

AI
BOOTCAMP
SECTION 02
PROJECT

AI 부트캠프 10기 박민경

“목차”

1. 문제 제시

2. 가설 설정

3. 모델 학습

4. 해석 및 결론

1. 문제 제시

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

📌 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. 가설 설정

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

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.)

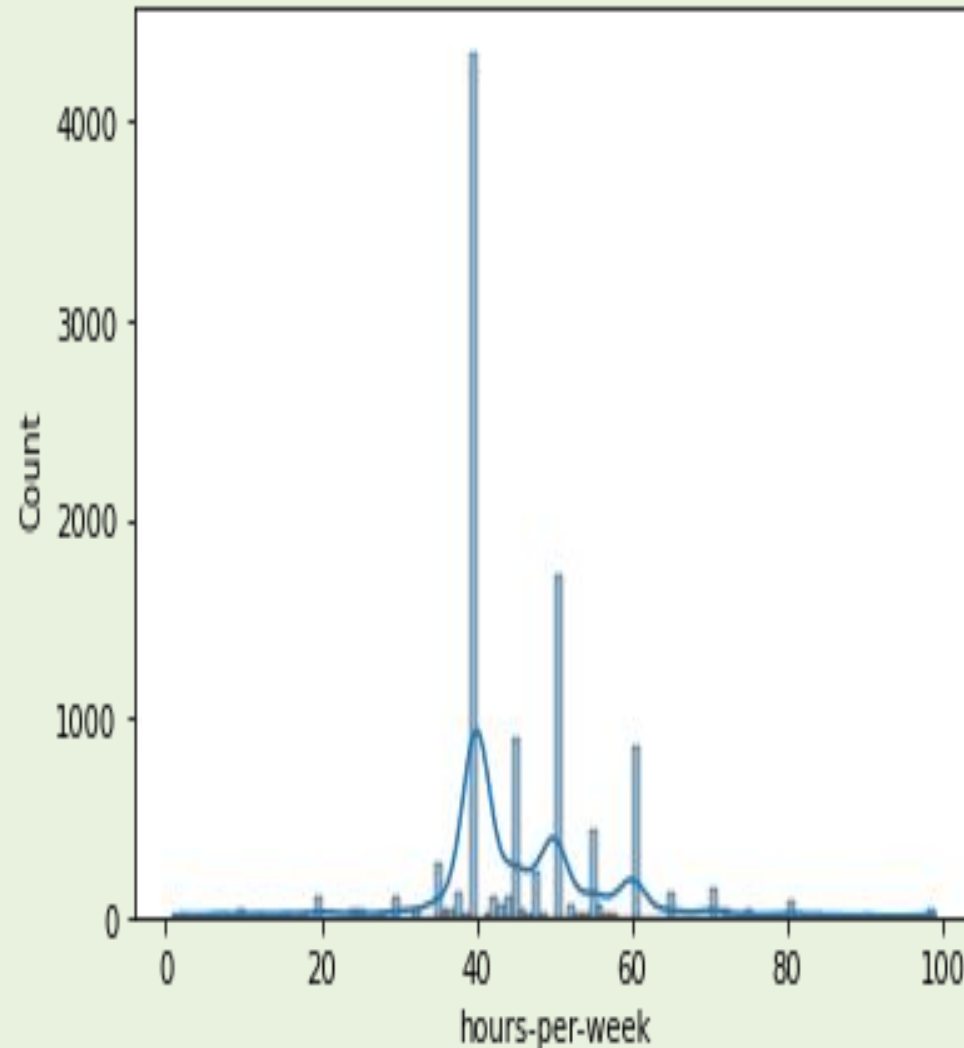
2. 가설 설정

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론



