

신용카드 사용자 연체 예측 AI

2조 머신러닝 프로젝트

김용삼 김성일 노지윤 박정환 윤태양 김혜리 안심향



INDEX

1. 분석 주제 이해
2. 데이터 이해 및 전처리
3. 데이터 탐색 (EDA)
4. 파생변수 생성
5. 모델링 비교 평가 및 선택
6. 결론 및 활용방안 (기대효과)
7. 프로젝트 후기



프로젝트 선정 기준

● 프로젝트 선정 기준 - 신용카드 사용자 연체 예측 AI

프로젝트 선정 기준

DAICON 커뮤니티 **대회** 교육 랭킹 더보기

신용카드 사용자 연체 예측 AI 경진대회
월간 데이콘 14 | 금융 | 정형 | Logloss

₩ 상금 : 100만원
🕒 2021.04.05 ~ 2021.05.24 17:59 [+ Google Calendar](#)
👤 2,427명 📅 마감

[대회안내](#) [데이터](#) [코드 공유](#) [토크](#) [리더보드](#) [제출](#)

머신러닝 대회

정형 데이터

쉬운 데이터 이해

분류 & 회귀

코드 조회 가능

다른 알고리즘 적용

1. 분석 주제 이해

1-1 분석 배경 - 비금융 데이터의 중요성

1. 분석 주제 이해

금융 판 혼드는 新신용평가

[이코노미조선]

非금융 빅데이터로 소외 고객 끌어안는다
디지털 전환, '대안 신용평가' 시장 키웠다

안소영 기자 김혜빈 이코노미조선 인턴기자

입력 2021.09.13 06:11



[Interview] 핀테크 베스트셀러 '디지털 뱅크' 저자 크리스 스키너 더파이낸서 대표

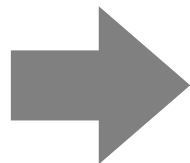
“은행은 돈이 아닌 데이터를 저장...핀테크가 신용평가 이끌 것”

411호 2021년 09월 06일



“ 은행은 돈이 아닌 데이터를 저장하는 곳,
핀테크가 신용평가 이끌 것 “

기업 & 고객



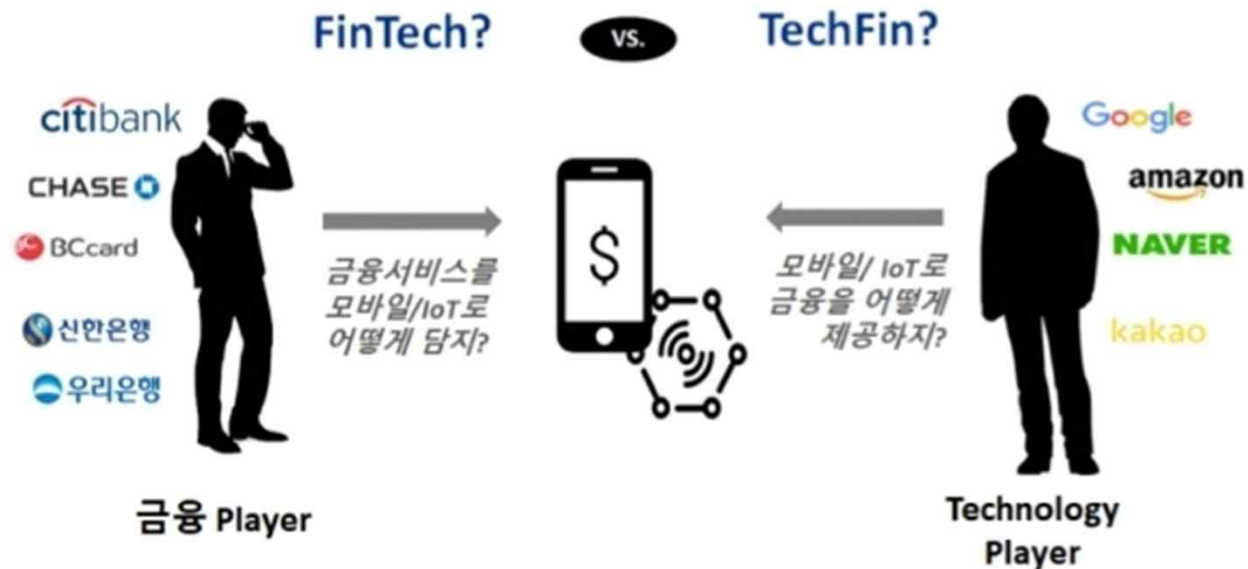
비금융 빅데이터 활용
전통적 평가 방식 보완하여 대출 시장 확대

1-1 분석 배경 - 핀테크

1. 분석 주제 이해

‘핀테크’ 란?

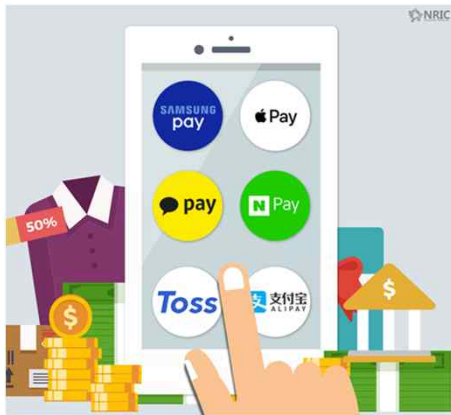
핀테크(Fin Tech)는 **금융(Finance)**과 **기술(Technology)**의 합성어.
최근 급속히 확산되고 있는 모바일뱅킹과 앱카드 등이 대표적.



1-1 분석 배경 - 국내 핀테크 기업들의 경쟁

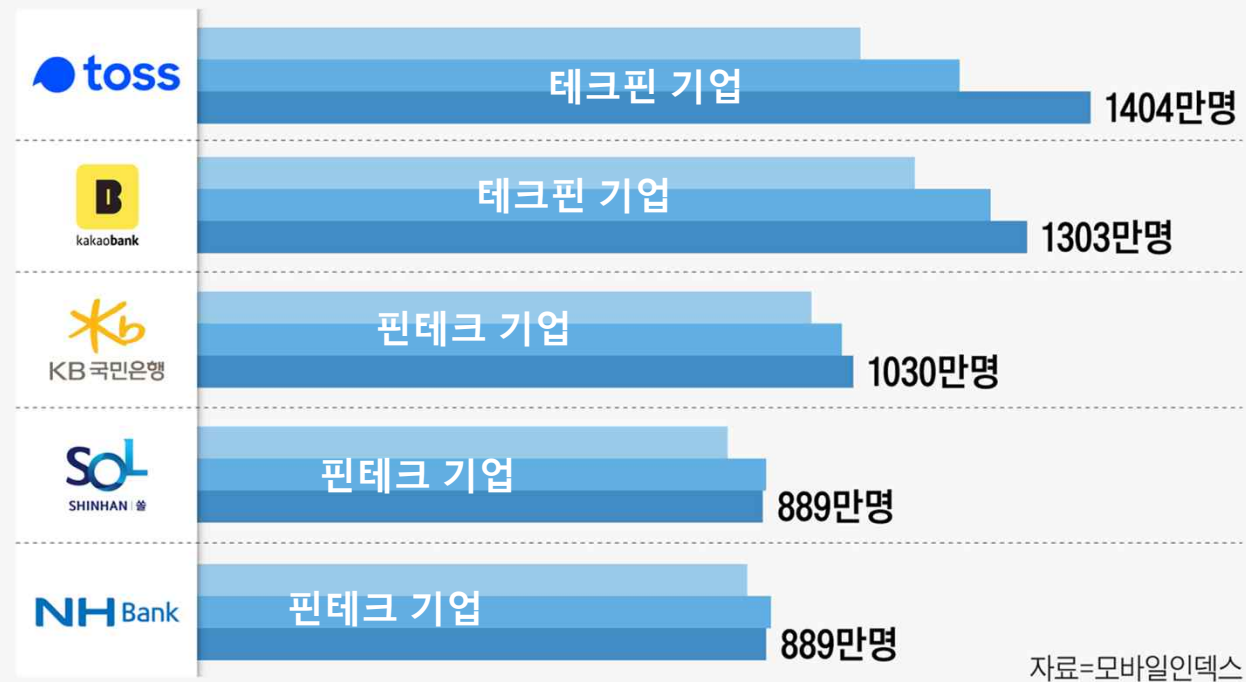
1. 분석 주제 이해

상위 핀테크 앱,
상위 은행 앱 보다
10.8배 더 높은
성과를 기록



2021년 주요 금융 업종 앱 월 사용자 수 비교

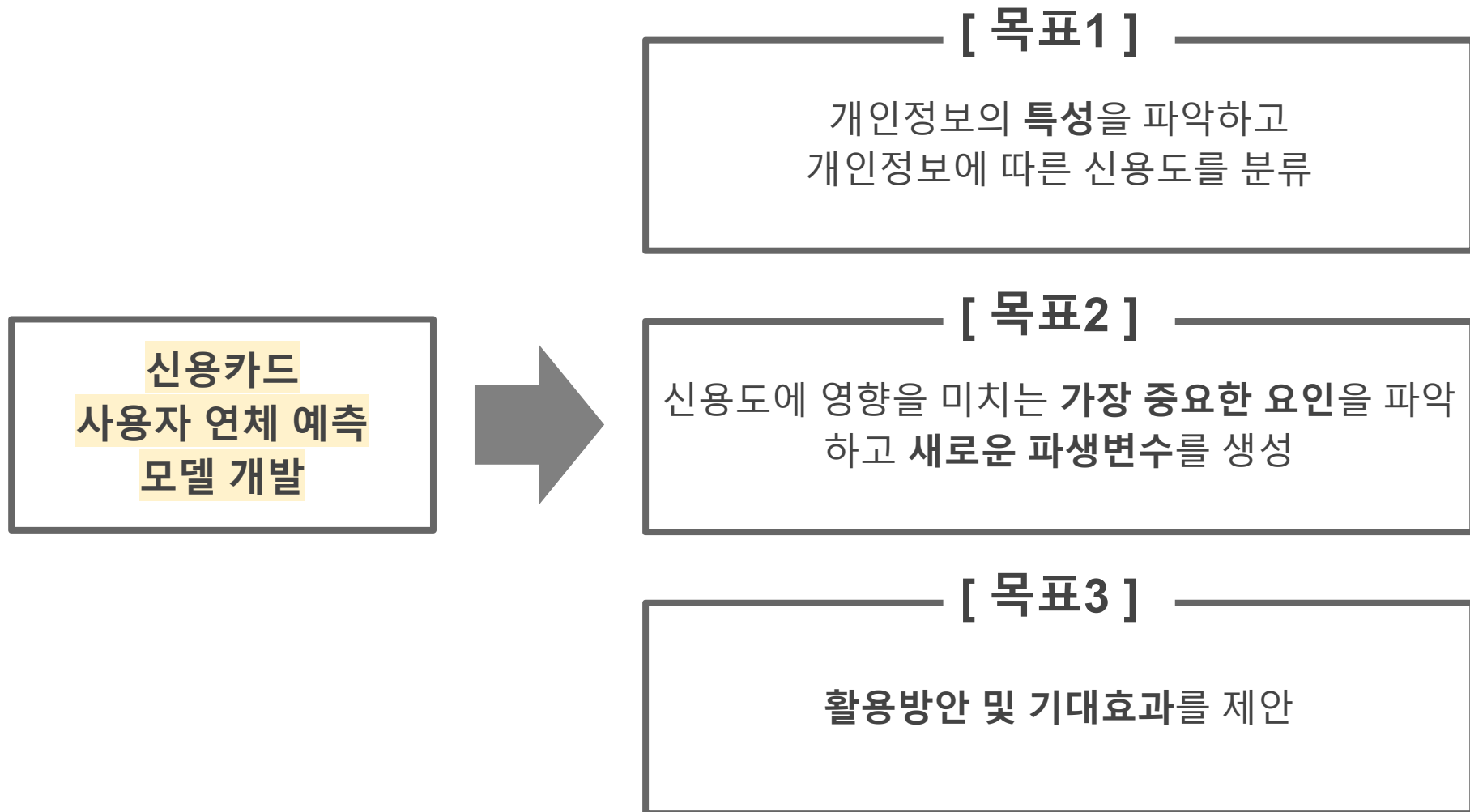
안드로이드OS와 IOS 월 사용자 기준 ■ 1월 ■ 3월 ■ 6월



“ 최신 머신러닝 알고리즘을 이용해서 등급을 산출하는 테크핀 ”

1-2 분석 목표

1. 분석 주제 이해



2. 데이터 이해 및 전처리

2-1 외부 데이터 활용 적합한가 (한국과 중국 문화의 차이)

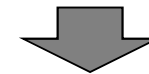
2. 데이터 이해 및 전처리

“ 신용카드를 부자를 위한 것이고,
모바일결제는 가난한 사람을 위해 만들어진 것이다.”



