Bank Loan modelling

Personal Loan classification problem

최종발표

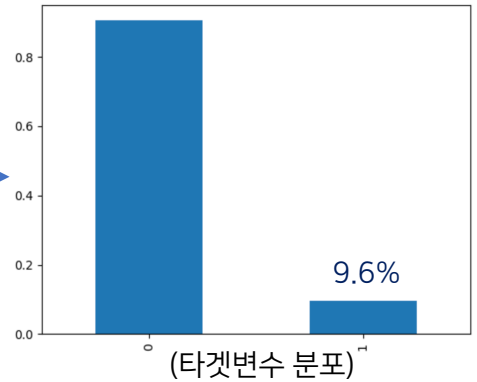
22512092 데이터사이언스학과 석사과정 황선진

목차

1 리뷰

- **분석 과제** : 은행의 대출 잠재고객을 파악하는 분류 모델을 개발
  - [데이터셋] 은행의 대출 마케팅을 실행한 고객 정보 및 해당 고객의 대출 실행 여부
  - [변수] 고객 인적사항·금융정보 12개 변수(ID 제외) · 대출 여부 1개 타겟 변수(불균형) / 5,000개 데이터 샘플

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online	CreditCard	Personal Loan
0	1	25	1	49	91107	4	1.6	1	0	1	0	0	0	0
1	2	45	19	34	90089	3	1.5	1	0	1	0	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...



(타겟변수 분포)

(변수 설명)

변수	ID 아이디	Age 나이	Experience 근무경력	Income 소득	Zipcode 우편번호	Family 가족 수	CCAvg 카드 소비액	Education 교육수준	Mortgage 모기지 금액	Securities Account 유가증권	CD Account 양도성 예금	Online 인터넷 뱅킹	CreditCard 신용카드	Personal Loan (Target) 개인대출
타입	범주형	수치형	수치형	수치형	범주형	수치형	수치형	범주형 (1/2/3)	수치형	범주형 (0/1)	범주형 (0/1)	범주형 (0/1)	범주형 (0/1)	범주형 (0/1)

✓ 이상치 처리 및 데이터 가공

연령층 범주형 변수 생성 | 이상치 처리(+3) | 이상치 처리(끝 0 추가)
시(카운티) 기준 변수 변환 | 연 기준 변환

※ 기존 변수 12개 + 파생변수 2개 데이터 분석 및 모델 학습

1 리뷰

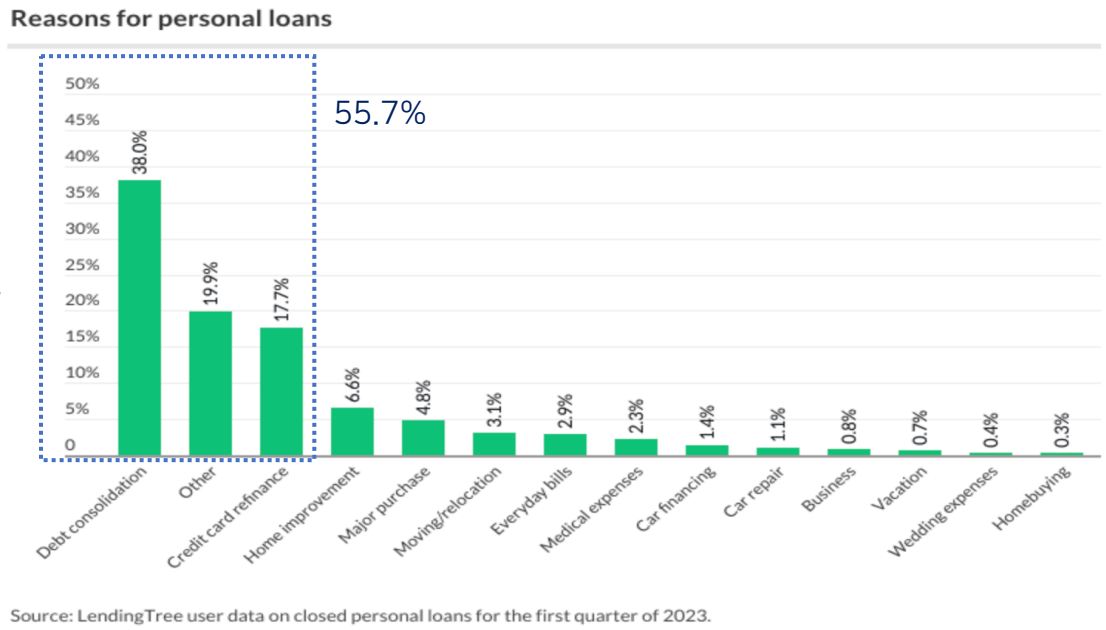
- [변수 분석] 대출 잠재고객을 구분하기 위한 중요 변수를 파악

  - (도메인 조사 기준) 총 부채 ↑ (* 현재 변수로 파악불가), 연 카드소비액 ↑ : 대출 가능성 ↑
  - (EDA 기준) 연소득 ↑, 연 카드소비액 ↑, 모기지론 ↑, 가족 수 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑
  - (DT 룰 기준) 연소득 ↑, 연 카드소비액 ↑, 가족 수 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑

✓ 도메인 조사

미국은 기존 부채를 통합·신용카드 금액을 재융자하는데 개인대출을 많이 활용
(출처 : LendingTree - 온라인 대출플랫폼)

