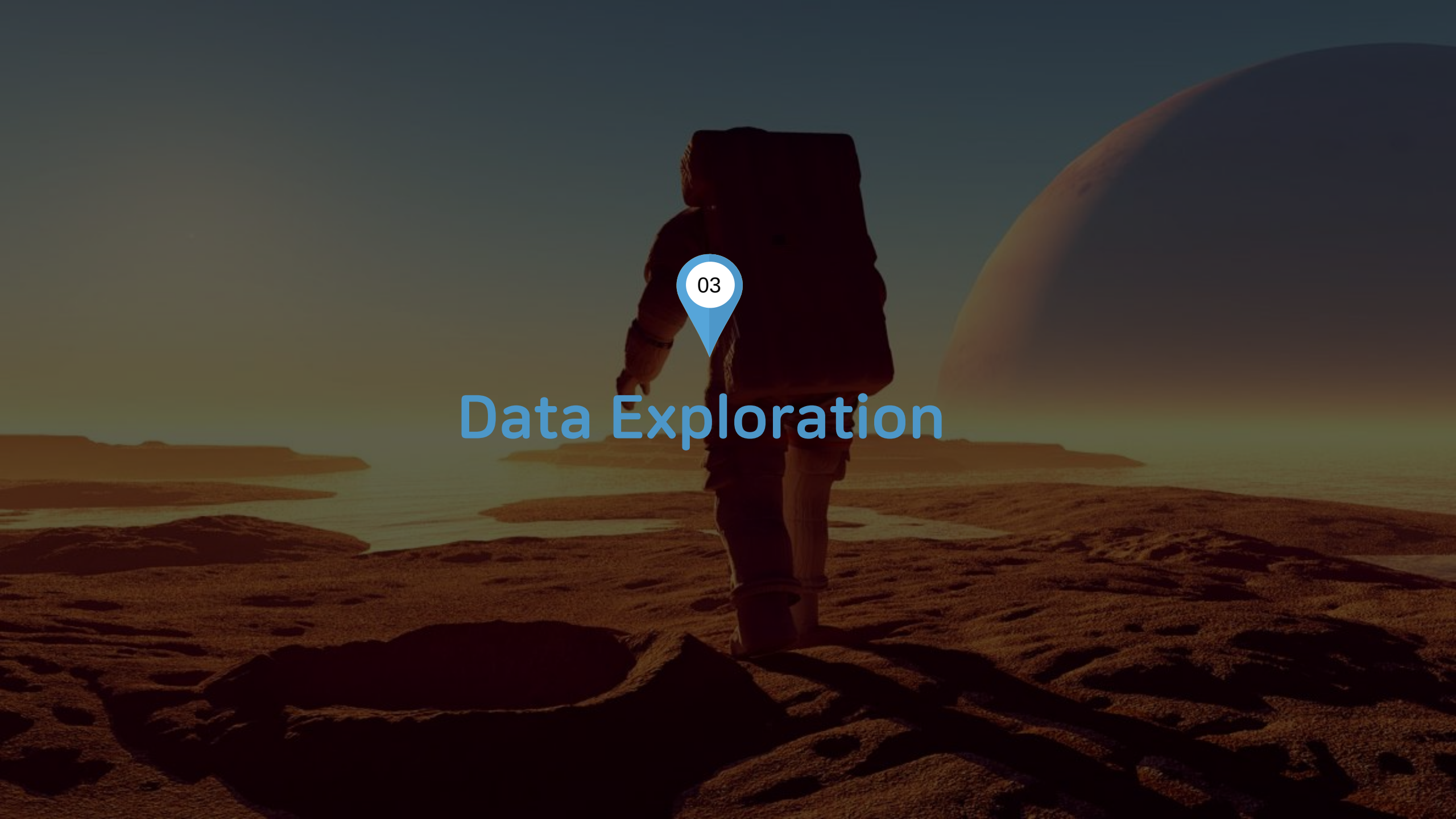
개인신용 관련 정보를 토대로 신용도를 평가하는 기관. 개인신용 관련 정보를 토대로 신용도를 평가하는 기관. 정보를 취합하고 평가하는 데 그치지 않고 직접 신용등급을 매기며, 이 신용등급은...

02

Home Credit Dataset

기업 소개 및 제공 데이터 정보

Client data	application_{train|test}	Train & Test data (ex) 성별, 자가유무, 자차유무, 자녀 수, 수입 등	124 columns	158MB /24MB
Credit Bureau	bureau	고객 신용 데이터 (ex) CB 기반 신용 상태, 신용 유지기간, 빚, 신용 유형	17 columns	162MB
	bureau_balance	월별 Balance 데이터 (ex) 월별 대출 상태	3 columns	358MB
Home Credit	POS_CASH_balance	HC기반 대출 정보 (ex) 신용점수 유지 기간,현재 계약 현황, 만기기한	8 columns	374MB
	credit_card_balance	기존 신용카드 대출 정보 (ex) 예전 신용 대출 계약 상태, 예전 채권 총액	22 columns	404MB
	previous_application	대출 status 데이터 (ex) 선급금, 이자, 지불 방법	25 columns	386MB
	installments_payments	대출 상환 관련 데이터 (ex) 대출 신청일, 할부 금액	7 columns	689MB
Total	**7 files**		**221 columns**	**2.49GB**