한국과 중국의 결제 문화 차이

공적 신용조회 시스템 無
위폐 범죄 多
외상 허용 불가
화폐폐 및 신용카드 사용 x
모바일 결제 문화 대중적

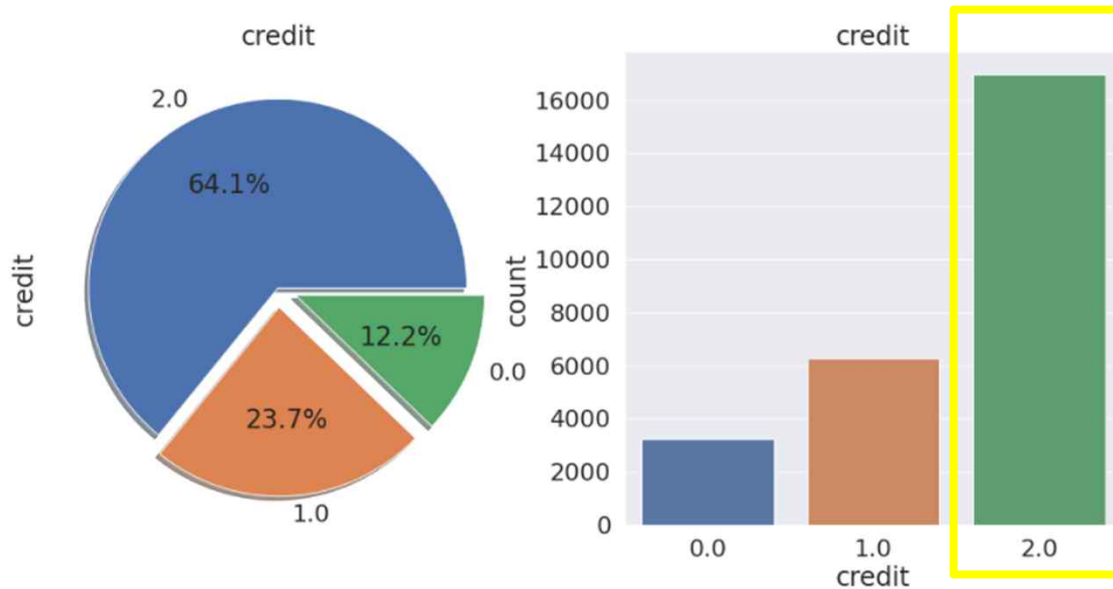


공공 데이터 활용 안함

2-2 데이터 이해 (1)

2. 데이터 이해 및 전처리

DAYS_BIRTH	begin_month	...	income_total	credit
-14431	-27	...	202500.0	0
-16954	-28	...	247500.0	1



train : (26457, 20)

test : (10000, 19)

[Target] Credit

사용자의 신용카드 대금
연체를 기준으로 **신용도**
(낮을 수록 높은 신용)

2-2 데이터 이해 (2)

2. 데이터 이해 및 전처리

Train 기초통계량

	child_num	income_total	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	work_phone	phone	email	family_size	begin_month
	26457.000000	2.645700e+04	26457.000000	26457.000000	26457.0	26457.000000	26457.000000	26457.000000	26457.000000	26457.000000
mean	0.428658	1.873065e+05	-15958.053899	59068.750728	1.0	0.224742	0.294251	0.091280	2.196848	-26.123294
std	0.747326	1.018784e+05	4201.589022	137475.427503	0.0	0.417420	0.455714	0.288013	0.916717	16.559550
min	0.000000	2.700000e+04	-25152.000000	-15713.000000	1.0	0.000000	0.000000	0.000000	1.000000	-60.000000
25%	0.000000	1.215000e+05	-19431.000000	-3153.000000	1.0	0.000000	0.000000	0.000000	2.000000	-39.000000
50%	0.000000	1.575000e+05	-15547.000000	-1539.000000	1.0	0.000000	0.000000	0.000000	2.000000	-24.000000
75%	1.000000	2.250000e+05	-12446.000000	-407.000000	1.0	0.000000	1.000000	0.000000	3.000000	-12.000000
max	19.000000	1.575000e+06	-7705.000000	365243.000000	1.0	1.000000	1.000000	1.000000	20.000000	0.000000

Test 기초통계량

	child_num	income_total	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	work_phone	phone	email	family_size	begin_month
count	10000.000000	1.000000e+04	10000.000000	10000.000000	10000.0	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	0.434700	1.850433e+05	-16020.466400	59776.690400	1.0	0.227600	0.296300	0.085600	2.202700	-26.272400
std	0.729102	1.015398e+05	4197.672887	138121.224504	0.0	0.419304	0.456648	0.279786	0.898272	16.348557
min	0.000000	2.700000e+04	-25152.000000	-15661.000000	1.0	0.000000	0.000000	0.000000	1.000000	-60.000000
25%	0.000000	1.215000e+05	-19483.250000	-3153.000000	1.0	0.000000	0.000000	0.000000	2.000000	-39.000000
50%	0.000000	1.575000e+05	-15606.000000	-1577.000000	1.0	0.000000	0.000000	0.000000	2.000000	-25.000000
75%	1.000000	2.250000e+05	-12539.000000	-410.000000	1.0	0.000000	1.000000	0.000000	3.000000	-12.000000
max	5.000000	1.575000e+06	-7489.000000	365243.000000	1.0	1.000000	1.000000	1.000000	7.000000	0.000000

2-2 데이터 이해 (3)

2. 데이터 이해 및 전처리

Categorical Features

[Binary]

gender : 성별
car : 차량 소유 여부
reality : 부동산 소유 여부
FLAG_MOBIL: 핸드폰 소유 여부
work_phone : 업무용 전화 소유 여부
phone : 전화 소유 여부
email : 이메일 소유 여부

[Multi]

child_num : 자녀수
family_size : 가족 규모
house-type : 생활 방식
occyp_type : 직업유형
income_type: 소득 분류
family_type : 결혼 여부

[Ordinal]

edu_type: 교육 수준
credit : 신용 기준(타겟)

2-2 데이터 이해 (4)

2. 데이터 이해 및 전처리

```
[18] # 자녀 수
data['child_num'].unique()

array([ 0, 1, 2, 3, 4, 5, 14, 19, 7])
```

```
[23] # income_total 생략 (연간 소득)
# 소득 분류
data['income_type'].unique()

array(['Commercial associate', 'Working', 'State servant', 'Pensioner',
       'Student'], dtype=object)
```

```
[25] # 교육 수준
data['edu_type'].unique()

array(['Higher education', 'Secondary / secondary special',
       'Incomplete higher', 'Lower secondary', 'Academic degree'],
      dtype=object)
```

```
[31] # 출생일
data['DAYS_BIRTH'].unique()

array([-13899, -11380, -19087, ..., -14560, -20910, -23013])
```

```
[45] # 업무 시작일
data['DAYS_EMPLOYED'].unique()

array([-4709, -1540, -4434, ..., -329, -2268, -616])
```

```
# 결혼 여부
data['family_type'].unique()

array(['Married', 'Civil marriage', 'Separated', 'Single / not married',
       'Widow'], dtype=object)
```

```
[29] # 생활 방식
data['house_type'].unique()

array(['Municipal apartment', 'House / apartment', 'With parents',
       'Co-op apartment', 'Rented apartment', 'Office apartment'],
      dtype=object)
```

```
[54] # 직업 유형
data['occyp_type'].unique()

array([nan, 'Laborers', 'Managers', 'Sales staff',
       'High skill tech staff', 'Core staff', 'Drivers', 'Medicine staff',
       'Accountants', 'Realty agents', 'Security staff', 'Cleaning staff',
       'Private service staff', 'Cooking staff', 'Secretaries',
       'HR staff', 'IT staff', 'Low-skill Laborers',
       'Waiters/barmen staff'], dtype=object)
```

```
[56] # 가족 규모
# 결혼 유무와 중복 가능성
data['family_size'].unique()

array([ 2., 3., 4., 1., 5., 6., 7., 15., 20., 9.])
```

```
[59] # 신용카드 발급 개월수
data['begin_month'].unique()

array([-6., -5., -22., -37., -26., -18., -41., -53., -38., -40., -51.,
       -60., -2., -14., -7., -35., -4., -13., -57., -47., -33., -30.,
       -20., -8., -39., -21., -19., -24., -48., -12., -10., -42., -29.,
       -3., -23., -25., -1., -15., -32., -59., -54., -34., 0., -27.,
       -45., -56., -46., -9., -44., -36., -43., -49., -11., -55., -58.,
       -28., -52., -17., -50., -16., -31.])
```

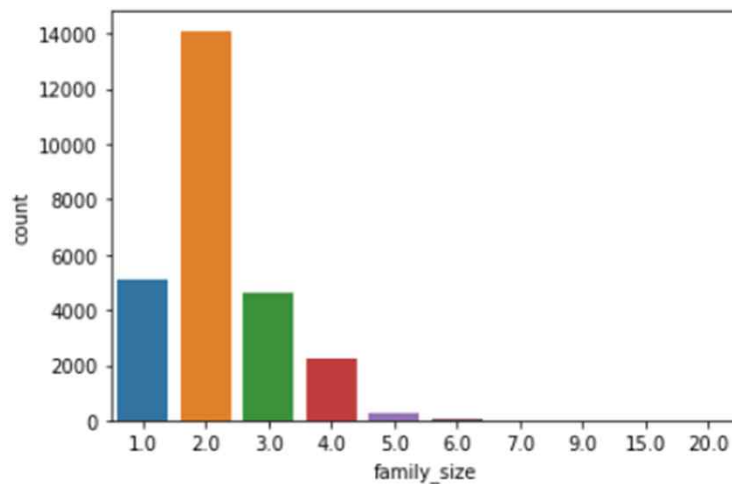
2-3 이상치 처리 (1)

2. 데이터 이해 및 전처리

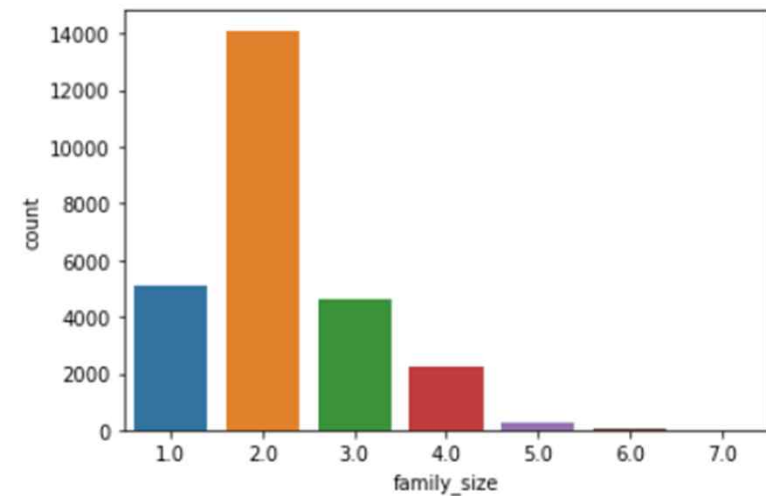
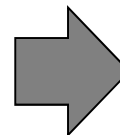
family_size : 가족 규모 > 7 제거

test data에는 없는 입력값

9, 15, 20인 데이터는 각각 한명의 동일인물인 데이터



[제거 전]