✓ Debt consolidation : 부채 통합
✓ Credit card refinance : 신용카드 재융자



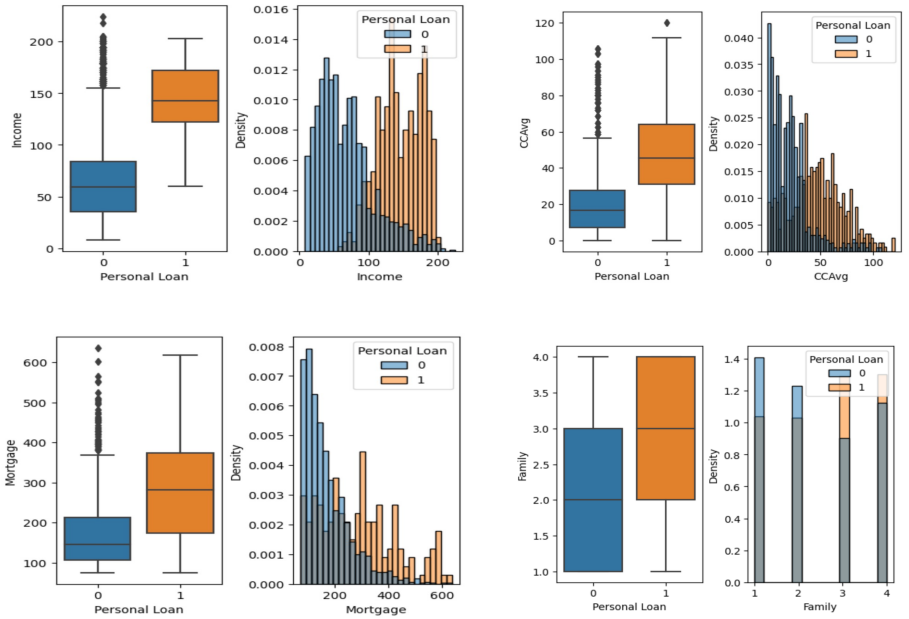
1 리뷰

- [변수 분석] 대출 잠재고객을 구분하기 위한 중요 변수를 파악

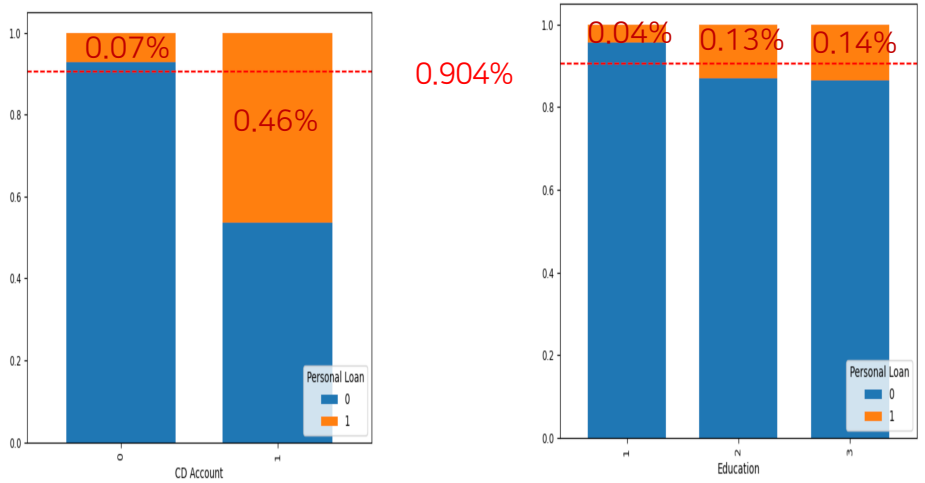
  - (도메인 조사 기준) 총 부채 ↑ (* 현재 변수로 파악불가), 연 카드소비액 ↑ : 대출 가능성 ↑
  - (EDA 기준) 연소득 ↑ , 연 카드소비액 ↑ , 모기지론 ↑ , 가족 수 ↑ , CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑
  - (DT 룰 기준) 연소득 ↑ , 연 카드소비액 ↑ , 가족 수 ↑ , CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑