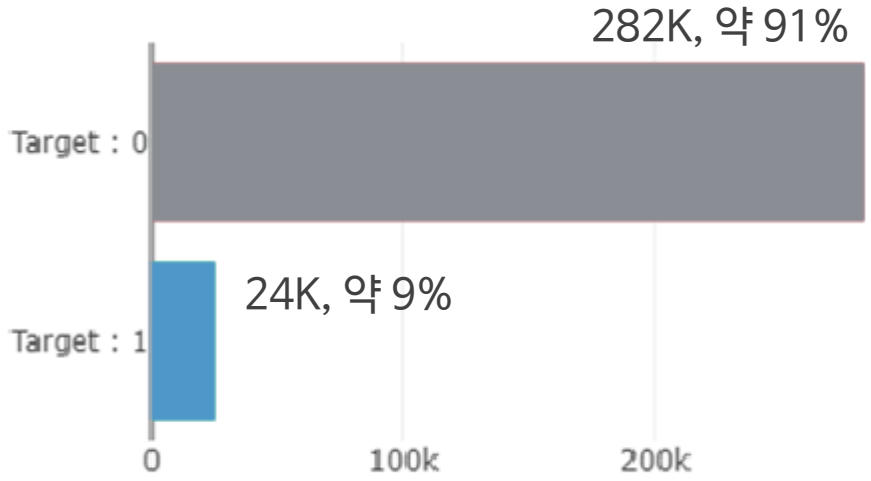

03
Data Exploration

03

Data Exploration

훈련 데이터 탐색 - Target

0. Target 변수



Target 0 : 주어진 기간 내
상환에 문제가 발생하지 않은 고객

Target 1 : 대출/분할 할부 상환 문제가 있는 고객

→ 정상 & 문제 고객 데이터 간 불균형 확인 가능

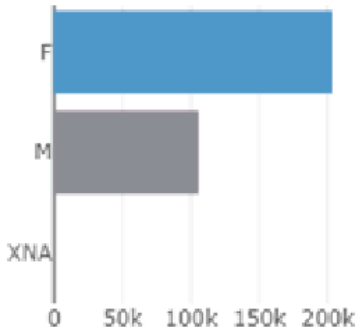
03

Data Exploration

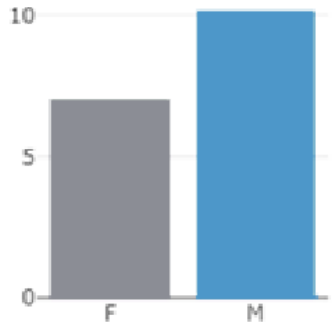
훈련 데이터 탐색 - Gender and Marriage

1. 대출 신청자 성별

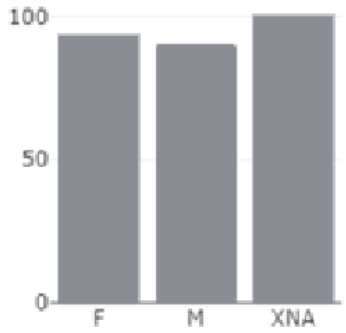
전체 성별 분포



체납자 성별 분포



정상고객 성별 분포

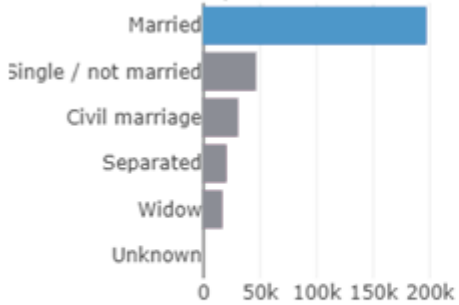


대출 신청자 비율	대출 상환 문제 비율
남성 << 여성	남성 > 여성

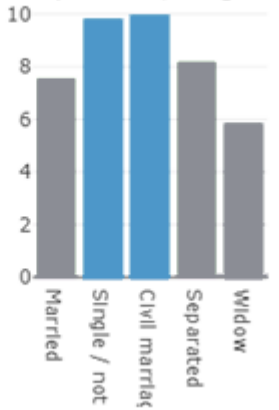
Variable name : CODE_GENDER

2. 대출 신청자의 결혼 형태

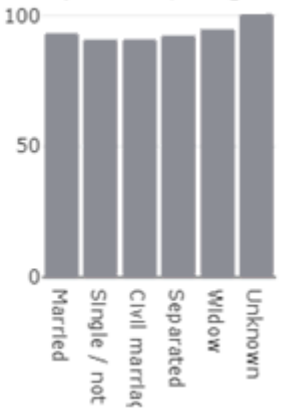
전체 결혼 상태



체납자 결혼 상태



정상고객 결혼 상태



가장 높은 신청자 유형	대출 상환 문제 비율
Married	Civil marriage & Single

Variable name : NAME_FAMILY_STATUS

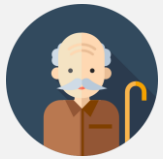
Data Exploration
훈련 데이터 탐색 - Age

TARGET
0
1

3. 대출 신청자 연령

연령별 체납자 비율

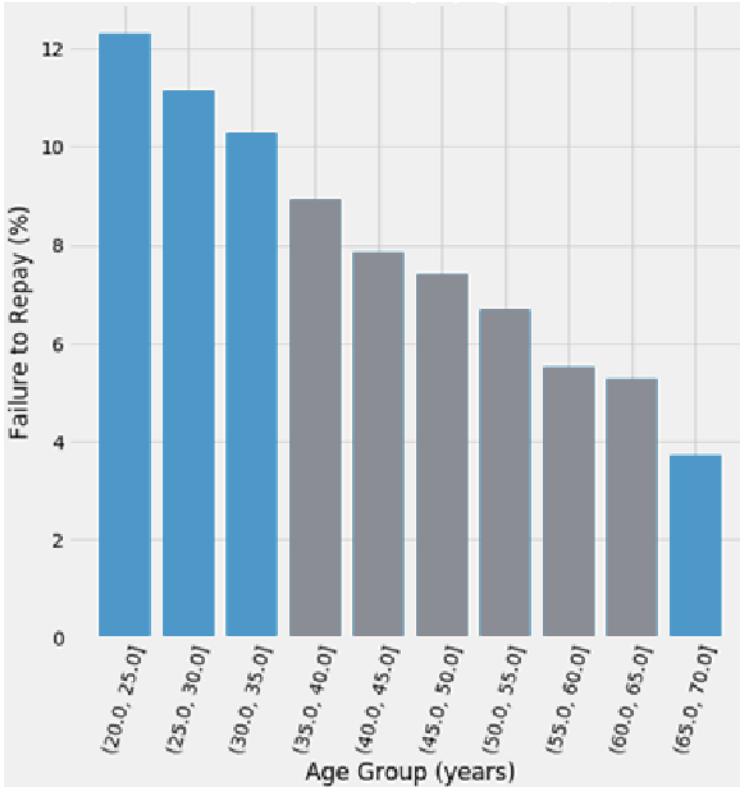
Over 10%
20 to 35



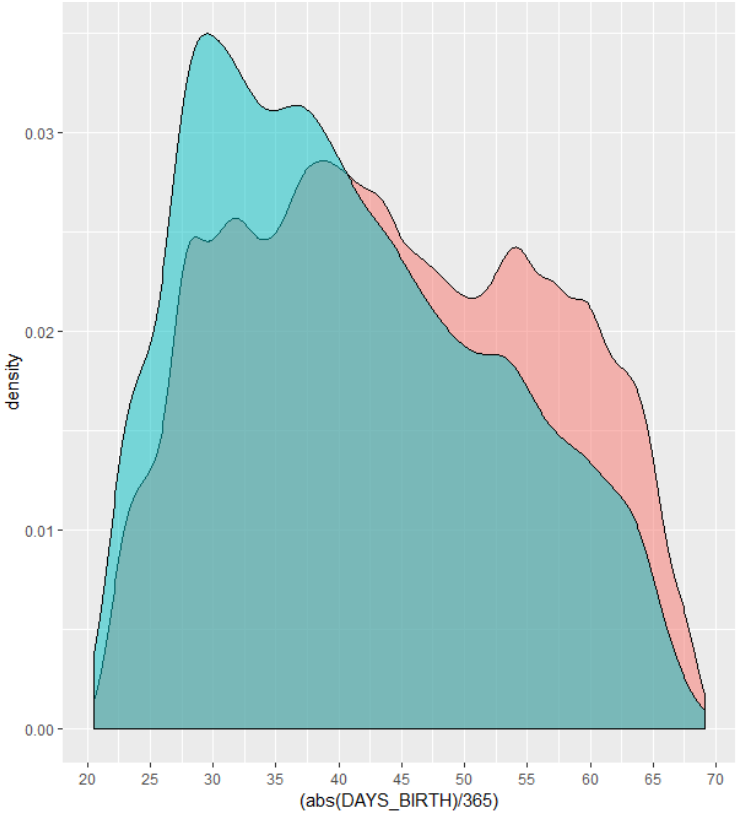
Under 4%
65 to 70

Variable name : DAYS_BIRTH

체납자 연령 분포 (percentage)



고객 상태별 연령 분포 (density)



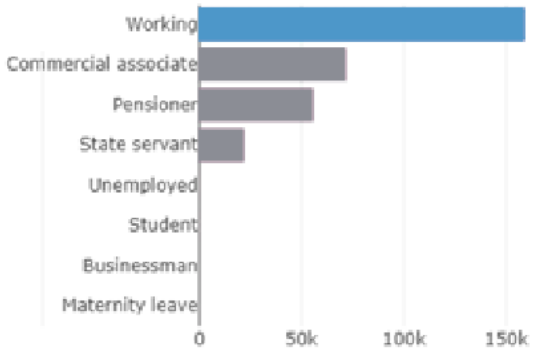
The plots show a clear trend!

Data Exploration
훈련 데이터 탐색 - Income types

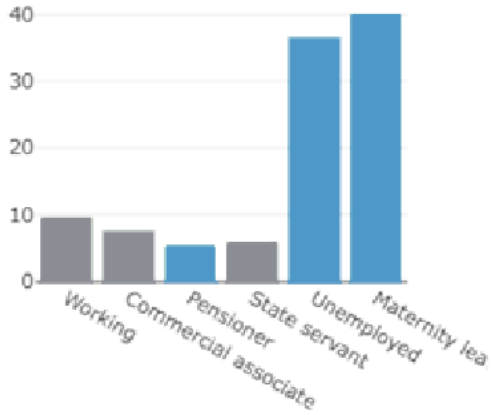
TARGET
0
1

4-1. 소득 형태

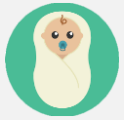
전체 소득 상태 분포



체납자 소득 상태 분포



소득 형태별 체납자 비율



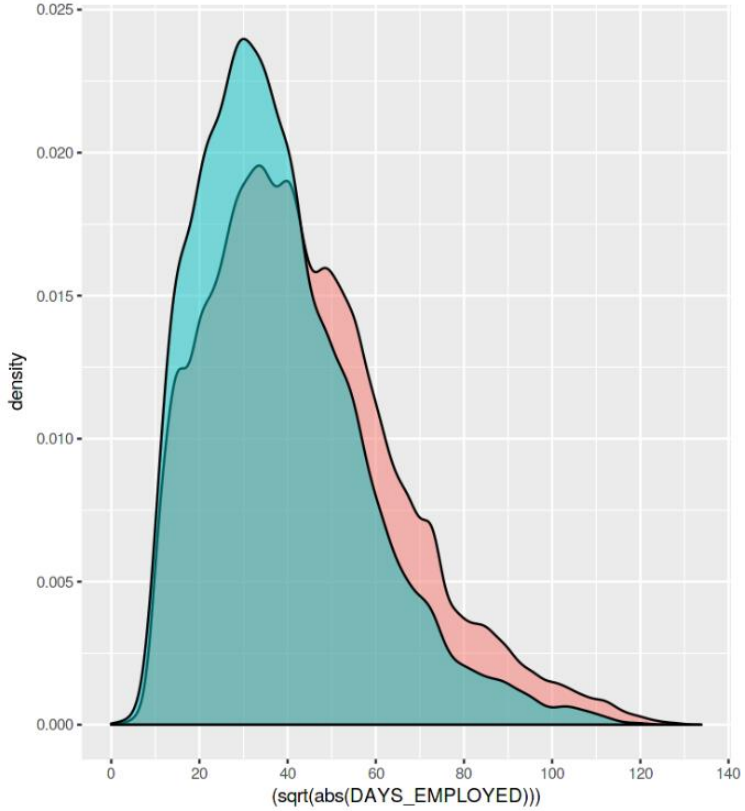
40% Maternity Leave

36% Unemployed

5% Pensioner

Variable name : NAME_INCOME_TYPE

4-2. 고객 상태별 근무 일수



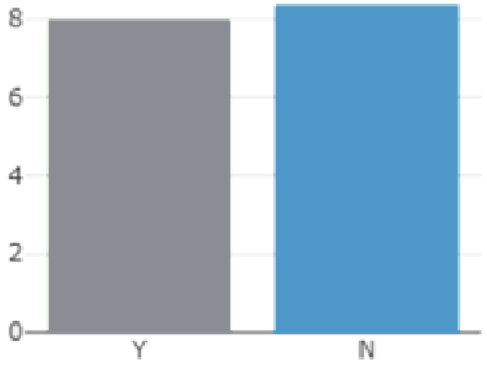
Variable name : DAYS_EMPLOYED

Data Exploration
훈련 데이터 탐색 - Own real estate or car

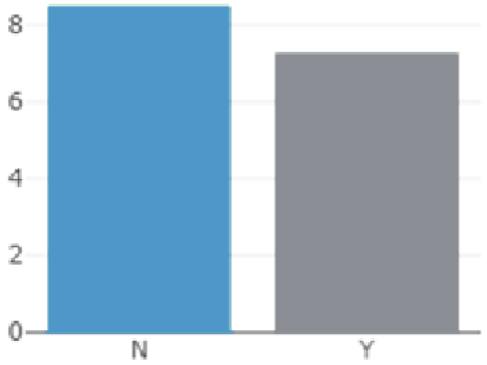
TARGET
0
1

5-1. 자가 및 자차 소유 정보

체납자 중 자가 소유 유무
FLAG_OWN_REALTY



체납자 중 자차 소유 유무
FLAG_OWN_CAR

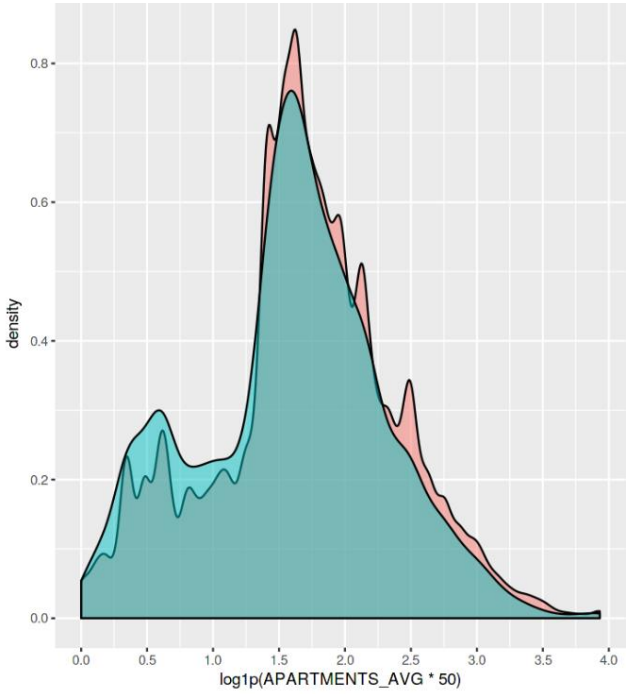


전체 소유 비율		체납자 특징
69%	34%	People who do not have
		 or 

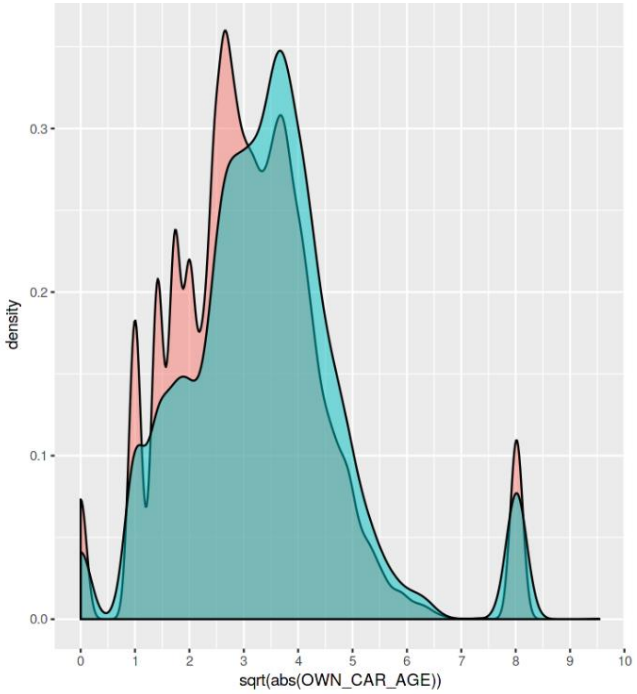
But there's not much difference
than we expected

