신용카드 연체 예측

빅 데 이 터 분 석 경 진 대 회

분린이

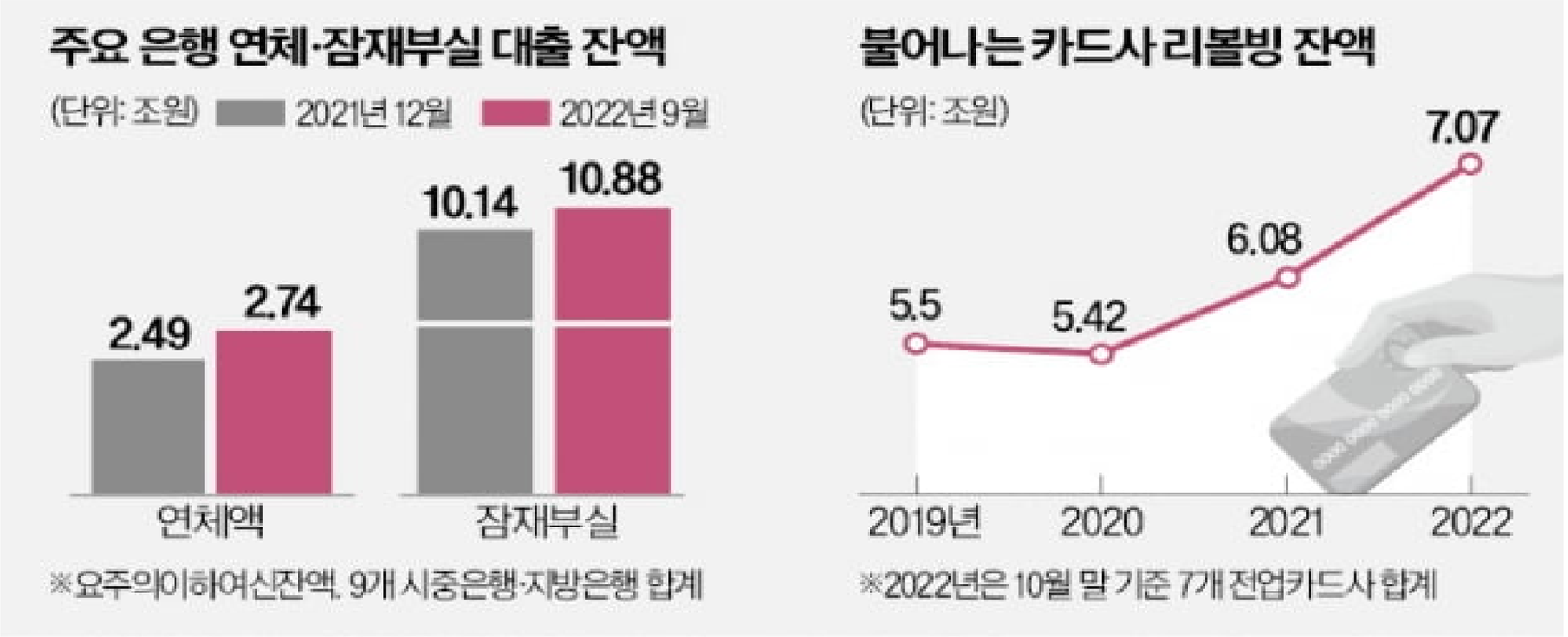
경영정보학과 5516206 최현성

경영정보학과 5517661 이민상

목차

01 개념 이해

배경 설명



리볼빙 증가는 부실로 연결될 수 있는 징후

01 개념 이해

가설 설정

- Credit Card Fraud Detection 데이터를 통해 분석
- 리볼빙 자산 증가가 카드사의 건전성 악화를 초래하는 지 여부에 주목

02 데이터 확인

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

데이터 확인

Application Data

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	...
100002	1	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	...
100003	0	Cash loans	F	N	N	0	270000.0	1293502.5	35698.5	...
100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...
100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	...
100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	...
...	...	...	...	...	...	...	...	...	...	...
456251	0	Cash loans	M	N	N	0	157500.0	254700.0	27558.0	...
456252	0	Cash loans	F	N	Y	0	72000.0	269550.0	12001.5	...
456253	0	Cash loans	F	N	Y	0	153000.0	677664.0	29979.0	...
456254	1	Cash loans	F	N	Y	0	171000.0	370107.0	20205.0	...
456255	0	Cash loans	F	N	N	0	157500.0	675000.0	49117.5	...

307511개의 행, 122개의 열

Previous Application

SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_
2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	1
2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	60
2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	11
2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	45
1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	33
...	...	...	...	...	...	...	
2300464	352015	Consumer loans	14704.290	267295.5	311400.0	0.0	26
2357031	334635	Consumer loans	6622.020	87750.0	64291.5	29250.0	8
2659632	249544	Consumer loans	11520.855	105237.0	102523.5	10525.5	10
2785582	400317	Cash loans	18821.520	180000.0	191880.0	NaN	18
2418762	261212	Cash loans	16431.300	360000.0	360000.0	NaN	36

