Credit Card Approval Prediction

위험 고객 관리를 위한 credit card 데이터 분석

Team1

박서영 박지혜 이민준 이상재

Credit Card Approval Prediction

INDEX

상황 가정

카드를 발급받고 쓰지않거나 연체가 지나치게 길어져
회사에 불이익을 주는 경우가 있다.
이에 예방 차원에서, 카드 발급 시 위험 고객을 분류를 위한 분석 의뢰가 들어왔다.

D	MONTHS_	STATUS
5001711	0	X
5001711	-1	0
5001711	-2	0
5001711	-3	0
5001712	0	C
5001712	-1	C
5001712	-2	C
5001712	-3	C
5001712	-4	C
5001712	-5	C
5001712	-6	C
5001712	-7	C
5001712	-8	C
5001712	-9	0
5001712	-10	0

Target data

Target

- 2개월 이상 연체된 경우
- 카드를 전혀 사용하지 않은 경우

→ Risk 값 == 1

- 그 외 나머지

→ Risk 값 == 0

Risk 환산

ID	STATUS		Risk
5001711	X		1
5001711	0		0
5001711	0		0
5001711	0		0

Mean		Final Risk
0.25	< 0.5	0

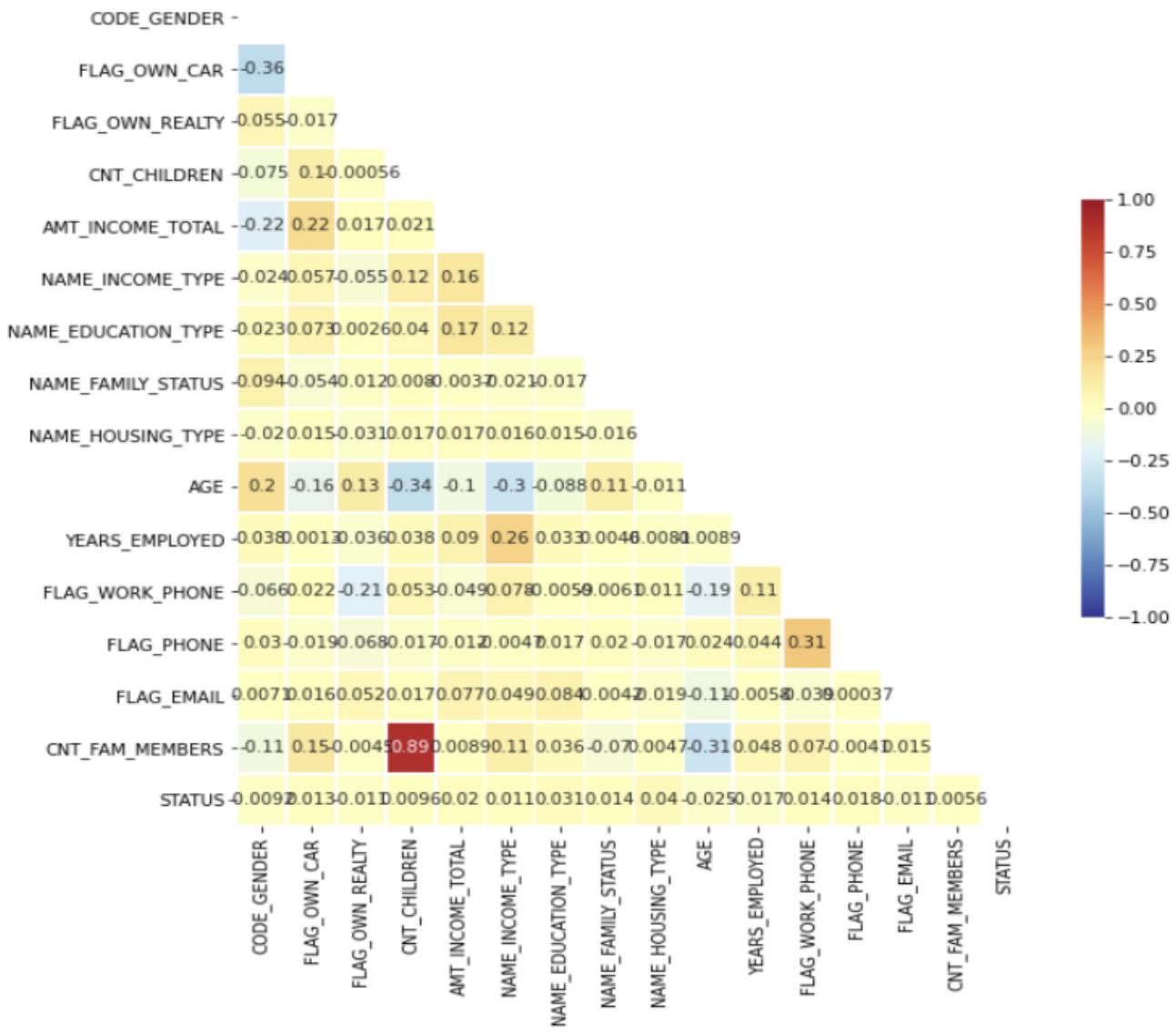
ID	STATUS		Risk
5001712	C		0
5001712	0		0
5001712	2		1
5001712	2		1
5001712	2		1

Mean		Final Risk
0.6	>= 0.5	1

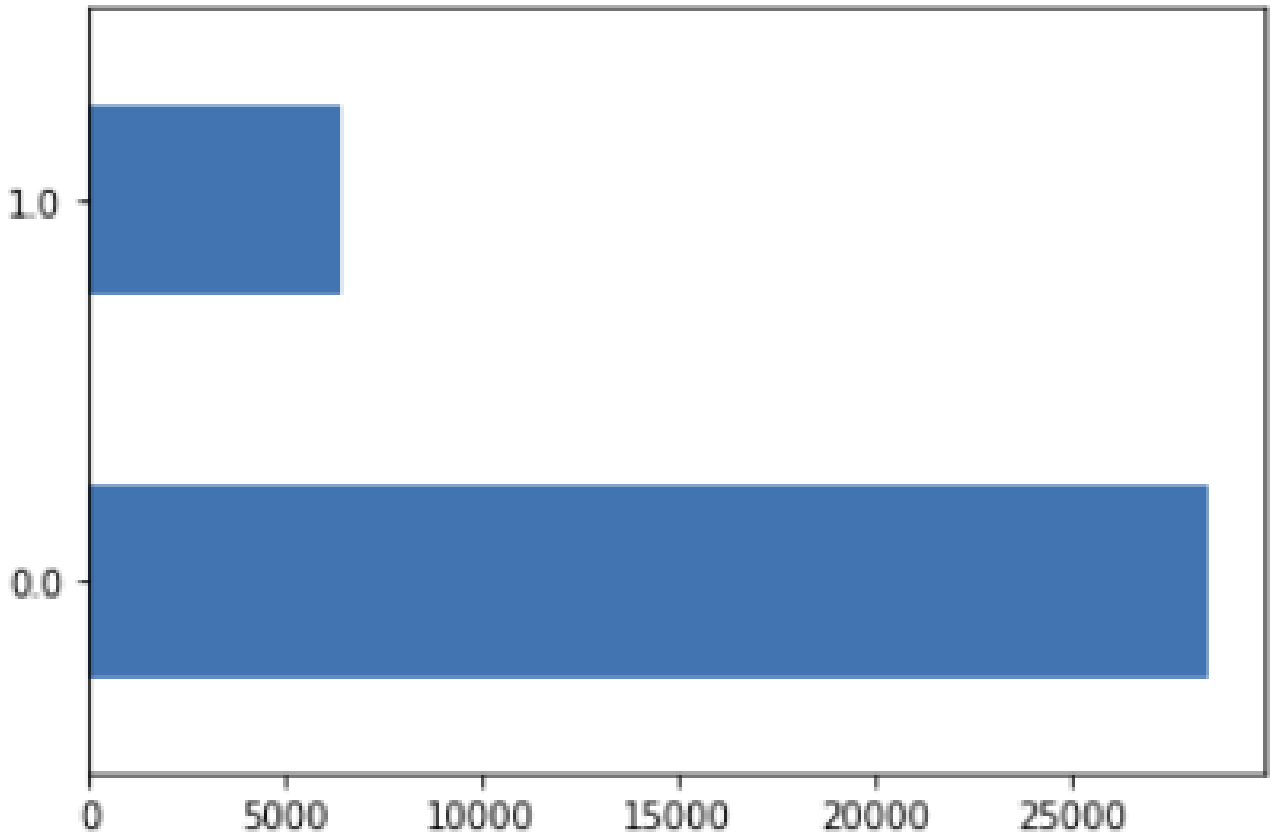
Features

- 성별
- 차 (유/무)
- 부동산 (유/무)
- 자녀 수
- 연봉
- 수입 종류 - 상인, 연금 수급자, 공무원, 학생, 그 외
- 가족 타입 – 결혼 여부, 따로 사는 경우, 미망인, 결혼으로 시민권 획득
- 거주 형태 - 쉐어아파트, 시립아파트, 오피스텔, 부모님집 거주
- 나이
- 근속년수
- 연락처
- 가족 인원수

