

AI
BOOTCAMP
SECTION 02
PROJECT

AI 부트캠프 10기 박민경

“목차”

1. 문제 제시 및 데이터셋 설명

2. 데이터 전처리 및 EDA

3. 가설 검증

4. 모델 학습 및 해석

5. 결론 및 회고

1. 문제 제시 및 데이터셋 설명

📌 문제 제시: 43912명의 결혼 여부, 성별, 인종, 교육 수준 등에 관련해 소득이 \$50000 넘는지 예측하기 ▶ “분류 문제”

📌 문제 선정 이유: 각자 가지고 있는 다른 배경이 소득에 어떤 영향을 주는지 알아보기 위해

📌 타겟 변수는 income_>50K

📌 age : 나이 workclass : 고용 형태 income_>50K: 소득이 \$50K 넘는지

fnlwgt : 사람 대표성을 나타내는 가중치 (final weight의 약자) education : 교육 수준

education_num : 교육 수준 수치 marital_status: 결혼 상태 occupation : 업종

relationship : 가족 관계 race : 인종 sex : 성별 native_country : 국적

capital_gain : 자본이익 capital_loss : 자본손실 hours_per_week : 주당 근무 시간

2. 데이터 전처리 및 EDA

- 📌 결측치와 중복값은 제거하여 처리
- 📌 capital gain과 capital loss를 총합해 'total capital'이라는 컬럼 추가 생성
- 📌 marital-status를 결혼 여부를 나타내도록 컬럼 값 조정
- 📌 education과 educational-num이 서로 연관되어 있으므로 education을 삭제하여 처리함

3. 가설 검증

더 많은 교육을 받은 사람이
소득 > \$50000일 가능성이 높다.



주당 근무 시간이 많을수록
소득 > \$50000일 것이다.

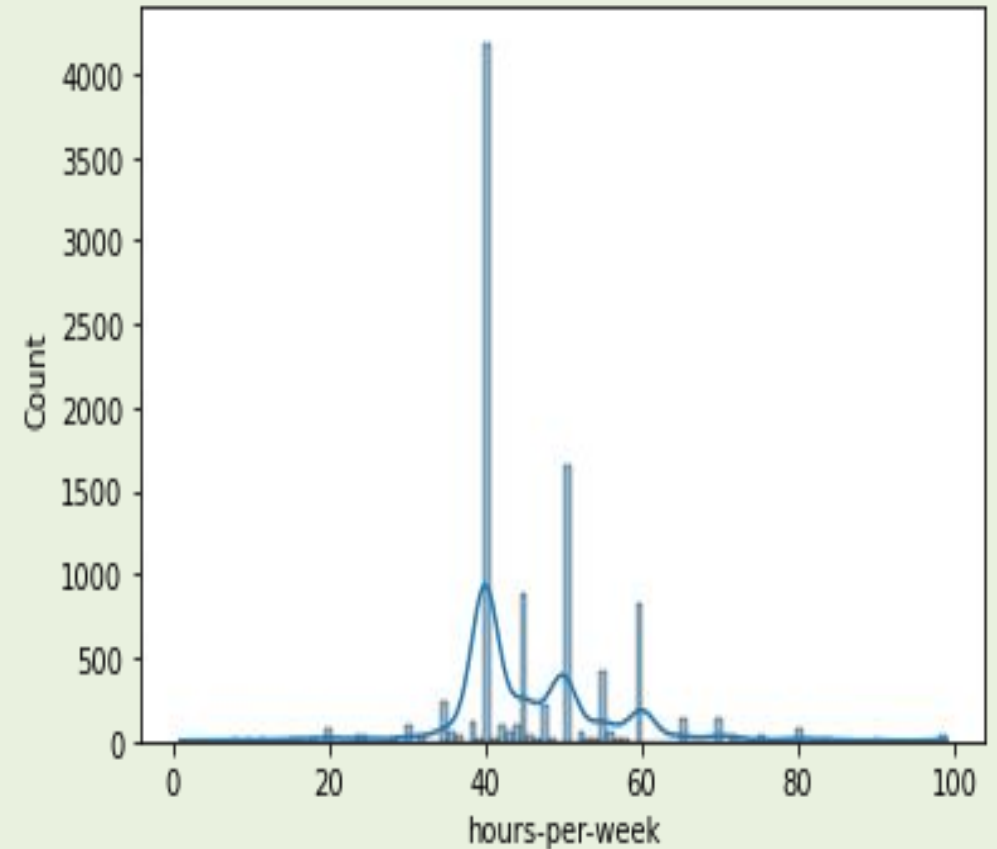


고용 형태가 소득에 영향을 줄
것이다.(일하지 않는 이는
소득 > \$50000일 가능성이 낮다.)