5-2. 고객 상태별 자가 및 자차 상태

고객 상태별 아파트 연식
APARTMENTS_AVG

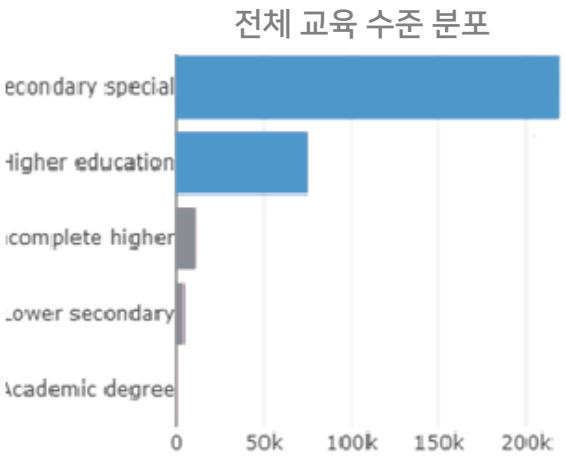


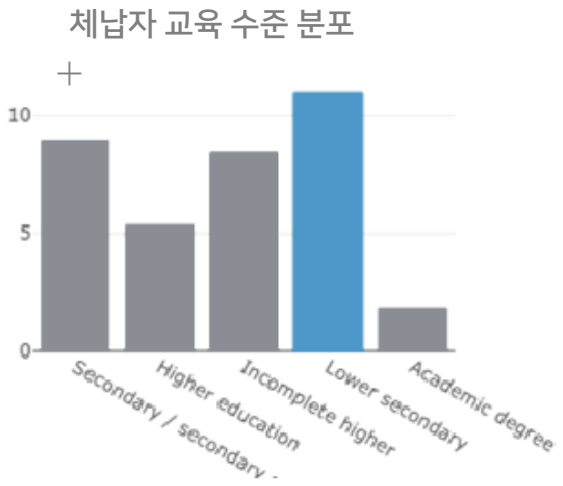
고객 상태별 차량 연식
OWN_CAR_AGE



Data Exploration
훈련 데이터 탐색 - Education and Housing Type

6. 고등 교육 수준

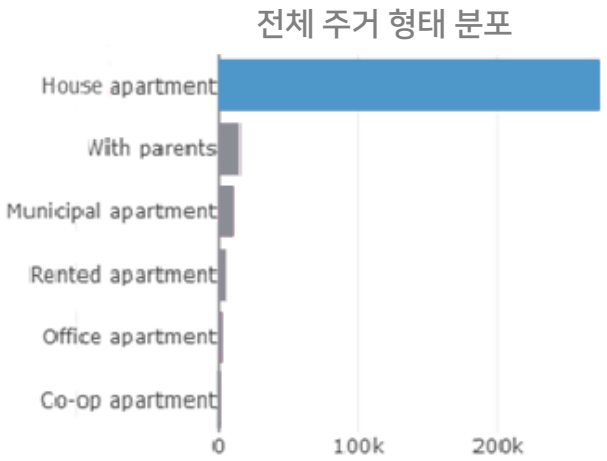


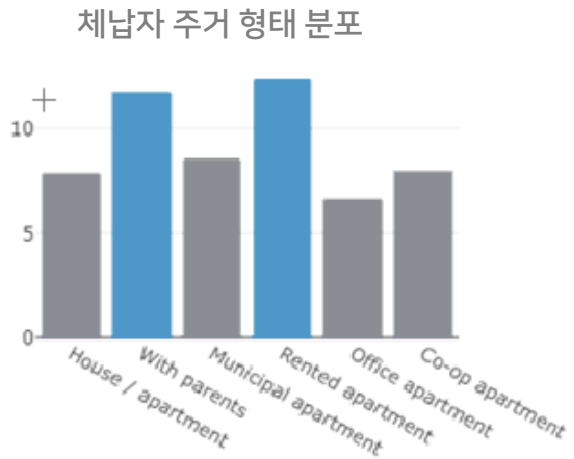


전체 교육 수준	체납자 교육 수준
71% & 24% Secondary & Higher education	**10.9%** Lower secondary

Variable name : NAME_EDUCATION_TYPE

7. 주거 형태





전체 주거 형태	체납자 주거 형태
88% House & Apartment	**12% & 11%** Ranted APT & With parents

Variable name : NAME_HOUSING_TYPE

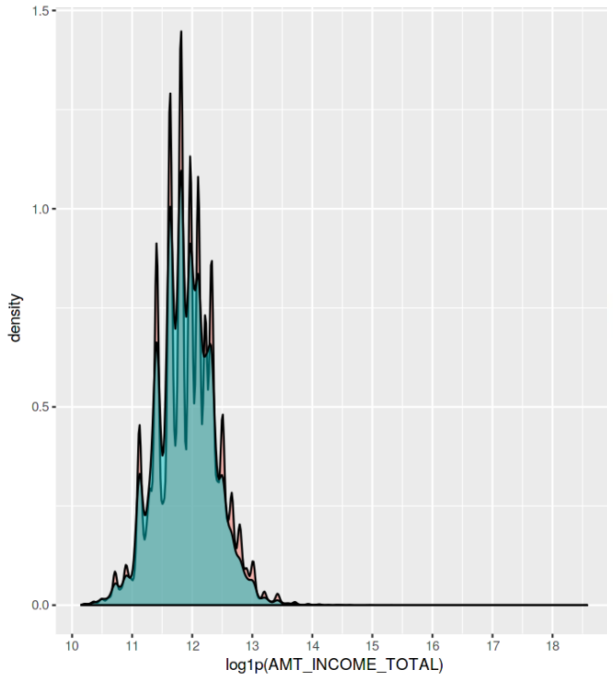
03

Data Exploration

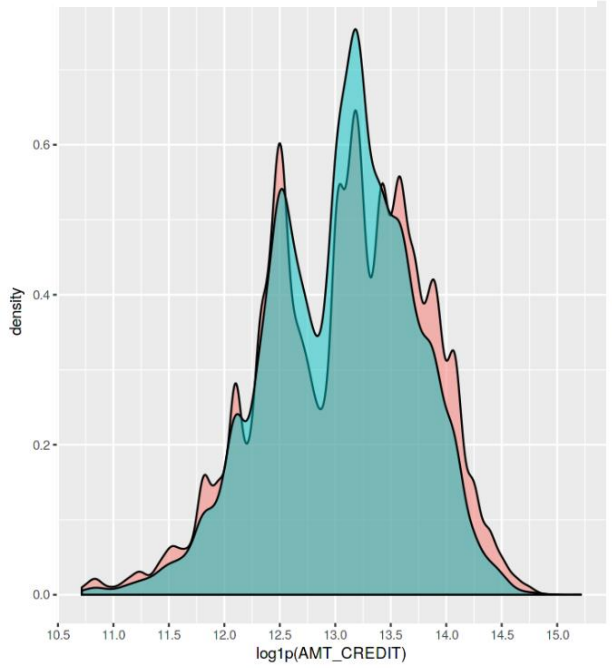
훈련 데이터 탐색 - 고객 정보 내 금액 관련 데이터

TARGET
0
1

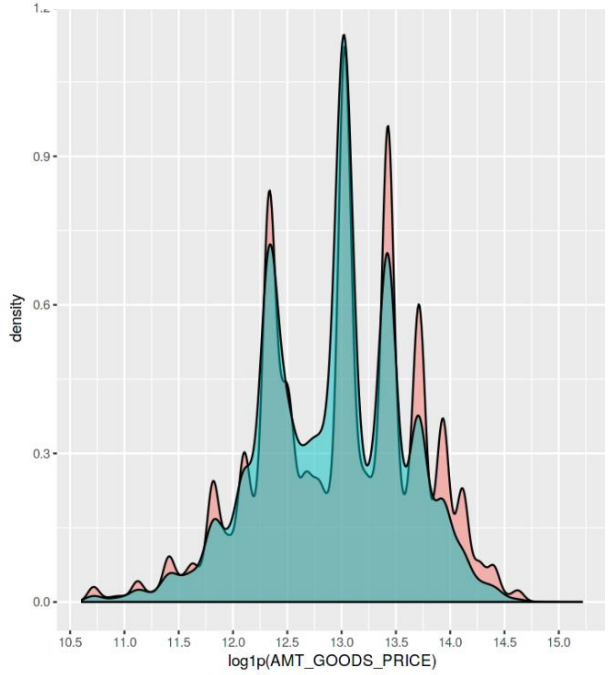
고객 유형별 전체 수익
AMT_INCOME_TOTAL



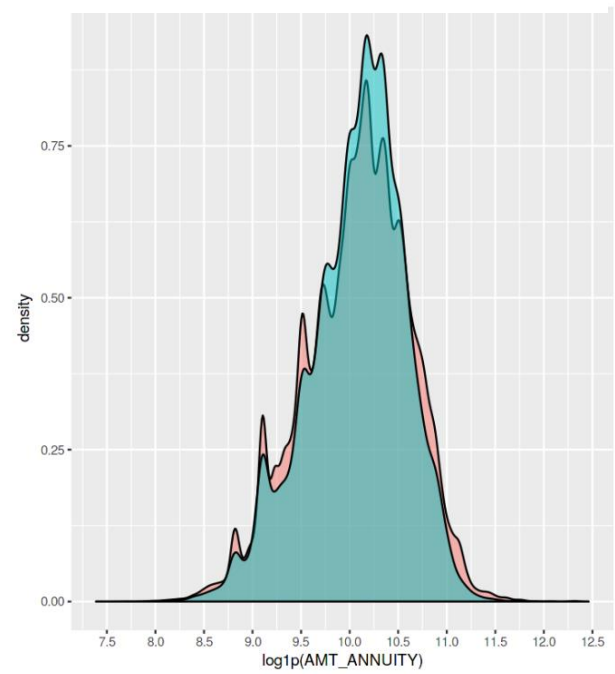
고객 유형별 현재 대출 금액
AMT_CREDIT



고객 유형별 대출 신청 금액
AMT_GOOD_PRICE



고객 유형별 담보 연금 금액
AMT_ANNUITY



고객 수입, 전체 대출 금액 등 금액 관련 변수들은
일부 체불자의 비율이 높아지는 구간이 있는 것 같다.