상관 매트릭스


CODE_GENDER
FLAG_OWN_CAR -0.36
FLAG_OWN_REALTY -0.055 -0.017
CNT_CHILDREN -0.075 0.1 -0.00056
AMT_INCOME_TOTAL -0.22 0.22 0.017 0.021
NAME_INCOME_TYPE -0.024 0.057 -0.055 0.12 0.16
NAME_EDUCATION_TYPE -0.023 0.073 0.0026 0.04 0.17 0.12
NAME_FAMILY_STATUS -0.094 -0.054 0.012 0.008 0.0037 -0.021 -0.017
NAME_HOUSING_TYPE -0.02 0.015 -0.031 0.017 0.017 0.016 0.015 -0.016
AGE 0.2 -0.16 0.13 -0.34 -0.1 -0.3 -0.088 0.11 -0.011
YEARS_EMPLOYED -0.038 0.0013 0.036 0.038 0.09 0.26 0.033 0.0046 0.0081 0.0089
FLAG_WORK_PHONE -0.066 0.022 -0.21 0.053 -0.049 0.078 0.0059 0.0061 0.011 -0.19 0.11
FLAG_PHONE 0.03 -0.019 0.068 0.017 -0.012 0.0047 0.017 0.02 -0.017 0.024 0.044 0.31
FLAG_EMAIL 0.0071 0.016 0.052 0.017 0.077 0.049 0.084 0.0042 0.019 -0.11 0.0058 0.039 0.00037
CNT_FAM_MEMBERS -0.11 0.15 -0.0045 0.89 0.0089 0.11 0.036 -0.07 0.0047 -0.31 0.048 0.07 -0.0041 0.015
STATUS -0.0092 0.013 -0.011 0.0096 0.02 0.011 0.031 0.014 0.04 -0.025 0.017 0.014 0.018 -0.011 0.0056
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
CNT_CHILDREN
AMT_INCOME_TOTAL
NAME_INCOME_TYPE
NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS
NAME_HOUSING_TYPE
AGE
YEARS_EMPLOYED
FLAG_WORK_PHONE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
STATUS
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

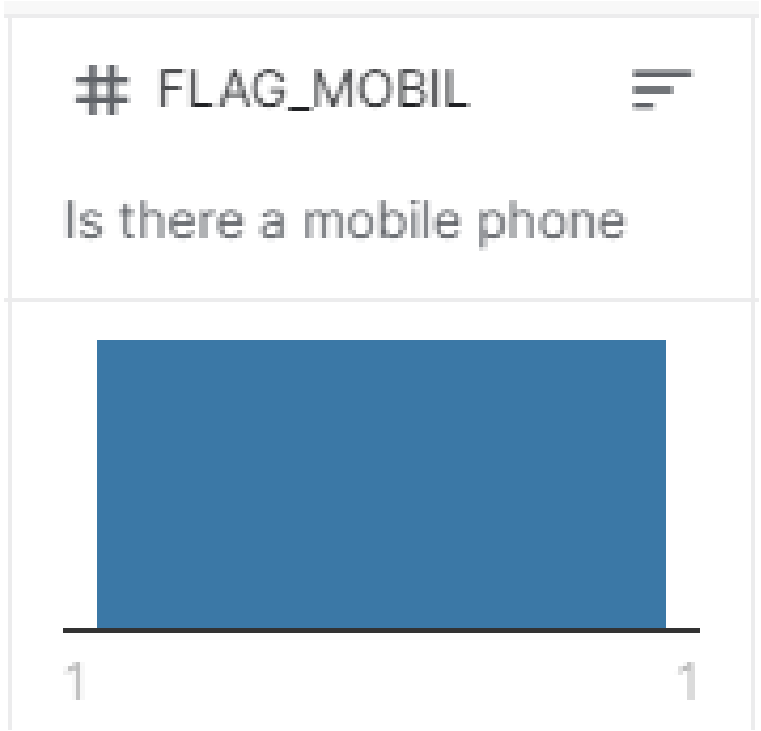
신용 상태 별 카운트



DAYS_EMPLOYED
-4542
-4542
-1134
-3051
-3051

YEARS_EMPLOYED
12
12
3
8
8

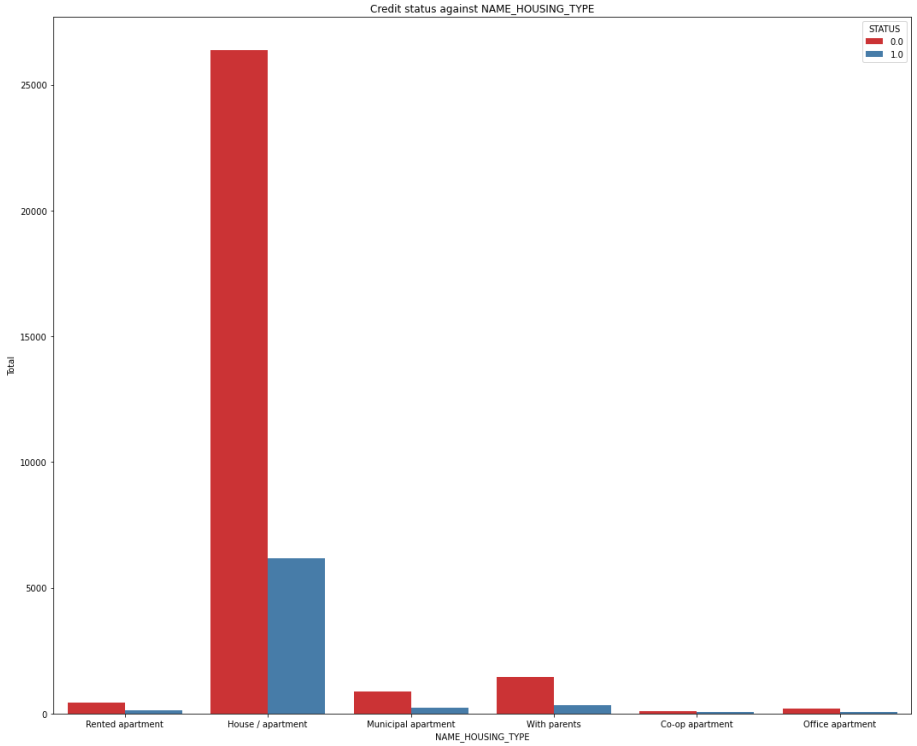
DAYS_BIRTH	AGE
-12005	32
-12005	32
-21474	58
-19110	52
-19110	52