3. 가설 검증

가설 1

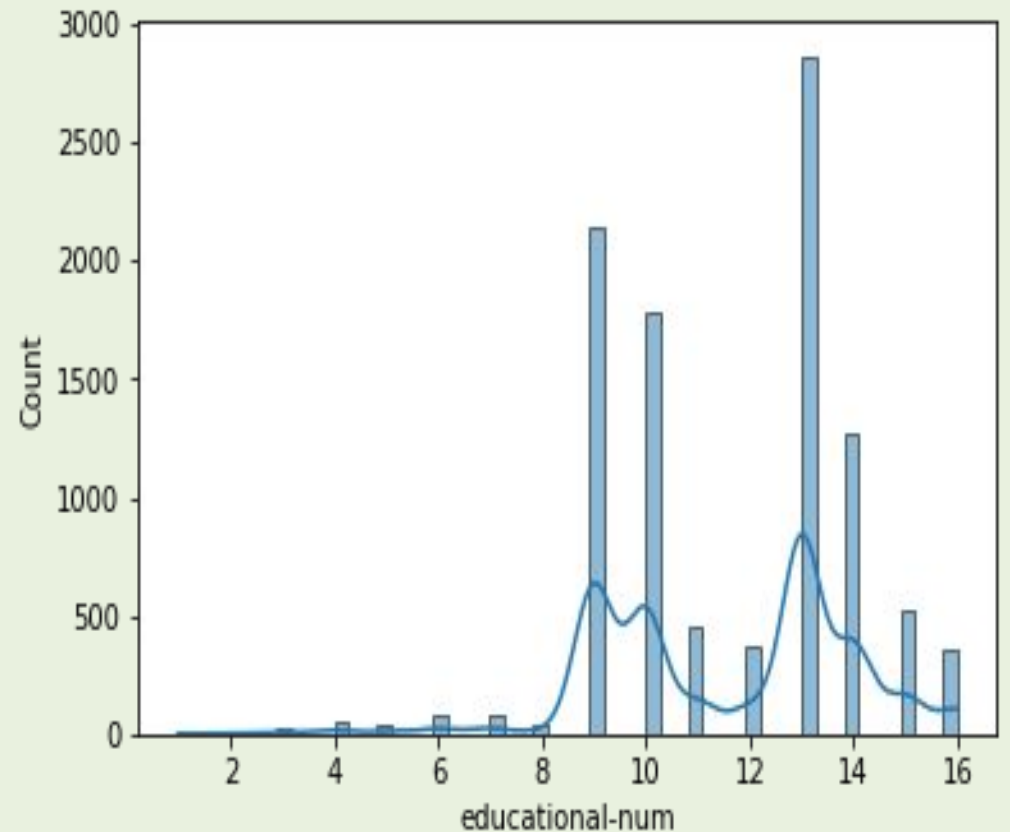
- 40~60시간이 가장 많이 분포되어 있다.
- 60시간 이상인 경우가 거의 없다는 점에서 해당 가설을 기각할 수 있다.
- 그러나 **주간 근무 시간이** 소득>\$50000에는 어떤 영향을 주고 있음을 확인할 수 있다.



3. 가설 검증

가설 2

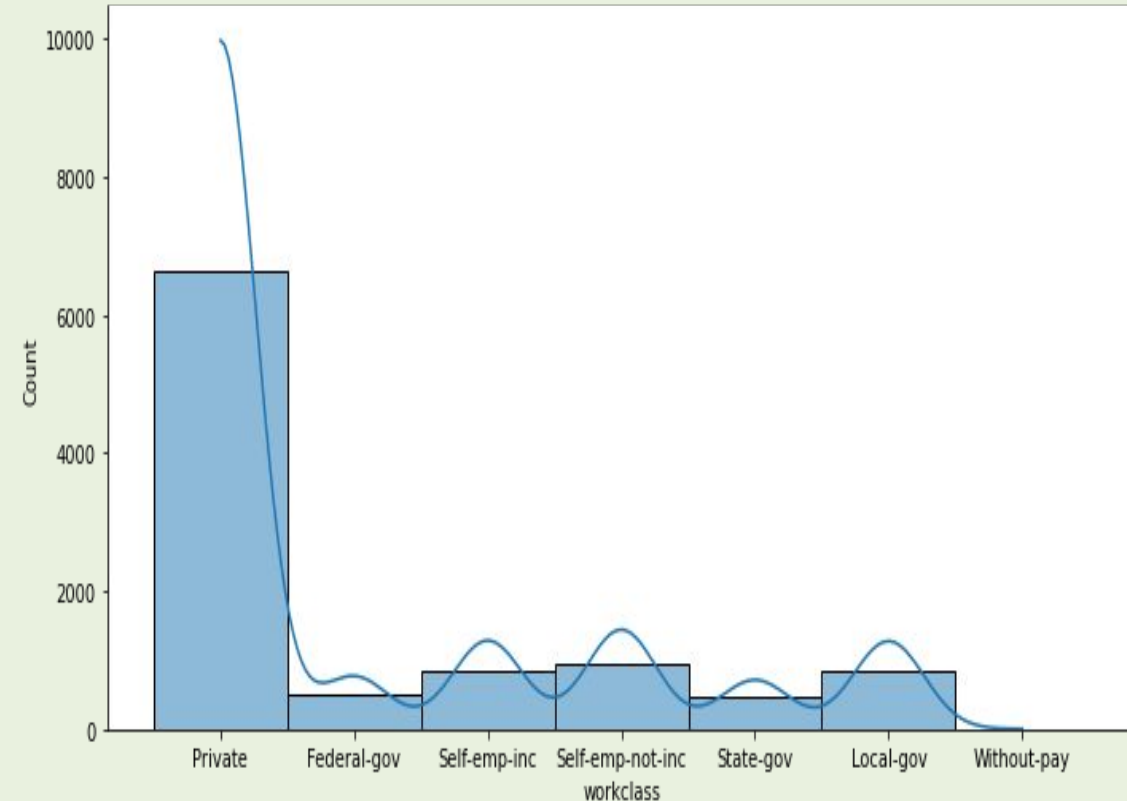
- 교육수준 수치가 9~14일 때 가장 많이 분포되어 있다.
- 15 이상인 경우 분포가 적다는 점에서 해당 가설을 기각할 수 있다.
- 그러나 일정 수준 이상의 교육을 받은 사람이 많이 분포했으므로 해당 변수가 타겟에 영향을 준다고는 볼 수 있다.