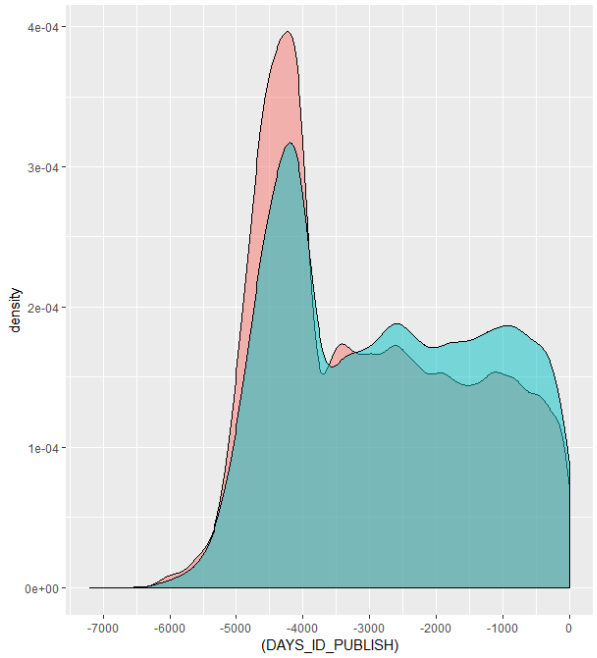
03

Data Exploration

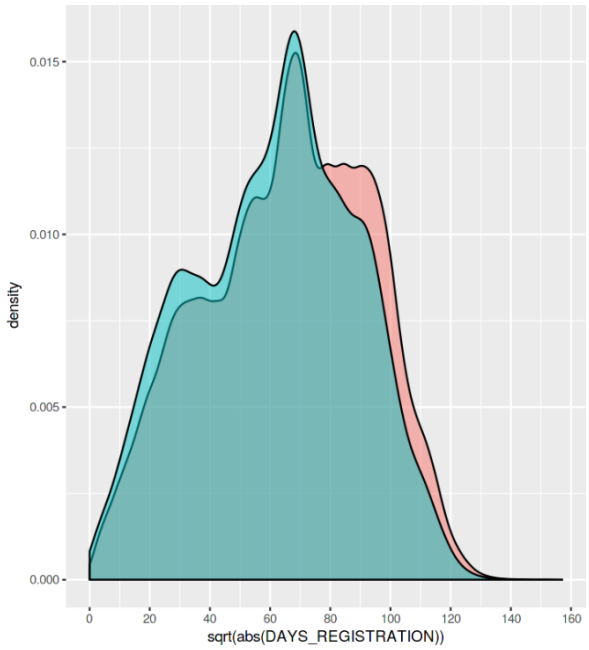
훈련 데이터 탐색 - 좀 더 소소한 개인 정보

TARGET
0
1

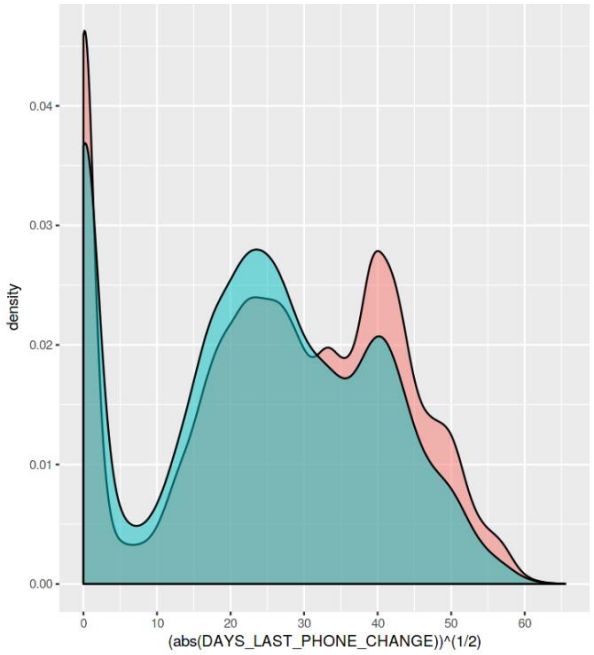
최근 신용 관련 자료 업데이트 일자
DAYS_ID_PUBLISH



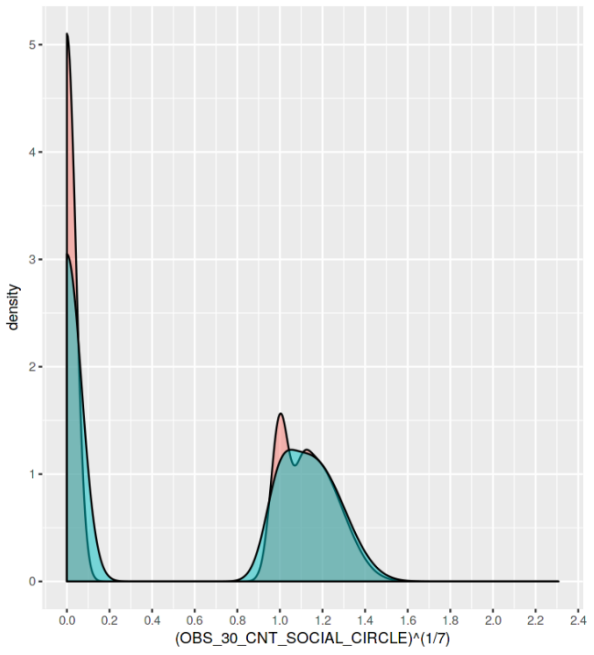
최근 개인 정보 업데이트 일자
DAYS_REGISTRATION



최근 휴대폰 변경 일자
DATS_LAST_PHONE_CHANGE



30일 기준 고객 사회적 환경 관찰 자료
CNT_SOCIAL_CIRCLE



체납자의 경우 정도가 약하나 정보가 최근에 변경된 경우가 더 많고,
사회적 환경 관찰 자료는 크게 유의미한 내용은 없는 것으로 보인다.

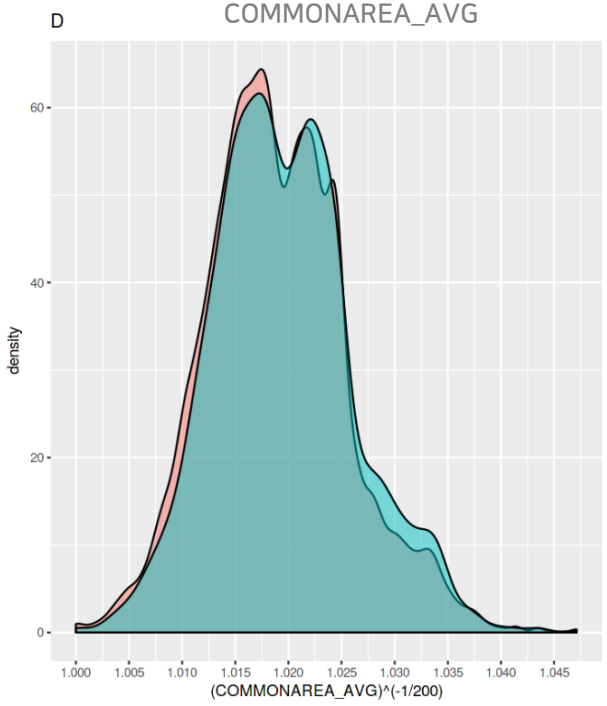
03

Data Exploration

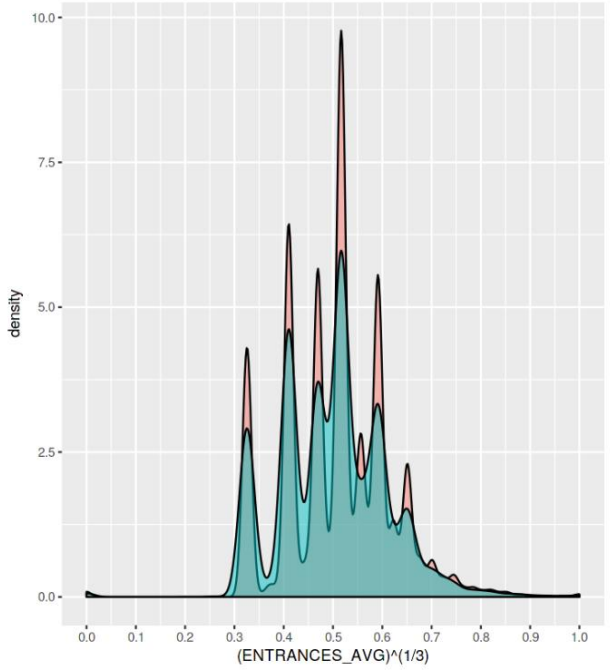
훈련 데이터 탐색 - 디테일한 주거 형태

TARGET
0
1

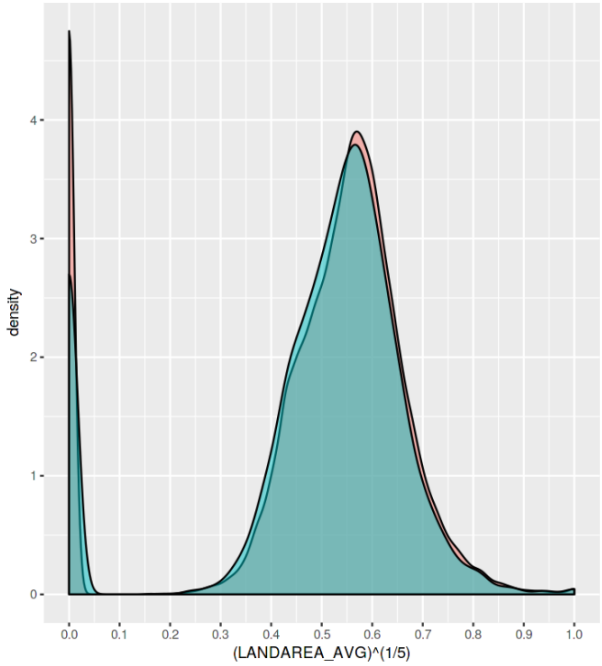
지형 전체 넓이
COMMONAREA_AVG



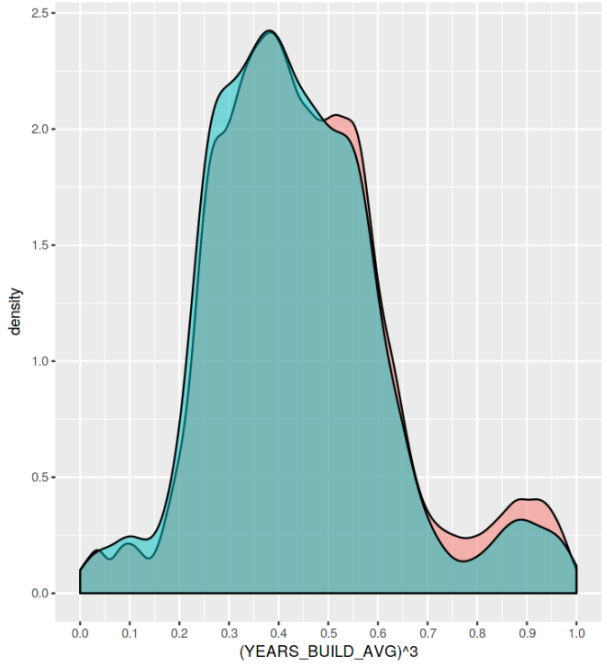
입구 넓이
ENTRANCES_AVG



집 전체 넓이
LANDAREA_AVG



건물 연식
YEARS_BUILD_AVG



고객 정보 접근에 있어 신선한 시도였지만, 그래프상 큰 차이는 없는 것으로 확인된다.

