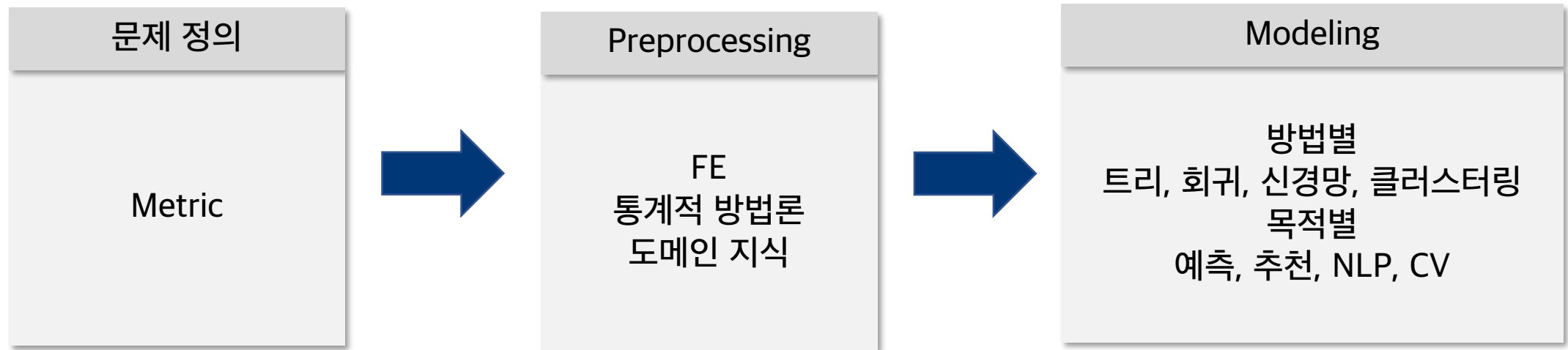


Feature engineering

머신러닝 문제해결 과정



FE의 중요성

실제로 프로젝트를 해보면
좋은 Feature만들기의 비중이 압도적임

이유는

모델링면에서 부족해도 FE에서 보완할 수 있기 때문

전처리 과정

Brainstorm	데이터 많이 보기, 다른 문제들 참고
FE	4가지 방법 + 센스
Select	Importance등 참고
Evaluate &Revise	FE가 적용된 새 데이터에 대한 모델 정확도

주의점

주로 쓰이는 방법들을 알려드리겠지만
사실 경험의 영역입니다

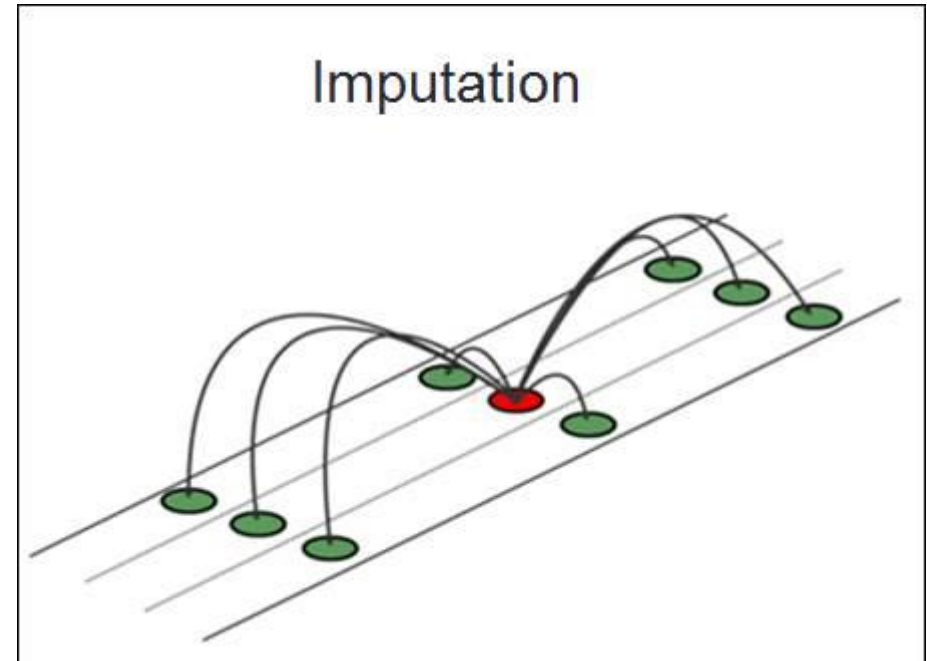
제가 지금 드리는 말씀이 와닿지 않아도 괜찮습니다
앞으로 많은 데이터를 보실테니까요

1. Imputing

WHY : 모델한테 Null값 넣으면 안됨

5%이상이면 Feature로 추가 (우연이 아님)


1. Numerical : 00이나 median
2. Categorical : mode, 없다면 'Null'
3. Random sampling : 임의로 뽑은 값



	col1	col2	col3	col4	col5		col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN	mean()	0	2.0	5.0	3.0	6.0 7.0
1	9	NaN	9.0	0	7.0		1	9.0	11.0	9.0	0.0 7.0
2	19	17.0	NaN	9	NaN		2	19.0	17.0	6.0	9.0 7.0