✓ EDA 기준

[수치형] 개인대출 고객(1)과 아닌 고객(0)을 구분한 데이터 시각화



[범주형] 개인대출 고객(1)과 아닌 고객(0)을 구분한 빈도표(정규화) 시각화

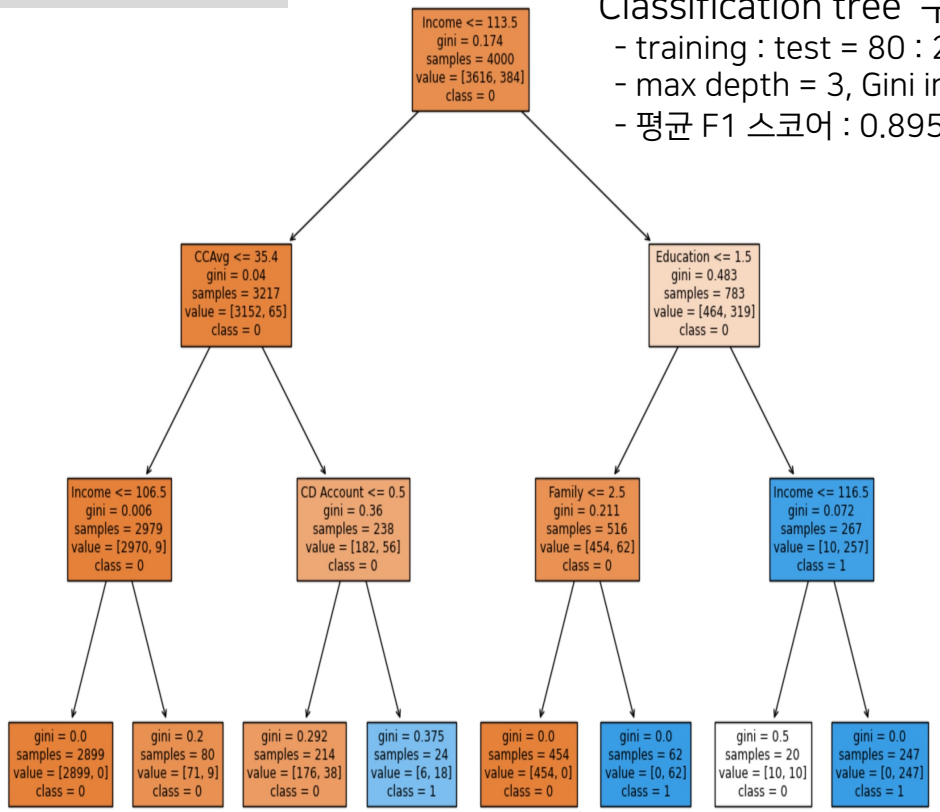


1 리뷰

- [변수 분석] 대출 잠재고객을 구분하기 위한 중요 변수를 파악

  - (도메인 조사 기준) 총 부채↑(* 현재 변수로 파악불가), 연 카드소비액 ↑ : 대출 가능성 ↑
  - (EDA 기준) 연소득↑, 연 카드소비액↑, 모기지론 ↑, 가족 수↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑
  - (DT 룰 기준) 연소득 ↑, 연 카드소비액 ↑, 가족 수 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑ * 연소득·교육수준이 중요변수

✓ DT 기준



(대표 트리 시각화)

Classification tree 구축
- training : test = 80 : 20 (5 fold 교차검증)
- max depth = 3, Gini index 기준
- 평균 F1 스코어 : 0.895

순번	룰 : y = 1 (개인대출 O)	Gini index	Cover age
1	(연소득 > 116.5) & (교육수준 : 석사·전문학위) * (해석) 전문직 고소득 고객층	0.0	0.062
2	(연소득 > 113.5) & (교육수준 : 학사학위) & (가족 수 : 3명 이상)	0.0	0.016
3	(연소득 =< 113.5) & (연 카드소비액 >35.4) & (CD 계좌 보유)	0.375	0.006

순번	룰 : y = 0 (개인대출 X)	Gini index	Cover age
4	(연소득 <= 113.5) & (연 카드소비액 <= 35.4) * (해석) 중·저소득 및 소비가 적은 고객층	0.006	0.745
5	(연소득 > 113.5) & (교육수준 : 학사학위) & (가족 수 : 2명 이하)	0.0	0.114
6	(연소득 =< 113.5) & (연 카드소비액 >35.4) & (CD 계좌 보유X)	0.292	0.054

[참고] (연소득 > 113.5·116.5) : 연소득 기준 상위 20%에 해당하는 값
(연 카드소비액 <= 35.4) : 연 카드소비액 기준 하위 20%에 해당하는 값

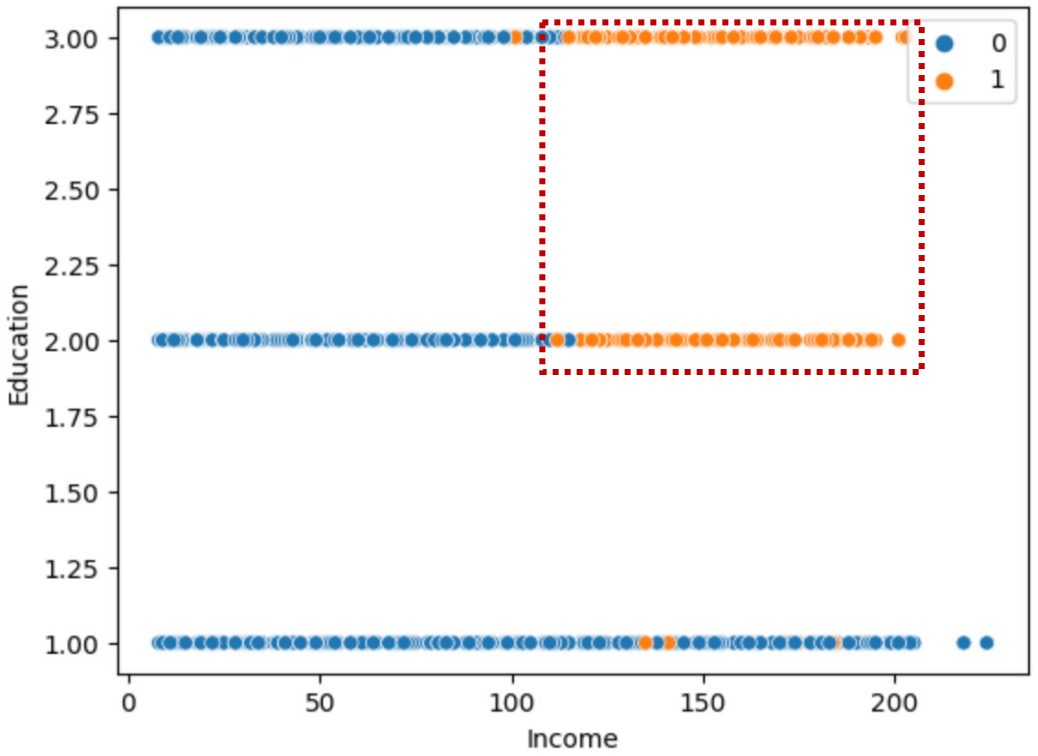
1 리뷰

- [변수 분석] 대출 잠재고객을 구분하기 위한 중요 변수를 파악

  - (도메인 조사 기준) 총 부채 ↑ (* 현재 변수로 파악불가), 연 카드소비액 ↑ : 대출 가능성 ↑
  - (EDA 기준) 연소득 ↑, 연 카드소비액 ↑, 모기지론 ↑, 가족 수 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑
  - (DT 룰 기준) 연소득 ↑, 연 카드소비액 ↑, 가족 수 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑ * 연소득·교육수준이 중요변수

