Part. 04

Ensemble Learning

중요변수 추출 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

I 중요 변수 추출 방법

- **선형회귀**
  - '다른 변수가 고정되어 있고 TV광고가 1단위 증가할 때, 매출이 0.046단위 증가한다.'
  - 변수의 유의성은 p-value를 통해 파악 가능

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

- **Decision Tree**
  - 독립 변수의 조건에 따라 종속변수를 분리 (비가 내린다 -> 축구를 하지 않는다)
  - 해석이 매우 용이

I 중요 변수 추출 방법

- **복잡한 모델은 해석이 쉽지 않음**
  - Bagging, RandomForest, Gradient Boosting, Neural Network 등은 모형에 대한 해석과 prediction에 대한 해석이 어려움

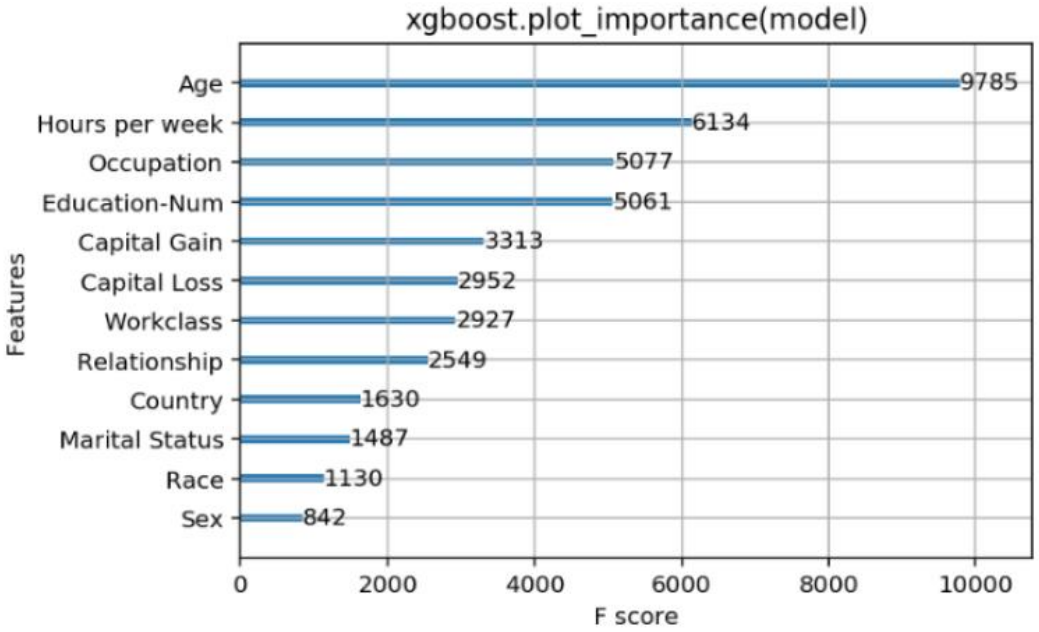
Accuracy가 낮고 설명하기 쉬운 모델 (Linear regression, DT)

vs

Accuracy가 높고 설명하기 어려운 모델 (Ensemble Learning, NN)

I 중요 변수 추출 방법

- **Xgboost의 feature importance**
  - Ensemble learning 모델들은 중요 feature를 추출 할 수 있는 알고리즘이 내장 되어있음
  - 은행 이용 고객 데이터에대해 수입이 50만달러가 넘는지(Target)를 예측하는데 중요한 feature들을 나타냄
  - Age, Hours per week 등이 중요한 feature라는 것을 알 수 있음



I 중요 변수 추출 방법

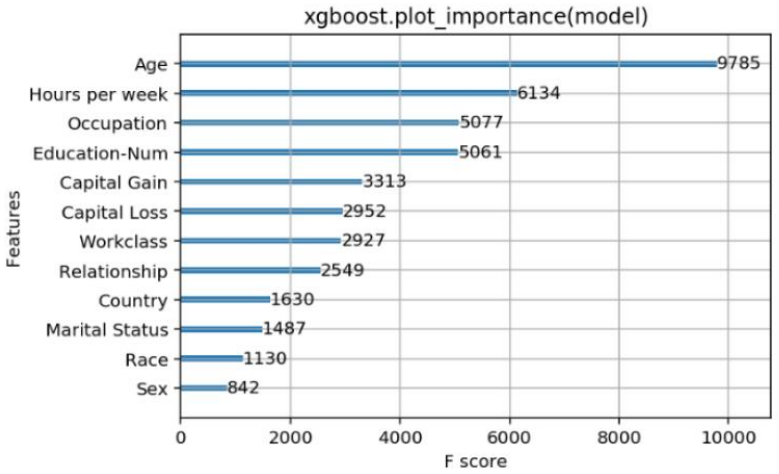
- **Xgboost의 feature importance 측정 기준**
  - Weight : 변수 별 데이터를 분리하는 데 쓰인 횟수
  - Cover : 변수 별 데이터를 분리하는 데 쓰인 횟수(해당 변수로 분리된 데이터의 수로 가중치)
  - Gain : Feature를 사용했을 때 줄어드는 평균적인 training loss