3. 가설 검증

가설 3

- never-work의 분포가 없으며, without-pay인 경우도 거의 존재하지 x
- 고용되지 않거나 임금을 받지 못하는 경우가 소득>\$50000의 분포에 거의 존재하지 않다는 점에서 기각할 수 없다고 할 수 있다.



4. 모델 학습 및 해석

Baseline 모델 설정

- 타겟이 binary 변수이며 분류 문제를 풀어야 하기 때문에 타겟 변수의 최빈값을 기본모델로 선정
- 기본모델 정확도: 0.752077

평가지표 설정

- 타겟이 불균형 클래스이기 때문에 정확도만 가지고 모델을 평가할 수 없으므로 **f1 스코어**를 함께 평가지표로 볼 것

불균형 클래스를 조정하기 위해 **class weight, scale_pos_weight**만 설정

logistic regression, random forest, xgbclassifier catboostclassifier 모델 학습

훈련/검증 **정확도와 f1 score, auc score**를 구함

4. 모델 학습 및 해석

✓ logistic regression

훈련 정확도 0.7967664409001529

검증 정확도 0.7913225848735089

Report

	precision	recall	f1-score	support
0	0.80	0.97	0.87	5718
1	0.73	0.27	0.39	1911
accuracy			0.79	7629
macro avg	0.76	0.62	0.63	7629
weighted avg	0.78	0.79	0.75	7629

f1 스코어 0.3919022154316272

auc점수 : 0.6172589465199269

✓ random forest

훈련 정확도 1.0

검증 정확도 0.8547647135928693

Report

	precision	recall	f1-score	support
0	0.88	0.94	0.91	5718
1	0.76	0.61	0.68	1911
accuracy			0.85	7629
macro avg	0.82	0.77	0.79	7629
weighted avg	0.85	0.85	0.85	7629

f1 스코어 0.679212507237985

auc점수 : 0.7745534084163974

✓ catboost

훈련 정확도 0.8667249289927901

검증 정확도 0.8400838904181414

Report

	precision	recall	f1-score	support
0	0.94	0.84	0.89	5718
1	0.64	0.83	0.72	1911
accuracy			0.84	7629
macro avg	0.79	0.84	0.81	7629
weighted avg	0.86	0.84	0.85	7629

f1 스코어 0.7233560090702948

auc점수 : 0.8382721560655904

4. 모델 학습 및 해석

📌 과적합을 상당 부분 해결하고 불균형 클래스를 조정해 f1 score가 더 높은 catboost 모델에 대해 RandomizedSearch CV로 하이퍼파라미터 조정해 2차 학습

✓ 파라미터 조정한 catboost 모델 성능

```
훈련 f1 score: 0.7790021426385063
검증 f1 score: 0.7227004984141367
Report
```

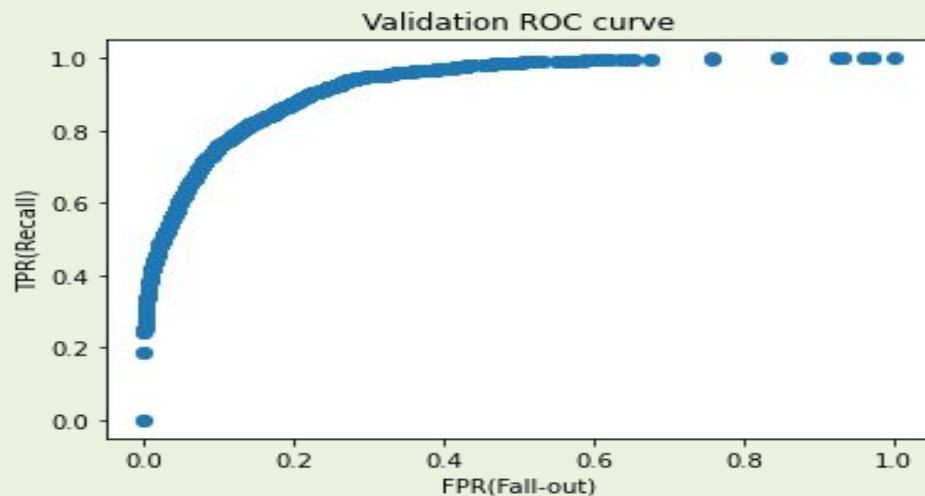
	precision	recall	f1-score	support
0	0.94	0.84	0.89	5718
1	0.64	0.83	0.72	1911
accuracy			0.84	7629
macro avg	0.79	0.84	0.80	7629
weighted avg	0.86	0.84	0.85	7629

```
auc점수 : 0.8379223834178113
```