✓ DT 기준

(연소득 <= 113.5) & (연 카드소비액 <= 35.4) → 대출 X



(연소득 > 116.5) & (교육수준 : 석사·전문학위) → 대출 O

2 모델 학습

- [모델 구현] 중요 변수 파악을 위한 분류 트리·랜덤포레스트·그래디언트 부스팅·로지스틱 회귀 4개 모델 선정
  - Grid Search를 통한 하이퍼파라미터 튜닝 적용(* 선형모델의 경우 불필요한 변수를 제거하며 성능 개선)
  - 데이터 불균형을 고려한 오버샘플링(SMOTE) 추가 적용

✓ (오버샘플링 수행하지 않은) 랜덤포레스트가 가장 성능 ↑ : 다양한 변수를 활용하여 일반화 성능 개선

* 평균 value(평균 std)

	분류 트리	랜덤포레스트	그래디언트 부스팅	로지스틱 회귀
F1	0.917(0.017)	**0.930(0.008)**	0.928(0.015)	0.706(0.024)
ROC-AUC	0.986(0.006)	**0.997(0.001)**	0.997(0.001)	0.957(0.010)
Accuracy	0.984(0.002)	**0.987(0.001)**	0.986(0.002)	0.949(0.004)

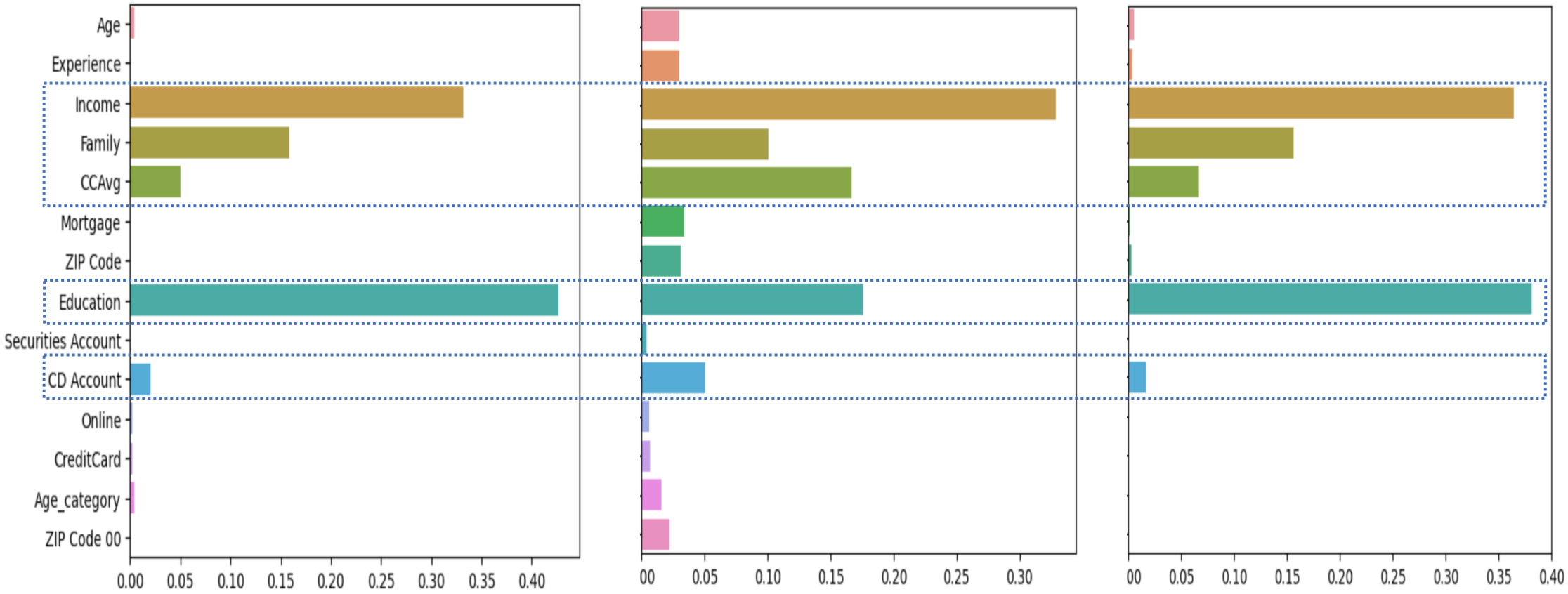
↓ ※ 오버샘플링 수행 : test set에 대한 precision 값이 낮아지면서 성능 개선 X

	분류 트리	랜덤포레스트	그래디언트 부스팅	로지스틱 회귀
F1	0.879(0.030)	0.884(0.034)	0.839(0.012)	0.545(0.035)
ROC-AUC	0.989(0.005)	0.993(0.002)	0.994(0.001)	0.927(0.014)
Accuracy	0.975(0.007)	0.976(0.007)	0.965(0.002)	0.861(0.015)

2 모델 학습

- [모델 해석] 대출 잠재고객 파악을 위한 중요 변수 파악
  - 트리 기반 모델(DT, RF, GB) : 연소득, 교육수준, 연카드 소비액, 가족수, CD 계좌 보유 * 방향성은 대출 여부와 양의 관계라고 추측