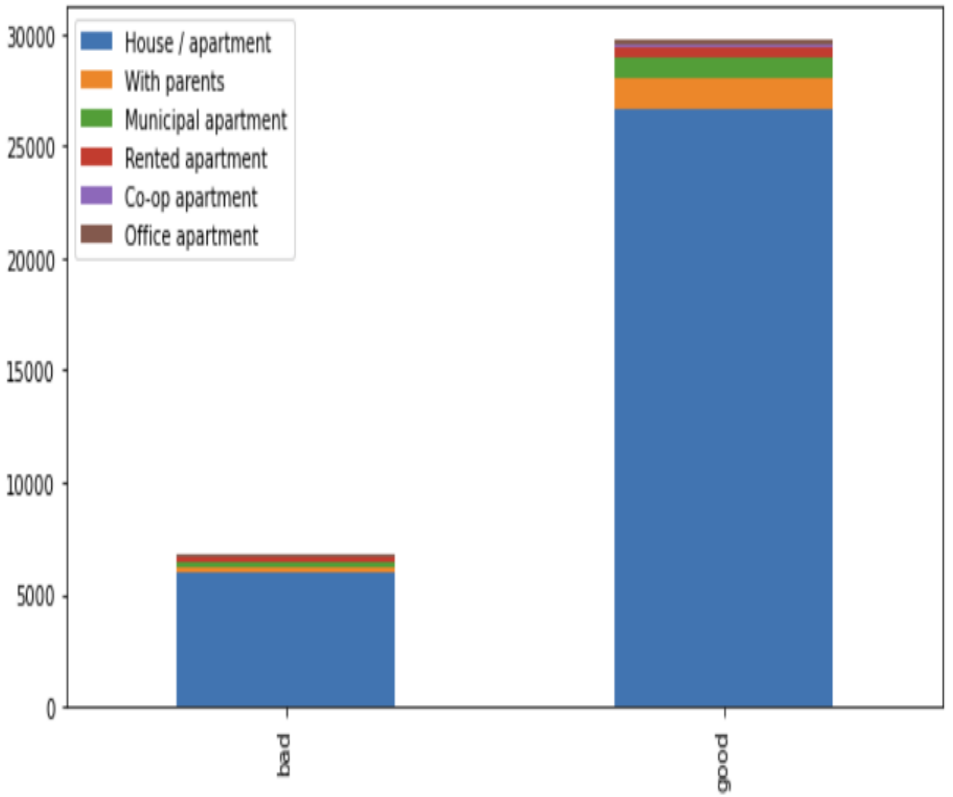


불필요한 컬럼 제거

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 36457 entries, 0 to 36456
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   36457 non-null  int64
 1   CODE_GENDER          36457 non-null  object
 2   FLAG_OWN_CAR         36457 non-null  object
 3   FLAG_OWN_REALTY      36457 non-null  object
 4   CNT_CHILDREN         36457 non-null  int64
 5   AMT_INCOME_TOTAL     36457 non-null  float64
 6   NAME_INCOME_TYPE     36457 non-null  object
 7   NAME_EDUCATION_TYPE  36457 non-null  object
 8   NAME_FAMILY_STATUS   36457 non-null  object
 9   NAME_HOUSING_TYPE    36457 non-null  object
 10  AGE                  36457 non-null  int64
 11  YEARS_EMPLOYED       36457 non-null  int64
 12  FLAG_WORK_PHONE      36457 non-null  int64
 13  FLAG_PHONE           36457 non-null  int64
 14  FLAG_EMAIL           36457 non-null  int64
 15  CNT_FAM_MEMBERS      36457 non-null  float64
 16  STATUS               36457 non-null  float64
dtypes: float64(3), int64(7), object(7)
memory usage: 5.0+ MB
```

NAME_HOUSING_TYPE





	NAME_HOUSING_TYPE	STATUS
0	Rented apartment	0.0
1	Rented apartment	0.0
2	House / apartment	1.0
3	House / apartment	1.0
4	House / apartment	1.0

	NAME_HOUSING_TYPE	STATUS
0	Co-op apartment	0.392857