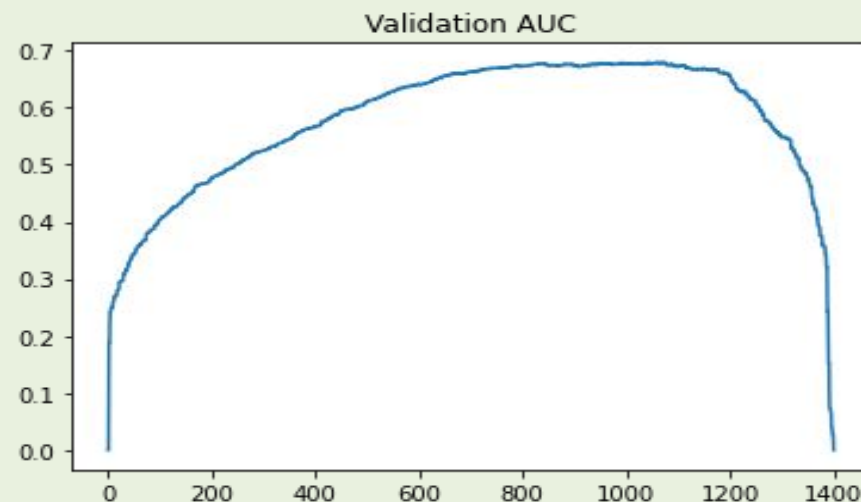
▶ 두 모델의 검증 f1 score와 auc score 모두 매우 근소한 차이를 가지는데, catboost 모델이 근소하게 평가지표들이 높은 수치를 보이고 있으므로 최종 모델로 catboost를 선택

4. 모델 학습 및 해석

📌 VAL ROC curve



📌 VAL AUC



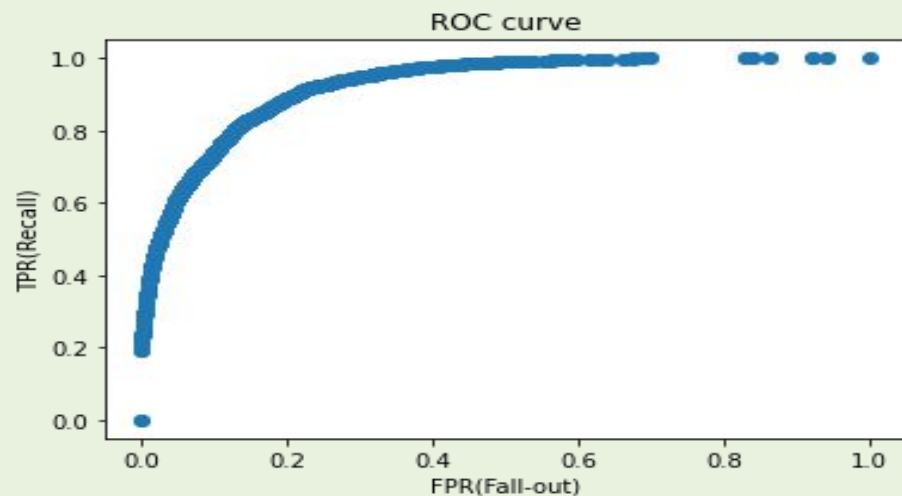
📌 VAL 평가지표: 최적 임계값=0.37478910754021555일 때

Report				
	precision	recall	f1-score	support
0	0.96	0.78	0.86	5718
1	0.58	0.90	0.70	1911
accuracy			0.81	7629
macro avg	0.77	0.84	0.78	7629
weighted avg	0.86	0.81	0.82	7629

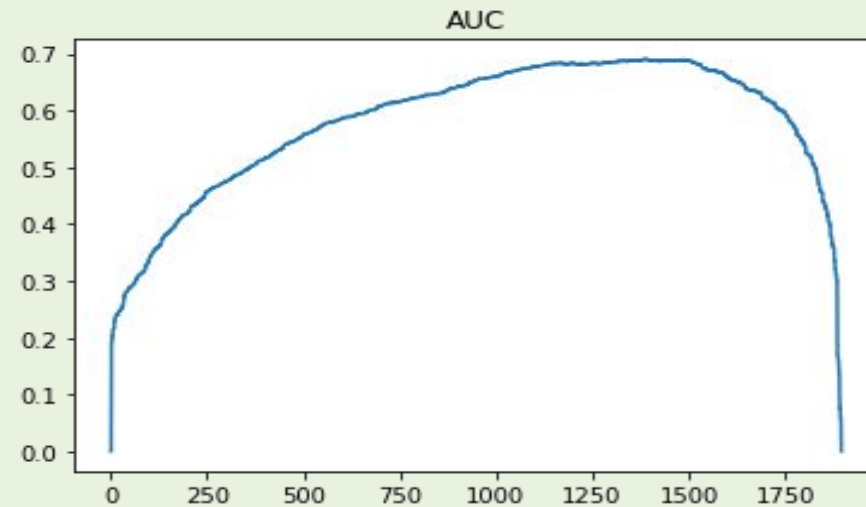
검정 정확도 0.8395595753047581
val_f1 스코어 0.7030451665644798
val_auc점수 : 0.8396728939376219