* 랜덤포레스트의 경우 다른 모델에 비해 다양한 변수를 사용



(결정 트리 - 피처 중요도)

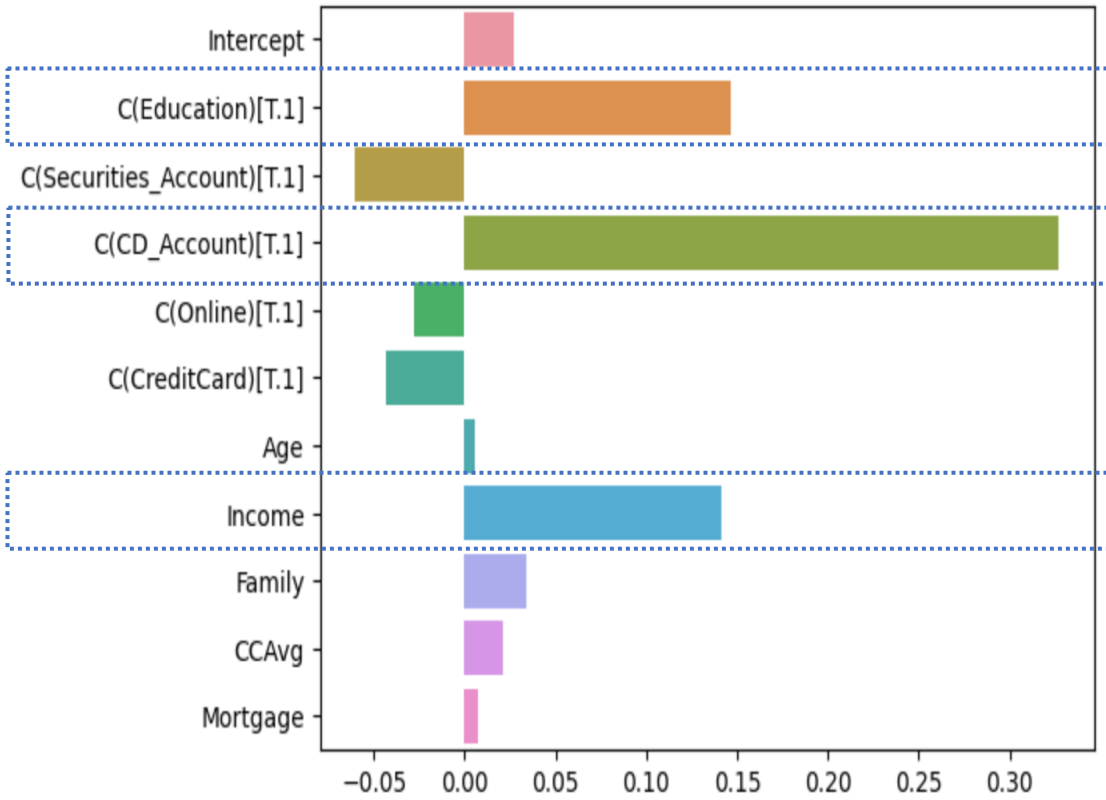
(랜덤포레스트 – 피처 중요도)

(그래디언트 부스팅 – 피처 중요도)

2 모델 학습

- [모델 해석] 대출 잠재고객 파악을 위한 중요 변수 파악
  - 로지스틱 회귀 모델 : 연소득 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑ * 모델의 적합도가 낮으므로 과대 해석 주의

※ 분석을 위해 불필요 변수 제외·수치형 변수에 대한 정규화(정규분포) 적용 ·교육수준 범주형 변수는 학사/석사이상(0/1)로 변환



(정규화 적용 – 회귀계수 시각화)

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0267	0.007	3.823	0.000	0.013	0.040
C(Education)[T.1]	0.1468	0.007	21.670	0.000	0.134	0.160
C(Securities_Account)[T.1]	-0.0604	0.011	-5.340	0.000	-0.083	-0.038
C(CD_Account)[T.1]	0.3271	0.016	20.853	0.000	0.296	0.358
C(Online)[T.1]	-0.0278	0.007	-4.124	0.000	-0.041	-0.015
C(CreditCard)[T.1]	-0.0438	0.007	-5.845	0.000	-0.058	-0.029
Age	0.0060	0.003	1.841	0.066	-0.000	0.012
Income	0.1416	0.004	31.989	0.000	0.133	0.150
Family	0.0341	0.003	10.325	0.000	0.028	0.041
CCAvg	0.0207	0.004	4.874	0.000	0.012	0.029
Mortgage	0.0069	0.003	2.074	0.038	0.000	0.013

(정규화 적용 – 회귀계수 통계)

3 종합 해석

- [중간 발표 - 변수 분석]

- (도메인 조사 기준) 총 부채↑(* 현재 변수로 파악불가), 연 카드소비액 ↑ : 대출 가능성 ↑
- (EDA 기준) 연소득↑, 연 카드소비액↑, 모기지론 ↑, 가족 수↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑
- (DT 기준) 연소득 ↑, 연 카드소비액 ↑, 가족 수 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑ * 연소득·교육수준이 가장 중요변수

- [모델 해석 – 변수 분석]

- 트리 기반 모델(DT, RF, GB) : 연소득, 연카드 소비액, 가족수, CD 계좌 보유, 교육수준 * 방향성은 대출 여부와 양의 관계라고 추측
- 로지스틱 회귀 모델 : 연소득 ↑, CD 계좌 보유, 교육수준 석사이상 : 대출 가능성 ↑ * 모델의 적합도가 낮으므로 과대 해석 주의