가설 1

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

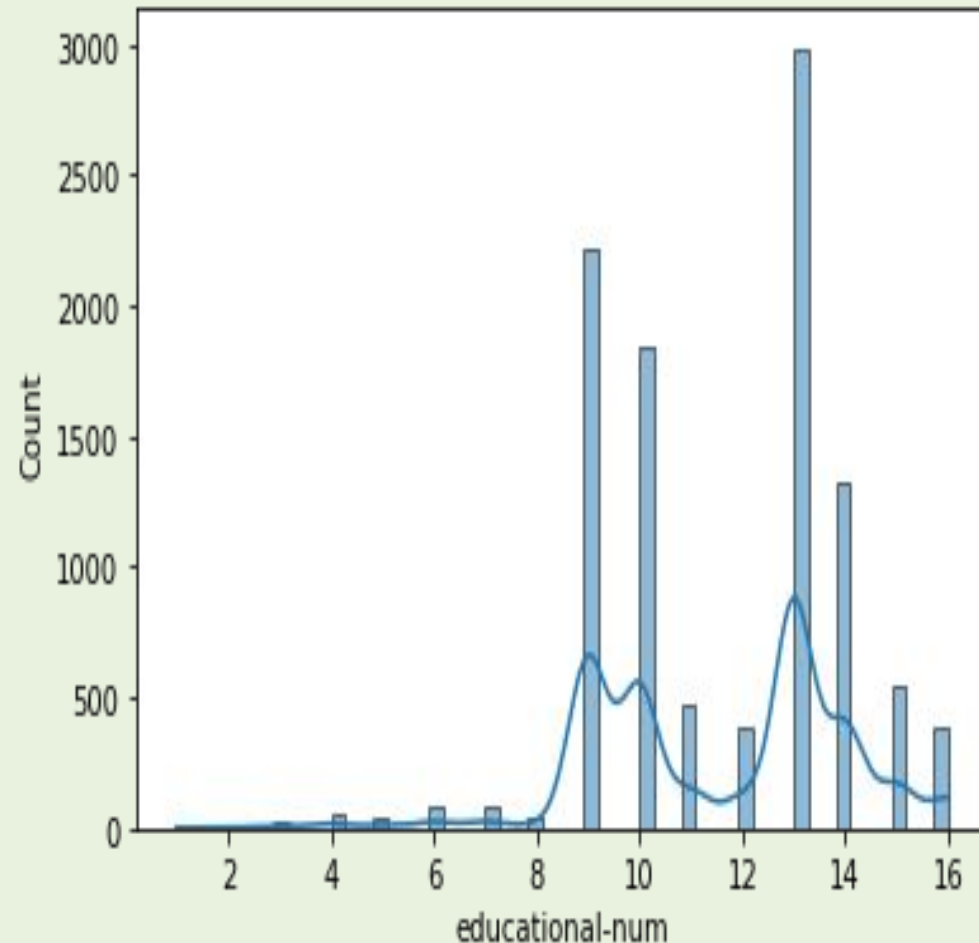
2. 가설 설정

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론



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.

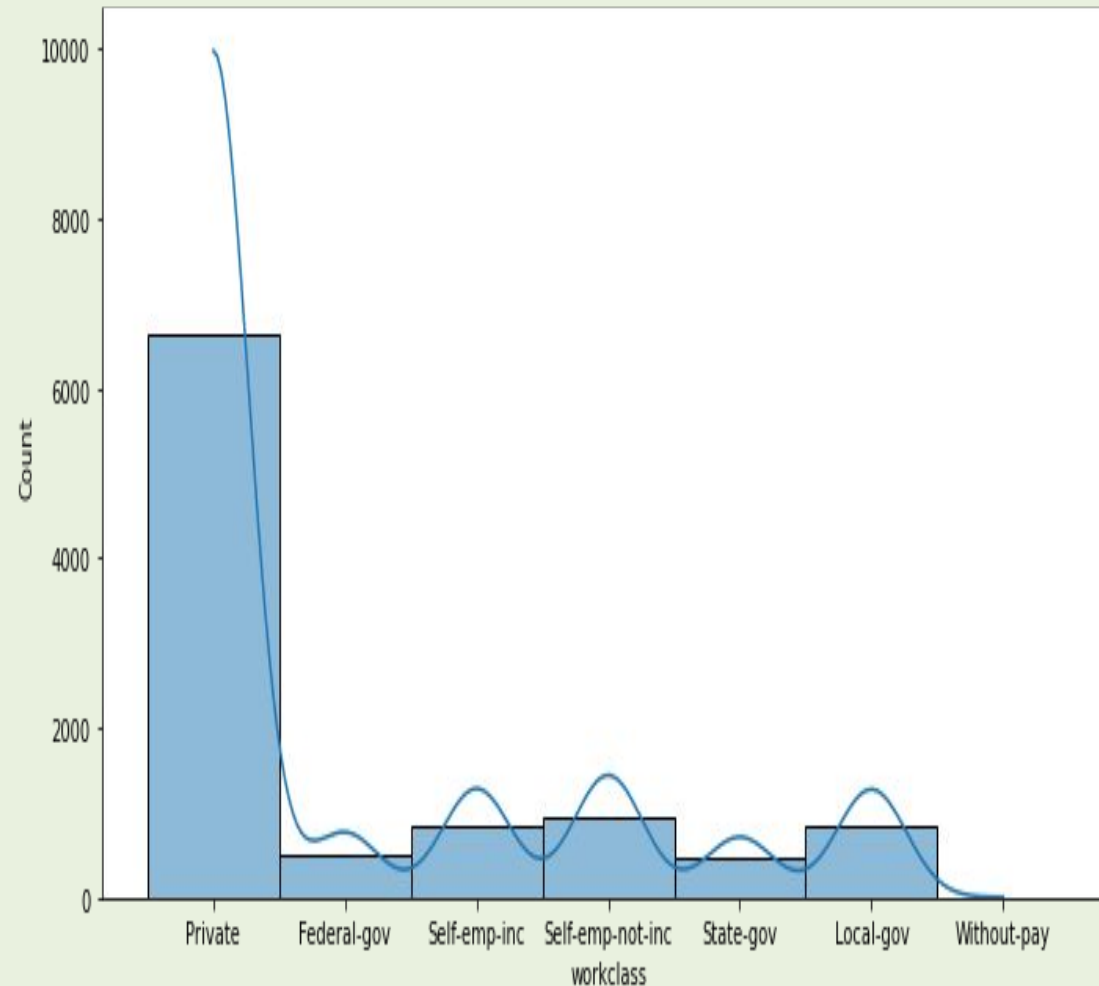
2. 가설 설정

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론



📌 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.