4. 모델 학습 및 해석

📌 TEST ROC curve



📌 TEST AUC



📌 TEST 평가지표: 최적 임계값=0.421078416524647일 때

Report				
	precision	recall	f1-score	support
0	0.95	0.80	0.87	7545
1	0.61	0.89	0.72	2627
accuracy			0.82	10172
macro avg	0.78	0.85	0.80	10172
weighted avg	0.86	0.82	0.83	10172

테스트 정확도 0.8394612662209988
test_f1 스코어 0.7235759739251901
test_auc점수 : 0.8452524290874472

4. 모델 학습 및 해석

모델 분석

- TEST 데이터셋에 대한 파라미터 조정된 Catboost 모델 성능: baseline 모델 성능보다 개선됨
- TEST 데이터의 f1 score는 훈련/검증 데이터의 f1 score와 큰 차이를 보이지 않았기 때문에 모델 학습에서 과적합이 발생하지 않았고, 학습 과정에서 일반화되었음을 알 수 있음
- 타겟 변수의 불균형 클래스 문제를 상당 부분 해결했음을 확인

5. 결론 및 회고

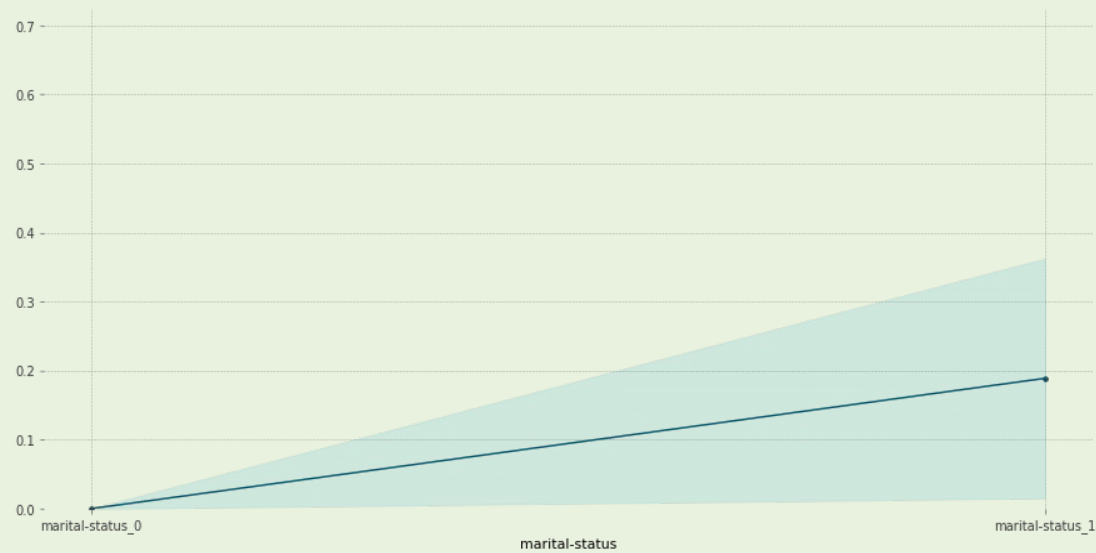
📌 순열 중요도를 통해 결혼 여부, 교육 수준, 나이, 자본 이익, 자본 총합계 순으로 개인의 소득이 \$50000를 넘는지 크게 작용하고 있음을 파악할 수 있음

Weight	Feature
0.0657 ± 0.0033	marital-status
0.0426 ± 0.0028	educational-num
0.0419 ± 0.0081	age
0.0367 ± 0.0044	capital-gain
0.0321 ± 0.0048	total_capital
0.0194 ± 0.0077	gender
0.0180 ± 0.0045	relationship
0.0160 ± 0.0078	hours-per-week
0.0140 ± 0.0045	occupation
0.0087 ± 0.0019	capital-loss
0.0047 ± 0.0031	workclass
0.0020 ± 0.0068	fnlwgt
0.0019 ± 0.0020	native-country
-0.0015 ± 0.0036	race