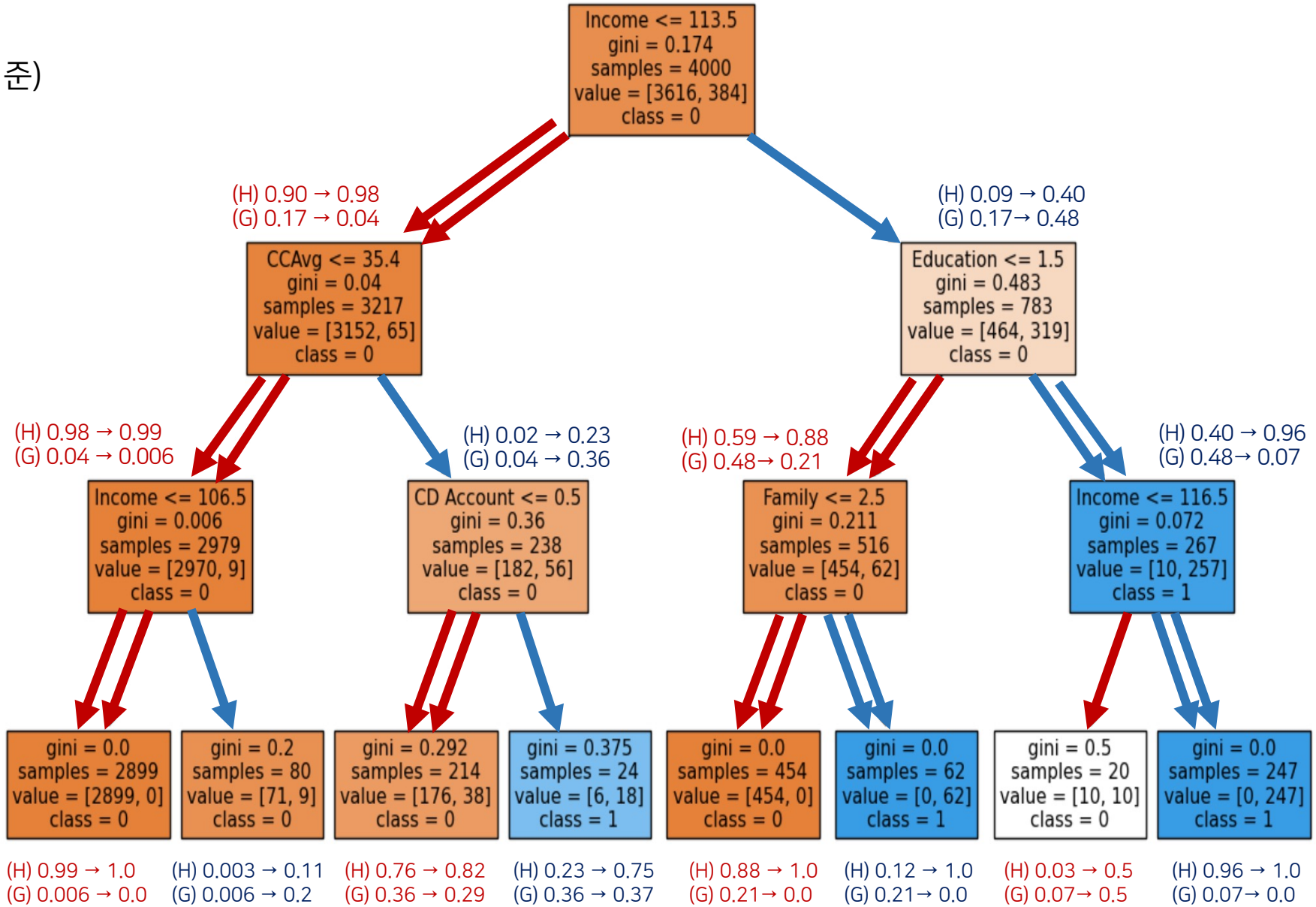
→ 중간 발표에서의 변수 분석과 모델 해석을 통한 변수 분석과 동일하다고 판단

✓ (중요변수) 연소득, 연 카드소비액, 교육수준, 가족 수, 양도성예금증서 계좌 여부

- 전문직 고소득 고객층이 개인대출을 받을 가능성이 높고, 중·저소득층에서 소비가 적은 고객은 대출 가능성이 낮다.
- 가족 수나 양도성예금증서 보유 여부가 잠재고객 판단에 활용될 수 있는 중요 변수가 될 수 있다.