[제거 후]

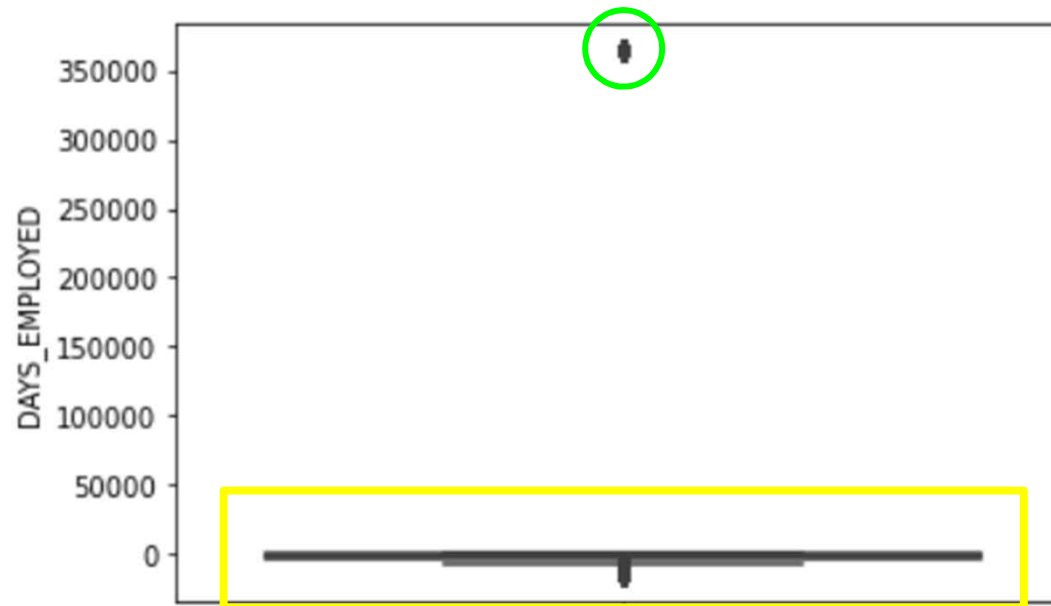
2-3 이상치 처리 (2)

2. 데이터 이해 및 전처리

DAYS_EMPLOYED : 업무시작일

데이터 수집 당시 (0)부터 역으로 셈
즉 -1은 데이터 수집일 하루 전부터 일을 의미
양수 값은 고용되지 않은 상태를 의미함

양수 값, 365243이 존재
양수 값, 365243의 income_type은 모두 "Pensioner", 즉 연금 수령자
전처리 : 365243을 0으로 바꿔서 진행



2-4 결측치 처리

2. 데이터 이해 및 전처리

[Train, Test 결측치]

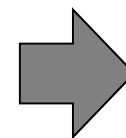
index	0	index	0
gender	0	gender	0
car	0	car	0
reality	0	reality	0
child_num	0	child_num	0
income_total	0	income_total	0
income_type	0	income_type	0
edu_type	0	edu_type	0
family_type	0	family_type	0
house_type	0	house_type	0
DAYS_BIRTH	0	DAYS_BIRTH	0
DAYS_EMPLOYED	0	DAYS_EMPLOYED	0
FLAG_MOBIL	0	FLAG_MOBIL	0
work_phone	0	work_phone	0
phone	0	phone	0
email	0	email	0
occyp_type	8171	occyp_type	3152
family_size	0	family_size	0
begin_month	0	begin_month	0
credit	0	dtype: int64	
dtype: int64			

[결측치 데이터의 소득유형]

Pensioner	0.543385
...	...
Student	0.000122

[소득유형 Pensioner의 직업유형]

NaN	0.997977
...	...
Managers	0.000225



결측치는 기존의 직업유형으로 대체하지 않고 **“NaN”**으로 대체

2-5 중복 데이터

2. 데이터 이해 및 전처리

DAYS_BIRTH	income_total	...	begin_month	credit
-14431	202500.0	...	-27	0
-14431	247500.0	...	-43	1

8755가지의 개인정보가 동일한 데이터가 존재

같은 개인정보 != 같은 credit
begin_month(발급월)에 따라 credit도 달라짐.

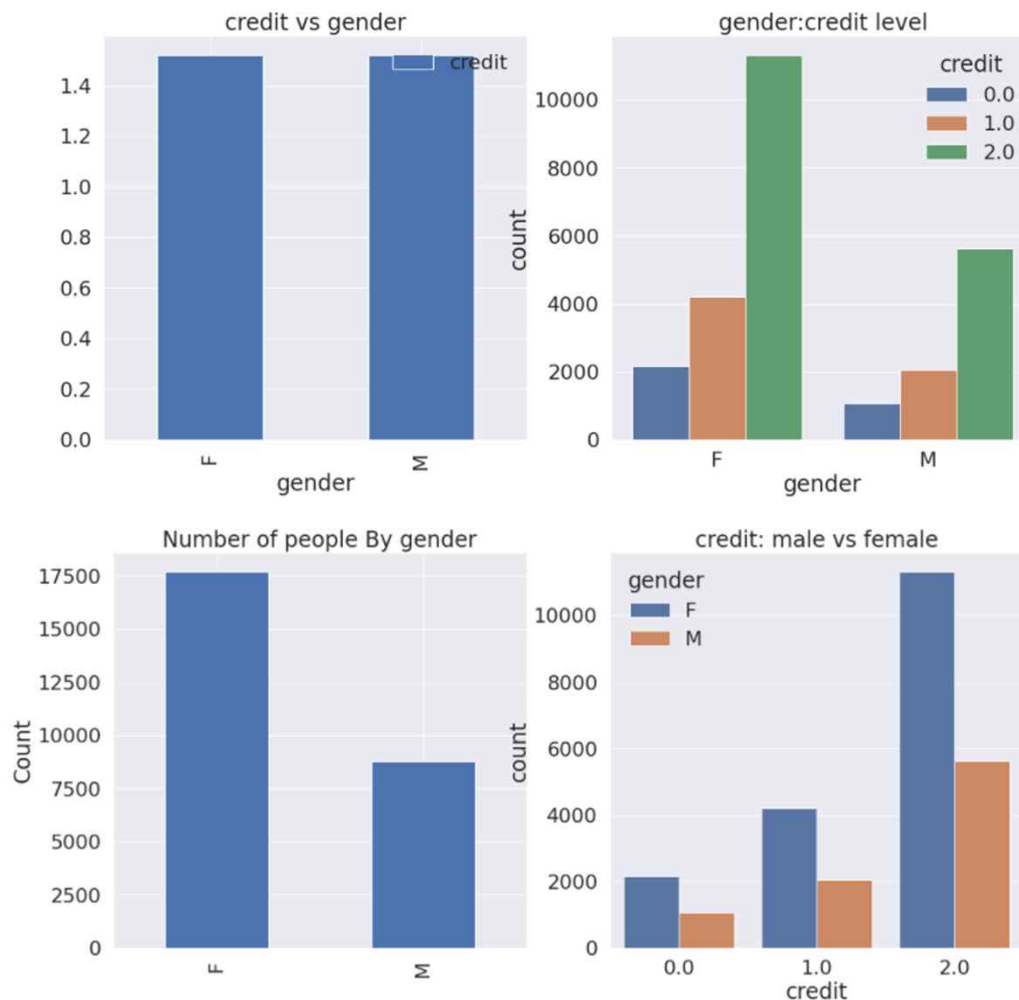
즉, 한 사람이 여러번 카드를 발급한 내역의 데이터.
발급할 때마다 항상 같은 신용도가 아니었기 때문이라고 추측
(이사람의 신용이? 아니면 신용척도가?)

※ 개인을 식별할 수 있는 변수를 만드는 것이 중요 ※

3. 데이터 탐색 (EDA)

3-1 변수별 신용도 EDA (1) - 비율 차이 X

3. 데이터 탐색 (EDA)



gender	F	M	All
credit			
0.0	2148	1074	3222
1.0	4220	2047	6267
2.0	11329	5639	16968
All	17697	8760	26457

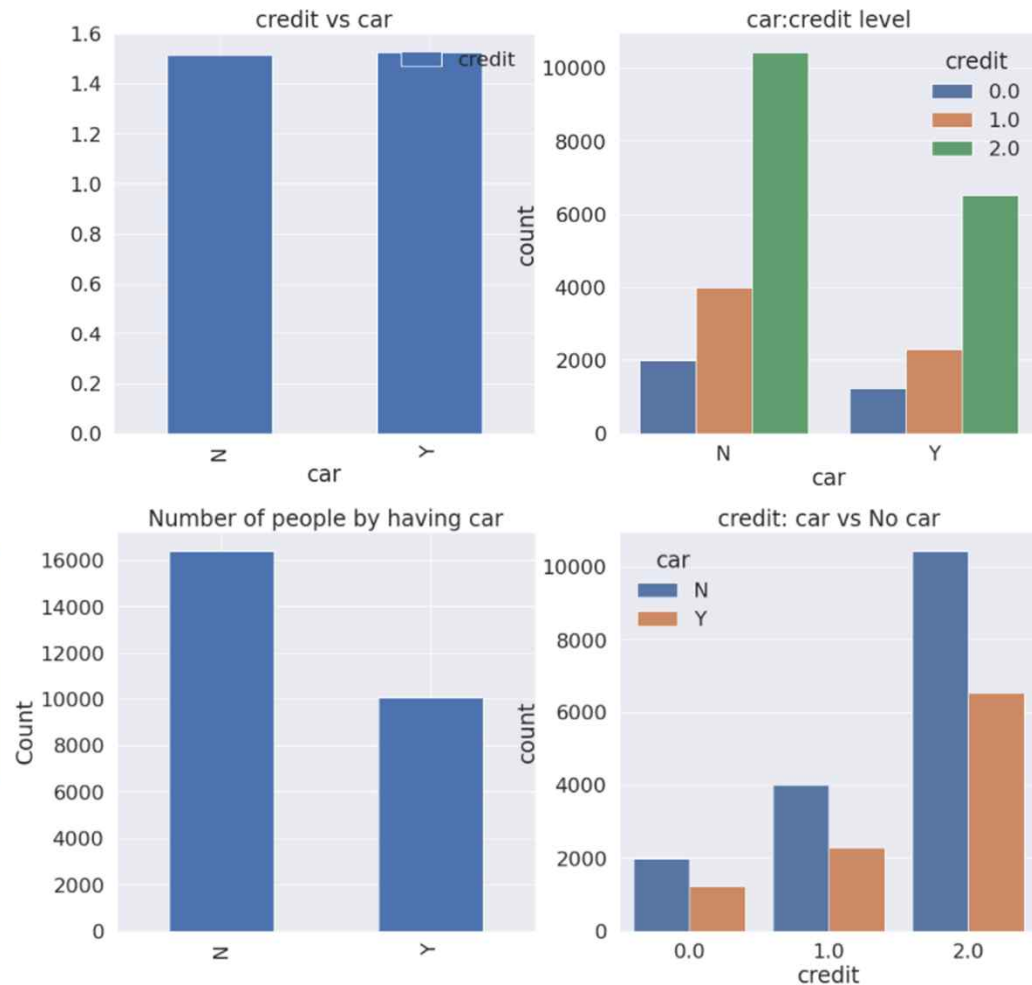
[gender vs credit]

3-1 변수별 신용도 EDA (2) - 비율 차이 X

3. 데이터 탐색 (EDA)