1	House / apartment	0.182715
2	Municipal apartment	0.205674
3	Office apartment	0.194656
4	Rented apartment	0.206957

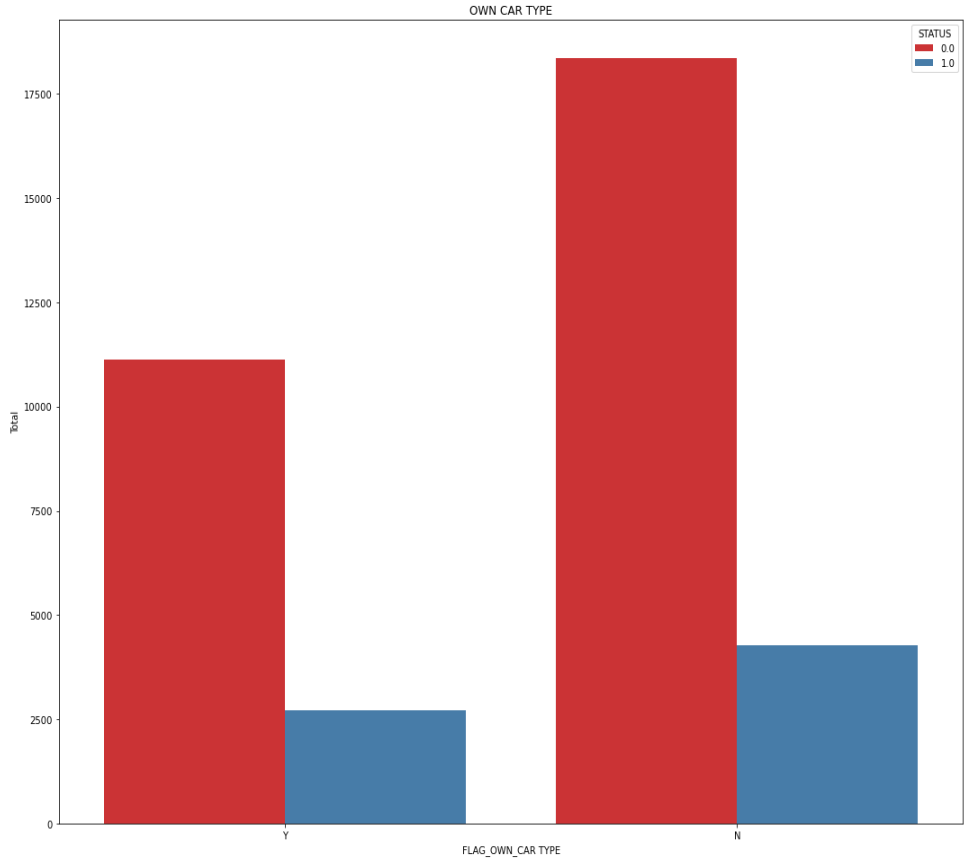
거주 타입 별 신용상태를 수치화

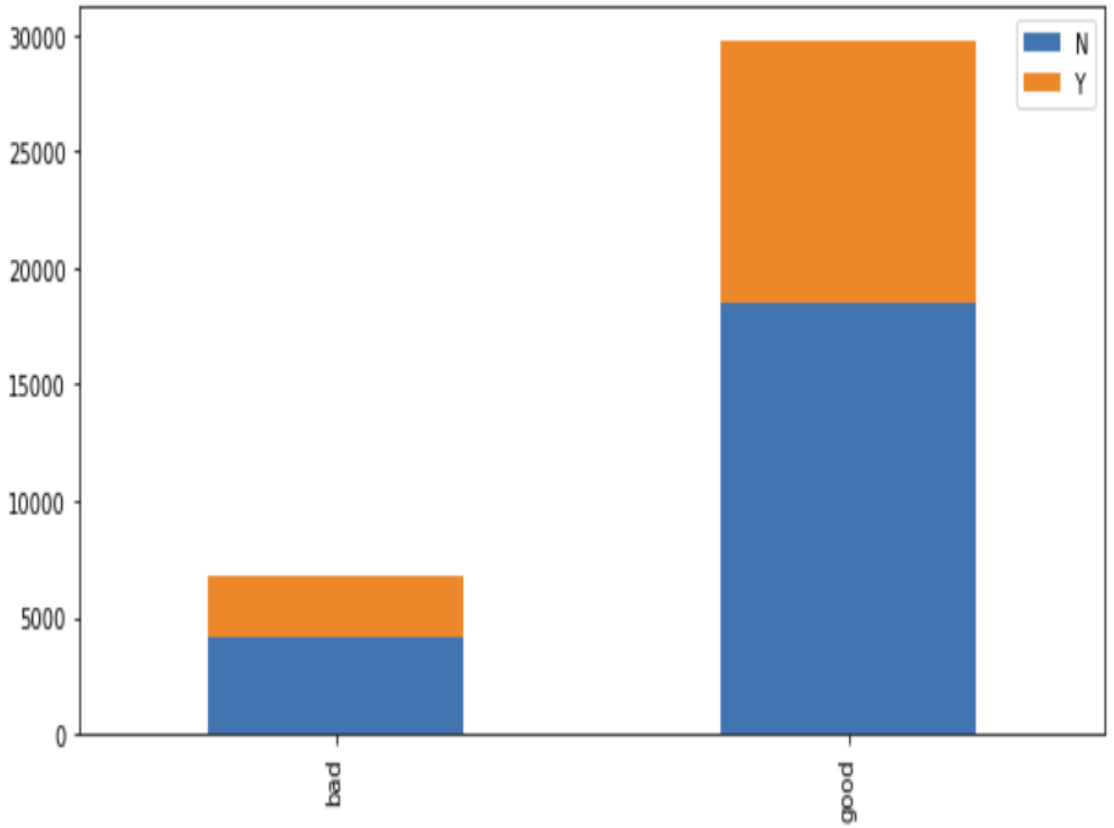
CODE_GENDER
M
M
M
F
F

CODE_GENDER
0
0
0
1
1

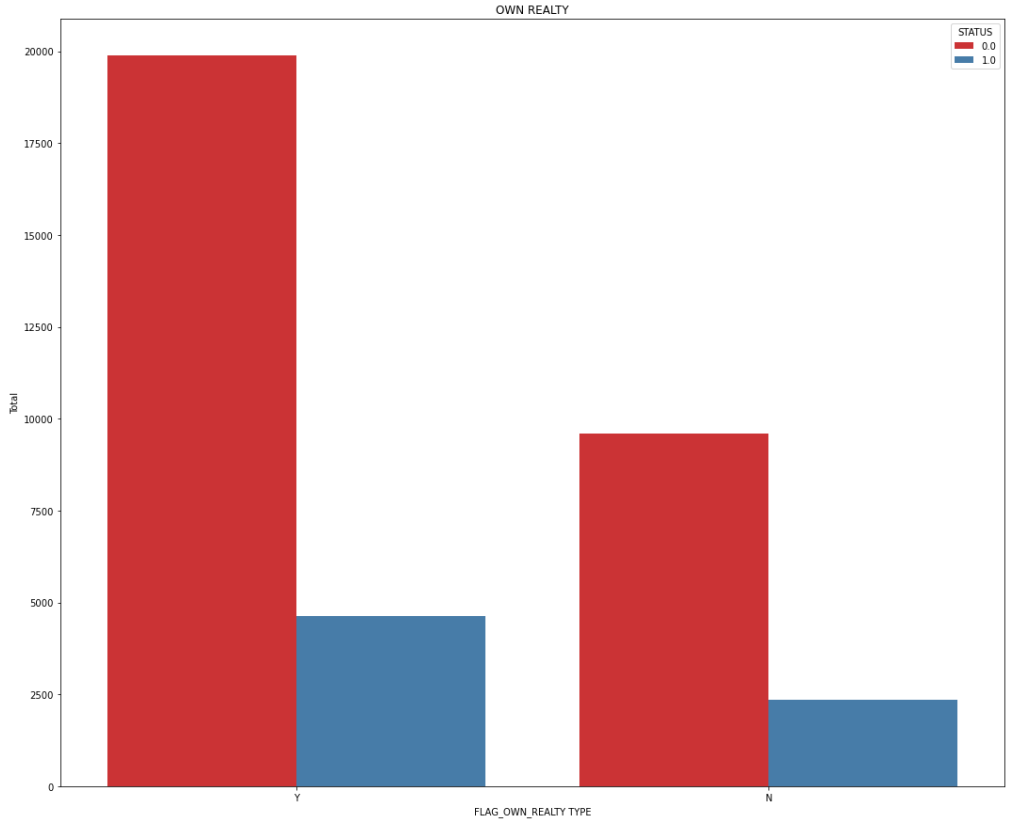
성별 변수 전처리

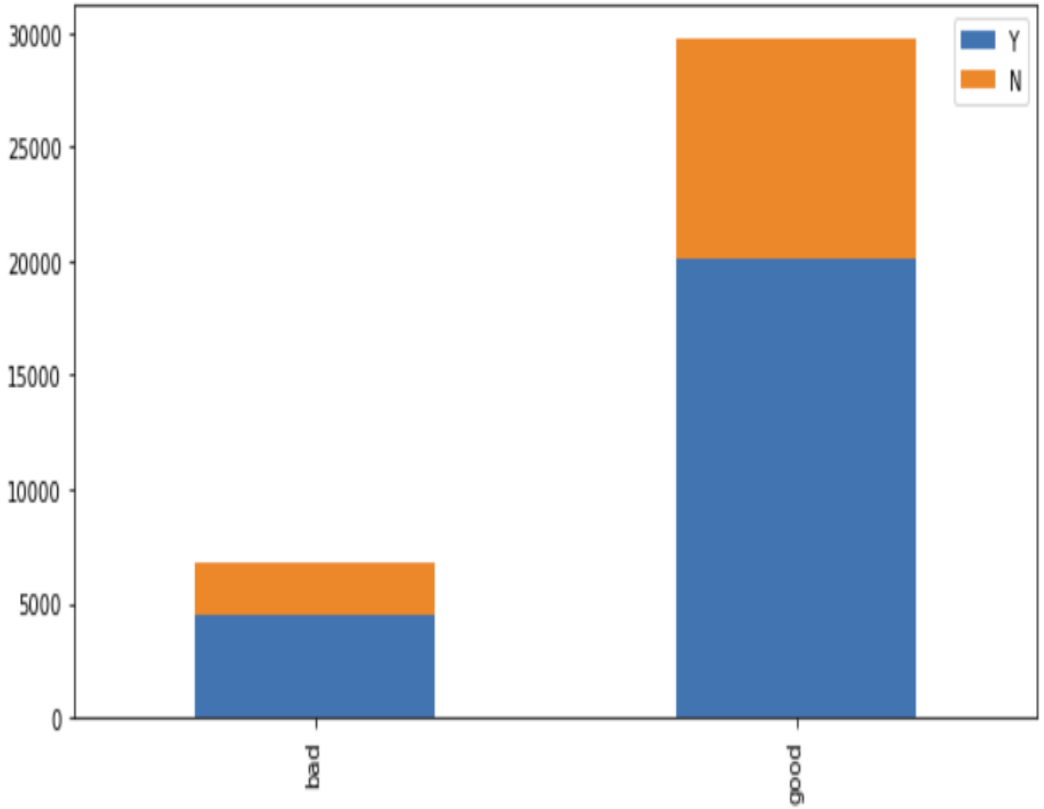
FLAG_OWN_CAR


OWN CAR TYPE
STATUS
0.0
1.0
Total
17500
15000
12500
10000
7500
5000
2500
0
Y
N
FLAG_OWN_CAR TYPE


30000
25000
20000
15000
10000
5000
0
N
Y
bad
good

FLAG_OWN_REALTY





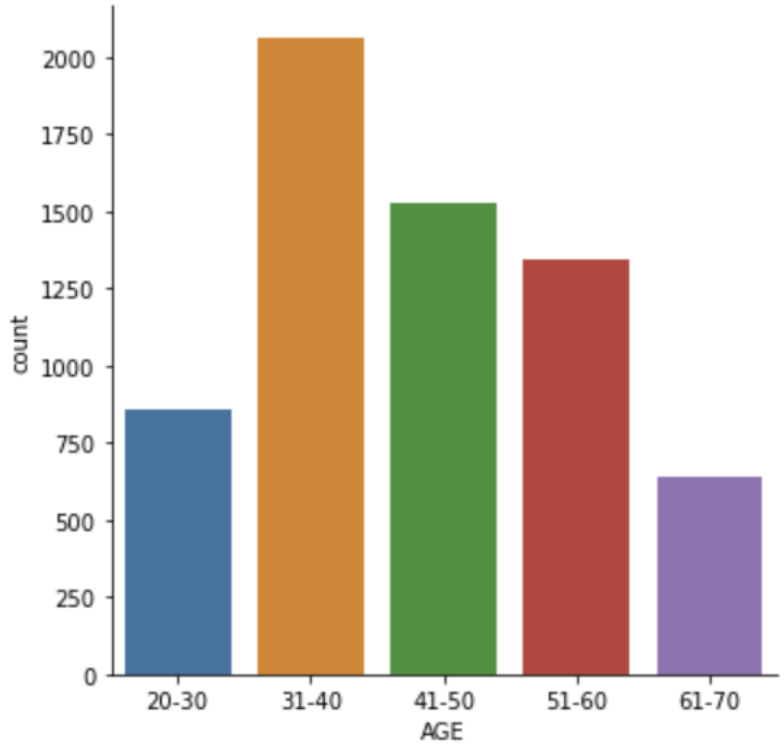
	ID	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	NAME_INCOME_TYPE	NAME_EDUCATION_TY
2	5008806	0	1	1	0	112500.0	0.182741	0.178
3	5008808	1	0	1	0	270000.0	0.188339	0.178
4	5008809	1	0	1	0	270000.0	0.188339	0.178
5	5008810	1	0	1	0	270000.0	0.188339	0.178
6	5008811	1	0	1	0	270000.0	0.188339	0.178
...	...	...	...	...	...	...	...	
36452	5149828	0	1	1	0	315000.0	0.182741	0.178
36453	5149834	1	0	1	0	157500.0	0.188339	0.202
36454	5149838	1	0	1	0	157500.0	0.178966	0.202
36455	5150049	1	0	1	0	283500.0	0.182741	0.178
36456	5150337	0	0	1	0	112500.0	0.182741	0.178

34928 rows × 17 columns

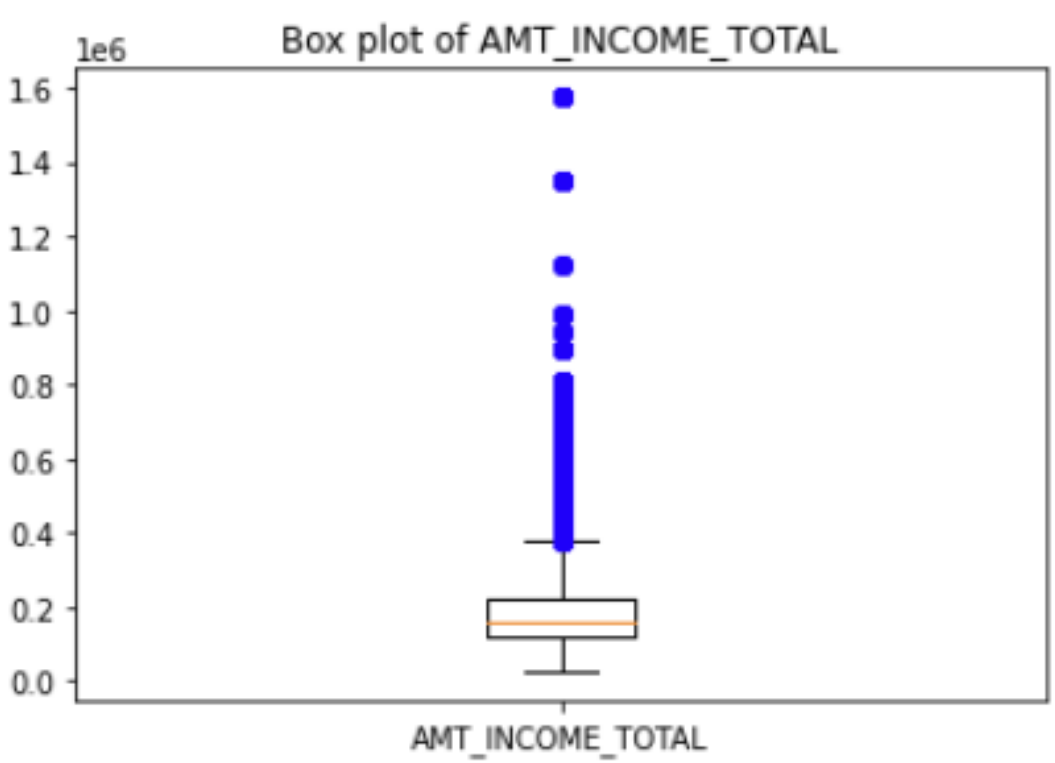
모든 변수의 관측값을 숫자벡터로 전처리

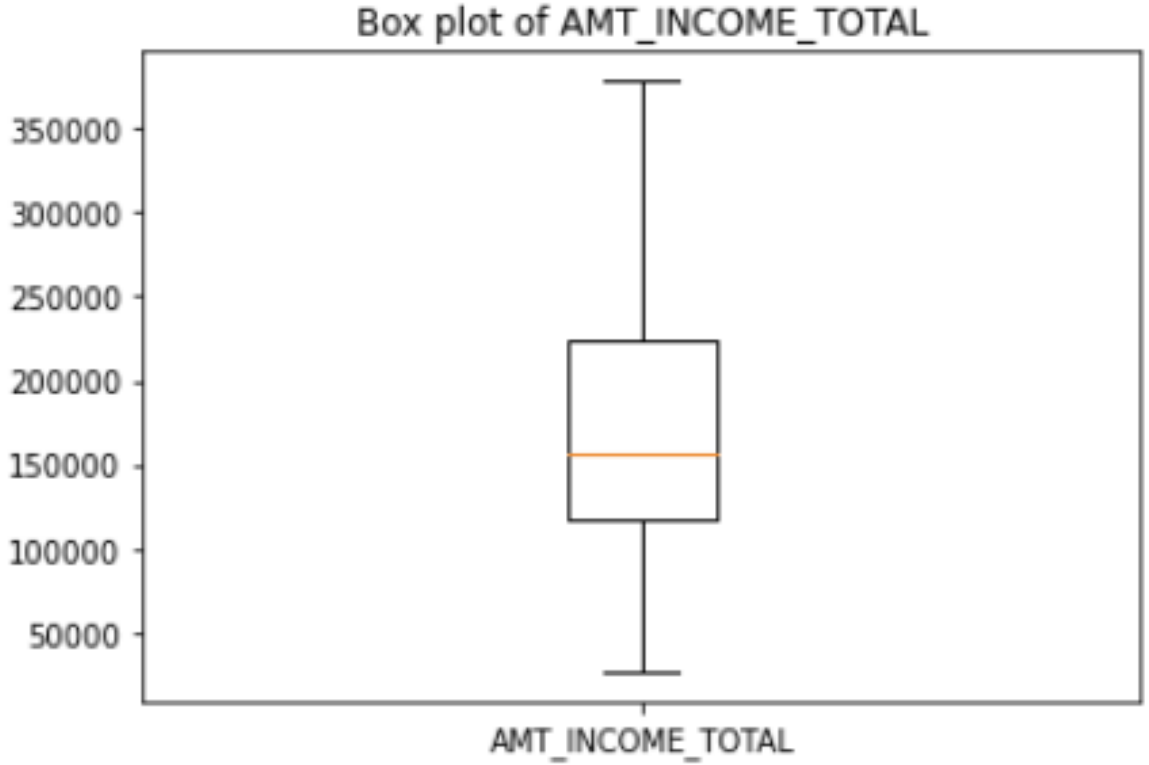
	AGE	STATUS