1670214개의 행, 37개의 열

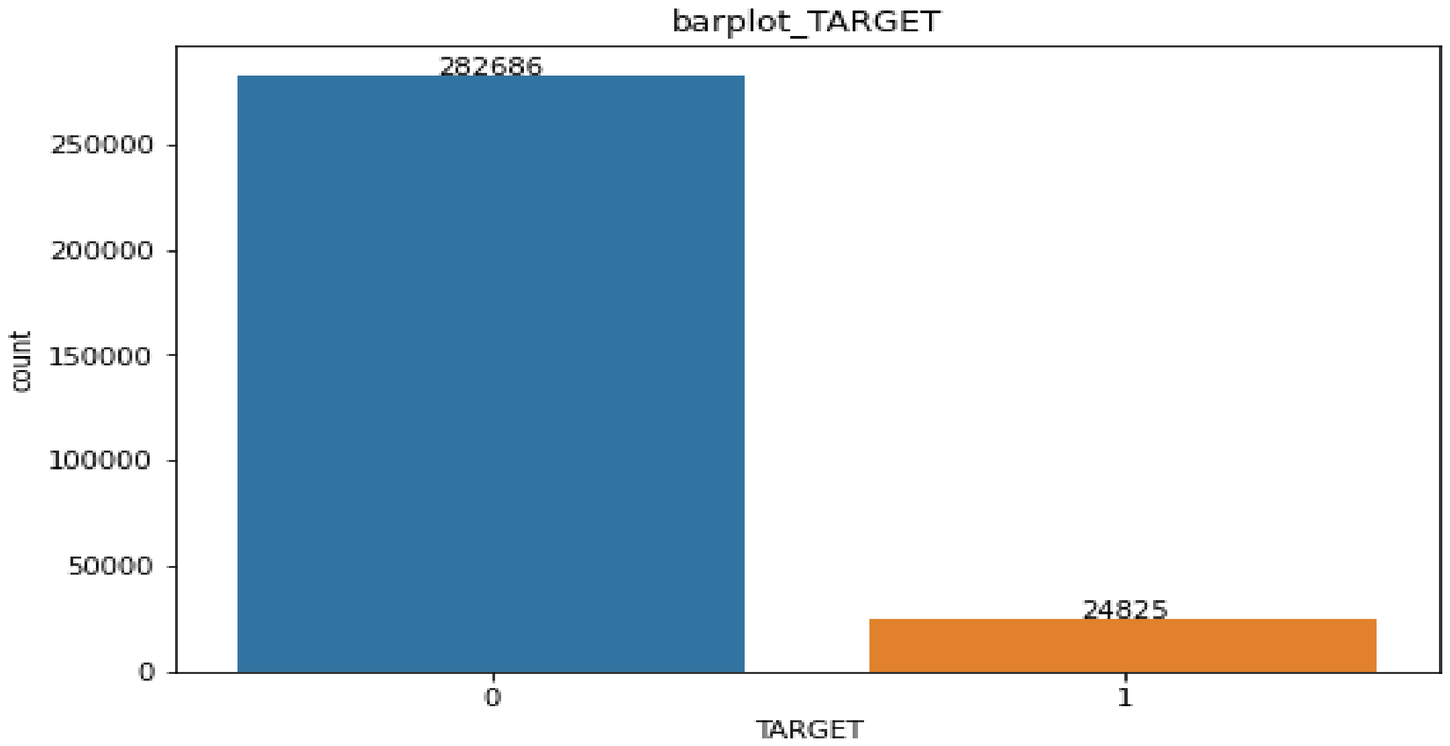
02 Application Data

EDA

01. 개념 이해

02 빅데이터 과제 분석

03. 결론



Target 값의 불균형 문제 해결 필요

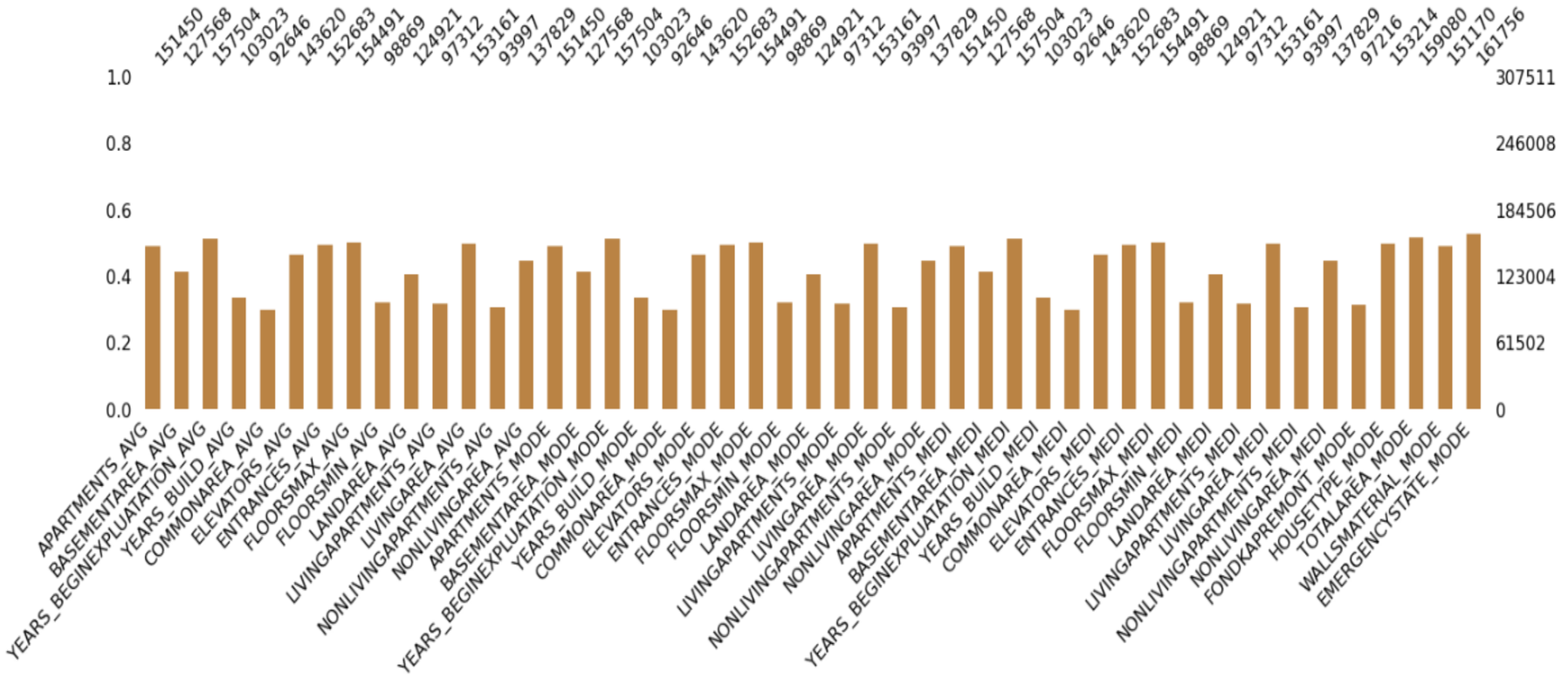
02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA