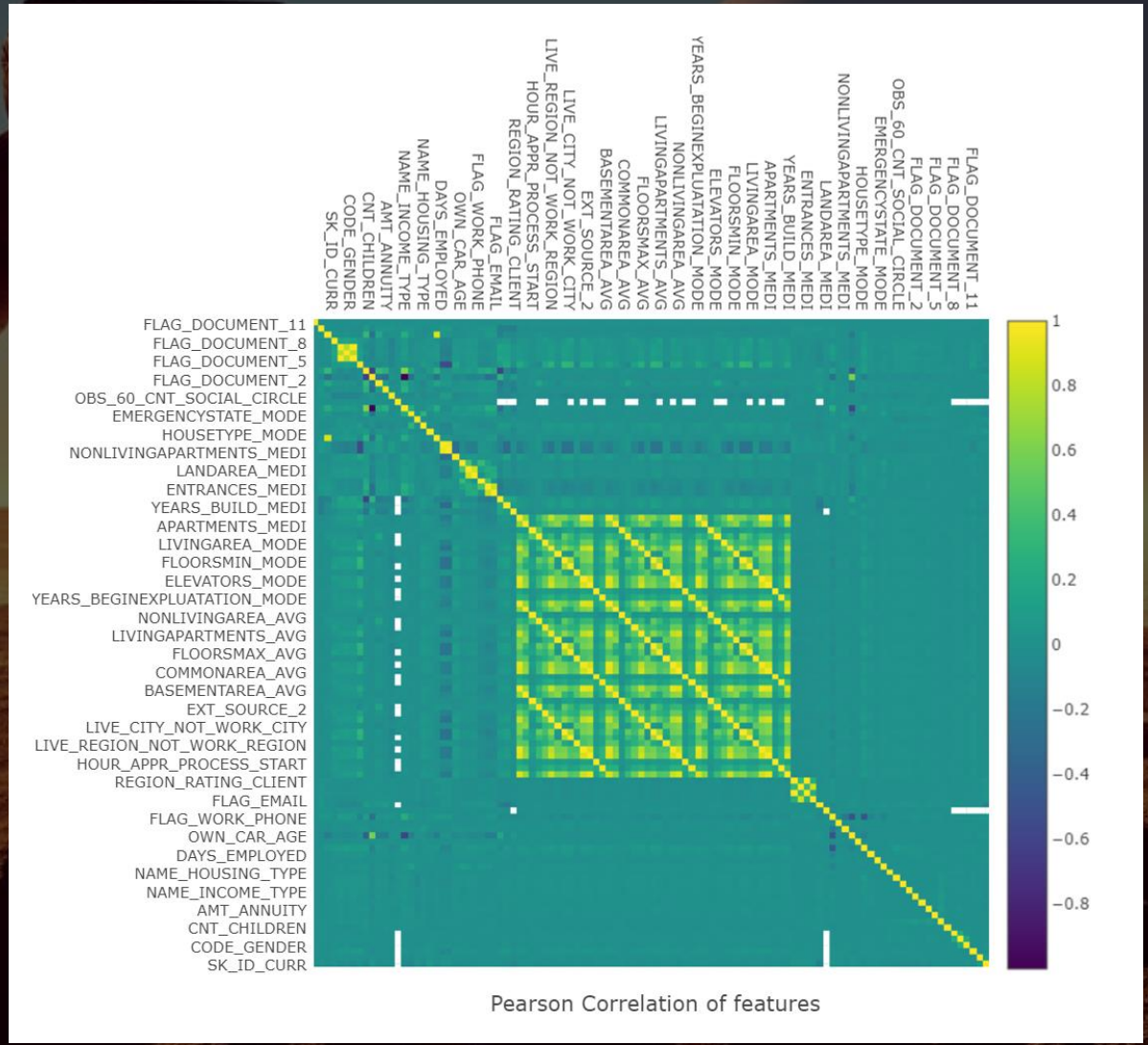
03

Data Exploration

Pearson Correlation of features

일부 변수간 높은 상관관계

→ 파생변수로 인한 것으로 확인되었음




04
Modeling

04

Modeling

Train & Validation set, Model selection

Kaggle dataset

train.csv

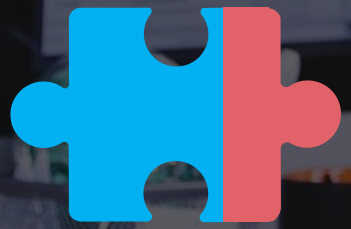
&

test.csv

Our new train/ validation set

train.csv

Random sampling (cv = 5fold) →



90% train set
10% validation set

Model selection

Logistic Regression

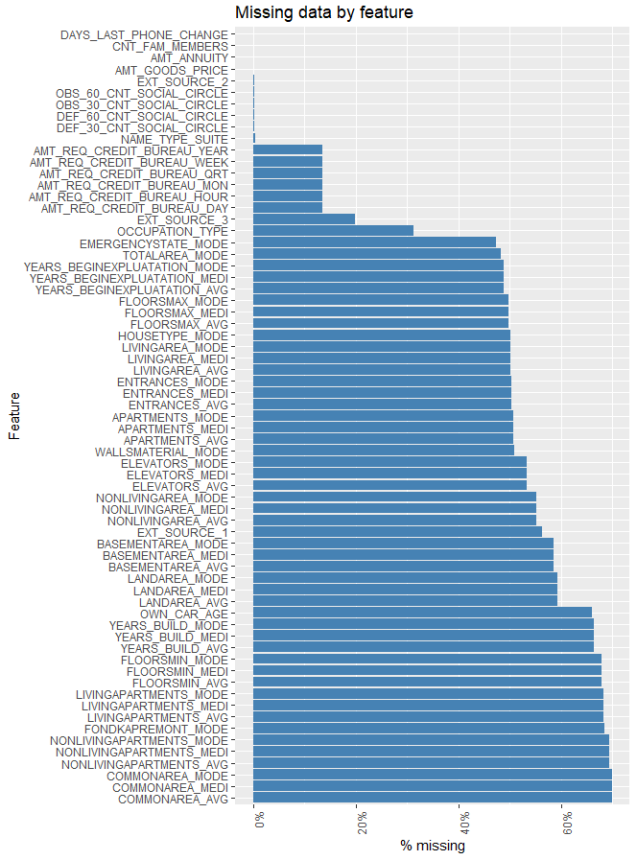
1. Logistic Regression
2. Regularized Logistic Regression

Bagging & Boosting

3. Logit Boosting
4. Logistic Bagging
5. XGBOOST

04

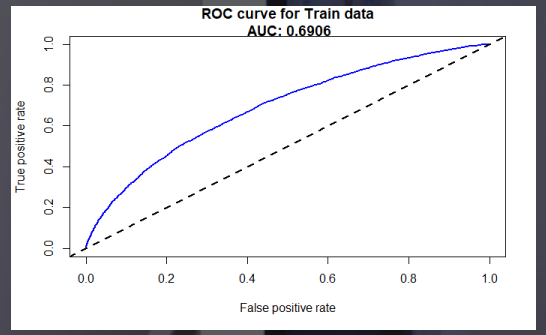
Modeling
Missing values



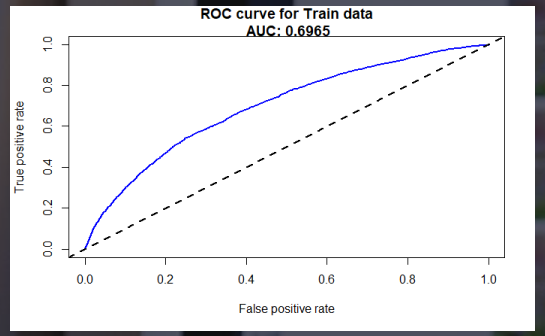
	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5

There are too many missing values!

Modeling
Missing values

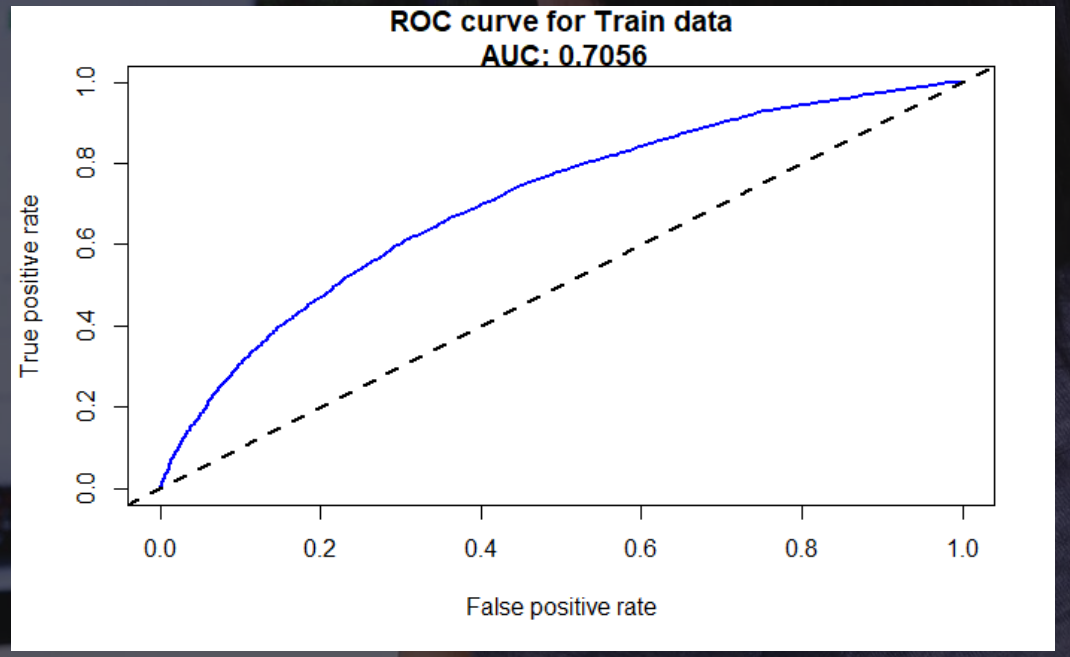


XGboost
(Remove NA values)
AUC : 0.6906



XGboost
(MICE package)
AUC : 0.6965

*** Integer -> NA=0,**
Character(factor) -> 'No value'

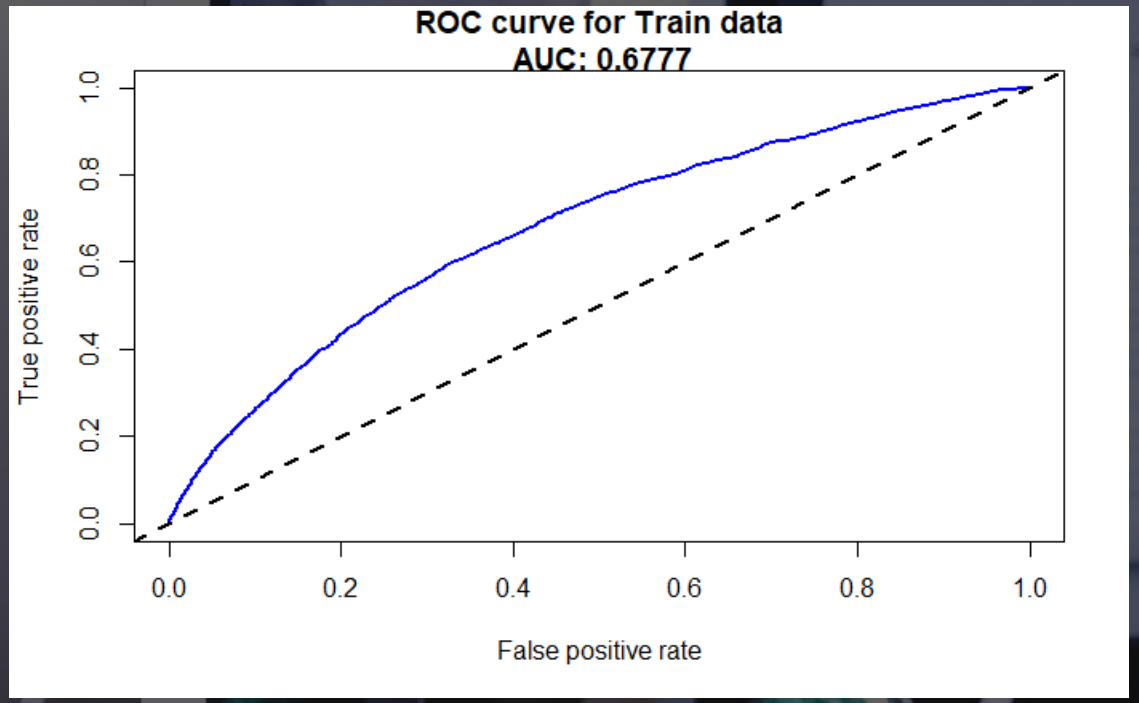


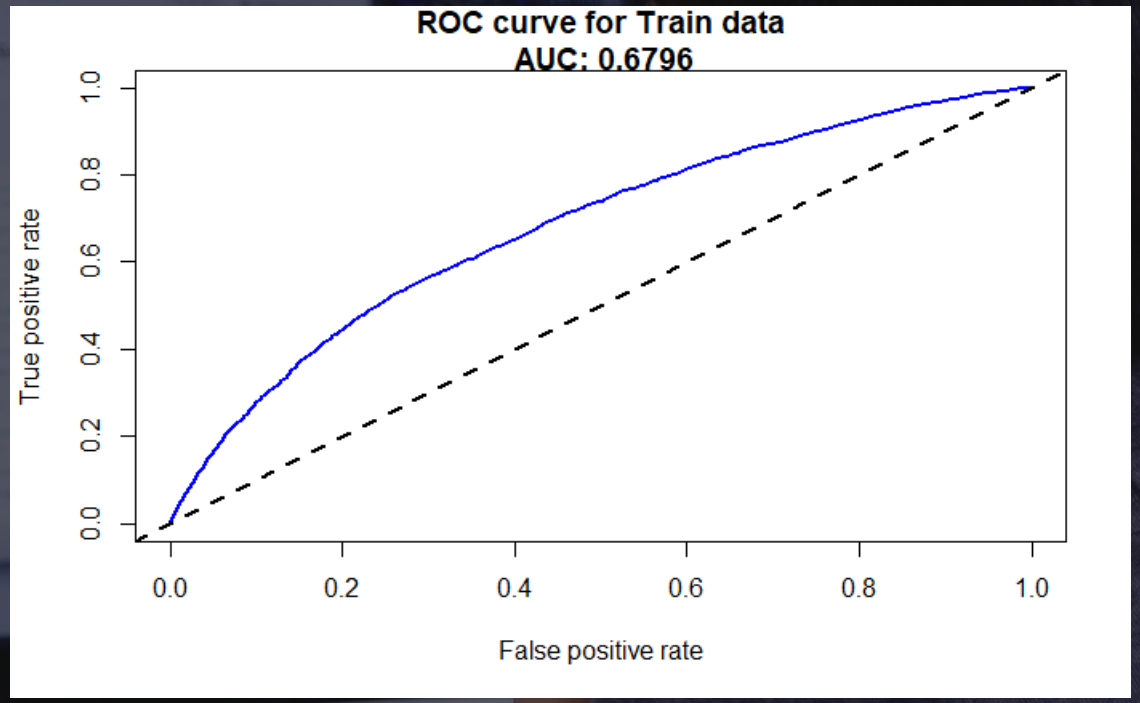
XGboost
(Replace NA values*)
AUC : 0.7056