4 Color DT

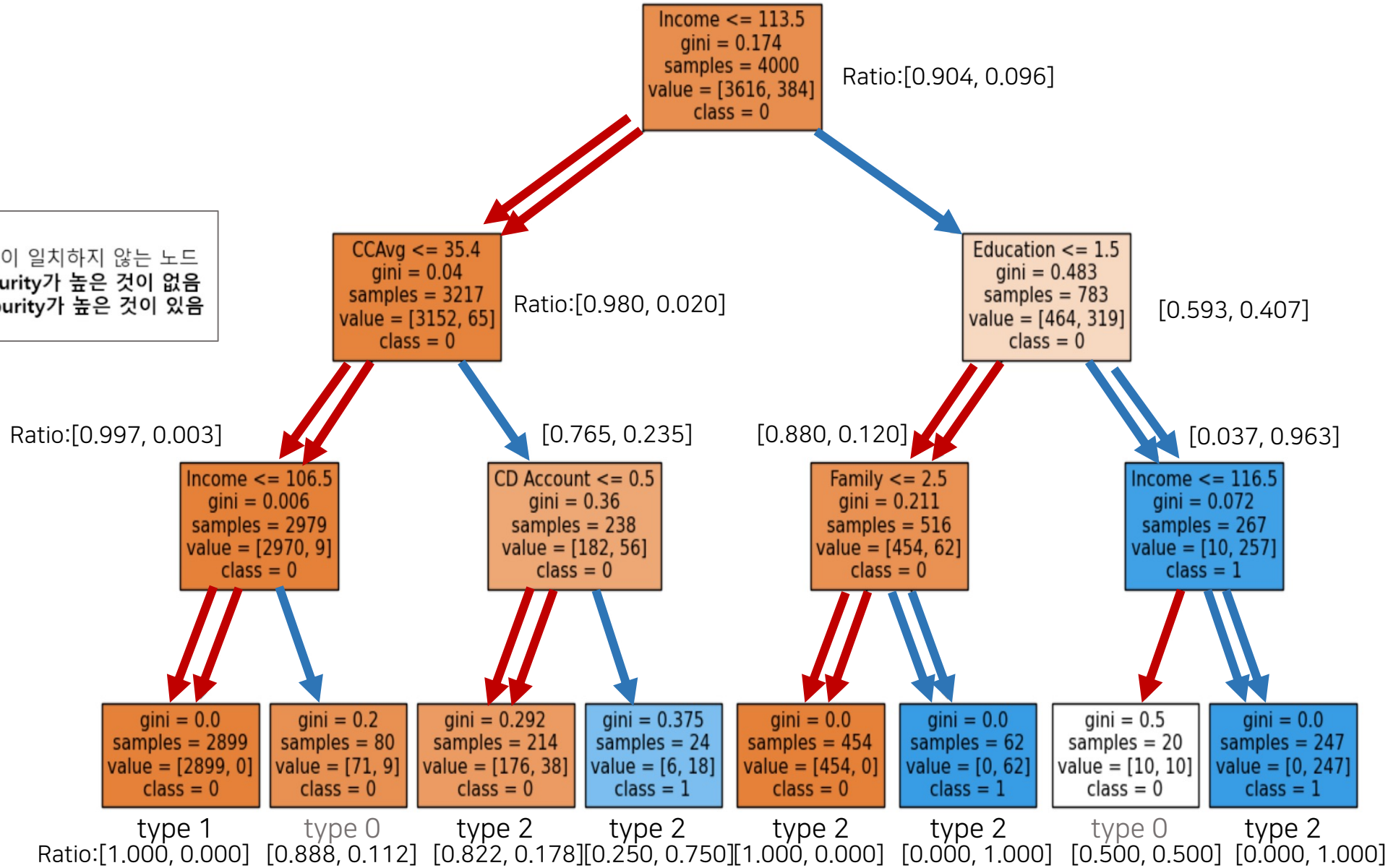
- Color DT 시각화
  - max depth = 3 (Gini index 기준)
  - F1 스코어 : 0.884



4 Color DT

- Color DT 분석
  - type 0 : 2개
  - type 1 : 1개
  - **type 2 : 5개**

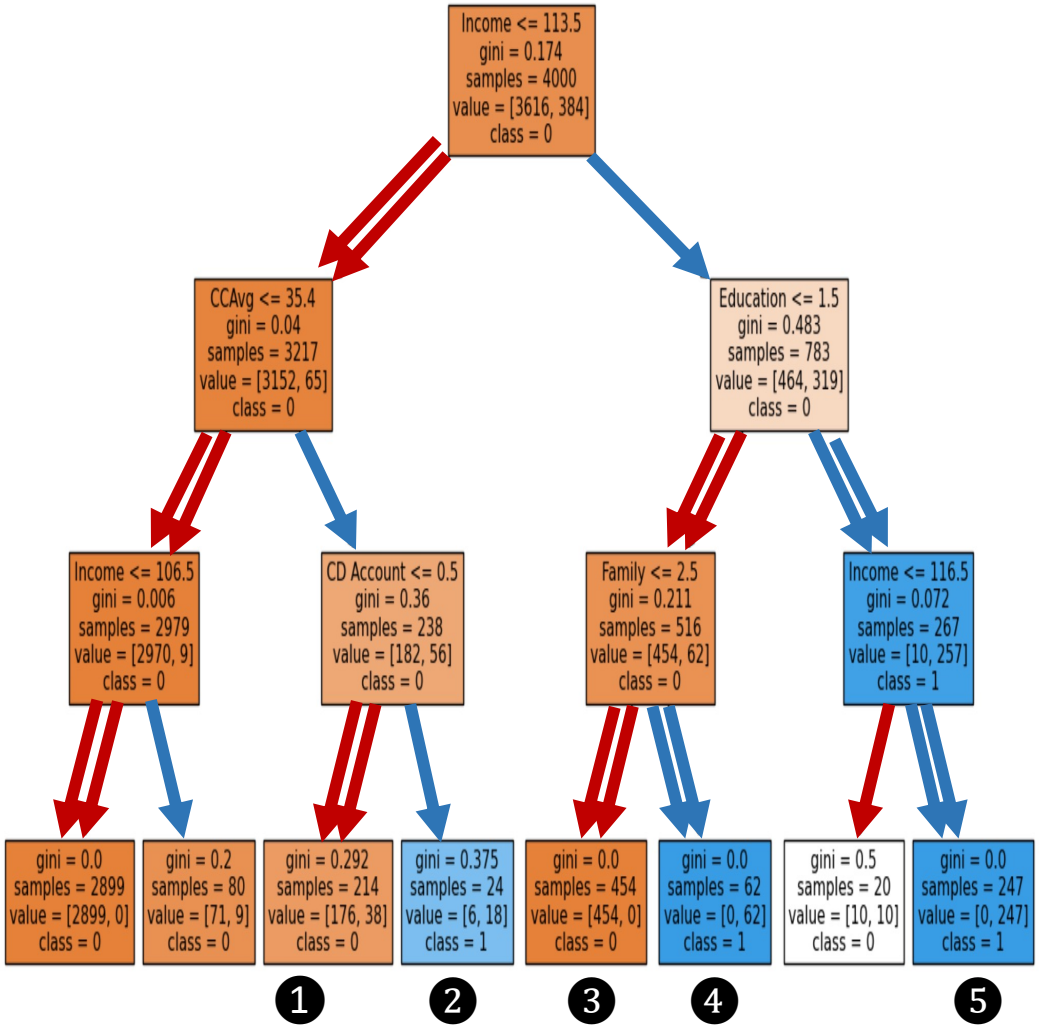
리프노드의 Type
Type 0: 리프노드 색상과 마지막 edge 색상이 일치하지 않는 노드
Type 1: Type 0이 아니면서, 선조 노드 중 **purity가 높은 것이 없음**
Type 2: Type 0이 아니면서, 선조 노드 중 **purity가 높은 것이 있음**