2. 가설 설정

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

Baseline

- 타겟이 binary 변수이며 분류 문제를 풀어야 하기 때문에 타겟 변수의 최빈값을 기준모델로 선정
- 기준모델 정확도: 0.760589

평가지표

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


3. 모델 학습


1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

 capital_gain과 capital_loss를 더해 자본 총합계 'total_capital' 컬럼 생성

 불균형 클래스를 조정하기 위해 class weight, scale_pos_weight만 설정

 logistic regression, random forest, xgbclassifier catboostclassifier 모델 학습

 훈련/검증 정확도와 f1 score, auc score를 구함

3. 모델 학습

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

✓ logistic regression

훈련 정확도 0.7973219165603935

검증 정확도 0.7979595554745855

Report

	precision	recall	f1-score	support
0	0.81	0.97	0.88	8358
1	0.71	0.26	0.38	2620
accuracy			0.80	10978
macro avg	0.76	0.61	0.63	10978
weighted avg	0.78	0.80	0.76	10978

f1 스코어 0.37905935050391937

auc점수 : 0.6127471691426964

✓ xgboost

훈련 정확도 0.874749498997996

검증 정확도 0.843960648569867

Report

	precision	recall	f1-score	support
0	0.95	0.84	0.89	8358
1	0.63	0.85	0.72	2620
accuracy			0.84	10978
macro avg	0.79	0.85	0.81	10978
weighted avg	0.87	0.84	0.85	10978

f1 스코어 0.7226809130645944

auc점수 : 0.8466888239817771

✓ random forest

훈련 정확도 0.9999392724843627

검증 정확도 0.8606303516123155

Report

	precision	recall	f1-score	support
0	0.89	0.94	0.91	8358
1	0.76	0.61	0.68	2620
accuracy			0.86	10978
macro avg	0.82	0.78	0.79	10978
weighted avg	0.85	0.86	0.86	10978

f1 스코어 0.6780303030303031

auc점수 : 0.7762750046123018

✓ catboost

훈련 정확도 0.8621789032610676

검증 정확도 0.8401348150847149

Report

	precision	recall	f1-score	support
0	0.95	0.83	0.89	8358
1	0.62	0.86	0.72	2620
accuracy			0.84	10978
macro avg	0.78	0.85	0.80	10978
weighted avg	0.87	0.84	0.85	10978

f1 스코어 0.7203187250996015

auc점수 : 0.8478447307420416

3. 모델 학습

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

📌 과적합 가능성과 불균형 클래스를 잘 조정해 f1 score가 좋은 **xgboost**, **catboost** 모델만 **RandomizedSearch CV**로 하이퍼파라미터를 조정해 2차 모델 학습 실행

✓ xgboost

```
훈련 f1 score: 0.7381201590956668
검증 f1 score: 0.7134478424801006
Report
      precision    recall  f1-score   support

     0       0.95      0.83      0.88      6256
     1       0.61      0.86      0.71      1978

 accuracy          0.83      8234
 macro avg       0.78      0.84      0.80      8234
weighted avg       0.87      0.83      0.84      8234

auc점수 : 0.8431292006185689
```

✓ catboost

```
훈련 f1 score: 0.7451420554191511
검증 f1 score: 0.7158250581026834
Report
      precision    recall  f1-score   support

     0       0.95      0.83      0.89      6256
     1       0.61      0.86      0.72      1978

 accuracy          0.84      8234
 macro avg       0.78      0.84      0.80      8234
weighted avg       0.87      0.84      0.84      8234

auc점수 : 0.8434117200975435
```

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

3. 모델 학습

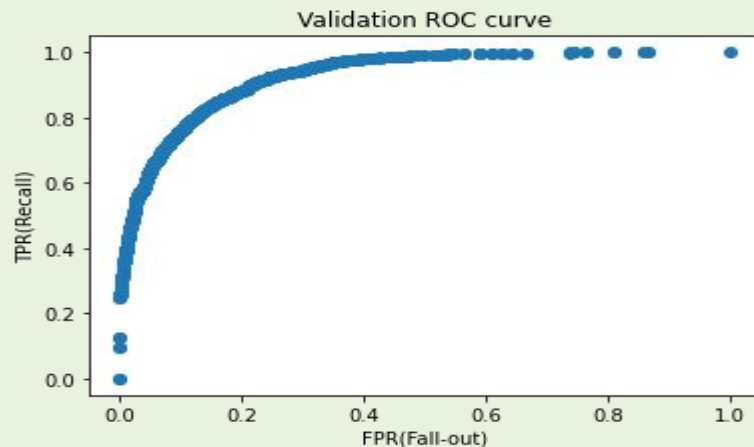
1 문제 제시

2 가설 설정

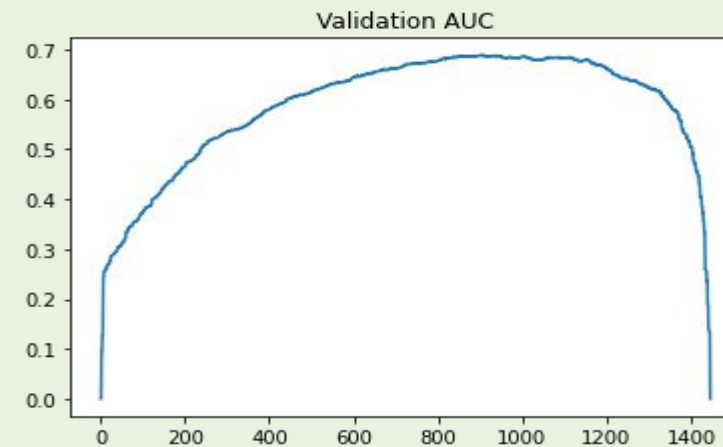
3 모델 학습

4 분석 및 결론

VAL ROC curve



VAL AUC



VAL 평가지표: 최적 임계값=0.5273530619303364일 때

Report				
	precision	recall	f1-score	support
0	0.95	0.84	0.89	6256
1	0.63	0.85	0.72	1978
accuracy			0.84	8234
macro avg	0.79	0.84	0.81	8234
weighted avg	0.87	0.84	0.85	8234

