Decision Tree를 이용한 소득수준 예측

목차

데이터셋 소개

1994년 미국 인구조사국 데이터베이스에서 추출된 데이터셋

미국에 사는 개인의 개인정보들이 담겨있다.

shape = (48842, 14)　14개의 변수 (5개는 int, 9개는 string)

target 변수인 소득수준은 string의 범주형 (연간소득 5만달러 기준)

데이터셋 소개

변수들 소개

숫자형 변수

나이, 교육연수, 자본소득(이자, 부동산 등)
자본손실, 주 노동시간

문자형 변수

직군, 교육수준, 기혼상태, 직업, 가족 내 역할
인종, 성별, 국적, 소득수준(target)

데이터 탐색 / 전처리

결측치

직군, 직업, 국적에 결측치가 존재한다.
각각 2799, 2809, 857개씩

중복값

6374개 존재한다.
우연히 겹친것이라고 판단되어서 드롭하지 않았다.
(미국에 사는 35세 기혼 백인 남성, 자영업자는 흔할 것)

교육연수와 교육수준

1대1 대응관계라 교육수준을 drop 하였다.

데이터 탐색 / 전처리

결측치 처리직군, 직업

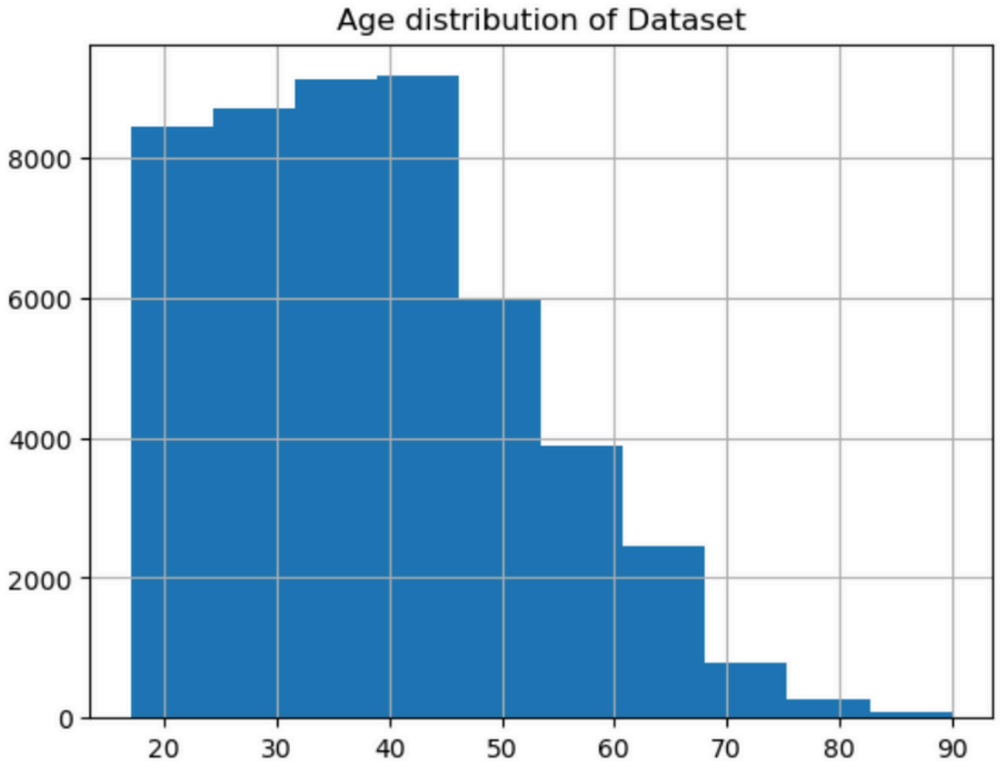
직업이 결측된 데이터들 중 거의 대부분 직군도 결측
단 10개의 데이터들만 직군이 Never-worked로 되어있다.