고객의 거주지에 대한 표준화된 정보 삭제 필요

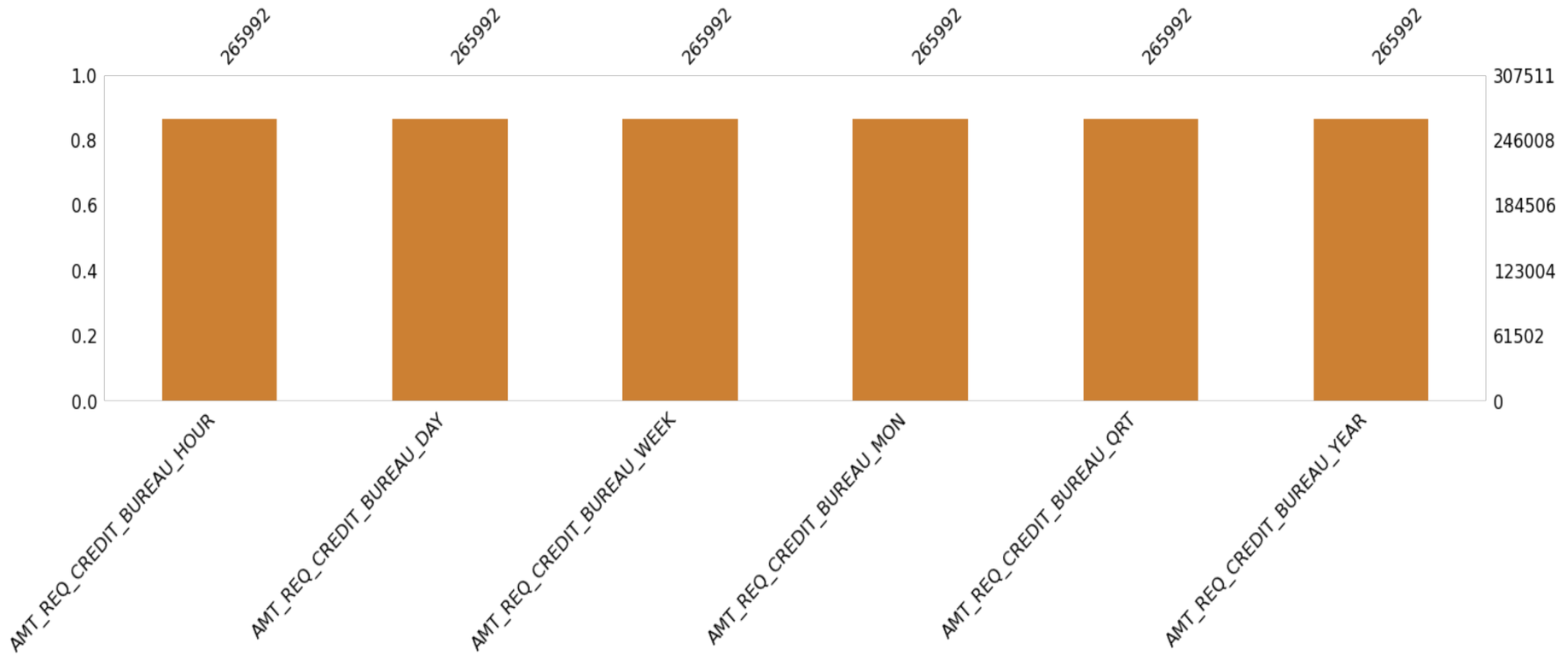
02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA



AMT_REQ_CREDIT_BUREAU 변수의 경우 결측치가 존재하나

데이터 개수 모두 265992개로 동일, 문의 횟수를 기준으로 통합

02 Application Data

EDA

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

new_ANNUITY_INCOME	new_CREDIT_INCOME	new_GOODS_INCOME
0.121978	2.007889	1.733333
0.132217	4.790750	4.183333
0.100000	2.000000	2.000000
0.219900	2.316167	2.200000
0.179963	4.222222	4.222222
...	...	...
0.174971	1.617143	1.428571
0.166687	3.743750	3.125000
0.195941	4.429176	3.823529
0.118158	2.164368	1.868421
0.311857	4.285714	4.285714

AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, AMT_INCOME_TOTAL

위 4개의 변수를 통해고객의 수입 대비 대출의 금전적 부담감에 대한 변수들을 생성

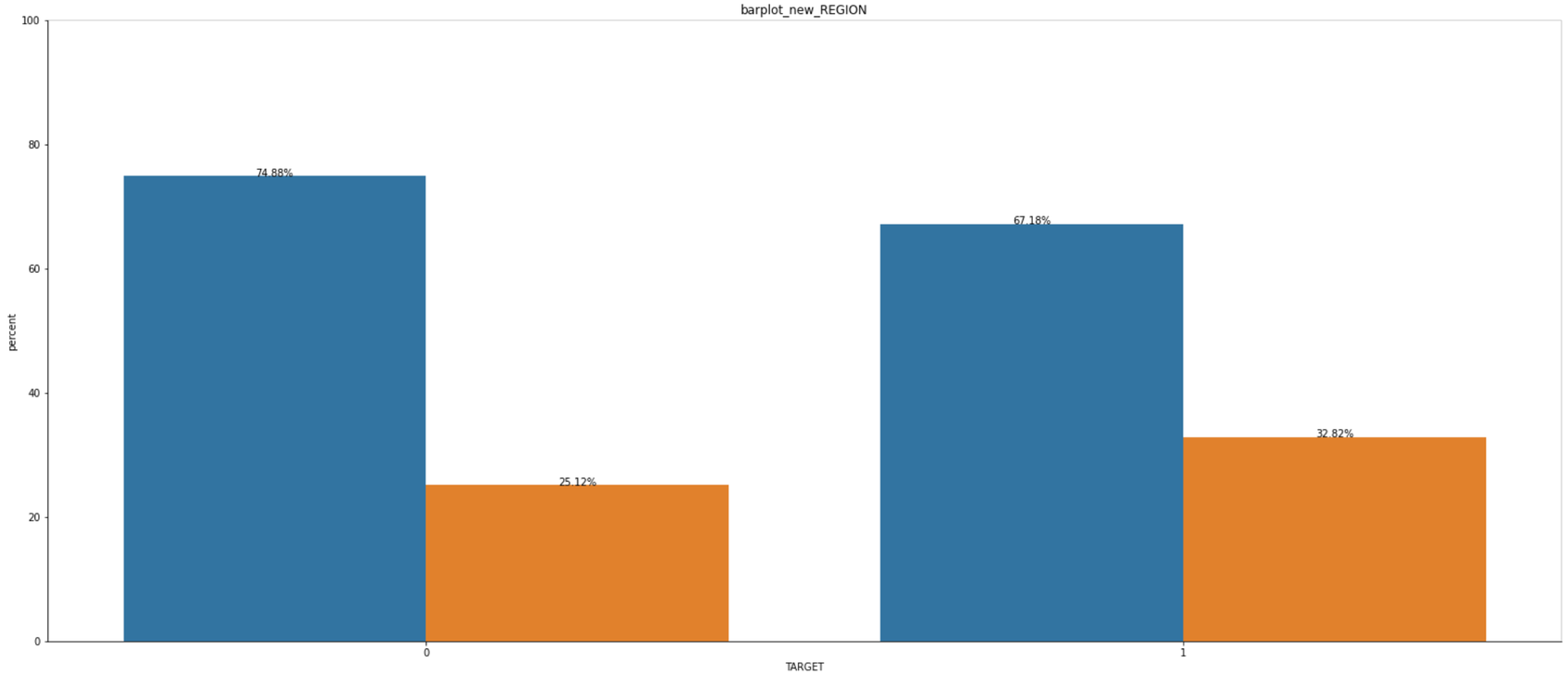
02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA



활동지역에 대한 정보에 차이점이 존재할 경우

TARGET이 1이 되는 비율이 높은 것을 확인

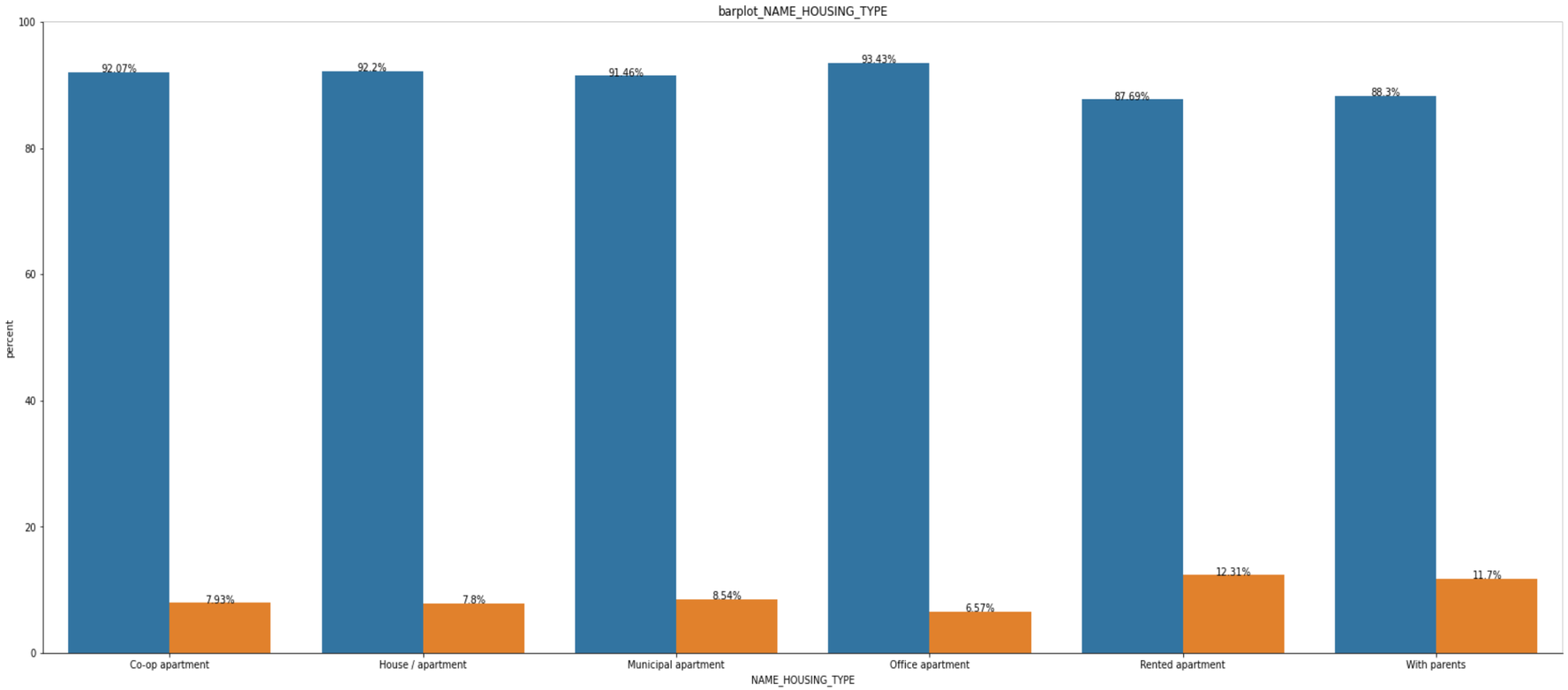
02 Application Data

EDA

01. 개념 이해

02 빅데이터 과제 분석

03. 결론



NAME_HOUSING_TYPE 변수를 통해 고객의 거주지에 대한 정보를 추출

고객이 안정적인 거주지를 확보하였는지 확인

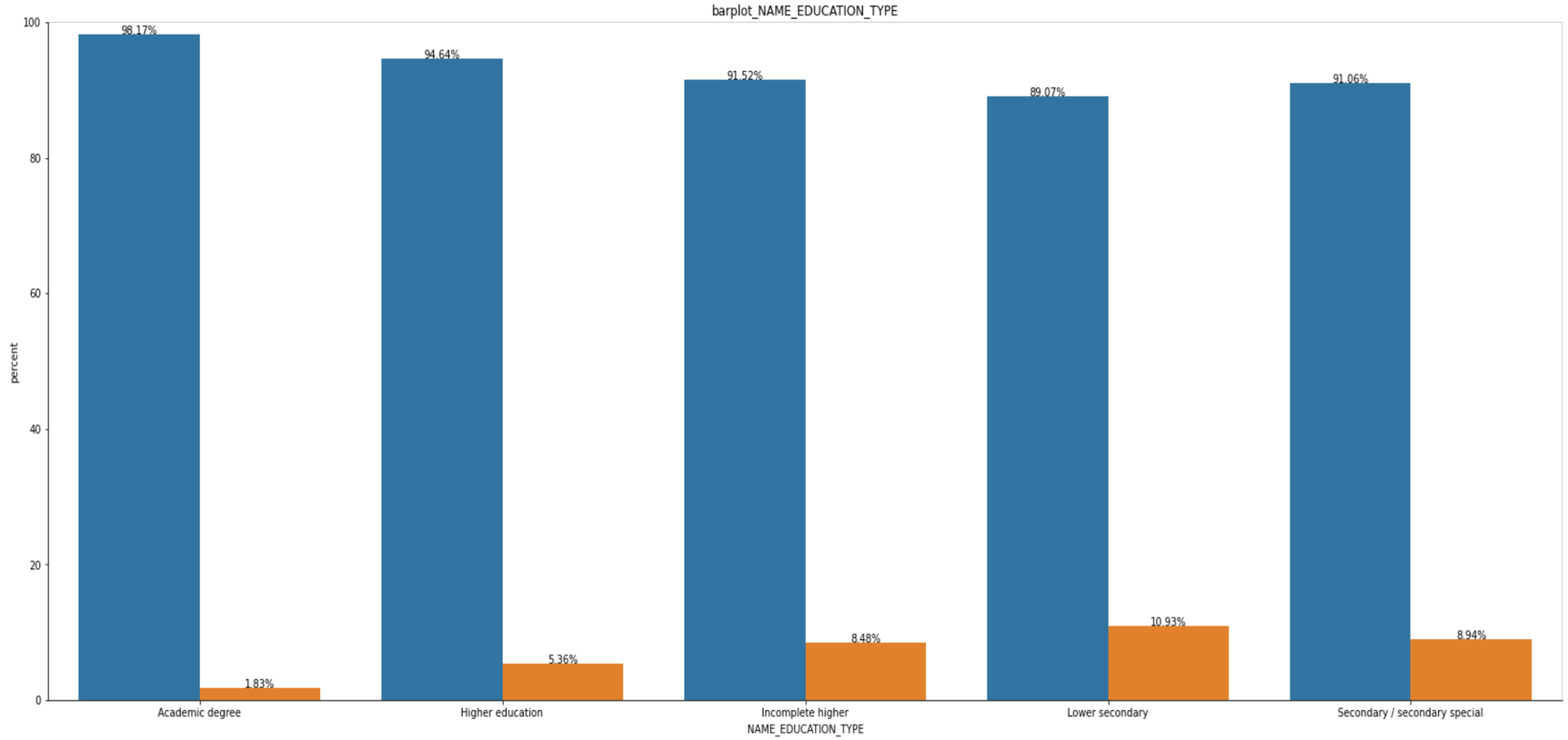
02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA



교육수준이 낮을수록 연체율이 높은 것으로 판단

고등교육을 기준으로 고등교육을 받은 고객과 받지 못한 고객을 나눔

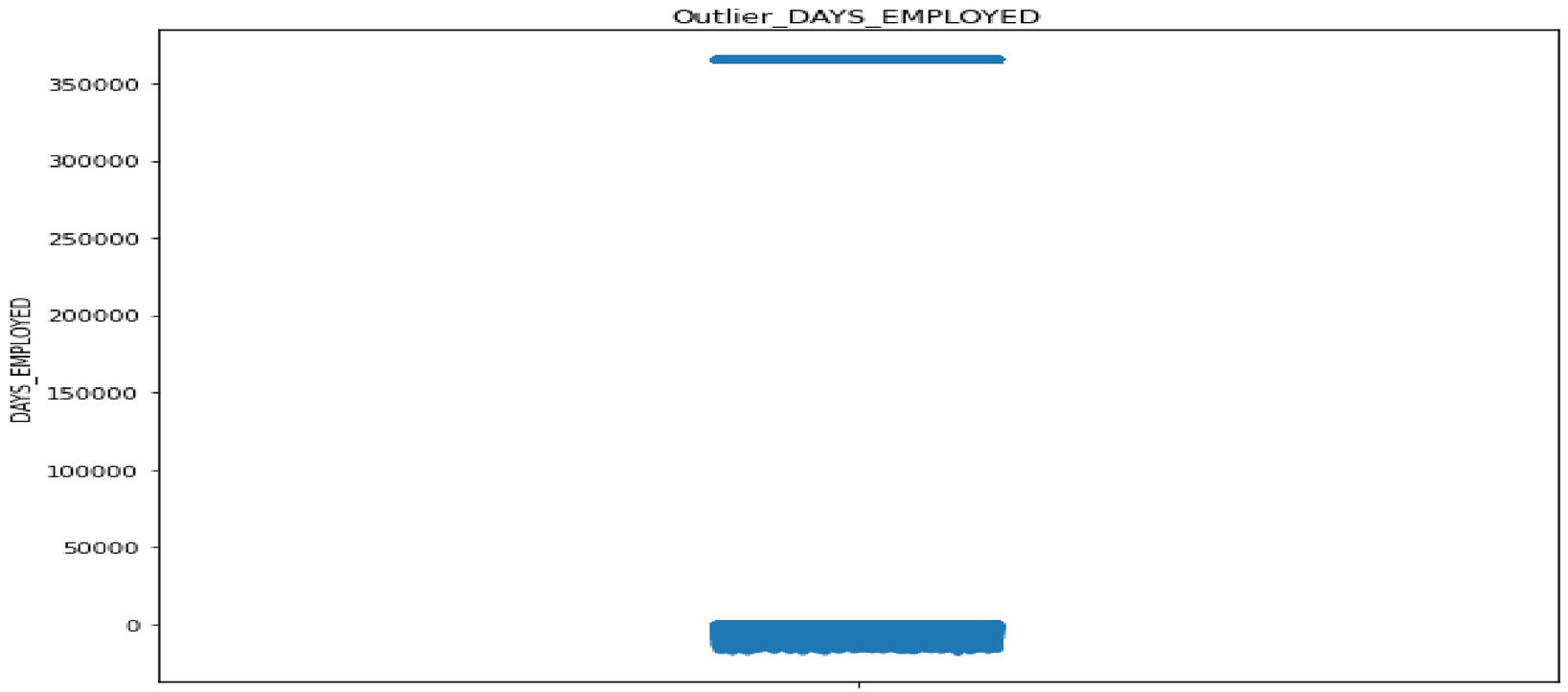
02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA



DAYS_EMPLOYED의 경우 이상치 다수 분포

DAYS_EMPLOYED의 구조를 보았을 때 365243이라는 수치는

단순한 이상치가 아닌 결측치 대신에 채워둔 값이라 판단

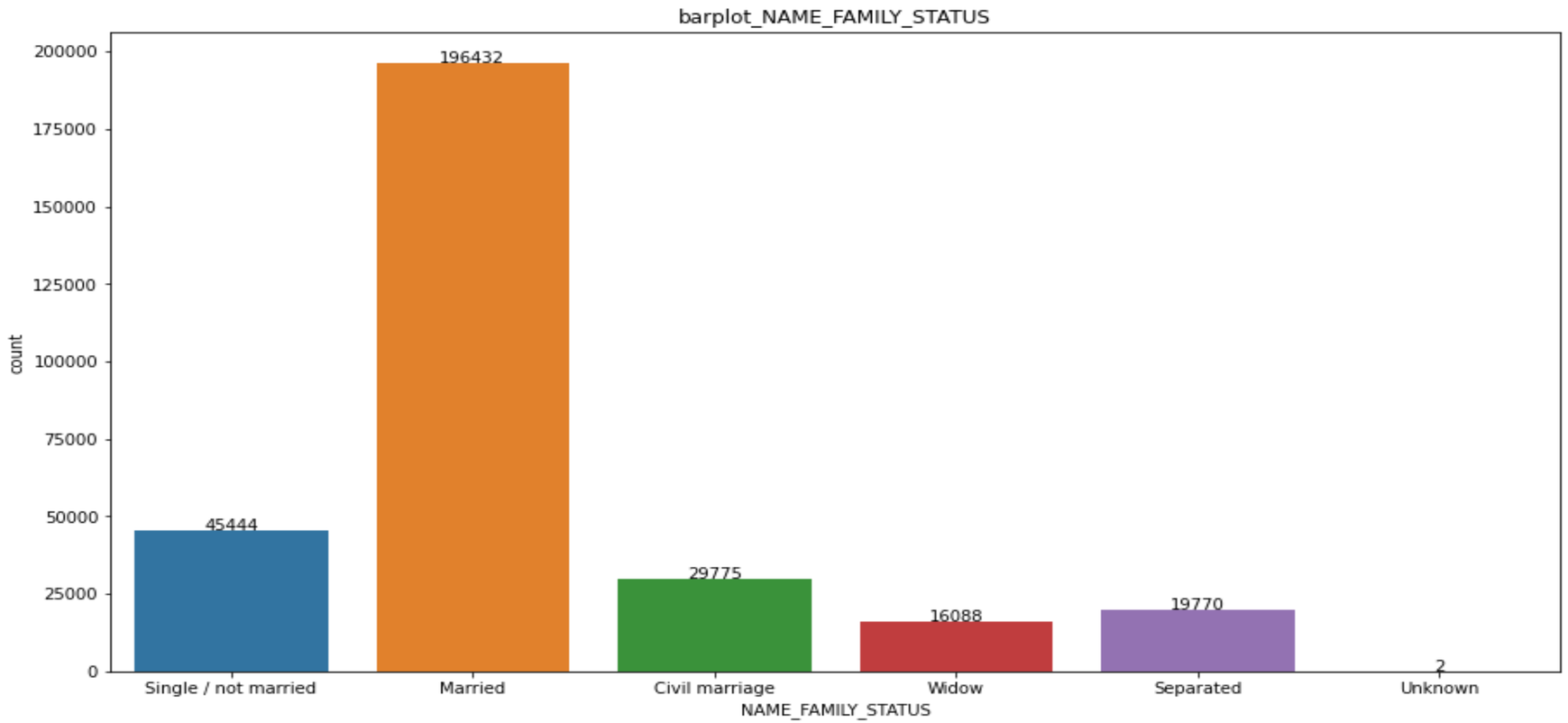
02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA



NAME_FAMILY_STATUS 변수를 통해 다양한 가족의 형태가 존재을 확인

CNT_CHILDREN와 CNT_FAM_MEMBERS을 활용하여 해당 가구의 경제활동인구를 추출

02 Application Data

01. 개념 이해

02 빅데이터 과제 분석

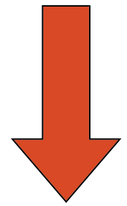
03. 결론

EDA

```
Data columns (total 20 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   SK_ID_CURR                   307509 non-null  int64
 1   TARGET                       307509 non-null  int64
 2   NAME_CONTRACT_TYPE           307509 non-null  object
 3   CODE_GENDER                  307509 non-null  object
 4   DAYS_BIRTH                   307509 non-null  int64
 5   DAYS_REGISTRATION            307509 non-null  float64
 6   REGION_RATING_CLIENT_W_CITY  307509 non-null  int64
 7   HOUR_APPR_PROCESS_START      307509 non-null  int64
 8   EXT_SOURCE_1                 134132 non-null  float64
 9   EXT_SOURCE_2                 306849 non-null  float64
 10  EXT_SOURCE_3                 246545 non-null  float64
 11  new_ANNUITY_INCOME           307497 non-null  float64
 12  new_CREDIT_INCOME            307509 non-null  float64
 13  new_GOODS_INCOME             307233 non-null  float64
 14  new_REGION                   307509 non-null  int64
 15  new_REQ_CREDIT               307509 non-null  int64
 16  new_HOME_TYPE                307509 non-null  int64
 17  new_EDU_LEVEL                307509 non-null  int64
 18  new_EMPLOYED_DAYS            252135 non-null  float64
 19  new_ECONOMIC_POPULATION      307509 non-null  float64
dtypes: float64(9), int64(9), object(2)
memory usage: 46.9+ MB
```

122개의 변수로 9개의 신규변수 생성

111개의 변수 삭제



20개의 변수들로 이루어진 application data 생성

02 Previous Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA

previous_SK_ID_CURR	previous_AMT_ANNUITY	previous_AMT_CREDIT	previous_AMT_GOODS_PRICE	previous_ID_COUNT
100001.0	3951.000000	23787.00	24835.500	1
100002.0	9251.775000	179055.00	179055.000	1
100003.0	56553.990000	484191.00	435436.500	3
100004.0	5357.250000	20106.00	24282.000	1
100005.0	4813.200000	20076.75	44617.500	2
...	...	...	...	...
456251.0	6605.910000	40455.00	40455.000	1
456252.0	10074.465000	56821.50	57595.500	1
456253.0	4770.405000	20625.75	24162.750	2
456254.0	10681.132500	134439.75	121317.750	2
456255.0	20775.391875	424431.00	362770.875	8