2	51-60	1.0
3	51-60	1.0
4	51-60	1.0
12	41-50	1.0
13	41-50	1.0
...	...	...
36432	51-60	1.0
36434	61-70	1.0
36437	31-40	1.0
36442	41-50	1.0
36452	41-50	1.0

6419 rows × 2 columns



나이구간 별 위험군에 속한 고객 카운트


Box plot of AMT_INCOME_TOTAL
1e6
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AMT_INCOME_TOTAL


Box plot of AMT_INCOME_TOTAL
350000
300000
250000
200000
150000
100000
50000
AMT_INCOME_TOTAL

이상치 제거 및 수치화

1. Logistic Regression

1) Logistic Regression Model

$$Log(Y/(1-Y)) = -0.0052X1-0.0226X2-0.0272X3-0.3639X4+0.1861X5+0.2076X6+0.3296X7+0.1101X8+0.9562X9-0.3174X10-0.5299X11+0.0523X12+0.0991X13-0.1656X14-0.3653X15$$

2) 위험군에 속할 확률(연체할 확률)

Y = exp(-0.0052 CODE_GENDER - 0.0226 OWN_CAR - 0.0272 OWN_REALTY - 0.3639 CHILDREN + 0.1861 INCOME + 0.2076 INCOME_TYPE + 0.3296 EDUCATION_TYPE + 0.1101 MARRIED + 0.9562 HOUSING_TYPE - 0.3174 AGE – 0.5299 YEARS_EMPLOYED + 0.0523 WORK_PHONE + 0.0991 PHONE - 0.1656 EMAIL - 0.3653 FAMILY) / 1+ exp(-0.0052 CODE_GENDER - 0.0226 OWN_CAR - 0.0272 OWN_REALTY - 0.3639 CHILDREN + 0.1861 INCOME + 0.2076 INCOME_TYPE + 0.3296 EDUCATION_TYPE + 0.1101 MARRIED + 0.9562 HOUSING_TYPE - 0.3174 AGE – 0.5299 YEARS_EMPLOYED + 0.0523 WORK_PHONE + 0.0991 PHONE - 0.1656 EMAIL - 0.3653 FAMILY)

* 데이터를 Scaling 하지 않았을 때 회귀계수
- 회귀계수 값이 너무 작음