검정 정확도 0.8358027690065581
val_f1 스코어 0.7217915590008613
val_auc점수 : 0.8445362219710938

3. 모델 학습

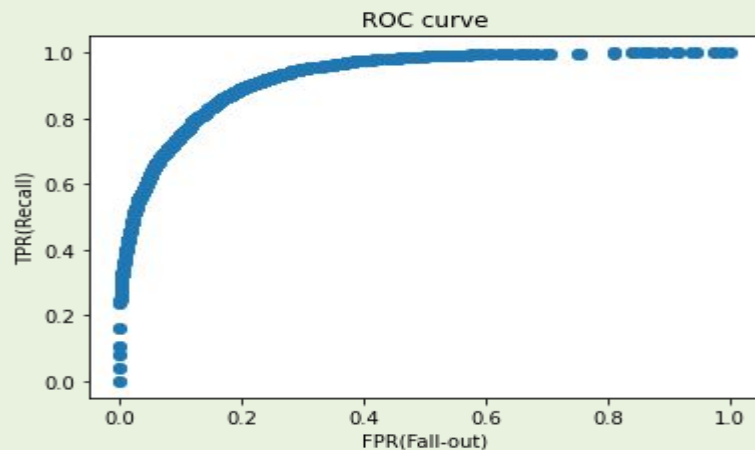
1 문제 제시

2 가설 설정

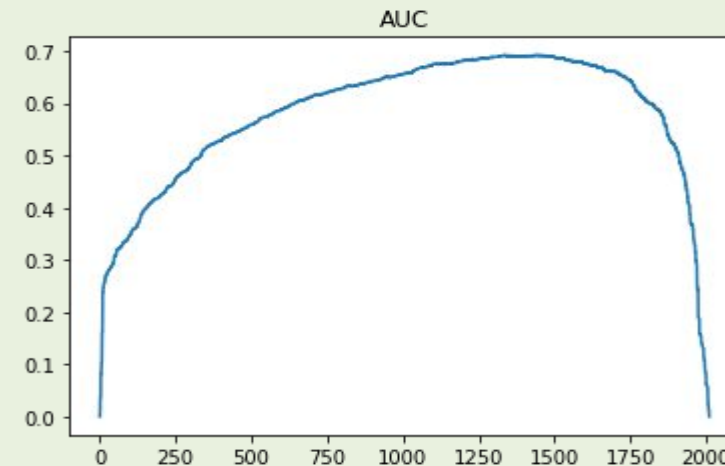
3 모델 학습

4 분석 및 결론

TEST ROC curve



TEST AUC



TEST 평가지표: 최적 임계값=0.45233375920960567일 때

Report				
	precision	recall	f1-score	support
0	0.96	0.81	0.87	8358
1	0.59	0.89	0.71	2620
accuracy			0.82	10978
macro avg	0.77	0.85	0.79	10978
weighted avg	0.87	0.82	0.83	10978

테스트 정확도 0.8360357077791948
test_f1 스코어 0.707242848447961
test_auc점수 : 0.846119729874381

3. 모델 학습

1 문제 제시

2 가설 설정

3 모델 학습

4 분석 및 결론

모델 분석

- 테스트 정확도는 baseline 정확도보다 향상
- f1 score는 훈련/검증 데이터의 f1 score와 큰 차이를 보이지 않아 어느 정도 일반화 성능을 갖추었음을 확인
- 타겟 변수의 불균형 클래스 문제를 상당 부분 해결
- 과적합이 발생하지 않고 일반화가 원활히 이루어졌음