5. 결론 및 회고

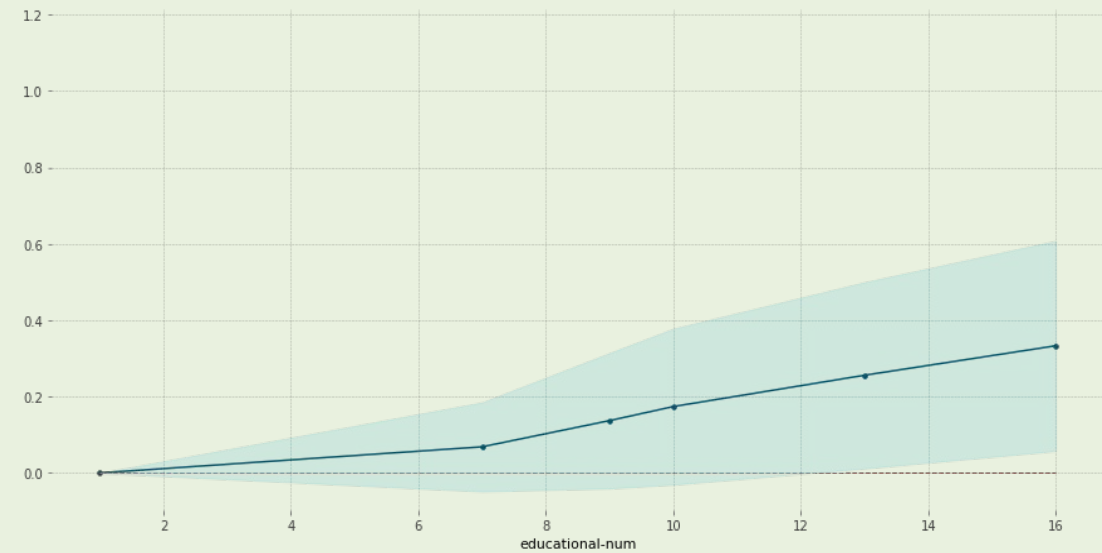
PDP for feature "marital-status"

Number of unique grid points: 2



PDP for feature "educational-num"

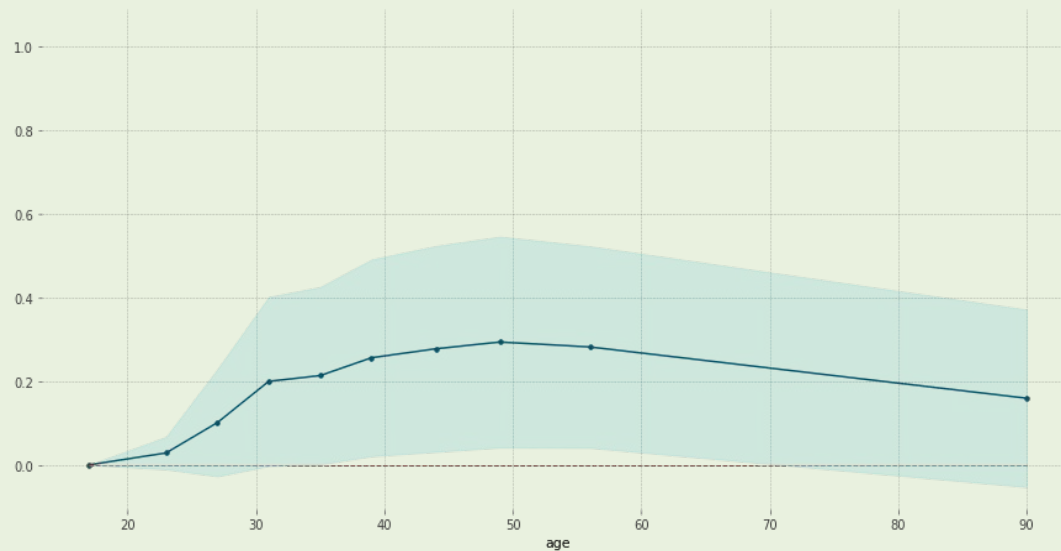
Number of unique grid points: 6



5. 결론 및 회고

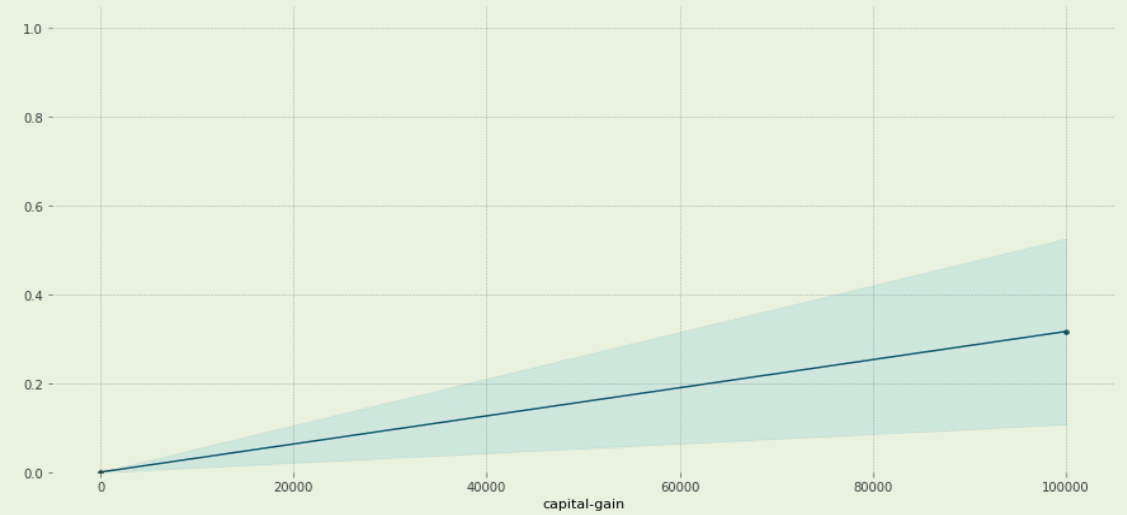
PDP for feature "age"