한 명의 고객에 대한 기록이 여러 행 존재

group by 함수를 이용하여 별도의 계산을 통해 한 고객에 대한 통계치를 하나씩 추출

02 Previous Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA

AMT_ANNUITY 외 2개의 대출금액에 대한 정보가 담긴 변수들의 평균값을 추출

- Previous Data에는 고객의 수입에 대한 정보가 존재하지 않음

→ **단순 평균값으로 집계**

한 고객에 대한 데이터의 개수를 count

- 과거에 고객이 몇 번 대출신청을 하였는지, 대출신청횟수 산출

→ **대출신청횟수 종속변수 예측에 영향을 줄 것이라 판단**

02 Previous Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA

대출신청금액과 대출허가금액이 따로 존재

- 신청시 기입한 금액과 실제로 수령받게 될 금액이 다름

→ **대출신청금액 대비 대출허가금액의 값을 추출하여 변수를 생성**

NAME_CONTRACT_STATUS 변수에 승인, 거절, 취소 값 존재

- 대출거절을 받은경우에 고객에게 어떠한 문제점이 있을 것이라 판단

→ **고객이 대출신청을 하였을때 어느정도 비율로 거절되는지를 확인**

02 Previous Data

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

EDA

DAYS_DECISION 기간에 따른 영향

- 금전적 도움의 필요성을 대출간격에 따라 판단

→ **새로운 대출을 받기까지 걸린 시간의 평균을 구해 새로운 변수 생성**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 338857 entries, 0 to 338856
Data columns (total 8 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   previous_SK_ID_CURR           338857 non-null  float64
 1   previous_AMT_ANNUITY          338377 non-null  float64
 2   previous_AMT_CREDIT           338857 non-null  float64
 3   previous_AMT_GOODS_PRICE      337793 non-null  float64
 4   previous_ID_COUNT             338857 non-null  int64
 5   previous_APPLICATION_TO_CREDIT  338857 non-null  float64
 6   previous_REFUSED_COUNT        338857 non-null  int64
 7   previous_DAY_DECISION         338857 non-null  float64
dtypes: float64(6), int64(2)
memory usage: 20.7 MB
```

→ 8개의 변수로 이루어진 Previous Data 생성

결측치 처리 및 샘플링

Imputation Using Deep Learning

- DataWig은 아마존이 개발한 OSS의 결손값 보완 라이브러리

→ **Deep Learning 기반 결측치 처리 알고리즘 Datawig 사용**

SMOTE+Tomek

- 생성된 데이터를 무조건 소수 클래스라고 하지 않고 분류 모형에 따라 분류
- 두 샘플 사이에 기타 관측치가 없을 경우 이를 Tomek links로 이후 그 중에서 다수 클래스에 속하는 데이터를 제외하는 방법

02 모델링

모델링

LightGBM RandomForest Catboost 사용

- LightGBM의 경우 결측치 포함 상태에서 모델 사용가능

→ **결측치 포함된 데이터, 결측치 채워진 데이터, 불균형 해소 데이터 3가지의 분석 방향 설정**

Precision

- 연체자를 잘 찾아내어 회피하는 것이 중요

→ **Precision를 통해 판단**

02 모델링

모델링

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

	결측치 포함 LGBM	결측치 처리 LGBM	불균형 해소 LGBM	결측치 처리 Random Forest	불균형 해소 Random Forest	결측치 처리 Catboost	불균형 해소 Catboost