4. 해석 및 결론

1 문제 제시

2 가설 설정

3 모델 학습

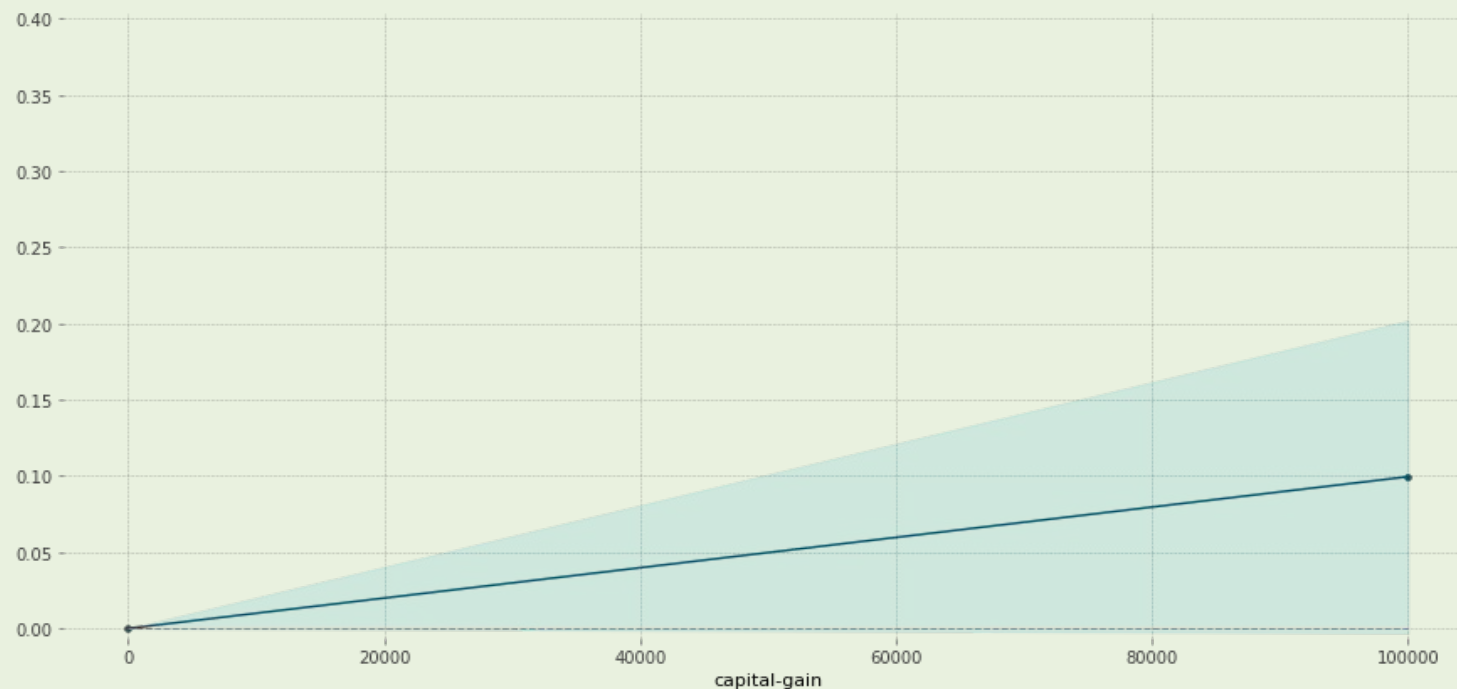
4 해석 및 결론

📌 순열 중요도

Weight	Feature
0.0734 ± 0.0105	marital-status
0.0474 ± 0.0086	age
0.0449 ± 0.0115	educational-num
0.0405 ± 0.0030	total_capital
0.0190 ± 0.0024	capital-gain
0.0180 ± 0.0068	occupation
0.0154 ± 0.0021	hours-per-week
0.0098 ± 0.0011	capital-loss
0.0086 ± 0.0025	relationship
0.0045 ± 0.0033	workclass
0.0017 ± 0.0033	fnlwgt
0.0013 ± 0.0056	gender
-0.0003 ± 0.0008	native-country
-0.0010 ± 0.0016	race

PDP for feature "capital-gain"