```
array([[-7.16359899e-11, -2.11536362e-11, -6.11366463e-11,
        -3.54933194e-11, -7.73645944e-06, -1.61627251e-11,
        -1.58885756e-11, -1.62631104e-11, -1.59904449e-11,
        -4.10934833e-09, -5.04720196e-10, -1.85144274e-11,
        -2.19337958e-11, -7.76231485e-12, -1.94173179e-10]])
```

3) 모델링 결과

- 위험군에 속할 확률에 가장 영향을 미치는 변수 : 'HOUSING TYPE', 'YEARS_EMPLOYED'
- 'HOUSING TYPE'에 따라 위험군에 속할 확률 증가
  부모님 집 > 일반(건물 소유) > 오피스텔 > 월세 > 시립 아파트 > 쉐어 하우스
- 'YEAR_EMPLOYED'에 따라 위험군에 속할 확률 감소

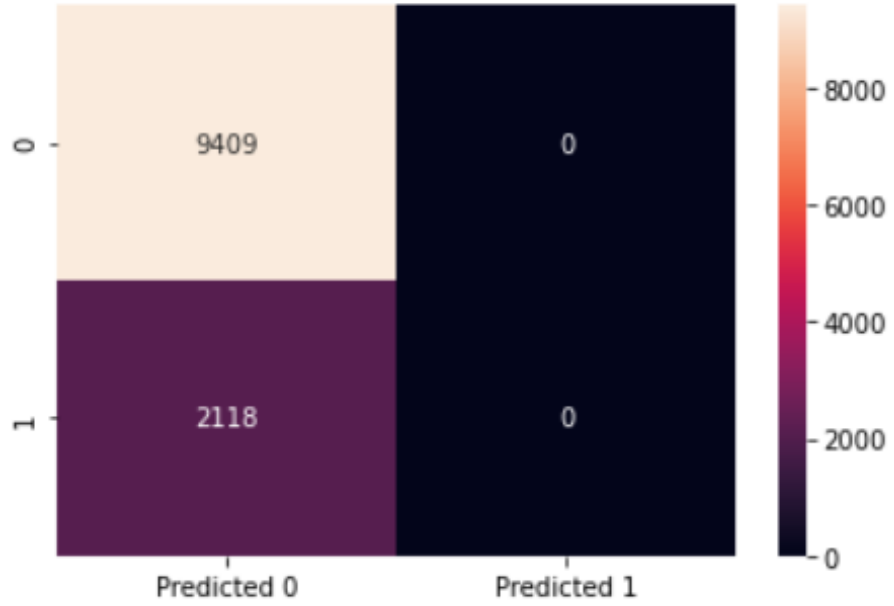
* 'HOUSING TYPE' 별 위험군에 속할 확률

Housing Type	가중치(스케일링)	가중치*회귀계수(0.9562)	가중치*회귀계수/0.5
쉐어하우스	1	0.9562	1.9124
일반	0.0378	0.03614436	0.07228872
시립아파트	0.1436	0.13731032	0.27462064
오피스텔	0.133	0.1271746	0.2543492
렌트 하우스(월세)	0.1506	0.14400372	0.28800744
부모님 집	0	0	0

- 'HOUSING TYPE'이 렌트 하우스일 경우가 부모님 집에 살 경우보다
  연체를 할 확률(위험군에 속할 확률)이 28% 높음

Logistic Regression

성능평가 – Confusion Matrix



	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	9409
1.0	0.00	0.00	0.00	2118
accuracy			0.82	11527
macro avg	0.41	0.50	0.45	11527
weighted avg	0.67	0.82	0.73	11527

- Accuracy(정확도) : 전체 데이터 중 맞게 예측한 데이터의 비율
모델 정확도 82%

- Precision(정밀도) : 예측한 것 중에 맞게 예측한 비율
Good client : 82%
Bad client : 0%

- Recall(재현율) : 실제값 중에 맞게 예측한 비율
Good client : 100%
Bad client : 0% -> 예측 실패

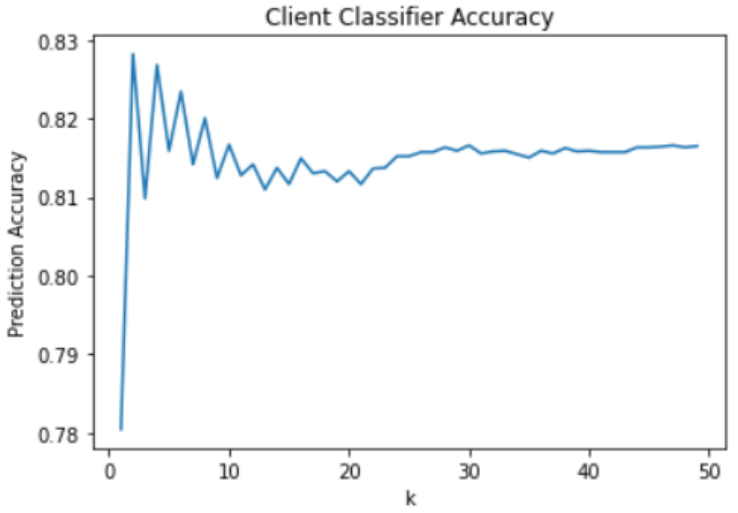
- 결론 : type2 error(위험군을 비위험군으로 예측)

- F1-score(정밀도와 재현율의 조화평균) :
Good client 분류의 성능이 좋다.

2. KNN

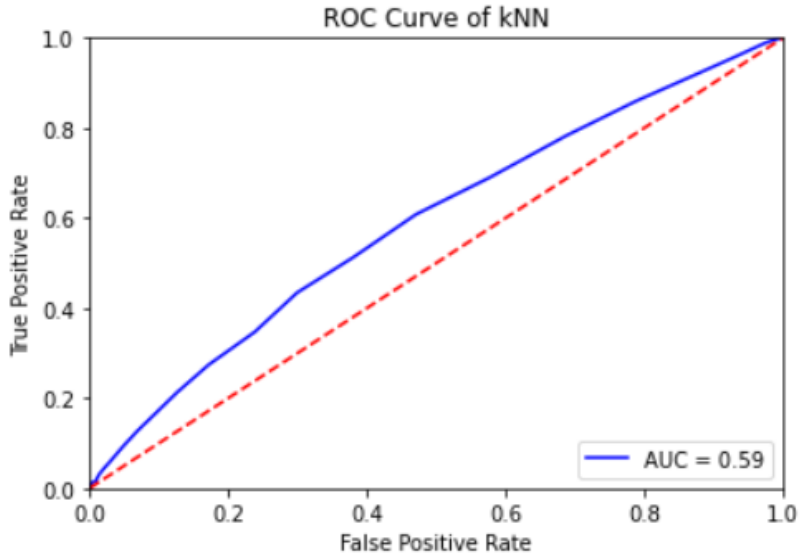
1) K값 변화에 따른 정확도

- k가 작으면 노이즈에 민감하게 반응
-> k =2 대신 k = 4를 채택



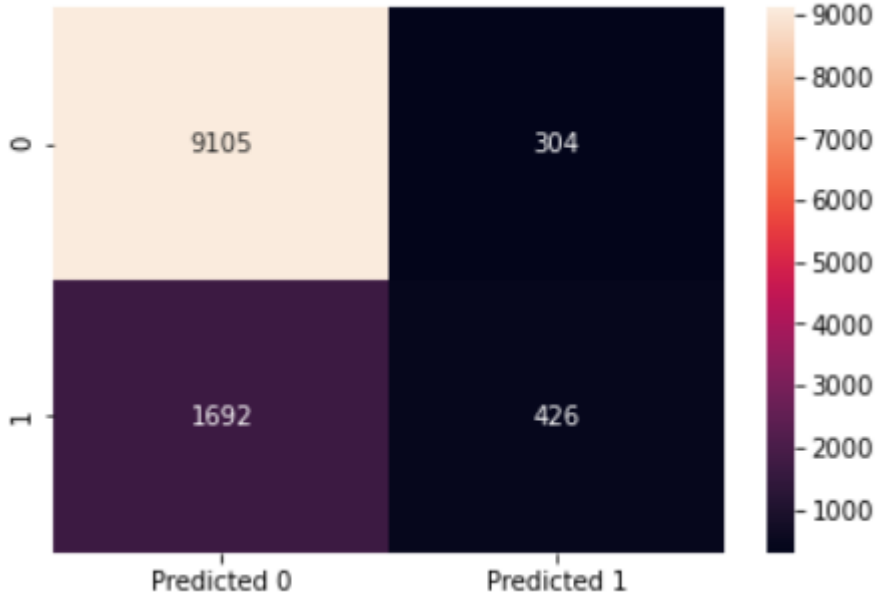
2) 성능평가 – ROC curve

- AUC = 0.5895
-> 값이 작으므로 성능이 좋지 않음



KNN

성능평가 – Confusion Matrix



	precision	recall	f1-score	support
0.0	0.84	0.97	0.90	9409
1.0	0.58	0.20	0.30	2118
accuracy			0.83	11527
macro avg	0.71	0.58	0.60	11527
weighted avg	0.80	0.83	0.79	11527