04

Modeling

Model Selection - Logistic Regression





1. Logistic Regression
AUC : 0.6777

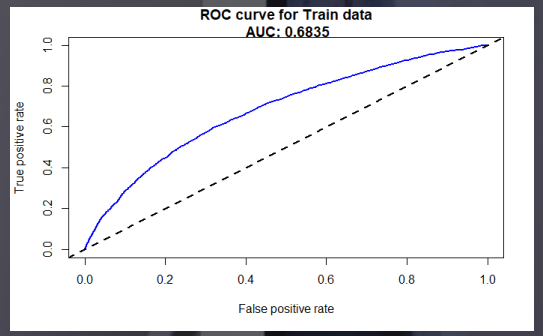
<

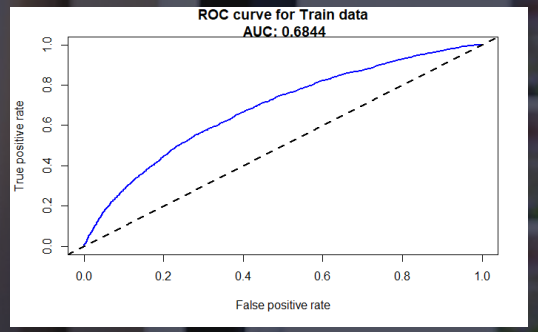
2. Regularized Logistic Regression
AUC : 0.6796

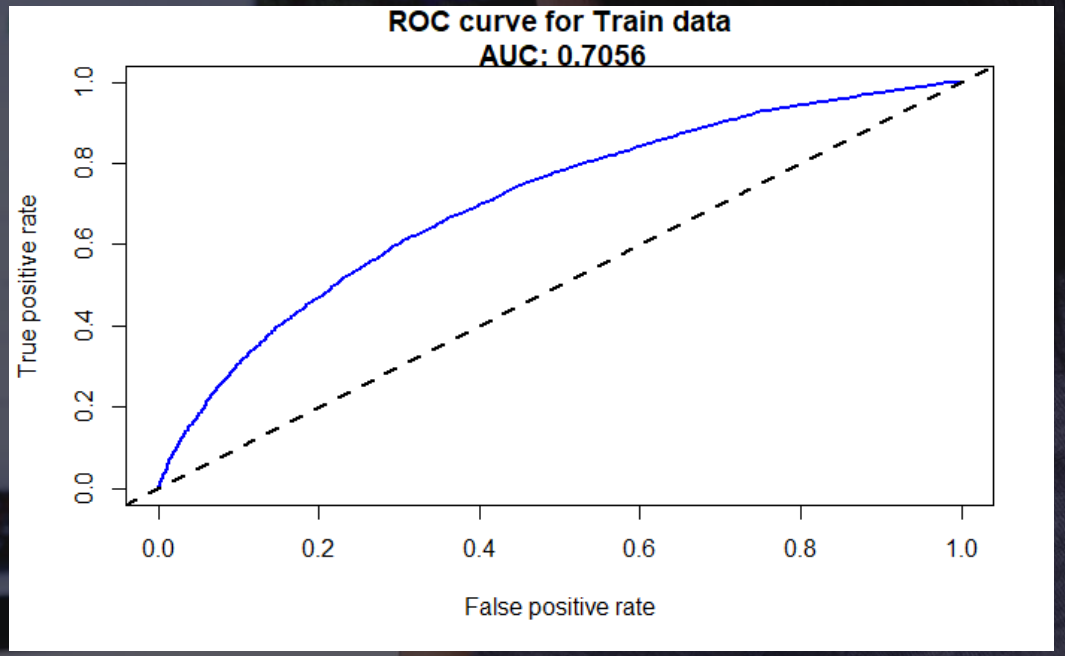
04

Modeling

Model Selection - Boosting & Bagging



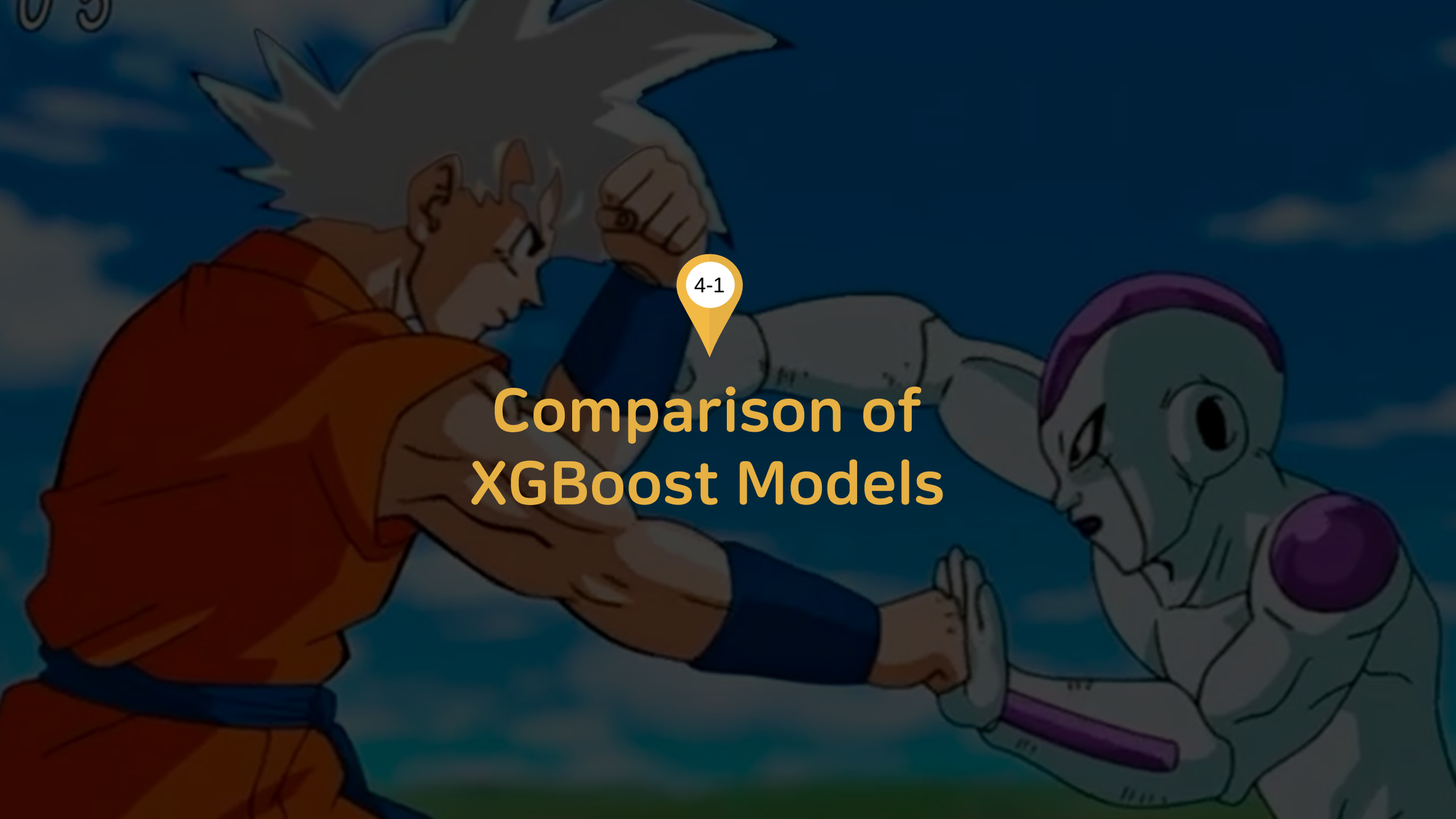




3. Logit Boosting
AUC : 0.6835

4. Logistic Bagging
AUC : 0.6844

5. XGBoost
AUC : 0.7056



Comparison of XGBoost Models

4-1

Comparison of XGBoost Models

EXT_SOURCE values

Value Description says….

EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source

EXT_SOURCE values(1~3) :
고객 정보가 아닌 Home credit 자체 정규화 된 신용 데이터로,

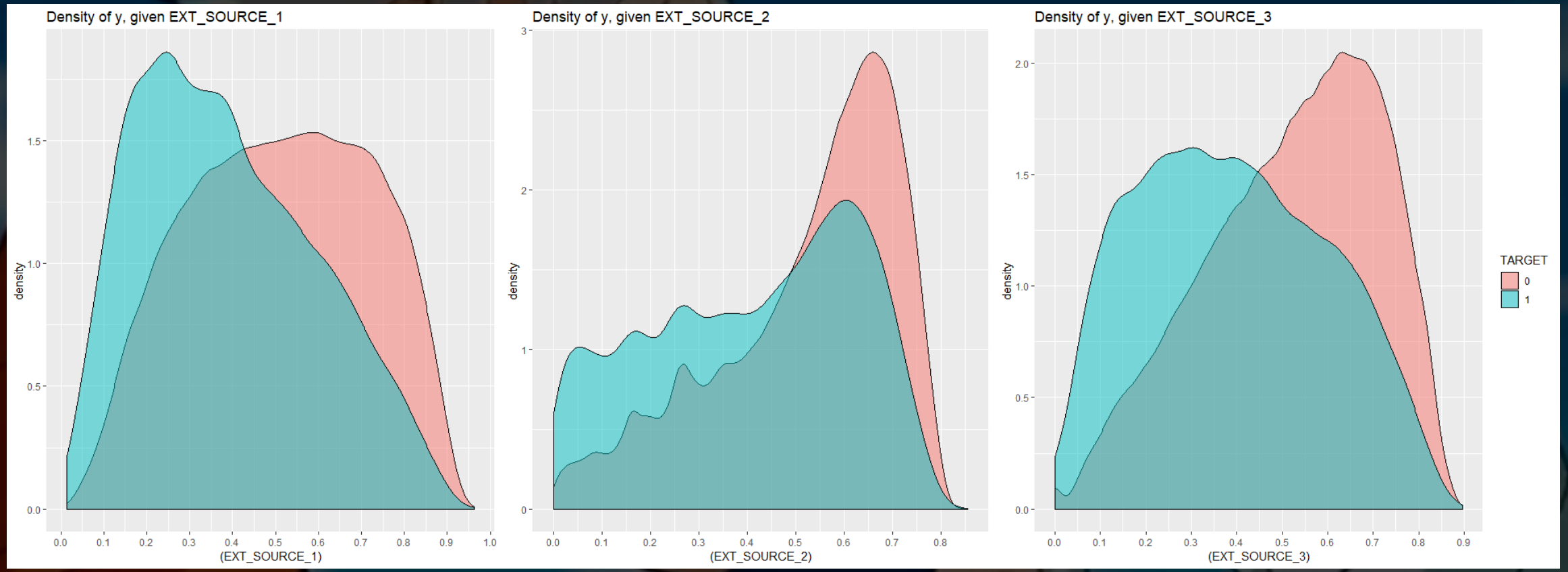
데이터를 구성하는 정보에 대한 설명이 전혀 없지만
예측력을 향상시키는데 매우 도움이 되는 변수
→ 실제 고객 평가 시 사용되는 지표일 가능성