Number of unique grid points: 2



4. 해석 및 결론

1 문제 제시

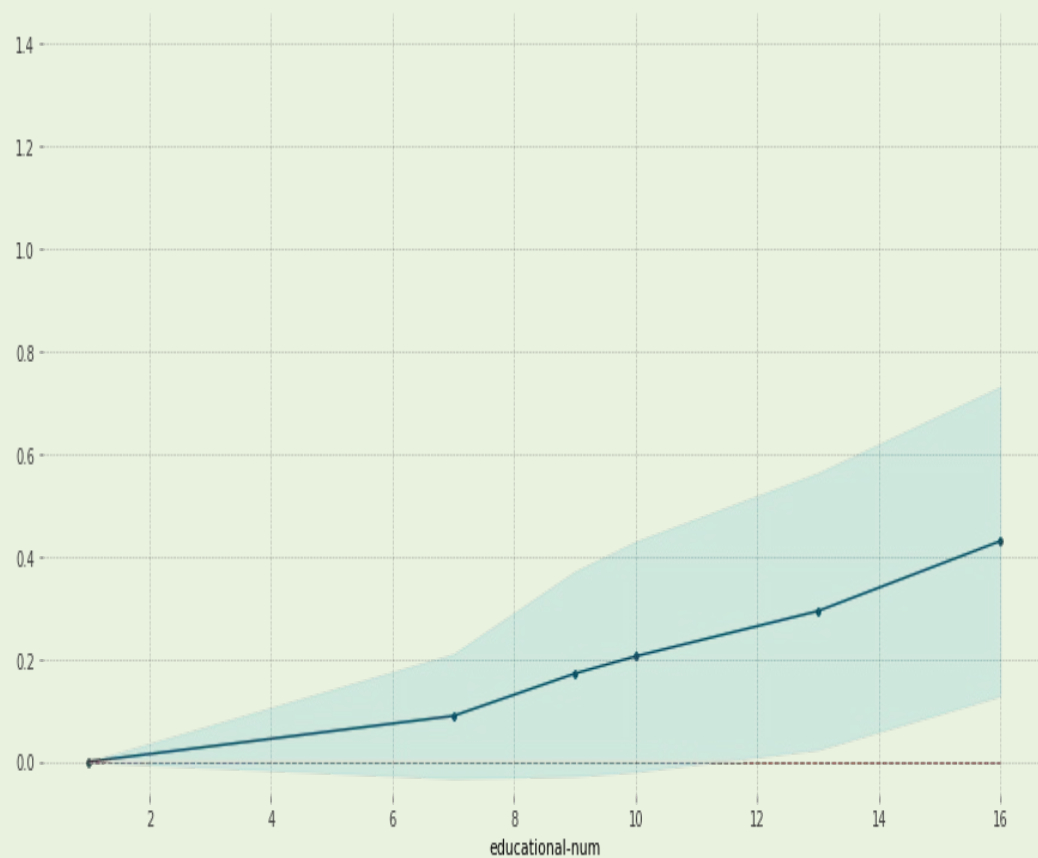
2 가설 설정

3 모델 학습

4 해석 및 결론

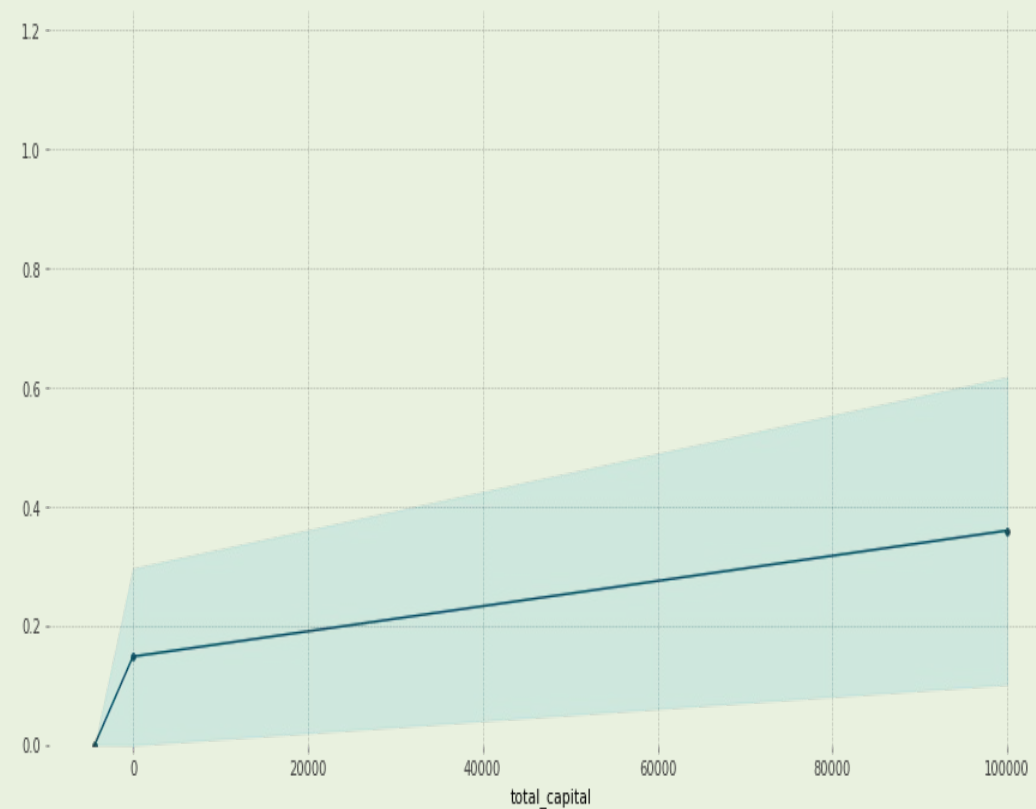
PDP for feature "educational-num"

Number of unique grid points: 6



PDP for feature "total_capital"