Number of unique grid points: 10



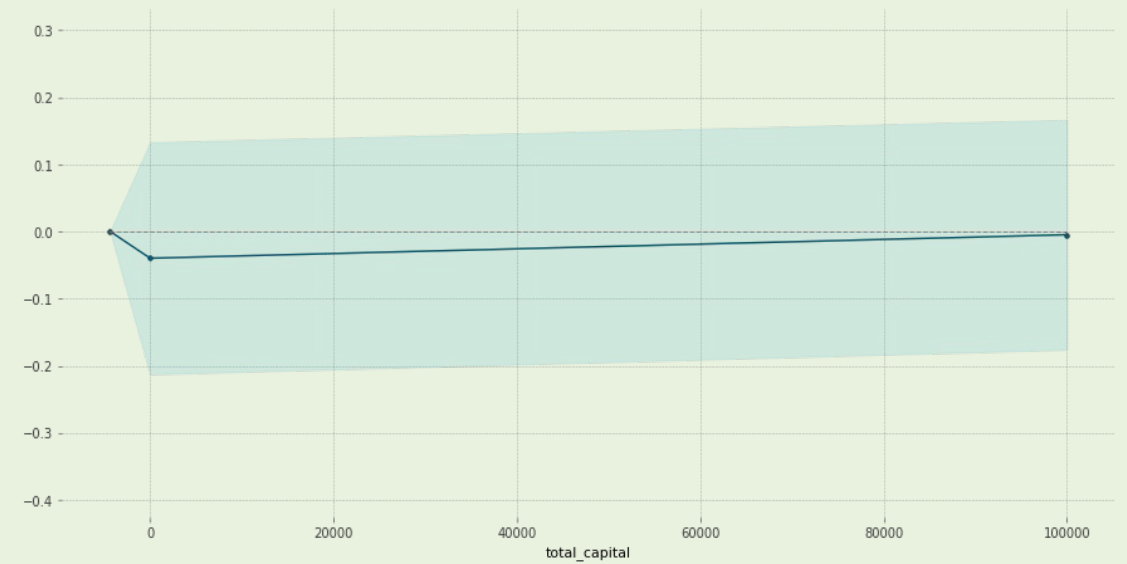
PDP for feature "capital-gain"

Number of unique grid points: 2



PDP for feature "total_capital"