4 Color DT

- Color DT 분석
  - type 2 리프노드 룰 - Irrelevant condition 제거 후 homogeneity 비교(* 5번 리프노드의 경우 기존 DT에 이미 반영)

- 1·2번 리프노드 : homogeneity 증가 / 3·4번 리프노드의 경우 homogeneity 소폭 감소 (* 모든 리프노드 : 커버리지 증가)



* value(샘플 수)

순번	리프노드 룰	예측 Class	homogeneity
1	(소득 <= 113.5) & **(카드소비액 > 35.4)** & (CD 계좌 X)	0	0.822 (214)
2	**(소득 <= 113.5)** & (카드소비액 > 35.4) & (CD 계좌 보유)	1	0.750 (24)
3	**(소득 > 113.5)** & (교육수준 : 학사) & (가족 수 : 2명 이하)	0	1.000 (454)
4	(소득 > 113.5) & **(교육수준 : 학사)** & (가족 수 : 3명 이상)	1	1.000 (62)

※ Irrelevant condition 제거

* 비율 검정

순번	underlying 룰	예측 Class	homogeneity	검정 p 값
1	(소득 <= 113.5) & ~~(카드소비액 > 35.4)~~ & (CD 계좌 X)	0	0.985 (3080)	0.000
2	~~(소득 <= 113.5)~~ & (카드소비액 > 35.4) & (CD 계좌 보유)	1	0.838 (99)	0.016
3	~~(소득 > 113.5)~~ & (교육수준 : 학사) & (가족 수 : 2명 이하)	0	0.989 (1064)	0.000
4	(소득 > 113.5) & ~~(교육수준 : 학사)~~ & (가족 수 : 3명 이상)	1	0.977 (180)	0.043

4 Color DT

- Color DT 인사이트
  - 불필요한 룰을 제거 → 기존 트리의 룰을 더 명료하게 이해할 수 있어서 **트리 모델 해석 및 설명**(변수 분석) **용이**

✓ 기존 DT - 룰

순번	룰 : y = 1 (개인대출 O)	homogeneity
1	(연소득 > 116.5) & (교육수준 : 석사·전문학위) * (해석) 전문직 고소득 고객층	1.000
2	(연소득 > 113.5) & (교육수준 : 학사학위) & (가족 수 : 3명 이상)	1.000
3	(연소득 =< 113.5) & (연 카드소비액 >35.4) & (CD 계좌 보유)	0.750

순번	룰 : y = 0 (개인대출 X)	homogeneity
4	(연소득 <= 113.5) & (연 카드소비액 <= 35.4) * (해석) 중·저소득 및 소비가 적은 고객층	0.996
5	(연소득 > 113.5) & (교육수준 : 학사학위) & (가족 수 : 2명 이하)	1.000
6	(연소득 =< 113.5) & (연 카드소비액 >35.4) & (CD 계좌 보유X)	0.822

✓ 컬러 DT - underlying 룰

룰 : y = 1 (개인대출 O)	homogeneity
(연소득 > 116.5) & (교육수준 : 석사·전문학위) * (해석) 전문직 - 고소득 고객층	1.000
(연소득 > 113.5) & (가족 수 : 3명 이상) * (해석) 고소득 – 자녀가 있는 고객층	0.977
(연 카드소비액 >35.4) & (CD 계좌 보유) * (해석) 소비가 일정규모 – CD계좌 보유 고객층	0.838

룰 : y = 0 (개인대출 X)	homogeneity
(연소득 <= 113.5) & (연 카드소비액 <= 35.4) * (해석) 중·저소득 - 소비가 적은 고객층	0.996
(교육수준 : 학사학위) & (가족 수 : 2명 이하) * (해석) 학사 – 자녀가 없는 고객층	0.989
(연소득 =< 113.5) & (CD 계좌 보유X) * (해석) 중·저소득 – CD 계좌 없는 고객층	0.985

[참고] (연소득 > 113.5·116.5) : 연소득 기준 상위 20%에 해당하는 값
(연 카드소비액 <= 35.4) : 연 카드소비액 기준 하위 20%에 해당하는 값

감사합니다

※ 참고 자료

- 미국 개인대출 조사
https://www.forbes.com/advisor/personal-loans/statistics/
https://www.lendingtree.com/personal/personal-loans-statistics/
- 캐글 노트북 참고
https://www.kaggle.com/code/farzadnekouei/imbalanced-personal-bank-loan-classification