Number of unique grid points: 3



4. 해석 및 결론

1 문제 제시

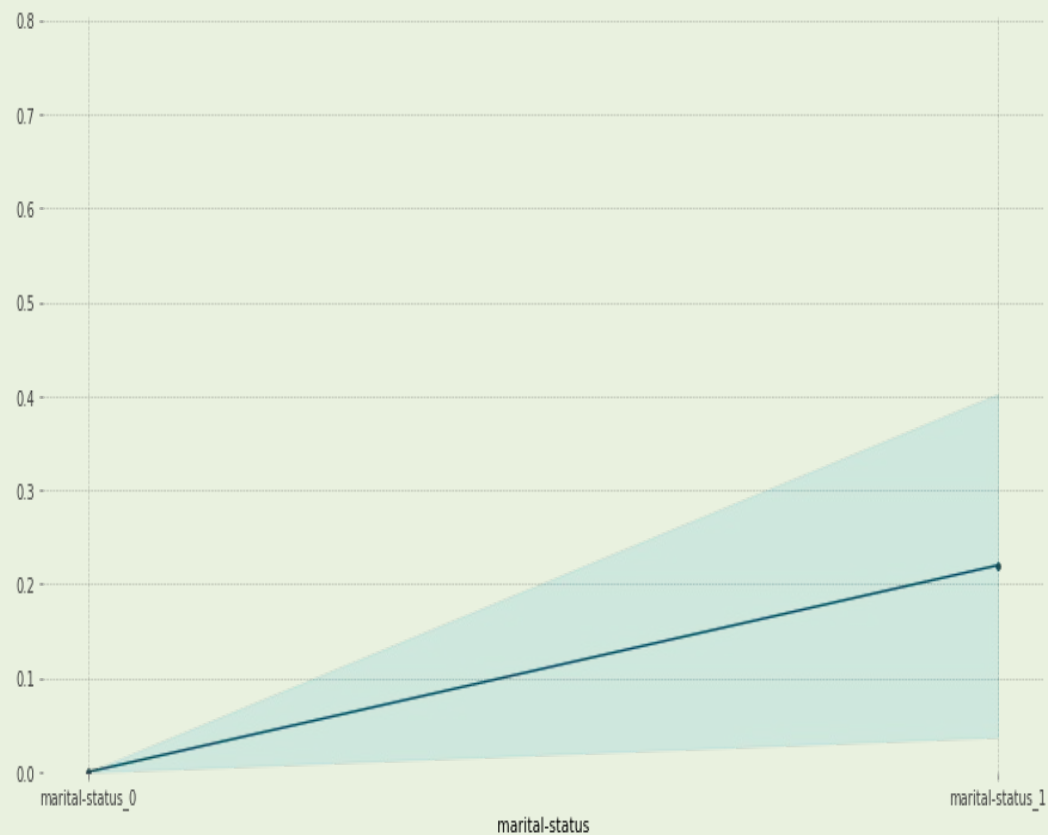
2 가설 설정

3 모델 학습

4 해석 및 결론

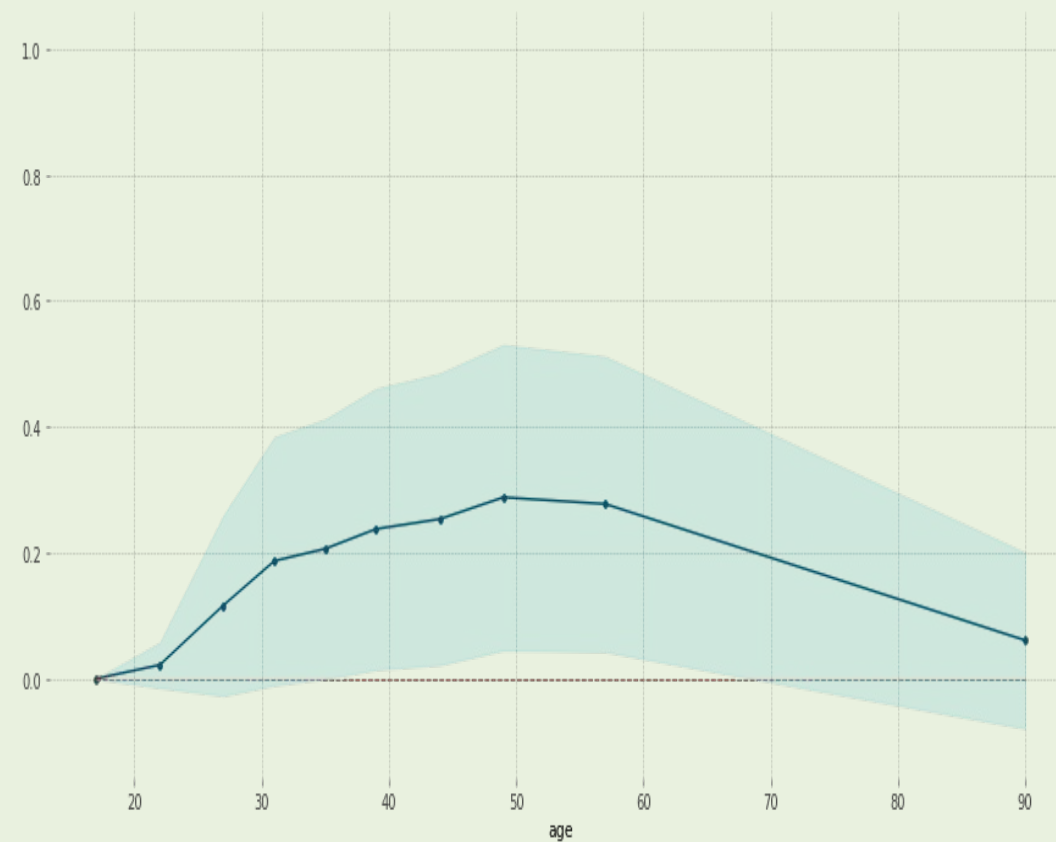
PDP for feature "marital-status"

Number of unique grid points: 2



PDP for feature "age"

Number of unique grid points: 10



4. 해석 및 결론

1 문제 제시

2 가설 설정

3 모델 학습

4 해석 및 결론

PDP plot

- capital gain, total capital이 많을수록 소득이 \$50k보다 많을 확률이 커짐
- educational num이 높을수록 소득이 \$50k보다 많을 확률이 커짐
- marital status=1일 때, 즉 결혼한 상태일 경우 소득이 \$50k보다 많을 확률이 커짐
- age는 20~50대에서는 나이가 많을수록, 50대 이상에서는 나이가 적을수록 소득이 \$50k보다 많을 확률이 커짐

가설 2: 더 많은 교육을 받은 사람이 소득>\$50000일 가능성이 높다.