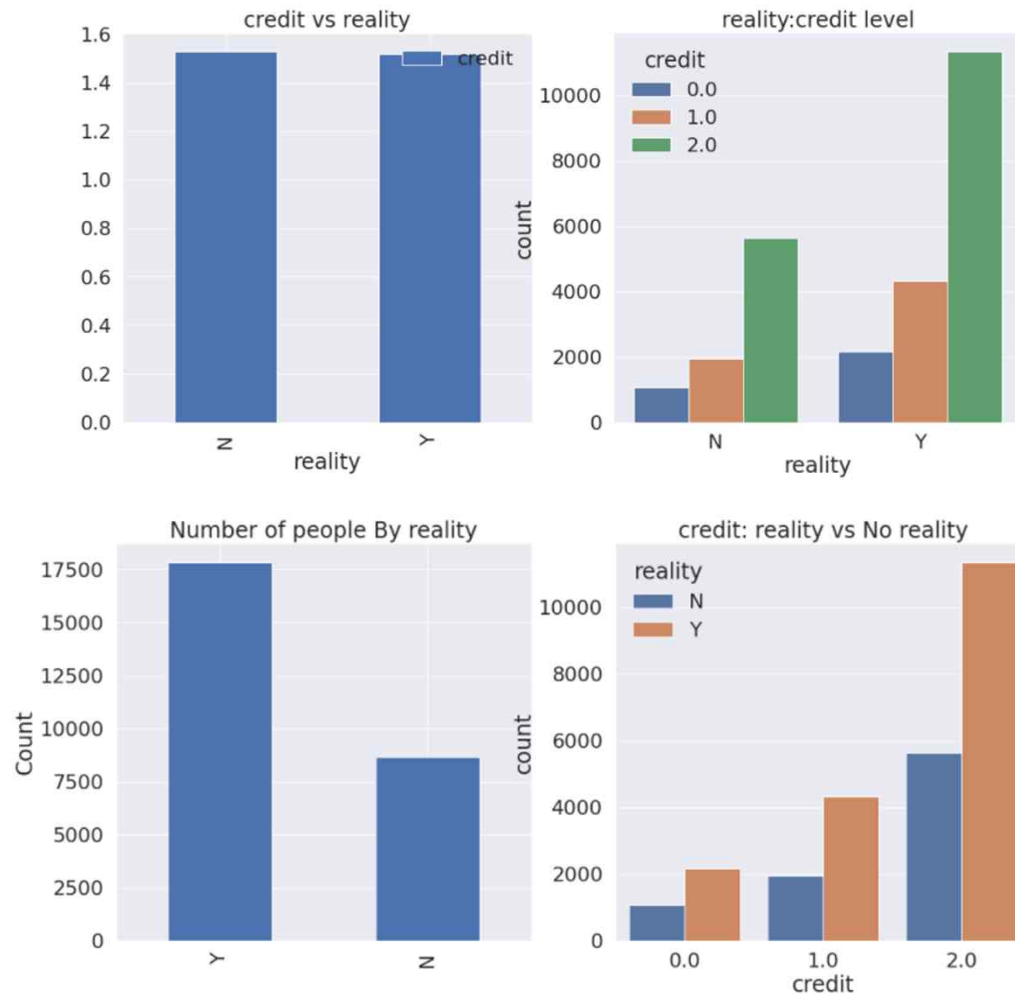
car	N	Y	All
credit			
0.0	1982	1240	3222
1.0	3990	2277	6267
2.0	10438	6530	16968
All	16410	10047	26457

[car vs credit]



3-1 변수별 신용도 EDA (3) - 비율 차이 X

3. 데이터 탐색 (EDA)

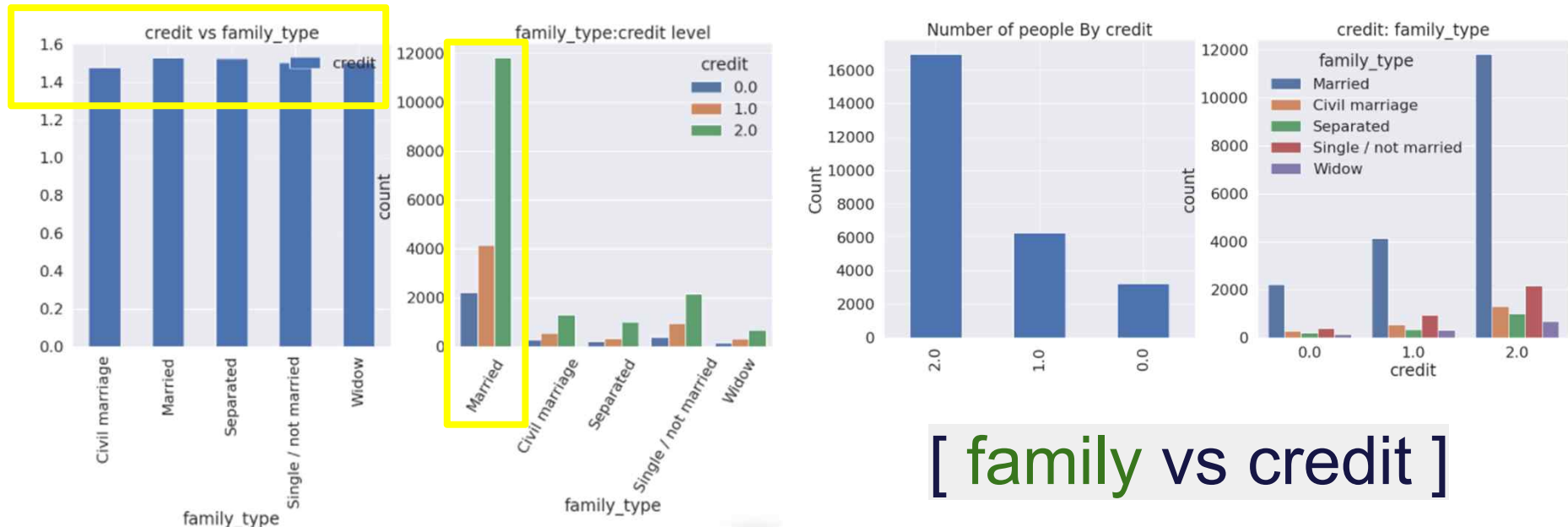


reality	N	Y	All
credit			
0.0	1064	2158	3222
1.0	1935	4332	6267
2.0	5628	11340	16968
All	8627	17830	26457

[reality vs credit]

3-1 변수별 신용도 EDA (4) - 비율 차이 O

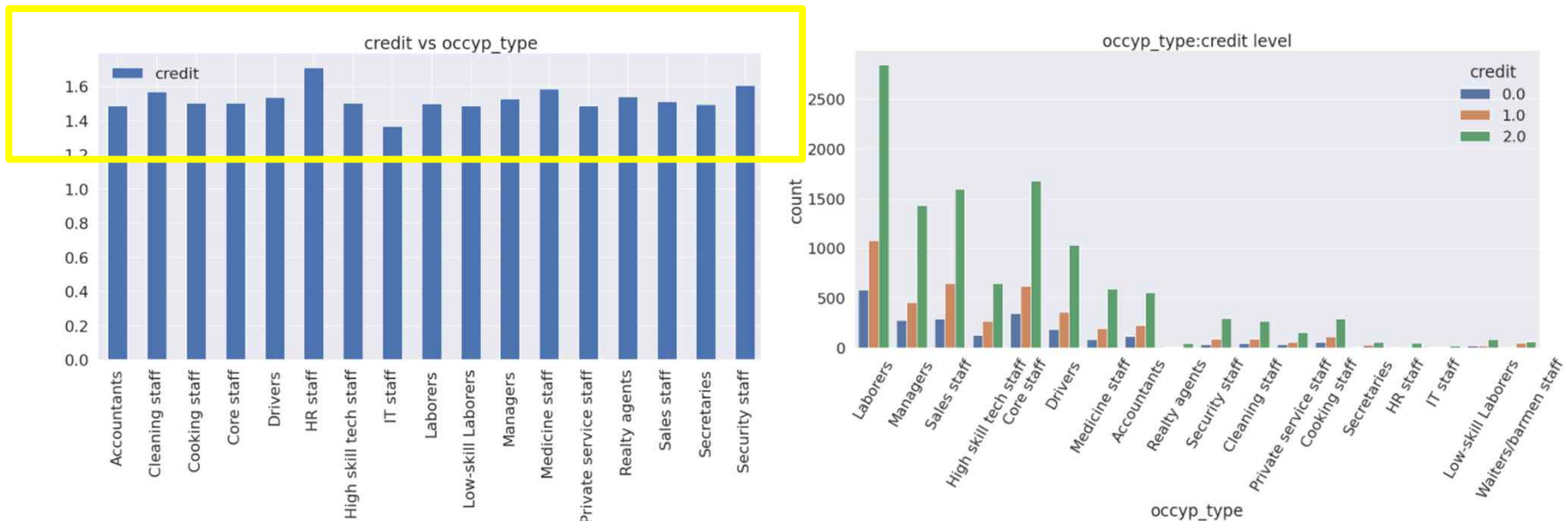
3. 데이터 탐색 (EDA)



family_type	Civil marriage	Married	Separated	Single / not married	Widow	All
credit						
0.0	288	2213	193	402	126	3222
1.0	539	4140	349	940	299	6267
2.0	1296	11843	997	2154	678	16968
All	2123	18196	1539	3496	1103	26457

3-1 변수별 신용도 EDA (5) - 비율 차이 O

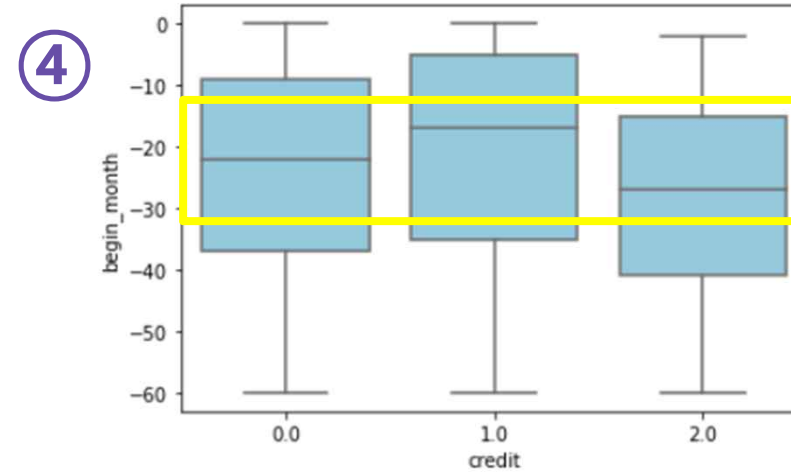
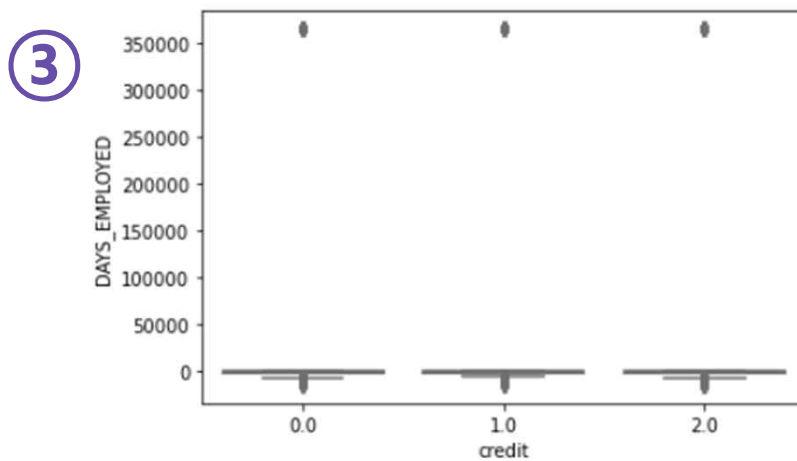
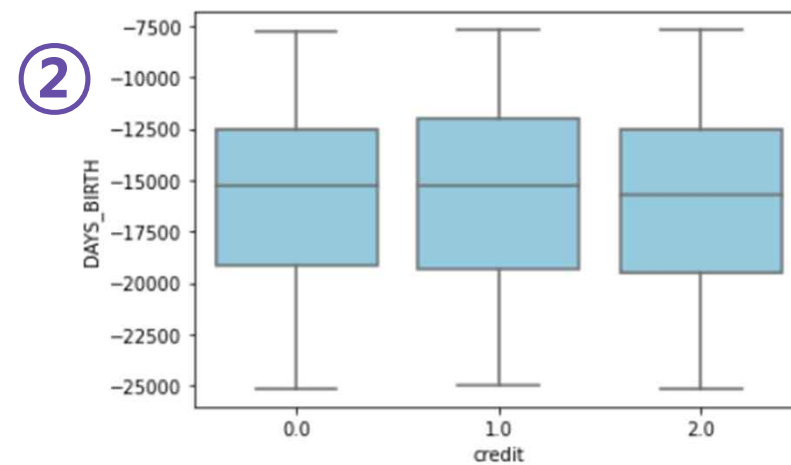
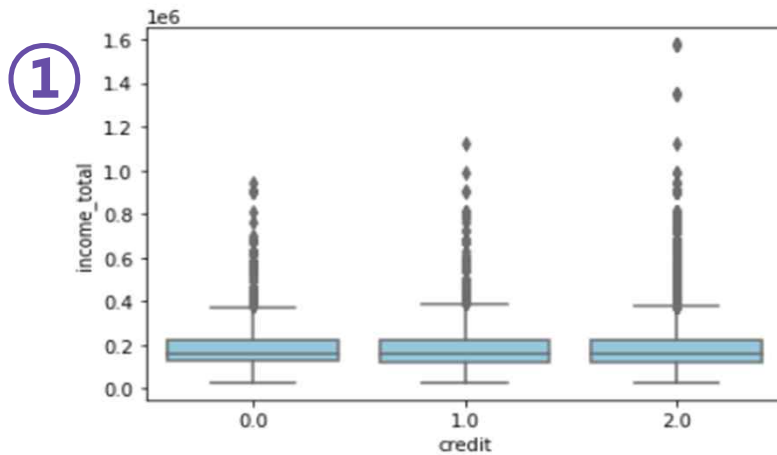
3. 데이터 탐색 (EDA)