I 중요 변수 추출 방법

- **Xgboost의 feature importance 의 문제점**

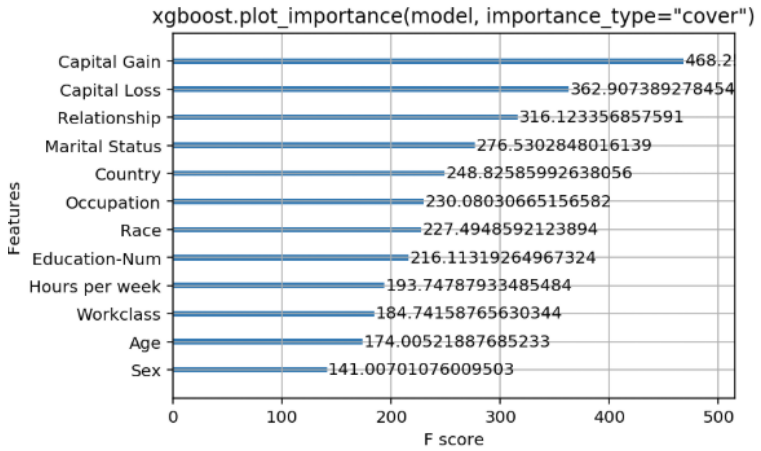
  변수 중요도 측정 기준별로 중요 변수가 상이함.

Weight



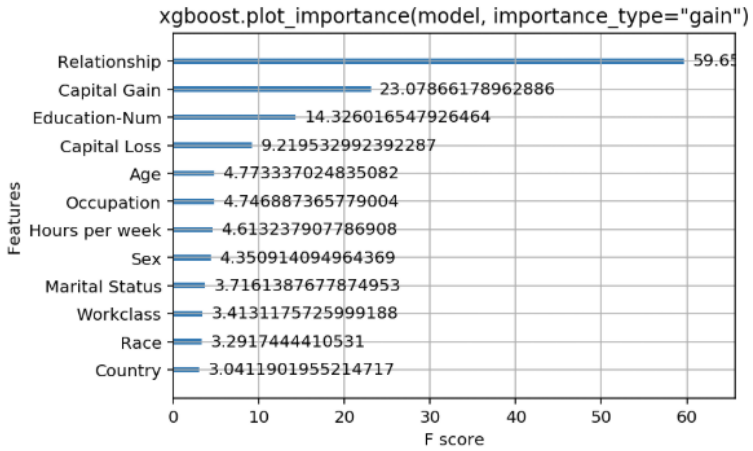
Age, Hours per week…

Cover



Capital Gain, Capital Loss…

Gain



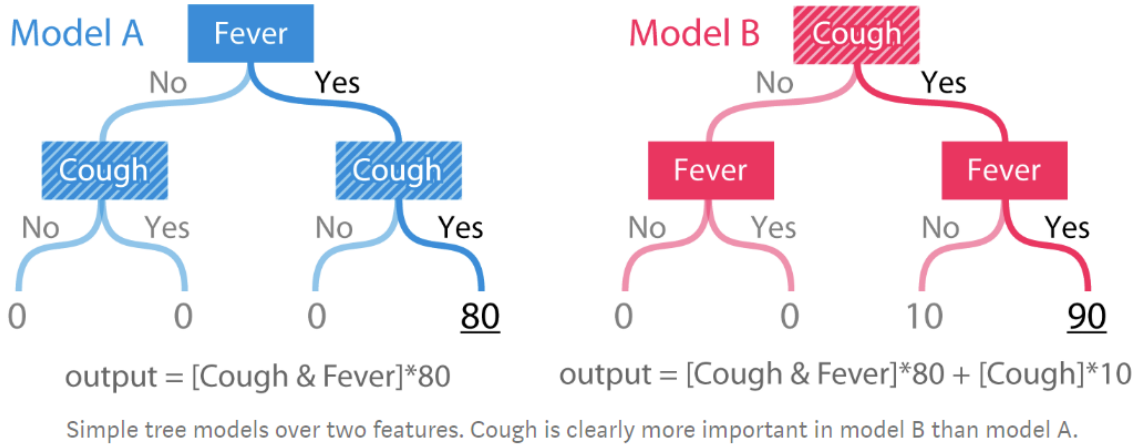
Relationship, Capital Gain…

I 중요 변수 추출 방법

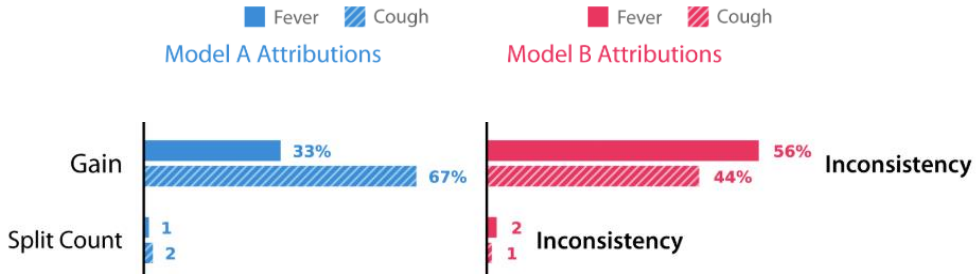
- **Feature importance의 좋고 나쁨의 기준**
  - Consistency : 특정 feature가 영향이 많이 가도록 모델을 수정하였다면, 중요도 측정 시 해당 feature의 중요도가 줄지 않아야함
  - Consistency가 없다면, 두 모델의 feature에 대해 비교가 힘들고 feature importance가 높다고 해서 중요하다라고 말하기 힘듦
  - 대부분의 feature importance지표는 inconsistency

I 중요 변수 추출 방법

- **Feature importance의 좋고 나쁨의 기준**
  - Model의 output = 특정 질병의 risk score
  - 두 모델은 거의 비슷한 모델 (cough 변수는 model A보다 B에서 더 중요한 역할을 함)
  - 변수중요도를 보면 각 모델별로 상이한 결과가 도출(inconsistency)



Simple tree models over two features. Cough is clearly more important in model B than model A.



Part. 04

Ensemble Learning

Shap value

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택