- 가설 2의 교육 수준이 타겟 변수에 영향을 준다는 것이 확인됨

4. 해석 및 결론

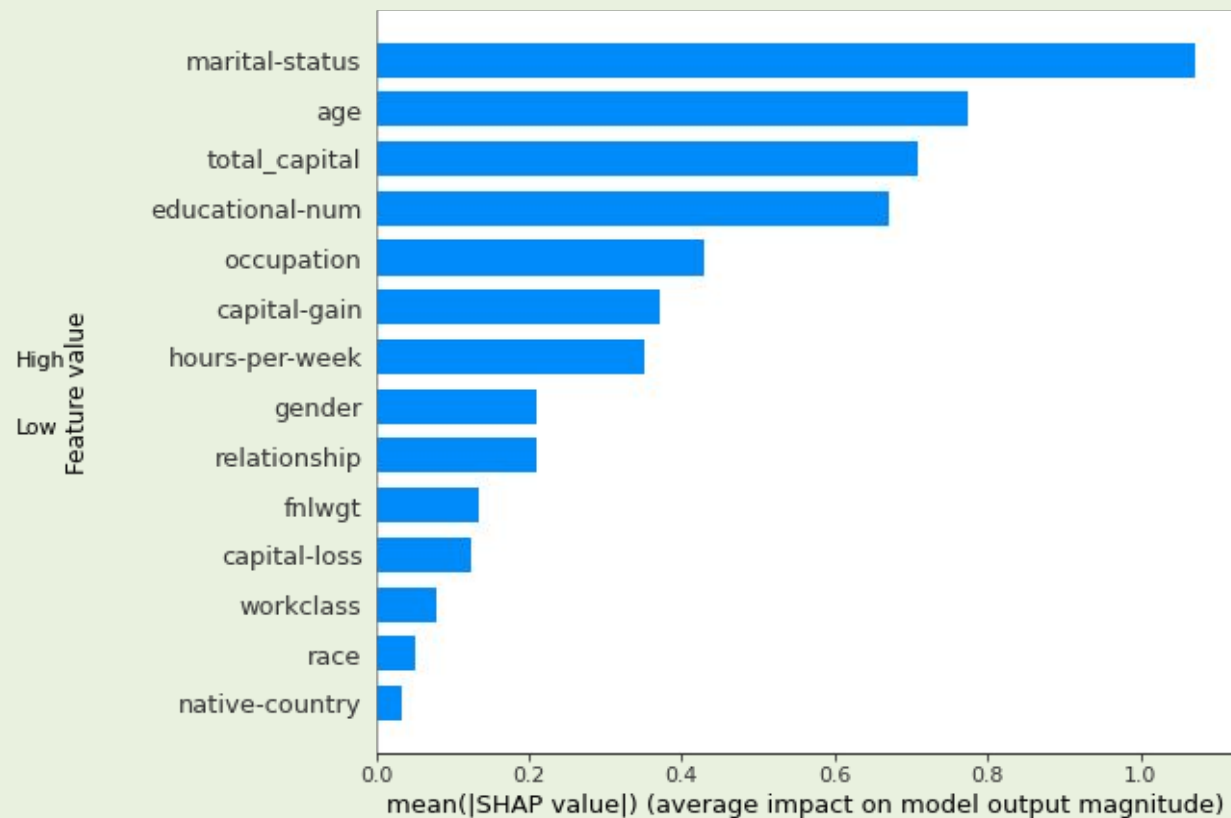
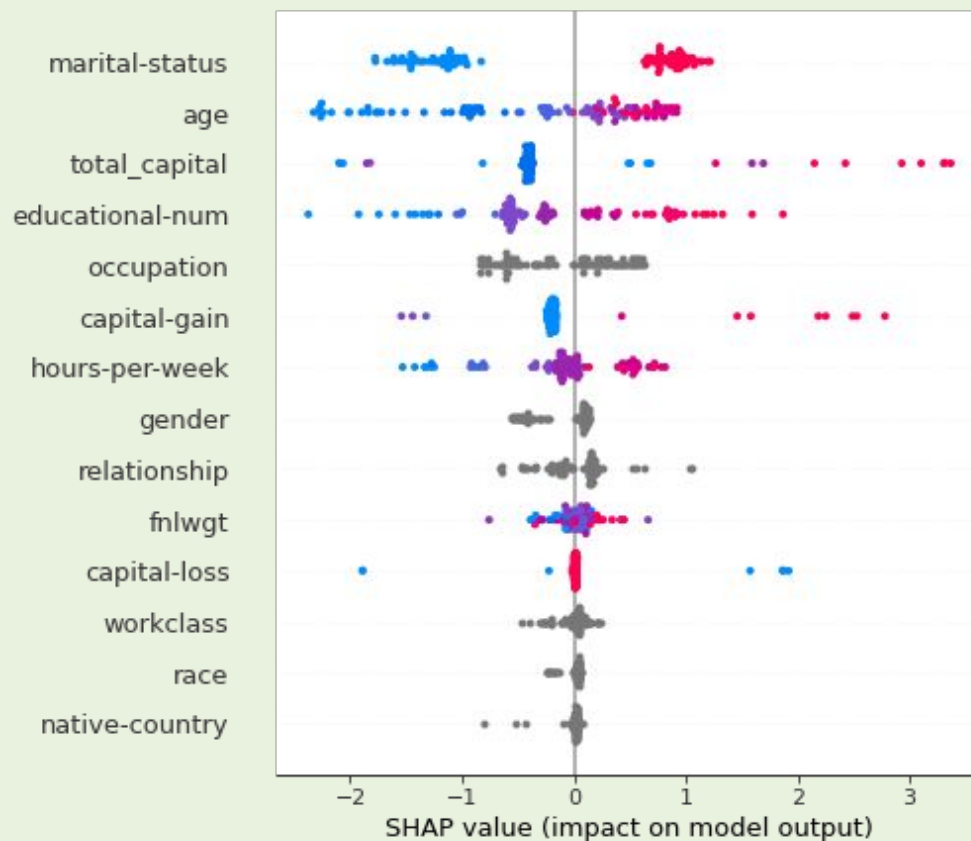
1 문제 제시

2 가설 설정

3 모델 학습

4 해석 및 결론

📌 SHAP Summary plot



4. 해석 및 결론

1 문제 제시

2 가설 설정

3 모델 학습

4 해석 및 결론

Shap summary plot

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

한계

- 카테고리를 디테일하게 전처리하지 못했다는 점



느낀 점



아쉬운 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 어떤 장점이 있는지 알고 직접 적용했다는 점에서 의미있다고 생각

감사합니다

THANK

YOU!