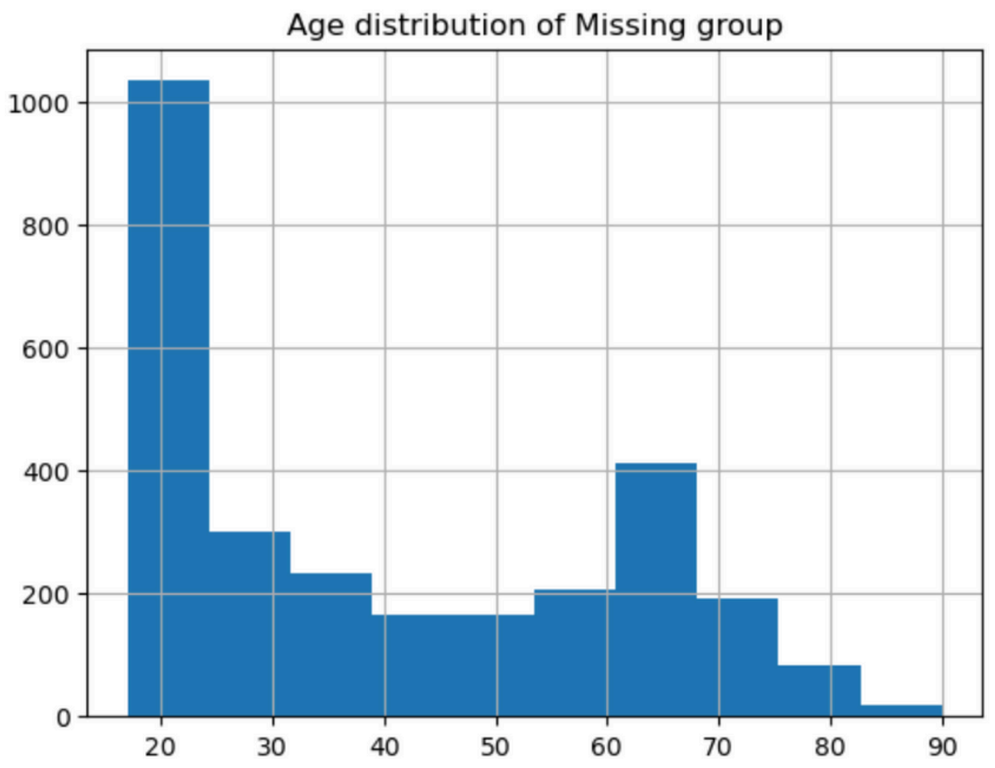
직군, 직업이 결측된 데이터들은 모두 무직자일까?

데이터 탐색 / 전처리

결측치 처리직군, 직업

직군, 직업이 결측된 데이터의 연령분포





데이터의 분포에 비해서 학생, 사회초년생 층 연령대인 20대 초반과
은퇴를 하는 나이대인 60대 이상 연령의 분포가 두드러지는 모습이다.

데이터 탐색 / 전처리

결측치 처리직군, 직업

실제 데이터고 추출 경로가 알려져있다보니까 그 당시 실업률과 비교해볼 수 있었다.
데이터의 레코드 중 91.3% 가량이 미국 국적이어서 전체 실업률을 미국의 실업률로 일반화하였다.


```
df.shape[0] * 0.055
```

```
2686.31
```

위의 숫자를 결측치의 수 2809와 비교해보니
얼추 비슷하다.

직군, 직업이 결측된 데이터들은 무직자라는 결론
새로운 범주를 만들어주는 것으로 결측치 처리

데이터 탐색 / 전처리

결측치 처리국적

857개, 1.8% 가량의 데이터의 국적이 결측되어있다.
정확한 이유는 모르겠지만 직업, 직군과 비슷하게 이유 있는 결측일 것이라고 생각되어서,
새로운 범주를 만들어줬다.

소득 수준(타겟)

5만달러 이하면 0, 아니면 1로 매핑해주었다.