- Accuracy(정확도) : 전체 데이터 중 맞게 예측한 데이터의 비율
모델 정확도 83%

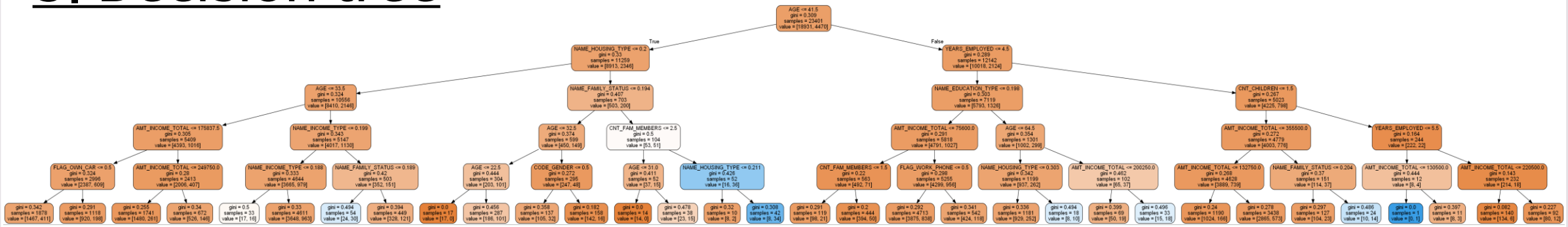
- Precision(정밀도) : 예측한 것 중에 맞게 예측한 비율
Good client : 84%
Bad client : 58%

- Recall(재현율) : 실제값 중에 맞게 예측한 비율
Good client : 97%
Bad client : 20%

- 결론 : type2 error(위험군을 비위험군으로 예측)

- F1-score(정밀도와 재현율의 조화평균) :
Good client 분류의 성능은 좋지만
Bad client 분류의 성능은 좋지 않음

3. Decision tree



1) 모델링 결과 :

- 나이가 41.5세보다 적은 사람
집의 형태가 Co-op apartment, Municipal apartment, Office apartment, Rented apartment인 사람 (HOUSING_TYPE > 0.2)
결혼 여부가 Separated인 사람 (FAMILY_STATUS > 0.194)
가족 수가 2.5명보다 많은 사람
집의 형태가 Co-op apartment, Municipal apartment, Rented apartment인 사람은 Bad client로 예측 (HOUSING_TYPE > 0.211)

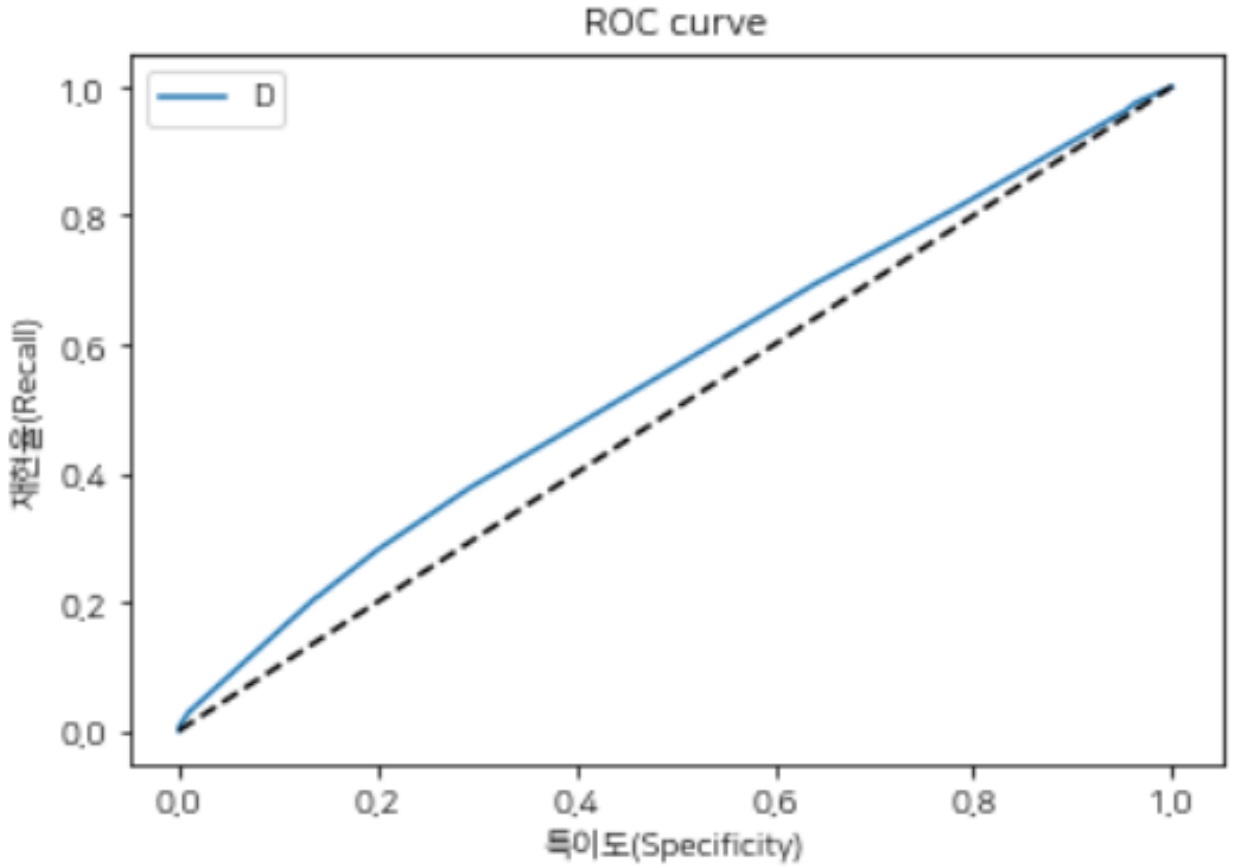
- 나이가 41.5세보다 많은 사람
근속 연수가 4.5년보다 많은 사람
자녀 수가 1.5명보다 많은 사람
근속 연수가 5.5년보다 적은 사람
총 수입이 130500위안보다 적은 사람은 Bad client로 예측

	NAME_HOUSING_TYPE	STATUS
0	Co-op apartment	0.392857
1	House / apartment	0.189843
2	Municipal apartment	0.213652
3	Office apartment	0.209924
4	Rented apartment	0.212174
5	With parents	0.181869

	NAME_FAMILY_STATUS	STATUS
0	Civil marriage	0.198302
1	Married	0.189796
2	Separated	0.209700
3	Single / not married	0.188238
4	Widow	0.194517

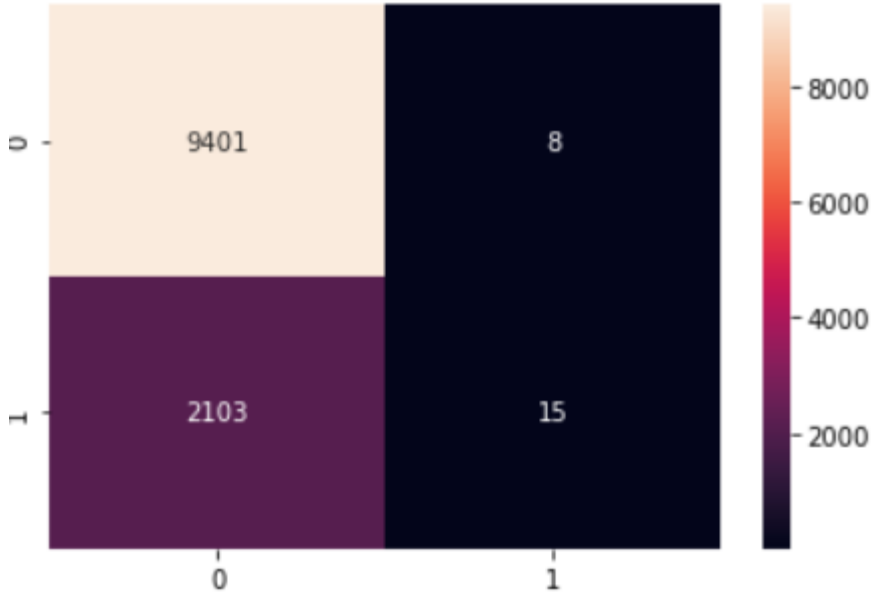
2) 성능평가 – ROC curve

- AUC = 0.5512
-> 값이 작으므로 성능이 좋지 않음



Decision tree

성능평가 – Confusion Matrix



	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	9409
1.0	0.65	0.01	0.01	2118
accuracy			0.82	11527
macro avg	0.73	0.50	0.46	11527
weighted avg	0.79	0.82	0.74	11527

- Accuracy(정확도) : 전체 데이터 중 맞게 예측한 데이터의 비율
모델 정확도 82%

- Precision(정밀도) : 예측한 것 중에 맞게 예측한 비율
Good client : 82%
Bad client : 65%

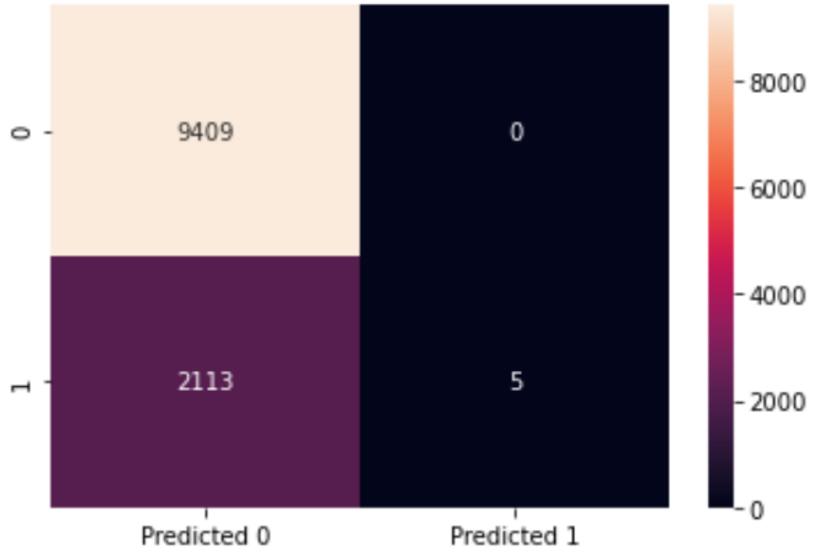
- Recall(재현율) : 실제값 중에 맞게 예측한 비율
Good client : 100%
Bad client : 1% -> 예측 실패

- 결론 : type2 error(위험군을 비위험군으로 예측)

- F1-score(정밀도와 재현율의 조화평균) :
Good client 분류의 성능은 좋지만