4-1

Comparison of XGBoost Models

EXT_SOURCE values - Target별 분포

TARGET
0
1

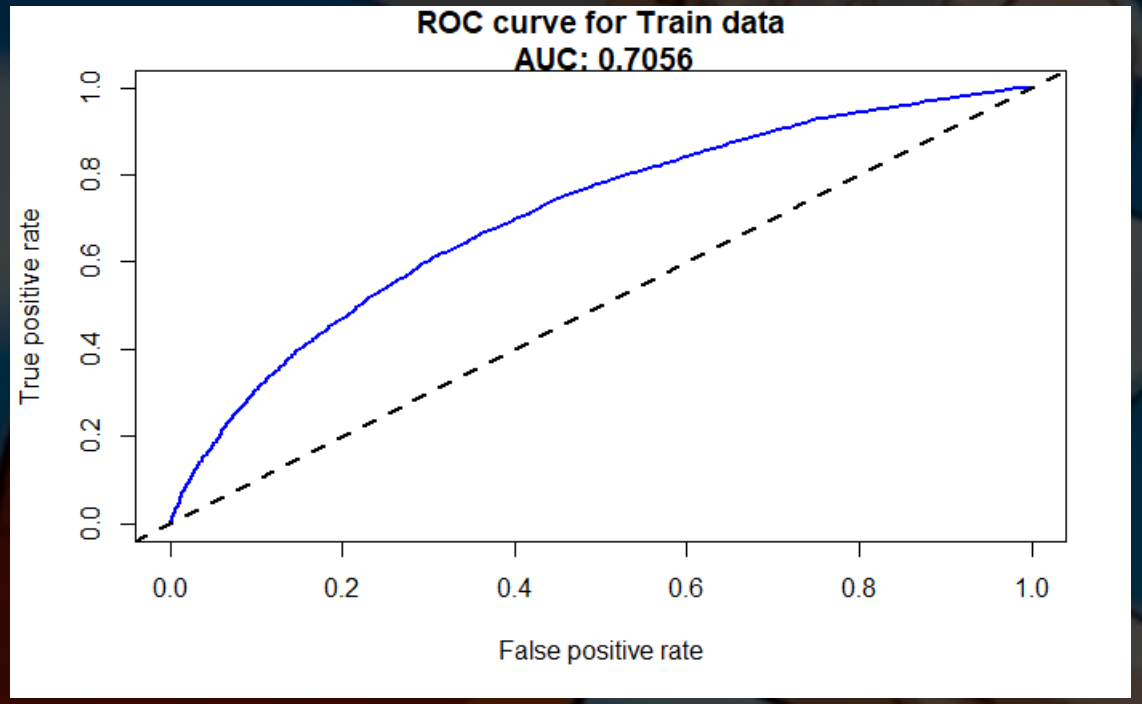


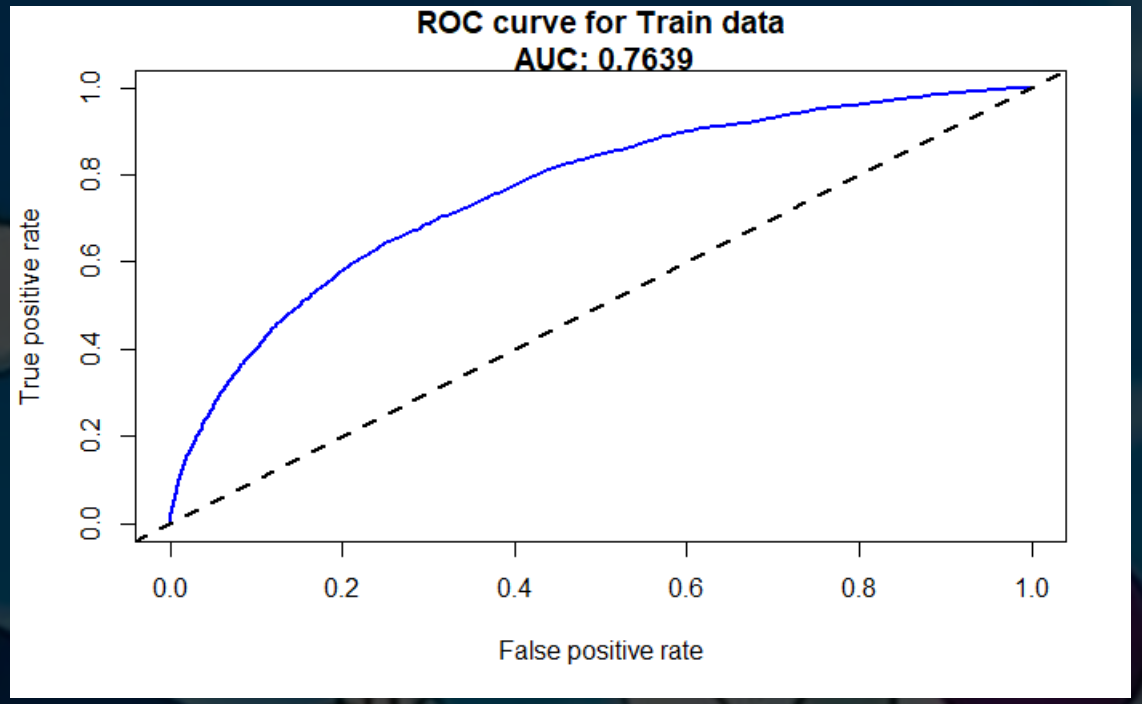
We sure they know something …

4-1

Comparison of XGBoost Models

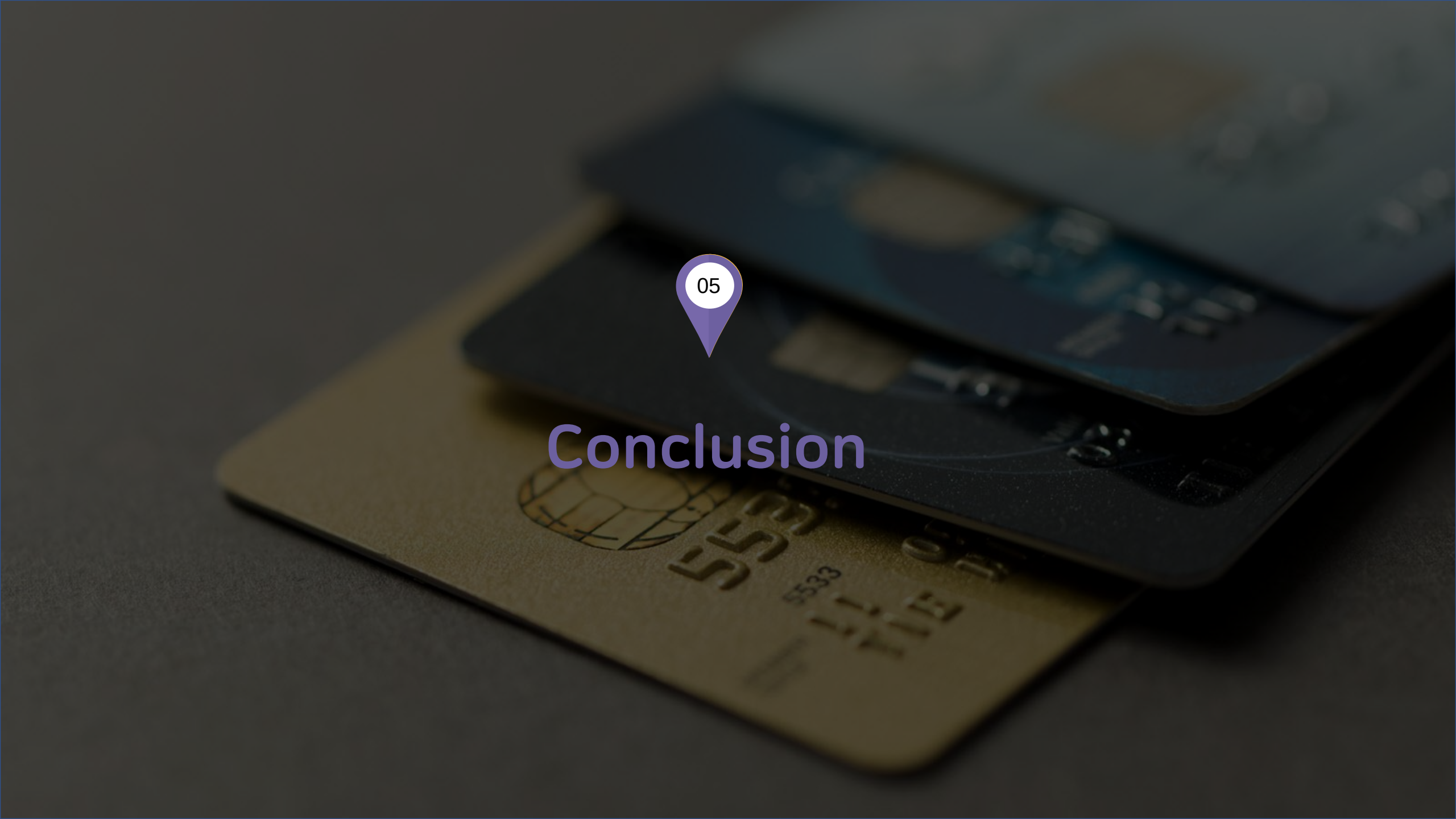
EXT_SOURCE values





XGBoost
(Exclude EXT_SOURCE values)
AUC : 0.7056

XGBoost
(Include EXT_SOURCE values)
AUC : 0.7639


05
Conclusion

05

Conclusion

Coefficient (L1, L2 규제 Elastic Net) - Regularized Logistic Regression

변수명	계수
CODE_GENDERM	-0.1733931
FLAG_OWN_CARY	0.1669117
AMT_CREDIT	-0.8405804
AMT_ANNUITY	-0.1027895
AMT_GOODS_PRICE	0.9639024
DAYS_BIRTH	-0.2049486
DAYS_ID_PUBLISH	-0.1085002
REGION_RATING_CLIENT_W_CITY	-0.1367554
DAYS_LAST_PHONE_CHANGE	-0.1683624
FLAG_DOCUMENT_3	-0.1067521
Bias	2.6002799

높은 계수를 가진 변수

전체 대출 금액과 대출 상품 가격의 계수가 가장 높게 나타났다.

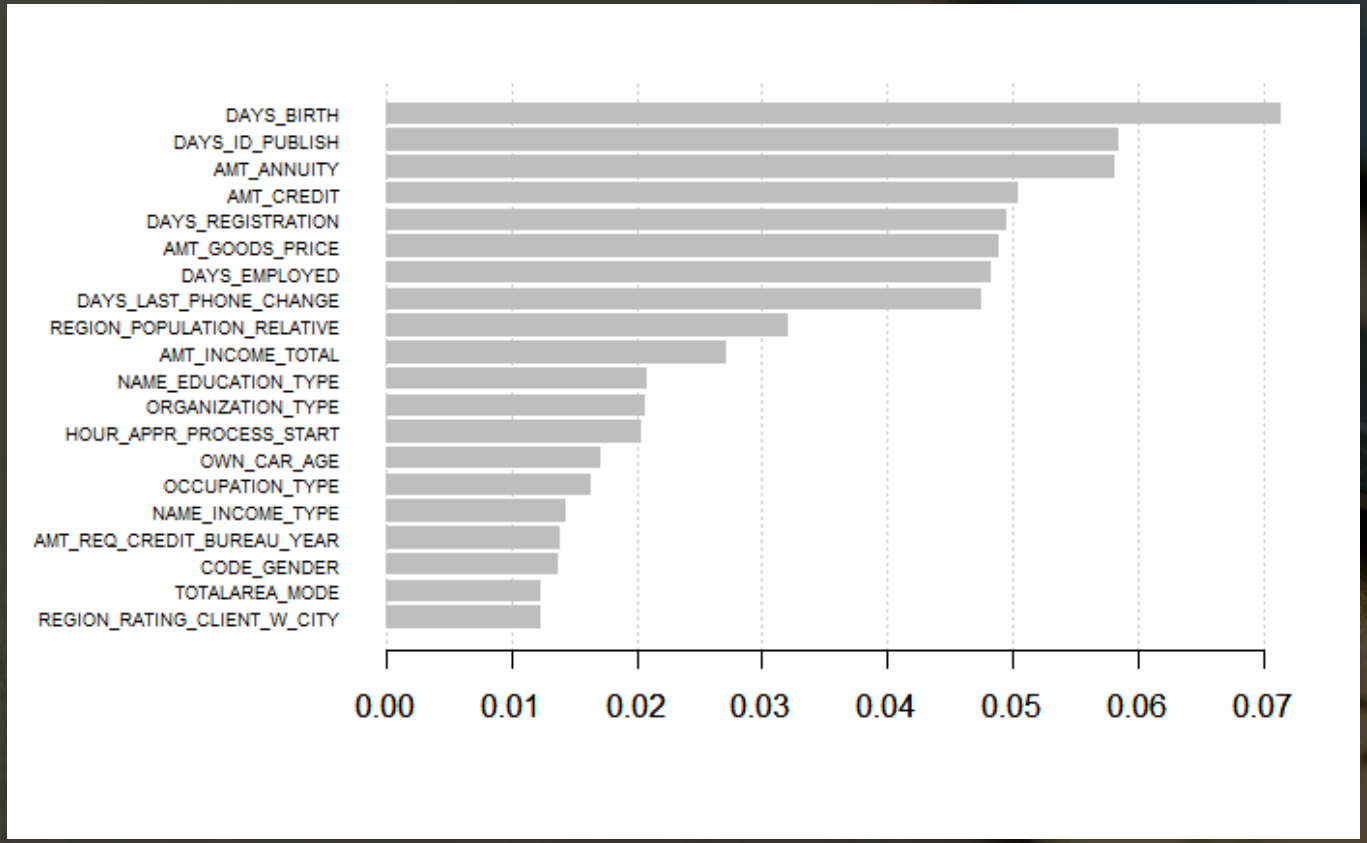
예상보다 높지 않았던 변수

상대적으로 20대 체납자 비율이 높은 편이기 때문에 연령대별로 차별성이 있다고 EDA에서 확인되었지만, 본 모델에서는 연령 정보가 큰 변수로 작용하지는 않았다.