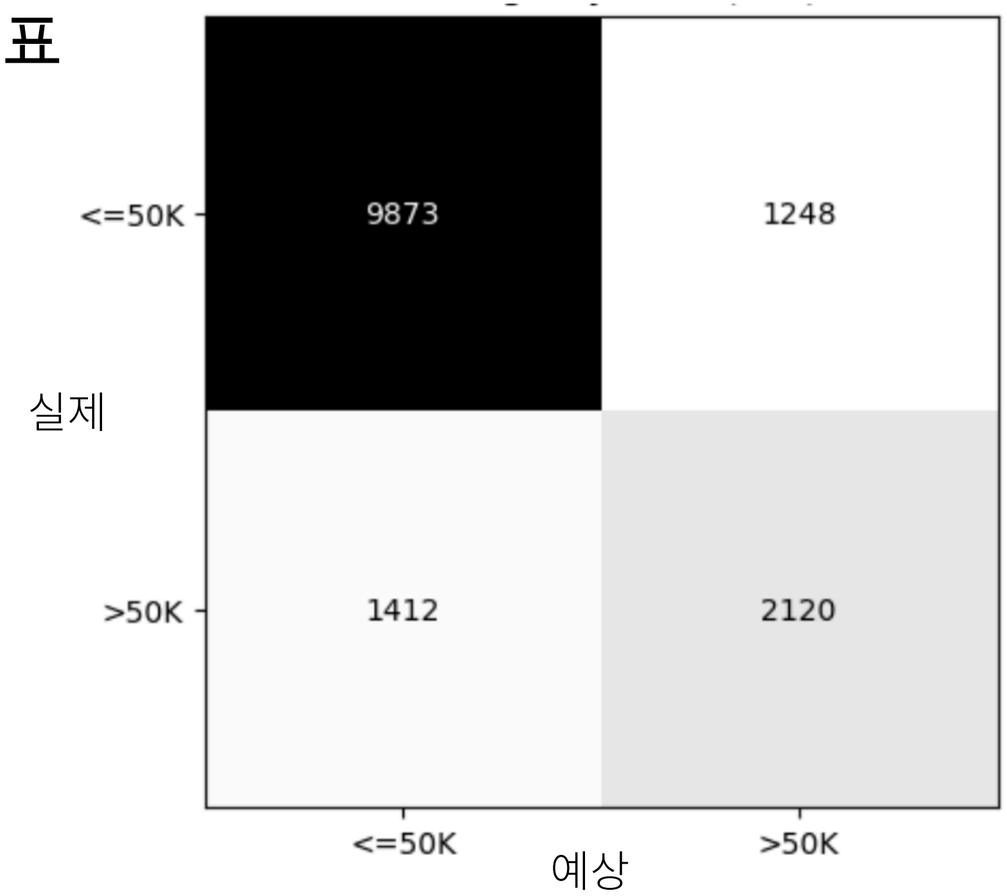
모든 범주형 피쳐

원 핫 인코딩을 해주었다.

모델 학습 및 평가

모델 정확도 : **0.818**

분할표



클래스 불균형 문제는 나온 모습이지만,
최대한 많은 사람에 대해 예측을 성공하는것이
중요한 문제라고 생각해서 무시함

모델 학습 및 평가 / 성능 향상

트리 깊이 : **54**, 리프노드 개수 : **6017**

train_score : 0.976
test_score : 0.818

과적합이 보인다. 트리 깊이, 트리 분할 기준 등
하이퍼파라미터 튜닝을 통해서 성능향상과 과적합 해소
GridsearchCV로 최적의 하이퍼파라미터를 찾은 결과

분류 기준 : log_loss , 트리 깊이 : 11

재적합 결과 **train_score : 0.873**
test_score : 0.853

일반적인 성능이 좋아진 모습

앙상블 모델과 비교

더 정확도가 높고 과적합에 강한 앙상블 모델들이 있는데 왜 의사결정나무?

앙상블 모델인 랜덤포레스트로 소득수준을 예측 해 보았다.

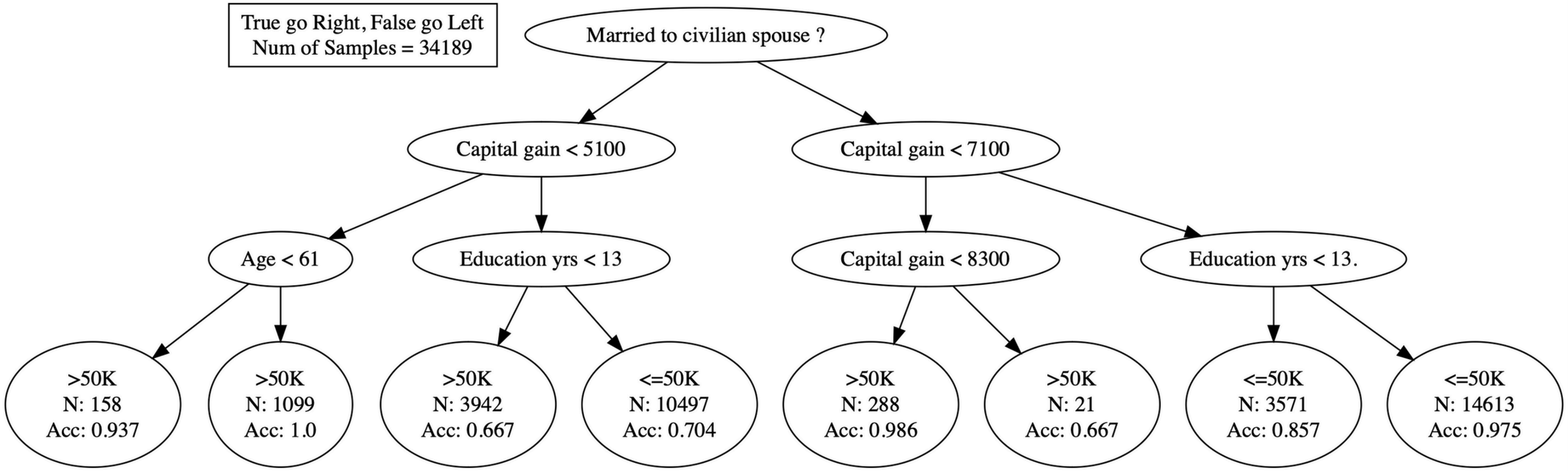
모델 정확도 : 0.864 , oob_score : 0.865

모델 정확도도 더 높고 과적합에도 강한 모습이다.

그럼에도 의사결정나무를 쓰는 이유를 찾아보니 설명력이 더 높다고 한다.

앙상블 모델과 비교

앙상블 모델은 변수 중요도까지밖에 못구하지만, 의사결정트리는 이런식으로 시각화도 가능하다.
깊이를 3으로 제한해도 꽤 높은 정확도를 보여주고 (정확도 약 0.844)
의사결정 과정이 직관적으로 시각화 되는 모습이다.



end of page