Bad client 분류의 성능은 좋지 않음

4. SVM

성능평가 – Confusion Matrix



	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	9409
1.0	1.00	0.00	0.00	2118
accuracy			0.82	11527
macro avg	0.91	0.50	0.45	11527
weighted avg	0.85	0.82	0.73	11527

- Accuracy(정확도) : 전체 데이터 중 맞게 예측한 데이터의 비율

모델 정확도 82%

- Precision(정밀도) : 예측한 것 중에 맞게 예측한 비율

Good client : 82%
Bad client : 100%

- Recall(재현율) : 실제값 중에 맞게 예측한 비율

Good client : 100%
Bad client : 0% -> 예측 실패

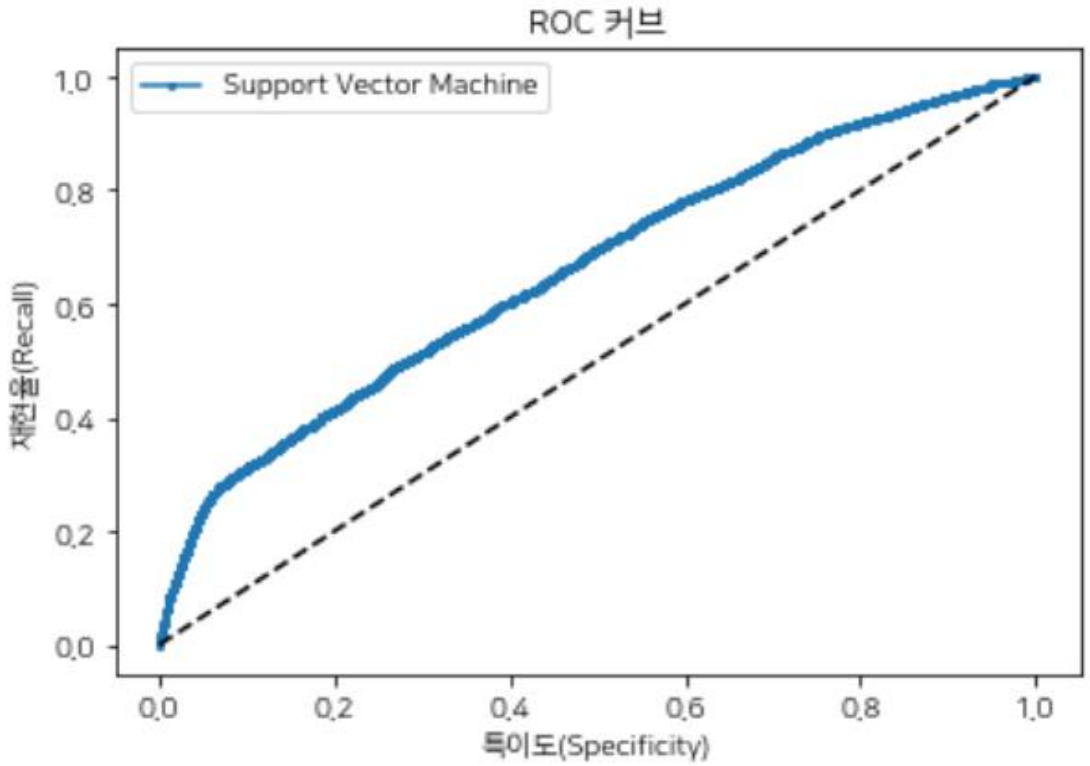
- 결론 : type2 error(위험군을 비위험군으로 예측)

- F1-score(정밀도와 재현율의 조화평균) :

Good client 분류의 성능이 좋다.

SVM ROC커브

- AUC = 0.6610
-> 값이 작으므로 성능이 좋지 않음



5. Ensemble Modeling

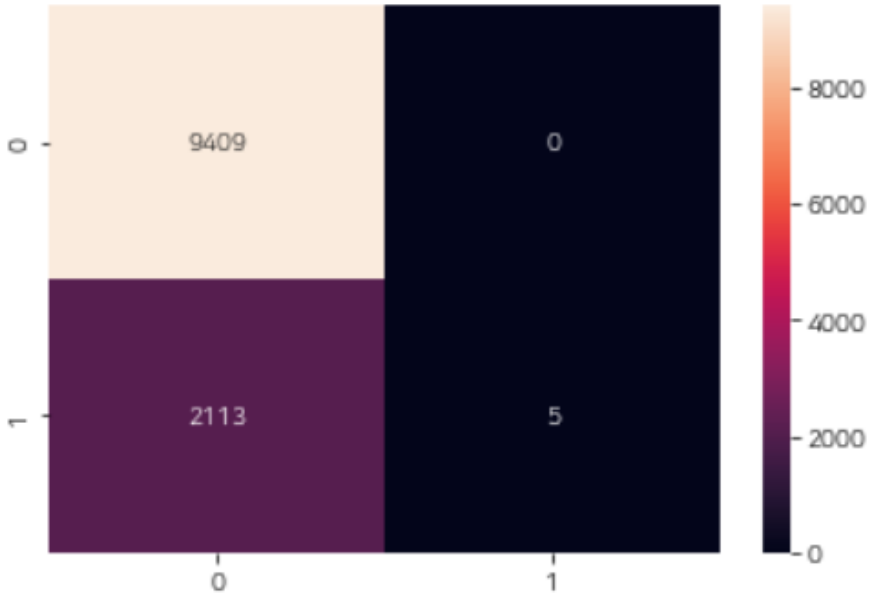
모델	정확도
Logistic Regression	0.8162
K neighbors Classifier	0.8268
Decision Tree Classifier	0.8168
SVM	0.8166
Voting Classifier	0.8162

Hard voting(Majority Voting) -> 과반수가 같은 오답을 제시할 경우
앙상블은 오답을 채택

KNN이 정확도가 가장 높았다.

Ensemble Modeling

성능평가 – Confusion Matrix



	precision	recall	f1-score	support
0.0	0.82	1.00	0.90	9409
1.0	0.00	0.00	0.00	2118
accuracy			0.82	11527
macro avg	0.41	0.50	0.45	11527
weighted avg	0.67	0.82	0.73	11527

- Accuracy(정확도) : 전체 데이터 중 맞게 예측한 데이터의 비율

모델 정확도 82%

- Precision(정밀도) : 예측한 것 중에 맞게 예측한 비율

Good client : 82%

Bad client : 0% -> 예측 실패

- Recall(재현율) : 실제값 중에 맞게 예측한 비율

Good client : 100%

Bad client : 0% -> 예측 실패

- 결론 : type2 error(위험군을 비위험군으로 예측)

- F1-score(정밀도와 재현율의 조화평균) :

Good client 분류의 성능이 좋다.

분석 결과 및 보완점

- 로지스틱 회귀분석 결과
  - 고객이 연체할 확률에 영향을 미치는 변수는 거주 형태와 근속연수이다.
  - 거주 형태가 렌트하우스일 경우가 부모님 집에 살 경우보다 연체를 할 확률(위험군에 속할 확률)이 28% 높다.

- 고객이 연체할 확률에 미치는 영향이 높은 변수
  - Decision Tree : 나이
  - Logistic Regression : 거주 형태

- 전체 모델의 정확도는 높으나 위험 고객을 찾아내는데 아쉬움이 있었다.

- 주어진 데이터 외에 좀 더 다양하게 컬럼을 추가하지 못하였다.

→ 추가자료 크롤링으로 컬럼 결합 및 해체 등을 시도해보고 모델링했다면 type 2 error를 줄이는데 더 효과적이었을 것 같다.

Thank you

Team1

박서영 박지혜 이민준 이상재