05

Conclusion

Full Model Feature Importance - Exclude EXT-SOURCE values



1. Credit or Company data

DAYS_BIRTH가
가장 중요한 변수임을 나타내고 있음

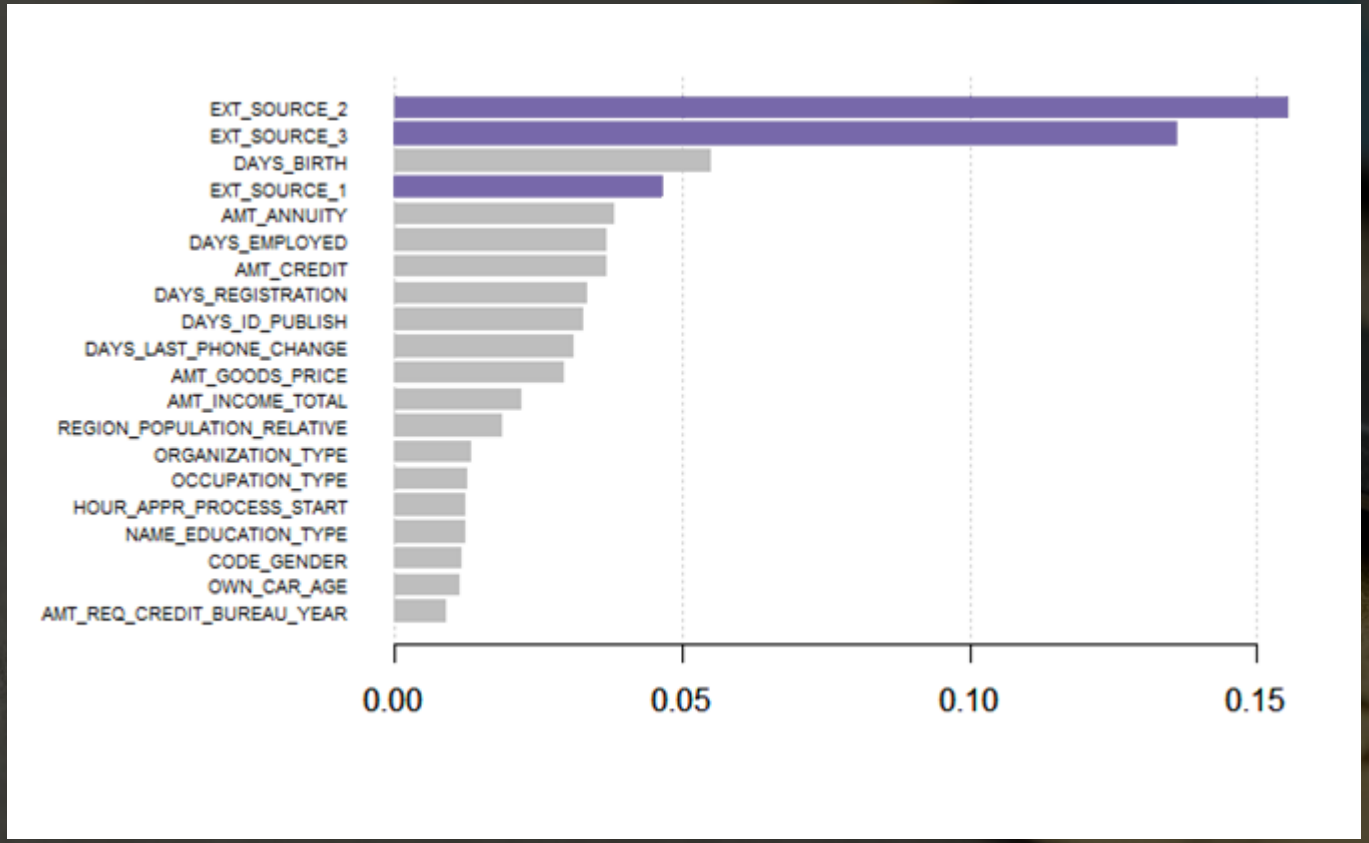
2. 중요한 개인 정보 변수

연령, 연금, 고용 현황 등 개인 정보 또한
상환 능력 평가 지표에 기여하고 있음을
확인할 수 있음

05

Conclusion

Full Model Feature Importance - Include EXT-SOURCE values



1. Credit or Company data

전체 모델에서 Home credit 신용 평가 지표 (EXT-SOURCE 1~3)이
중요한 변수로 작용하고 있음

2. 중요한 개인 정보 변수

연령, 연금, 고용 현황 등 개인 정보 또한 상환 능력 평가 지표에 기여하고 있음을 확인할 수 있음

05

Conclusion

Confusion matrix

1. Exclude EXT-SOURCE values

```
     cutoff  error rate sensitivity specificity    f1 score
     0.1400      0.1714      0.3660      0.8681      0.2516
        pred
response     0     1
       0 24595  3736
       1  1535   886
```

2.Include EXT-SOURCE values

```
     cutoff  error rate sensitivity specificity    f1 score
     0.1600      0.1463      0.3953      0.8928      0.2984
        pred
response     0     1
       0 25295  3036
       1  1464   957
```

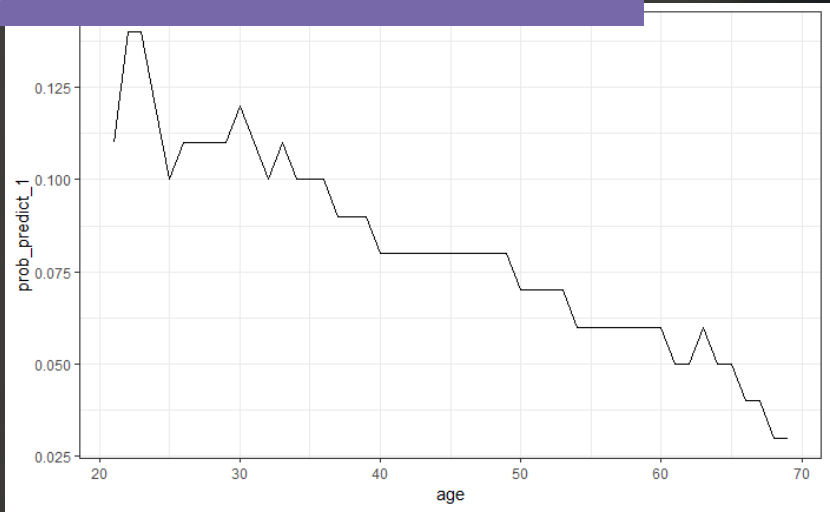
F1 score를 가장 높게 하는
cutoff를 찾아 예측을 한 결과에 해당하는 cross table

05

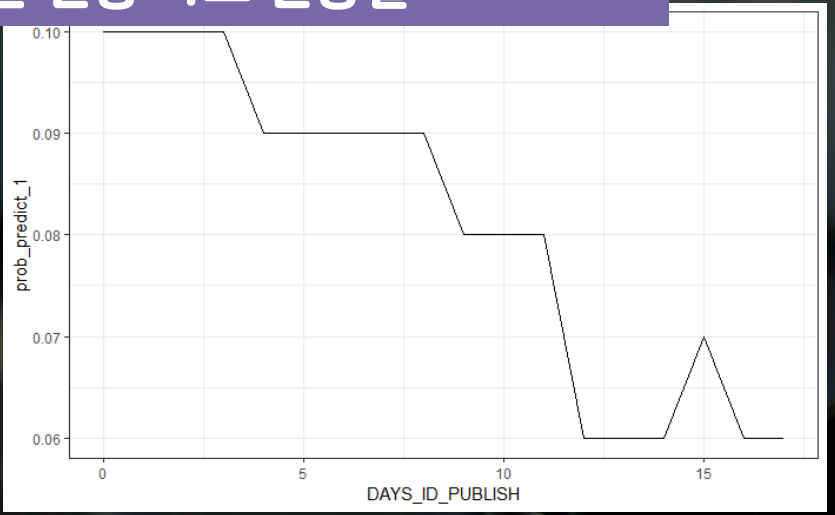
Conclusion

중요한 변수들이 끼치는 확률에 대한 영향

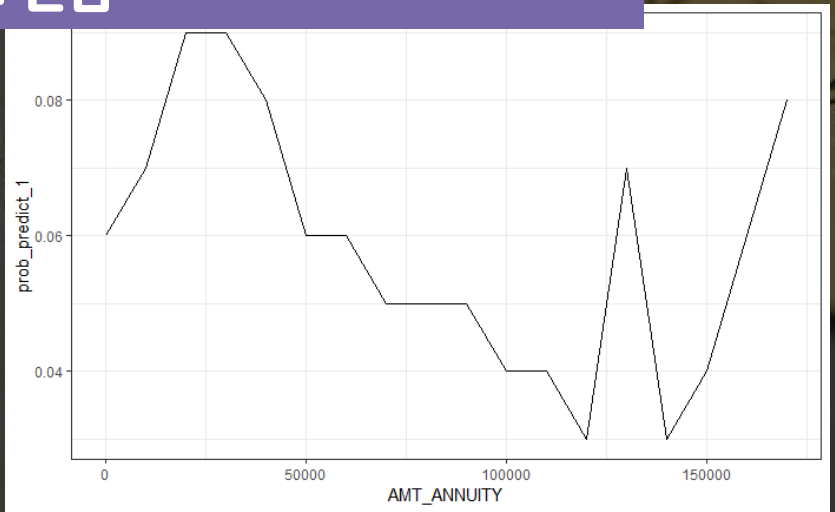
연령



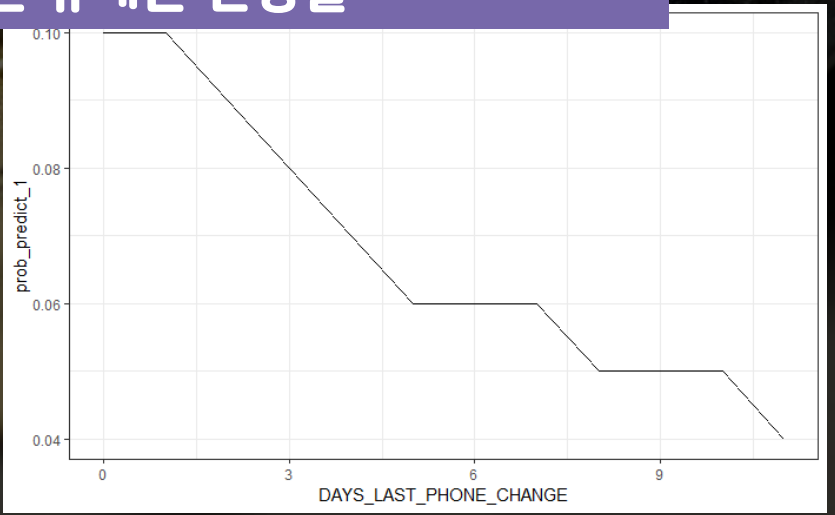
최근 신용 자료 변동일



담보 연금



최근 휴대폰 변경일



05

Conclusion
Final

결론

1. 개인정보 데이터를 통한 모델링 후 연체여부에 중요한 개인정보 변수 확인
2. Full model과 고객 개인정보만 사용한 model의 차이
3. 두 결과를 통해 현재 금융업에서 연체여부를 판단할 때
   어느 부분을 새롭게 체크해야 하는지 알 수 있음.

후속 연구 제안

1. 결측치에 대한 통계적 접근
2. 주성분분석
3. 반응변수 비대칭성 해결을 위한 Case sampling 시도

Any Questions?

개인 신용도 예측 변수 분석

서울대학교 빅데이터 아카데미
2018-3 고급 빅데이터 분석 기법
BA 노은선 이현호 최의관