accuracy	91.89	92.60	92.72	92.15	93.29	91.89	89.60
precision	88.98	91.59	93.04	91.27	93.40	90.25	90.00
recall	91.89	92.60	92.72	92.15	93.29	91.89	89.60
F1-score	88.24	90.07	92.71	88.81	93.29	88.06	89.57

02 최적화

모델별 Hyperparameter 최적화

Random Forest	n_estimators	[100]
	max_depth	[6, 8, 10]
	min_samples_leaf	[8, 12, 18]
	min_samples_split	[8, 16, 20]
Light GBM	n_estimators	[50, 100, 150]
	max_depth	[-1, 8, 10, 12]
	min_samples_leaf	[8, 12, 16, 20]
	num_leaves	[31, 62, 93, 124]

02 모델링

01. 개념 이해

02 빅데이터 과제 분석

03. 결론

모델별 Hyperparameter 최적화

RandomForest

- Hyperparameter 튜닝 후 성능 떨어짐

→ Default 값으로 설정했을 때 가장 좋은 성능

Lightgbm

- max_depth : -1 min_samples_leaf : 8 n_estimators : 150 num_leaves : 124

Accuracy:92.72% Precision:93.04%
Recall:92.72% F1-score:92.71%

→

Accuracy:94.29% Precision:94.50%
Recall:94.29% F1-score:94.29%

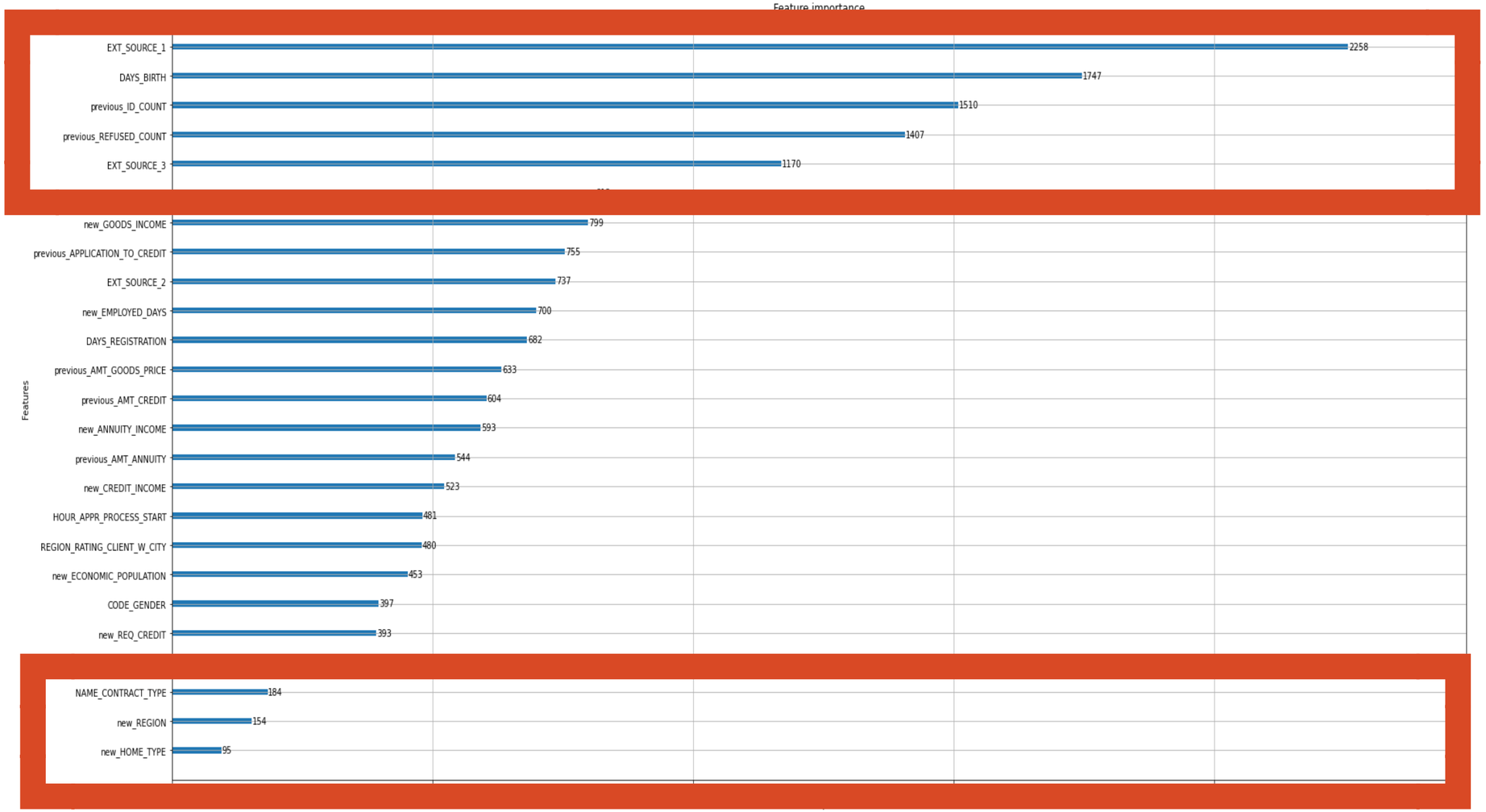
03 변수중요도

01. 개념 이해

02 빅데이터 과제 분석

03 결론

변수 중요도



03 모델 해석

01. 개념 이해

02 빅데이터 과제 분석

03 결론

모델 해석

카드사의 건전성 악화 초래 여부를 결정하는 요인 변수중요도로 파악

상위 5개의 변수	하위 3개의 변수
EXT_SOURCE_1	NAME_CONTRACT_TYPE
DAYS_BIRTH	new_HOME_TYPE
previous_ID_COUNT	new_REGION
previous_REFUSED_COUNT	
EXT_SOURCE_3	

→ 관련 변수의 수치를 통해 취약점 파악 후 새로운 해결책 구축 필요

03 상위 5개

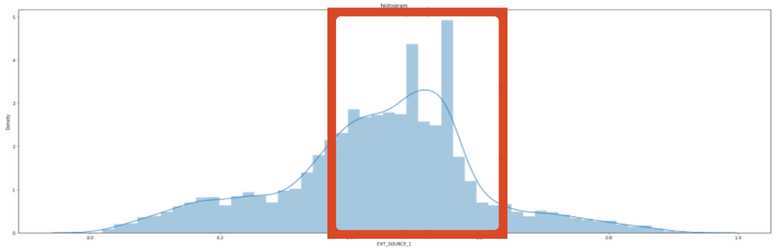
01. 개념 이해

02 빅데이터 과제 분석

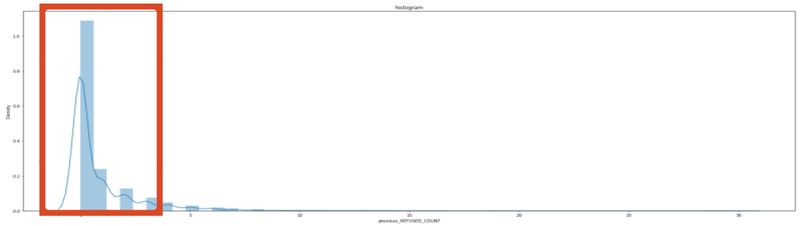
03 결론

변수 중요도

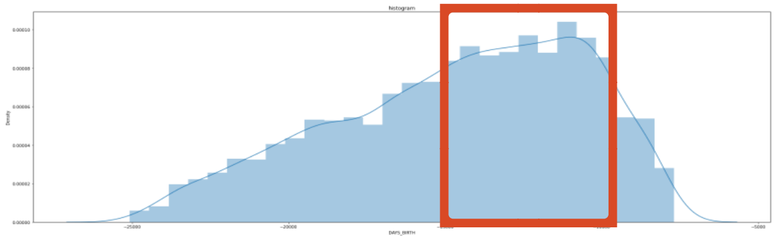
EXT_SOURCE_1



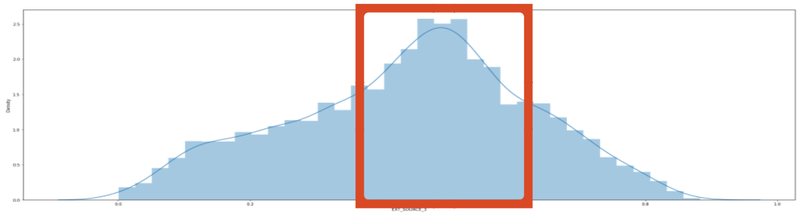
previous_REFUSED_COUNT



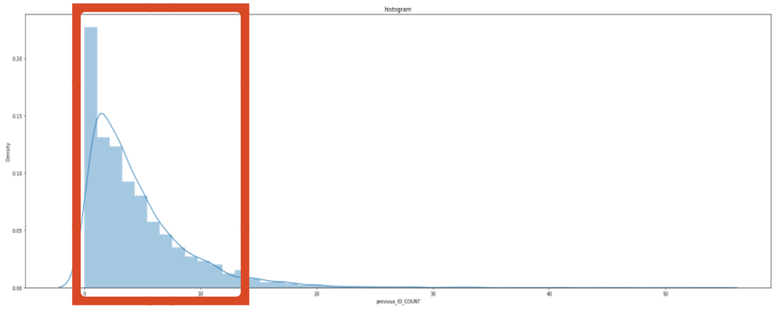
DAYS_BIRTH



EXT_SOURCE_3



previous_ID_COUNT



03 모델 해석

01. 개념 이해

02 빅데이터 과제 분석

03 결론

모델 해석

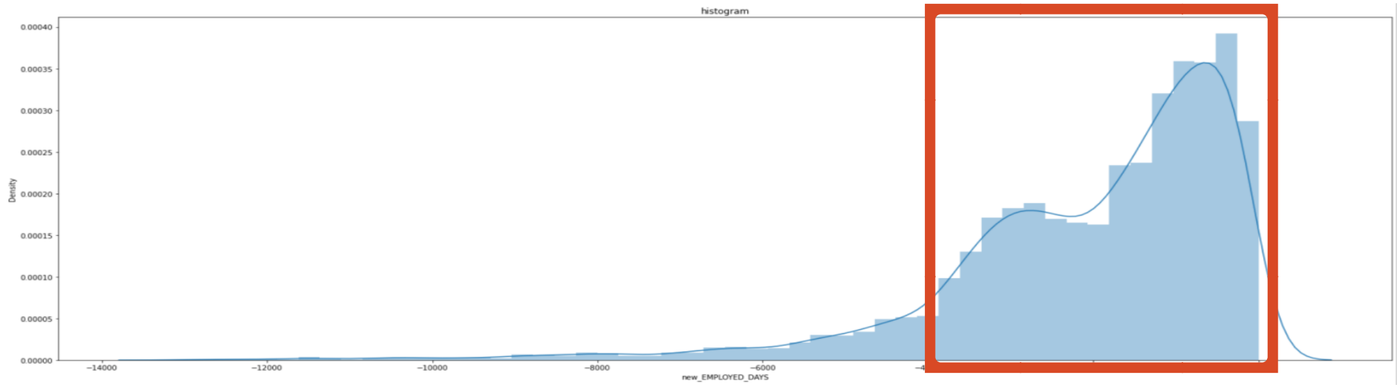
Feature Importance를 통한 결과 해석

연체자 예측실패가 높은 구간 : 20대 후반 - 30대 후반

- 이러한 고객의 금전적배경은 안정적이지않으며 외부의 개입,

  단순한 변심등의 불확실한 변수들이 너무 많은 것으로 보임

EMPLOYED_DAYS



- 고용일을 판단할 수 있는 EMPLOYED_DAYS 통해서 고용일이 얼마되지 않은

  사람은 예측율이 낮은 것을 확인할 수 있음

03 모델 해석

01. 개념 이해

02 빅데이터 과제 분석

03 결론

모델 해석

Feature Importance를 통한 결과 해석

특정 대상에 대한 집중적인 조사 필요

- 안정적인 정착 여부를 역학조사를 통해 파악

→ **카드사의 건전성 악화 초래 여부를 결정하는 요인 파악 및 해결 가능**

- 연체자 회피를 보다 더 높일 수 있음
- 광범위한 조사를 줄여 비용 절감 실현 가능

→ **연체자 회피, 조사비용 절감으로 인한 금융기업의 손실의 최소화**

감사합니다

분린이