Number of unique grid points: 3



5. 결론 및 회고

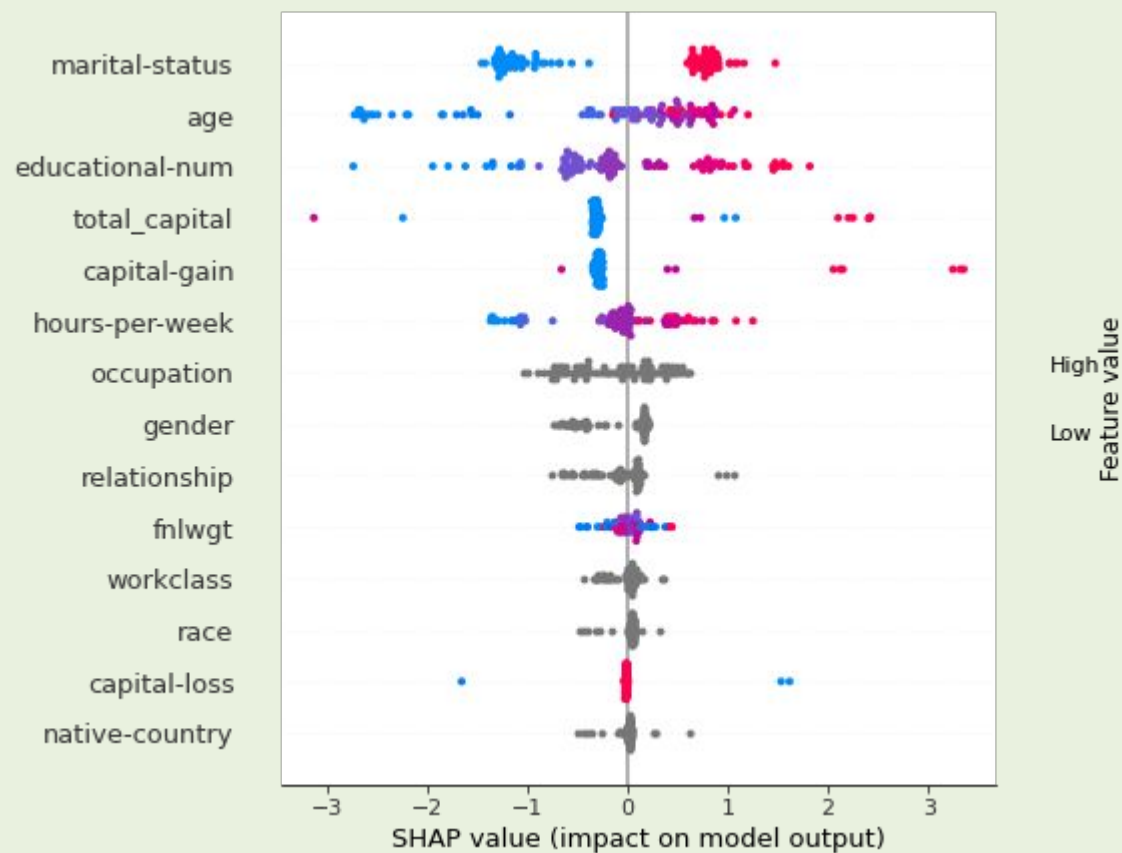
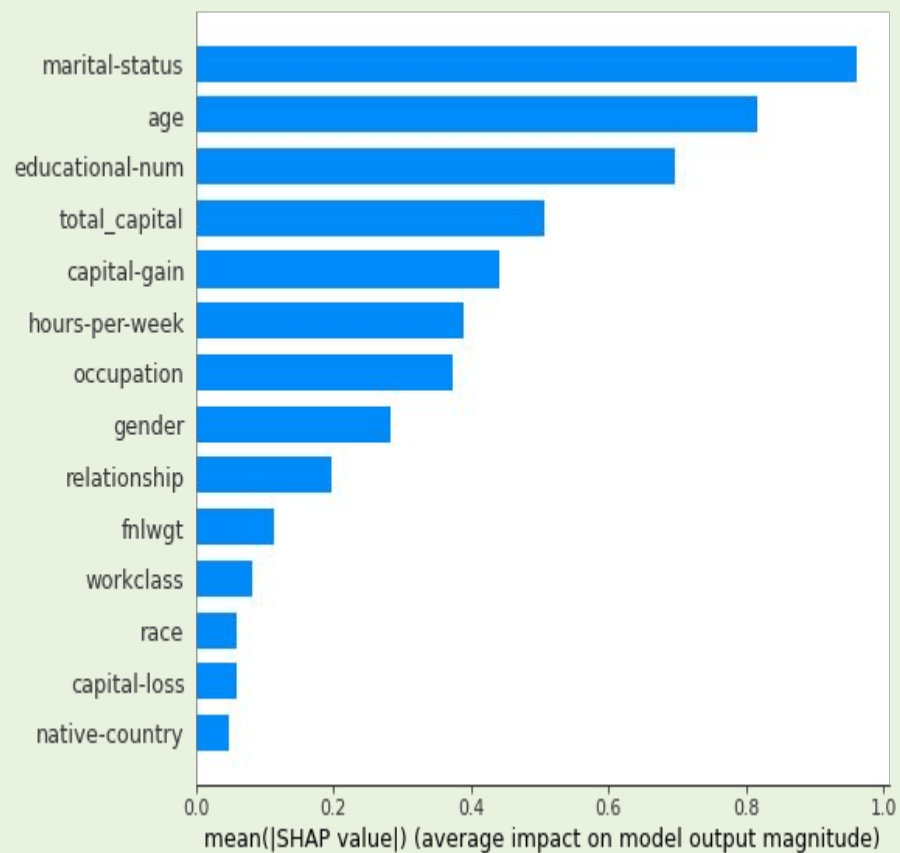
PDP plot

- capital gain, total capital: 자본이익/자본 총합계가 많을수록 소득이 \$50k보다 많을 확률이 커짐
- educational num: 교육수준이 높을수록 소득이 \$50k보다 많을 확률이 커짐
- marital status=1일 때, 즉 결혼한 상태일 때 소득이 \$50k보다 많을 확률이 커짐
- age는 20~50대에서는 나이가 많을수록, 50대 이상에서는 나이가 적을수록 소득이 \$50k보다 많을 확률이 커짐

 가설 2의 '더 많은 교육을 받은 사람이 소득>\$50000일 가능성이 높다.'에서 교육 수준이 타겟 변수에 영향을 주고 있다는 분석이 타당하다는 것을 확인

5. 결론 및 회고

SHAP Summary plot



5. 결론 및 회고



Shap summary plot

- positive/negative를 떠나 타겟 변수가 가장 크게 영향을 주는 특성은 marital status, age, educational-num, total capital 순
- 결혼 관계(marital status), 나이(age), 교육수준(eduacational-num), 자본총합계(total capital), 자본이익(capital-gain)은 특성값이 작을수록 타겟에 negative한 영향을 주고, 특성값이 클수록 positive한 영향을 줌

5. 결론 및 회고

아쉬운 점

- 카테고리를 디테일하게 전처리하지 못했다는 점
- feature engineering을 더 다양하게 시도하고 싶었지만, 시간 분배에 실패해 하지 못한 점
- 카테고리형 변수를 인코딩하는 과정에서 다양한 인코딩 방법을 사용하지 못한 점

 그러나 catboost 모델의 장점에 대해 알고 직접 구현했다는 점에서 의미있었음

감사합니다

THANK

YOU!