[occyp_type vs credit]

3-2 신용도에 따른 연속형 변수 분포

3. 데이터 탐색 (EDA)



4. 파생 변수 생성

4-1 파생변수 생성

4. 파생 변수 생성

경진대회 1위를 수상한 코드의 파생변수 참고
더 나은 파생변수를 위한 고민

‘sincerity’

‘income_total’, ‘DAYS_BIRTH’, ‘DAYS_EMPLOYED’의 결합

‘punctuality’

기타 정보(성별, 재산 유무 등)의 활용

```
# before_EMPLOYED: 고용되기 전까지의 일수
df['before_EMPLOYED'] = df['DAYS_BIRTH'] - df['DAYS_EMPLOYED']
df['income_total_befofEMP_ratio'] = df['income_total'] / df['before_EMPLOYED']
```

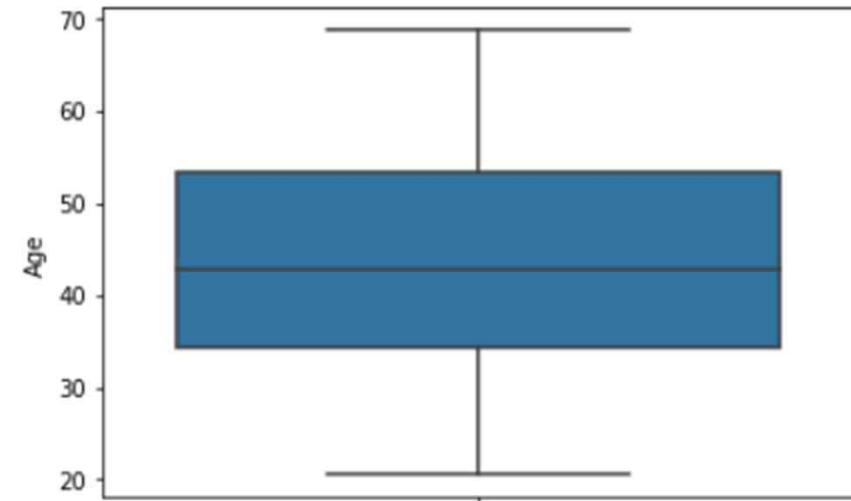
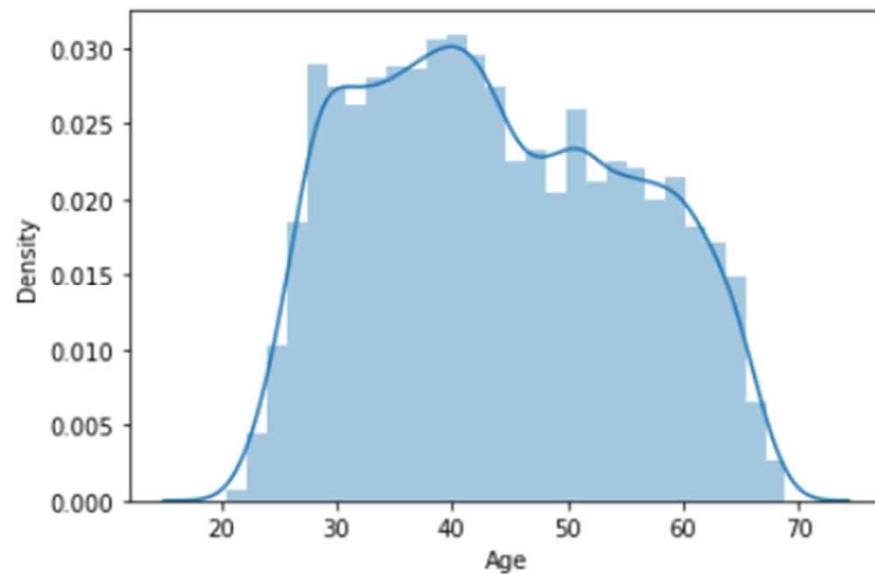
```
#ability: 소득/(살아온 일수+ 근무일수)
df['ability'] = df['income_total'] / (df['DAYS_BIRTH'] + df['DAYS_EMPLOYED'])
```

4-2 Age 변수 추가

4. 파생 변수 생성

[나이(Age) 변수 생성]

- 고객의 나이 나타냄
- 음수로 표현된 DAY_BIRTH 컬럼의 값을 절대값을 취해서 양수로 변경
- 365일로 나누어 나이로 표현

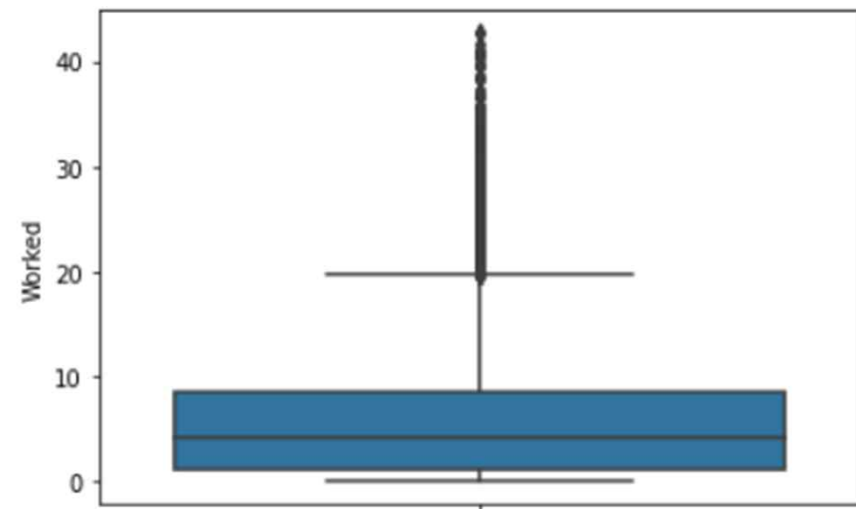
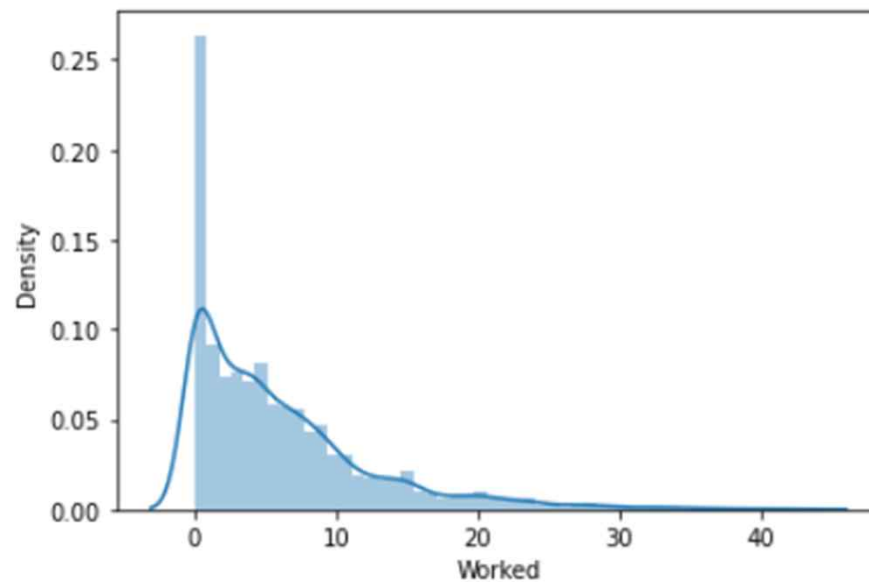


4-3 Worked 변수 추가

4. 파생 변수 생성

[업무기간(Worked) 변수 생성]

- 고객의 업무기간을 추측
- DAY_EMPLOYED를 365로 나누어 년수(소수점 1자리)로 표현
- 해당 컬럼의 값이 양수는 0으로 음수는 양수로 변경

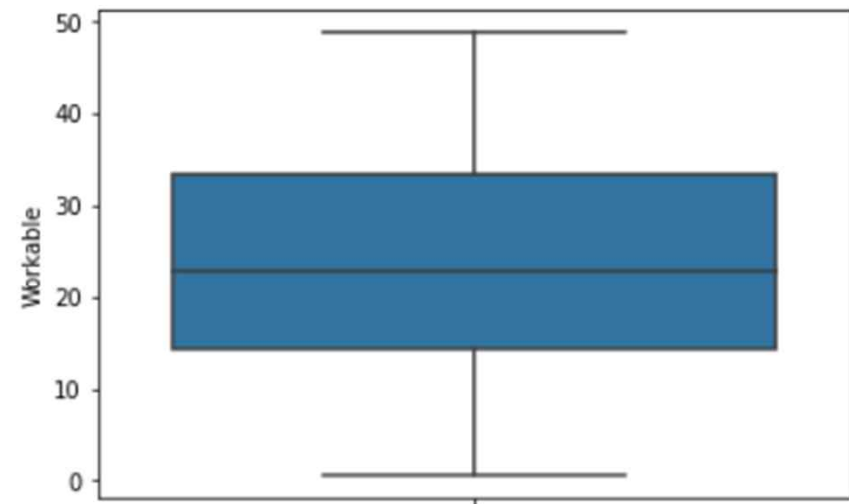
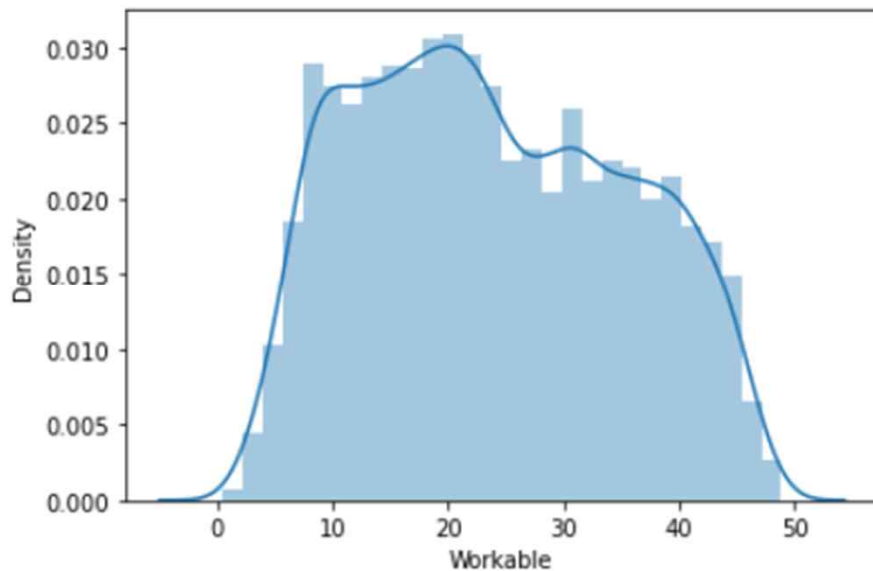


4-4 Workable 변수 추가

4. 파생 변수 생성

[업무가능연수(Workable) 변수 생성]

- 일반적으로 20살부터 일한다고 가정
- 나이(Age)에서 20을 빼서 해당나이에 근무 가능 년수를 추측함
- 업무가능연수(Workable)에서 업무기간(Worked)를 빼서 실제 실업기간을 추측함

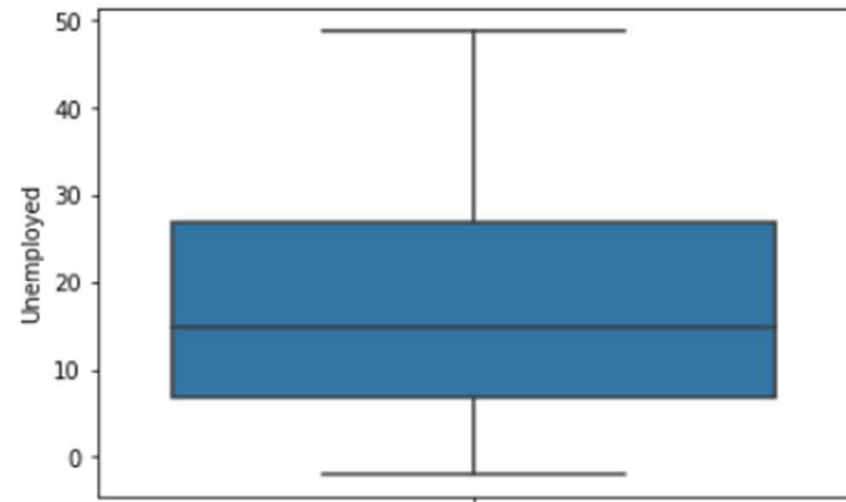
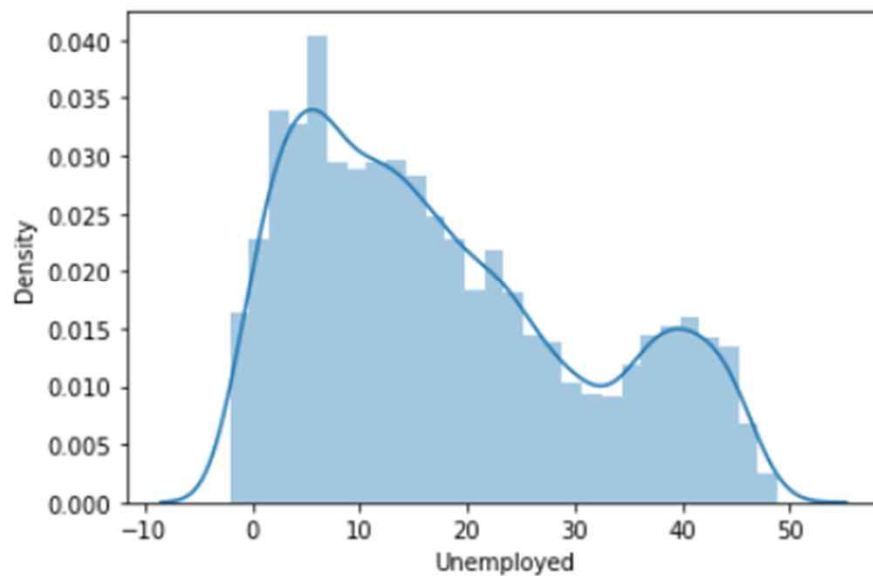


4-5 Unemployed 변수 추가

4. 파생 변수 생성

[실업기간(Unemployed) 변수 생성]

- 업무가능년수(Workable)에서 업무기간(Worked)를 빼서 실업기간을 추측함



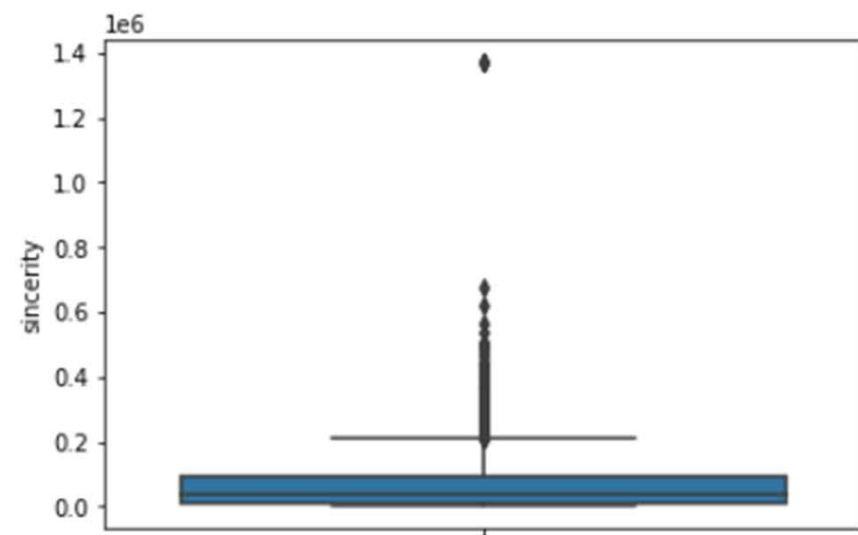
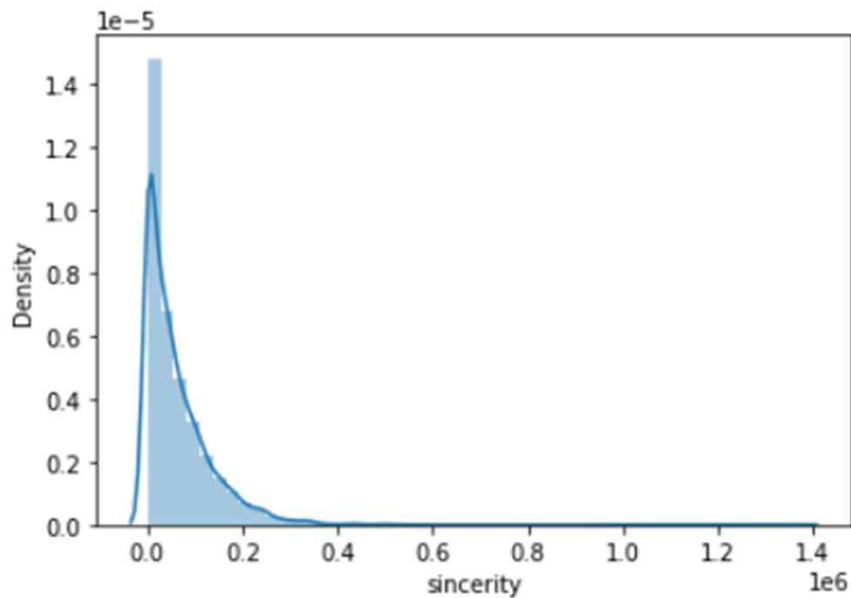
4-6 Sincerity 변수 생성

4. 파생 변수 생성

[성실성(sincerity) 변수 생성]

- 1위 코드에서 ability 칼럼을 보완함 -> 나이와 업무기간이 능력에 반비례 관계
- 업무가능기간 동안 실제 업무시간과 소득이 많을수록 성실도를 높게 추측
- 즉 일을 할 수 있을때 쉬지 않고 일한 사람들의 분별력을 얻기 위한 파생변수

```
df['sincerity'] = ( df['income_total'] * df['Worked'] / df['Workable'] ).round(2)
```



4-7 punctuality 변수 생성

4. 파생 변수 생성

[시간엄수(punctuality) 변수 생성]

- 시간엄수 네이밍: 납입기일을 잘 지킨다는 의미
- 자녀가 없고, 자동차 없고, 성별은 여성, 교육수준이 고등교육에 미치지 못할 경우에 신용도(credit)가 나쁘다고 (잔인하게)가정
- 임의로 각 값에 중요도를 부여
- 만점(216) 가정 -> 자녀수(5), 자동차 보유(2), 성별(2), 교육수준(3)

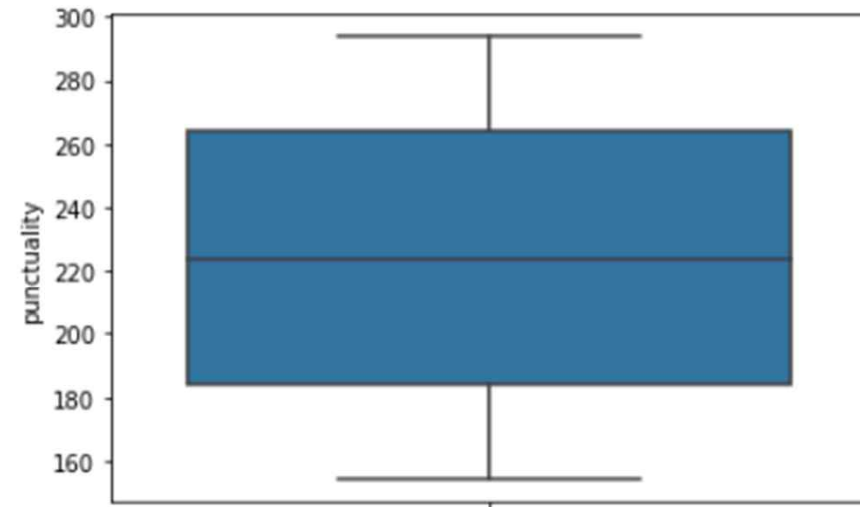
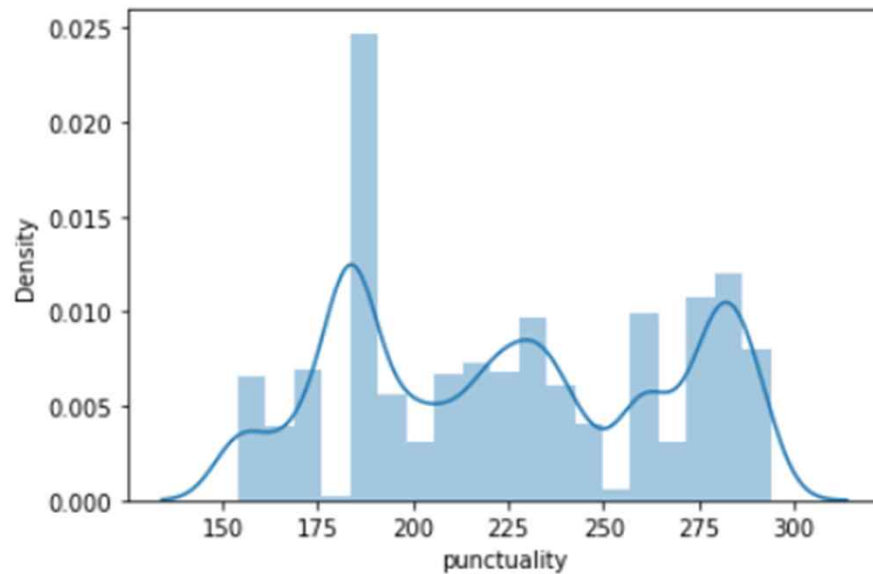
[시간엄수(punctuality) 변수에 추가 점수 부여]

- 가족유형(family_type)와 차량 유무(Car)에 따른 추가 점수 부여
- 가중치 맞추기 위해 10을 곱함

4-8 punctuality 변수 생성

4. 파생 변수 생성

```
df['punctuality'] = 216 - (6-df['child_num']) * (3-df['car']) * (3-df['gender']) * (4-df['edu_type'])  
df['punctuality'] = df['punctuality'] + (df['family_type'] * df['car'] * 10)
```



4-9 income_mean, ID 변수 생성

4. 파생 변수 생성

[income_mean]

가족 수에 따라 소비가 다르게 발생할 것이라는 가정
income_total을 family_size로 나눔

[ID]

중복 이용 개인 식별 변수
각 컬럼의 값을 더해서 고유한 한 사람을 파악

```
df['ID'] = \
df['child_num'].astype(str) + '_' + df['income_total'].astype(str) + '_' + \
df['DAYS_BIRTH'].astype(str) + '_' + df['DAYS_EMPLOYED'].astype(str) + '_' + \
df['work_phone'].astype(str) + '_' + df['phone'].astype(str) + '_' + \
df['email'].astype(str) + '_' + df['family_size'].astype(str) + '_' + \
df['gender'].astype(str) + '_' + df['car'].astype(str) + '_' + \
df['reality'].astype(str) + '_' + df['income_type'].astype(str) + '_' + \
df['edu_type'].astype(str) + '_' + df['family_type'].astype(str) + '_' + \
df['house_type'].astype(str) + '_' + df['occyp_type'].astype(str)
```

4-10 Cluster

4. 파생 변수 생성

Number of Categorical features: 12

['gender', 'car', 'reality',
'income_type', 'house_type', 'occyp_type', 'edu_type', 'family_type',
'work_phone', 'phone', 'email', 'ID']

Number of Numerical features: 11

['child_num', 'income_total', 'family_size', 'begin_month', 'Age', 'Worked', 'Workable',
'Unemployed', 'sincerity', 'income_mean', 'punctuality']

- 수치형 변수 사용하여 **K-means Clustering**

4-11 실제 모델링에 사용된 변수

4. 파생 변수 생성

gender	car	reality	child_num	income_total	income_type	edu_type	family_type	house_type	work_phone	phone	email	occyp_type
1	1	1	-0.589850	12.218505	1	3	4	1	0	0	0	1
1	1	2	0.795036	12.419174	1	1	2	2	0	0	1	2

생성 파생변수 9개

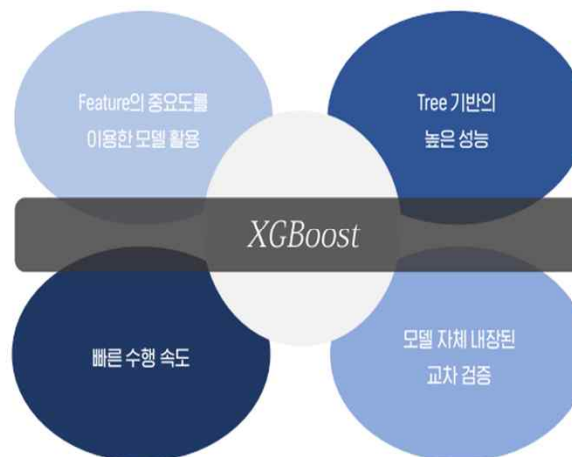
family_size	begin_month	credit	Age	Worked	Workable	Unemployed	sincerity	income_mean	ID	punctuality	cluster
-0.216262	-1.215242	1.0	-0.488305	1.058911	-0.488305	-0.937490	1.095506	0.002062	1	0.157702	19
0.897294	-1.275633	1.0	-1.087676	-0.280783	-1.087676	-0.802469	0.402356	-0.254157	2	-1.172746	32

5. 모델링 비교 평가 및 선택

5-1 후보 모델 선정

5. 모델링 비교 평가 및 선택

1. 다수의 **범주형 변수**가 존재
 1. 독립변수와 종속변수간의 선형적인 관계보다는 **비선형적인 관계**를 보임
 1. **높은 성능, 학습 속도**
- > **트리기반**의 알고리즘이 적절할 것으로 판단!!



5-1 후보 모델 선정

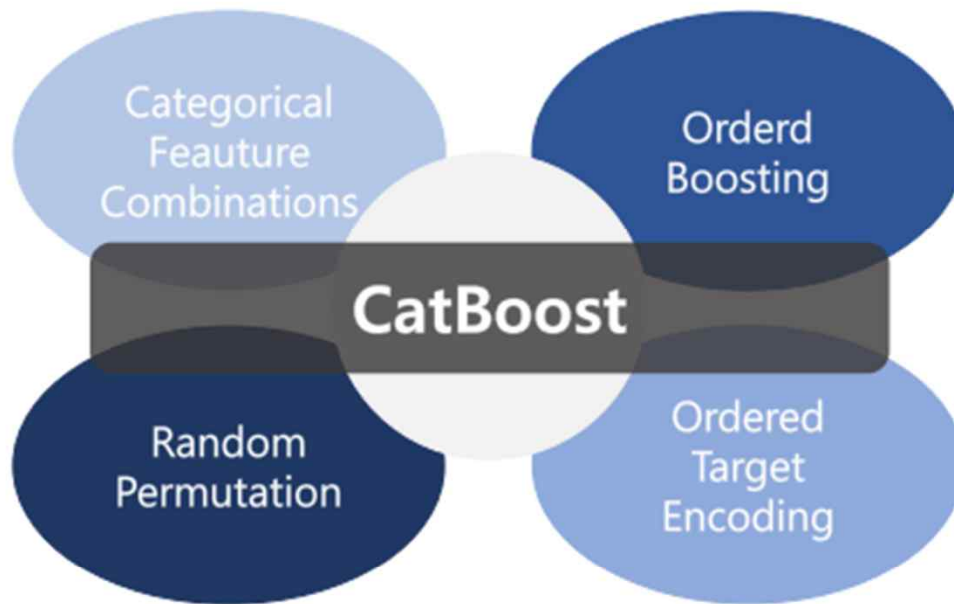
5. 모델링 비교 평가 및 선택



모델	Catboost	Random Forest	Light GBM	XGBoost
Log loss 성능	0.66449	0.69569	0.68475	0.69555
장점	높은 성능 예측 과적합 감소 범주형 변수 자동전처리 튜닝 간소화	앙상블 알고리즘 중 빠른 속도 룰이 명확 시각화 가능	XGBosst보다 빠른 학습 성능비슷 더적은메모리사용 범주형 변수 자동변환 최적 분할	병렬인 CPU 환경에서 GBM 대비 빠른 학습 과적합 규제 교차검증 수행 결손값 자체 처리 조기 중단 기능
단점	Sparse한 matrix 처리 못함 수치형인경우 학습 느림	하이퍼파라미터 너무 많음 튜닝을 위한 시간 소모	데이터가 작으면 과적합 발생	과한 반복학습 단축시 예측 성능이 안된 상태에서 학습 종료될 수 있음

후보 모델 선정

5. 모델링 비교 평가 및 선택

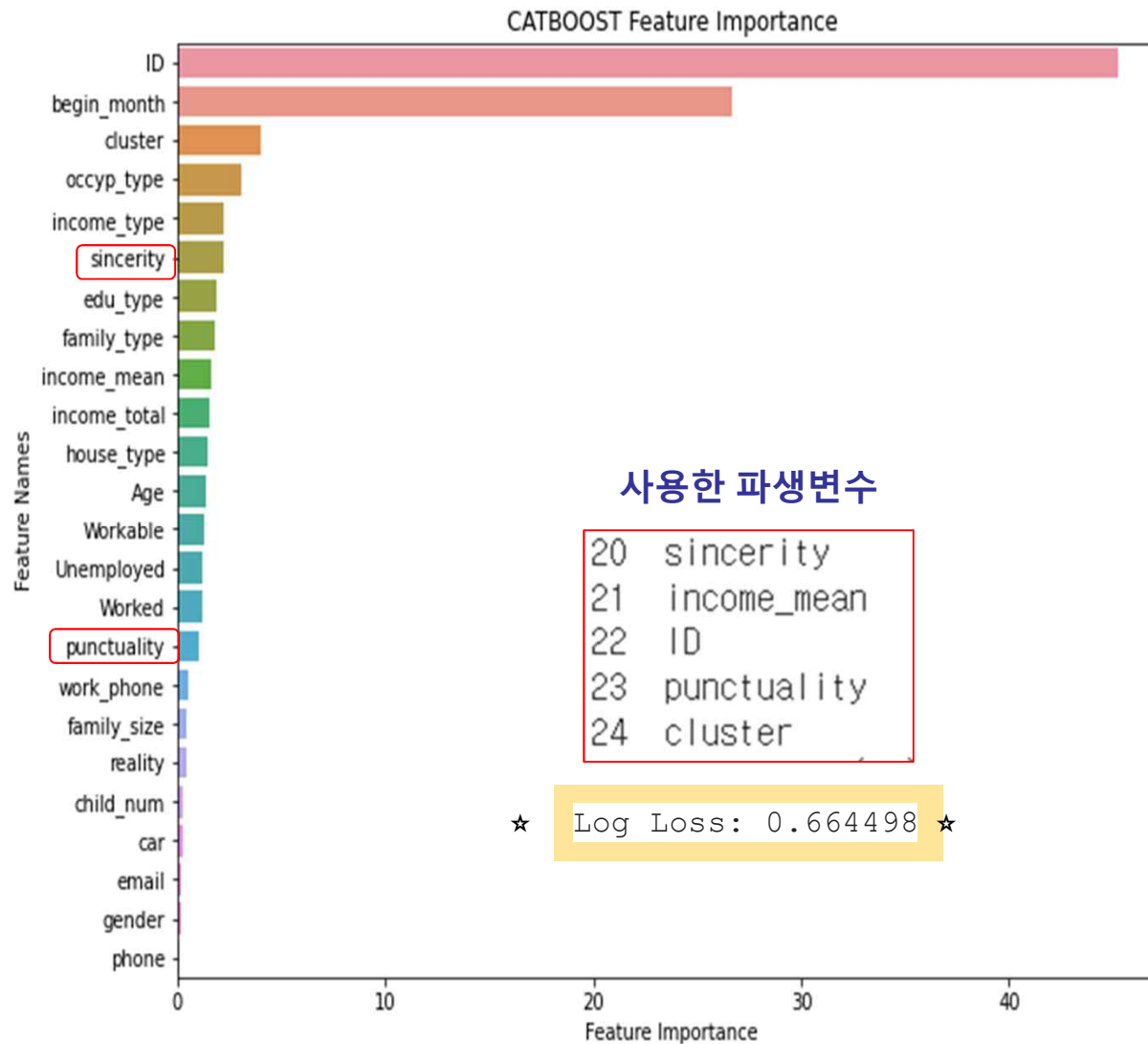


[K-Fold 교차검증(15 fold)]

1		...		
	2	...		
		...		
		...	14	
		...		15

CatBoost is a high-performance open source library for gradient boosting on decision trees

5. 모델링 비교 평가 및 선택



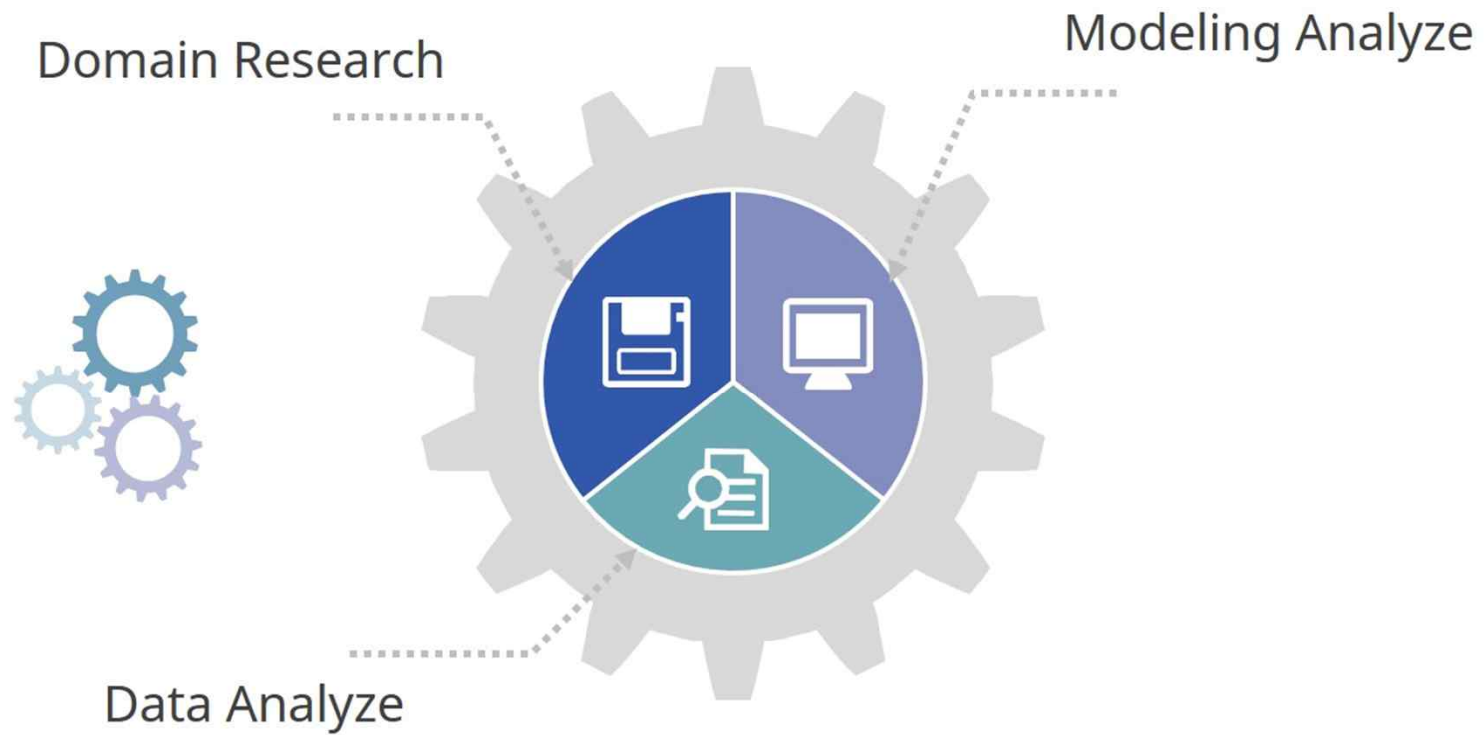
경진대회 1위 파생변수

19	ability
20	income_mean
21	ID
22	before_EMPLOYED
23	income_total_befofoEMP_ratio
24	cluster

Log Loss: 0.664545

6. 결론 및 활용방안

6-1 결론

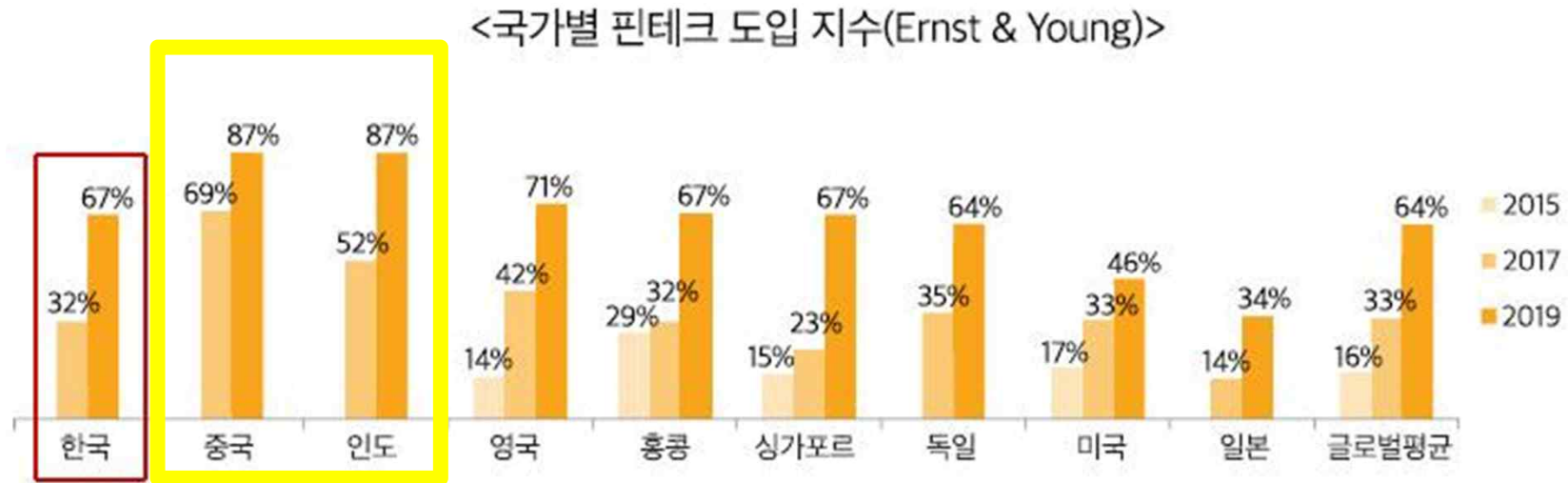


6-2 기업 입장에서의 활용 방안 (1)

6. 결론 및 활용방안



6-2 기업 입장에서의 활용 방안 (2) - 글로벌 시장 개척



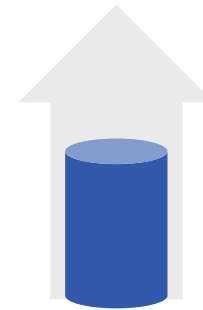
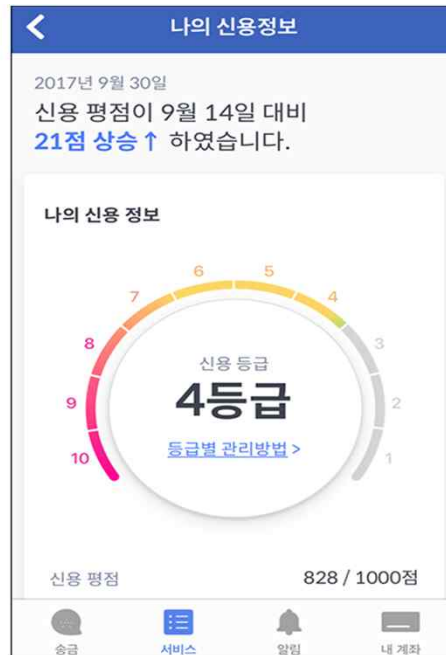
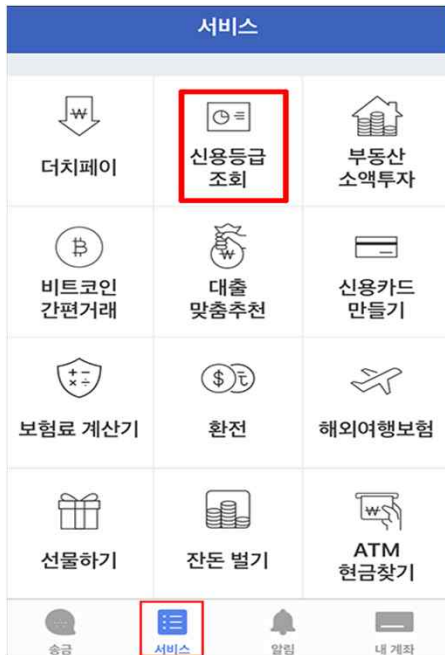
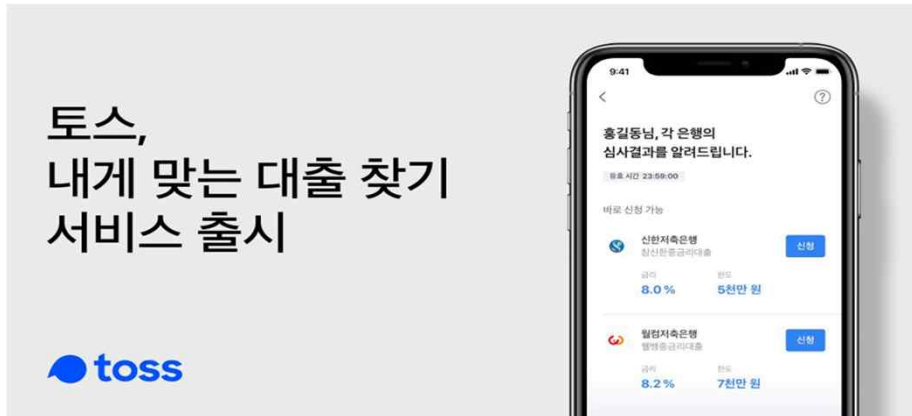
* EY(Ernst&Young) - 2019 Fintech Adoption Index (27개국 27,000명 대상 온라인 인터뷰 실시)

글로벌 시장 공략

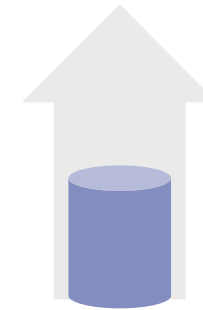
중국 뿐만 아니라 인도나 베트남 같은 신흥 시장에서 신규 핀테크 & My Data 비즈니스 가능
(ex. Toss App 의 인도 진출 시나리오)

6-3 고객 입장에서의 활용 방안

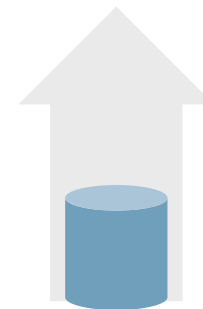
6. 결론 및 활용방안



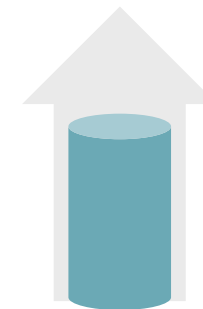
맞춤형
대출 서비스 이용



금융권 소외
계층 해소



소득 불평등
감소



신용정보
조회가능

7. 프로젝트 후기

7-1 프로젝트 후기 - “가장 훌륭한 선수는 팀이다”

7. 프로젝트 후기

[역할분담] 일 2회 회의	기술리더	중국 도메인 분석	EDA & 파생변수	모델링 비교분석
	윤태양, 노지윤	김성일, 김용삼	김혜리, 박정환	안심향



데이터 관련	프로젝트 경험	팀워크
<ul style="list-style-type: none"> 한국과 중국의 문화차이 외부 데이터 활용 신용도의 유연성 (한국 10) 변수 추가(요일, 거래내역) 교육유형 이해 어려움 	<ul style="list-style-type: none"> Dacon과 Kaggle 문제 풀이 다양한 머신러닝 모델 활용 전반적 데이터 분석 경험 다양한 논리를 가지고 시도 단기간 도메인 지식 축적 	<ul style="list-style-type: none"> 다양한 연령대의 팀원들과의 소통 존중과 협력 그리고 이해 각기 역량과 이해력 차이 극복 적절한 역할분담을 통한 목표달성



THANK YOU

FOR WATCHING