2. Encoding

User	City
1	Roma
2	Madrid
1	Madrid
3	Istanbul
2	Istanbul
1	Istanbul
1	Roma



User	Istanbul	Madrid
1	0	0
2	0	1
1	0	1
3	1	0
2	1	0
1	1	0
1	0	0

One hot encoding example on City column

WHY

모델한테 Categorical 넣으면 안됨

1. Categorical하면 원핫 인코딩

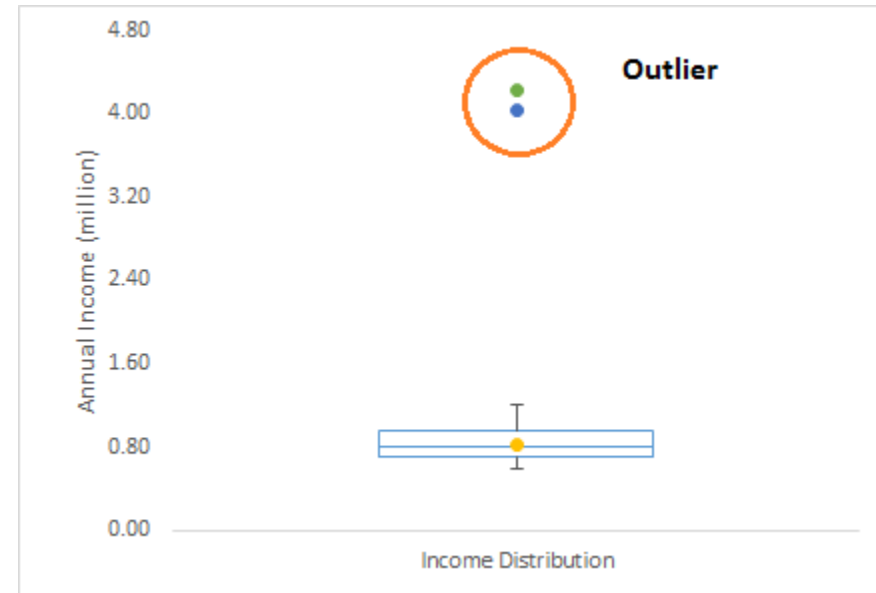
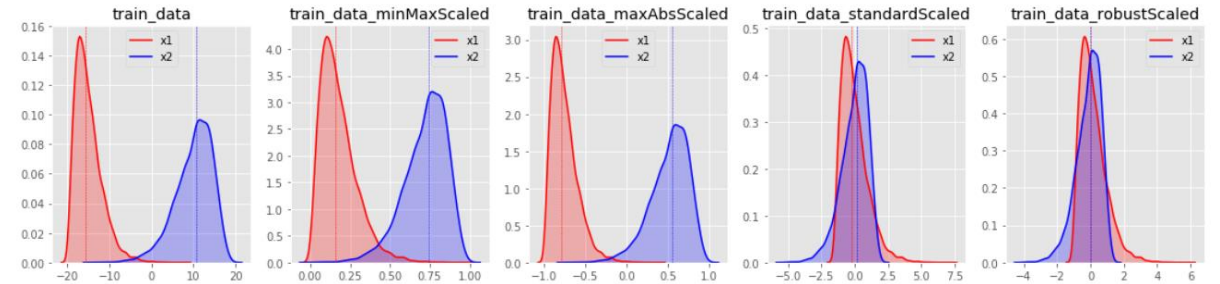
2. 순서있다면 레이블 인코딩 가능

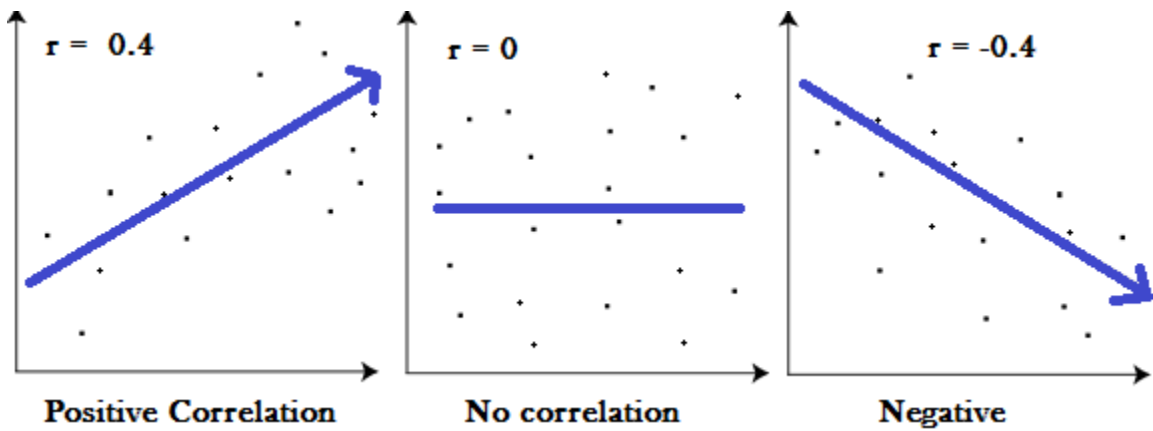
3. Outlier removing

WHY

모델 오버피팅 방지

Boxplot그리고 IQR밖의 인스턴스 제거





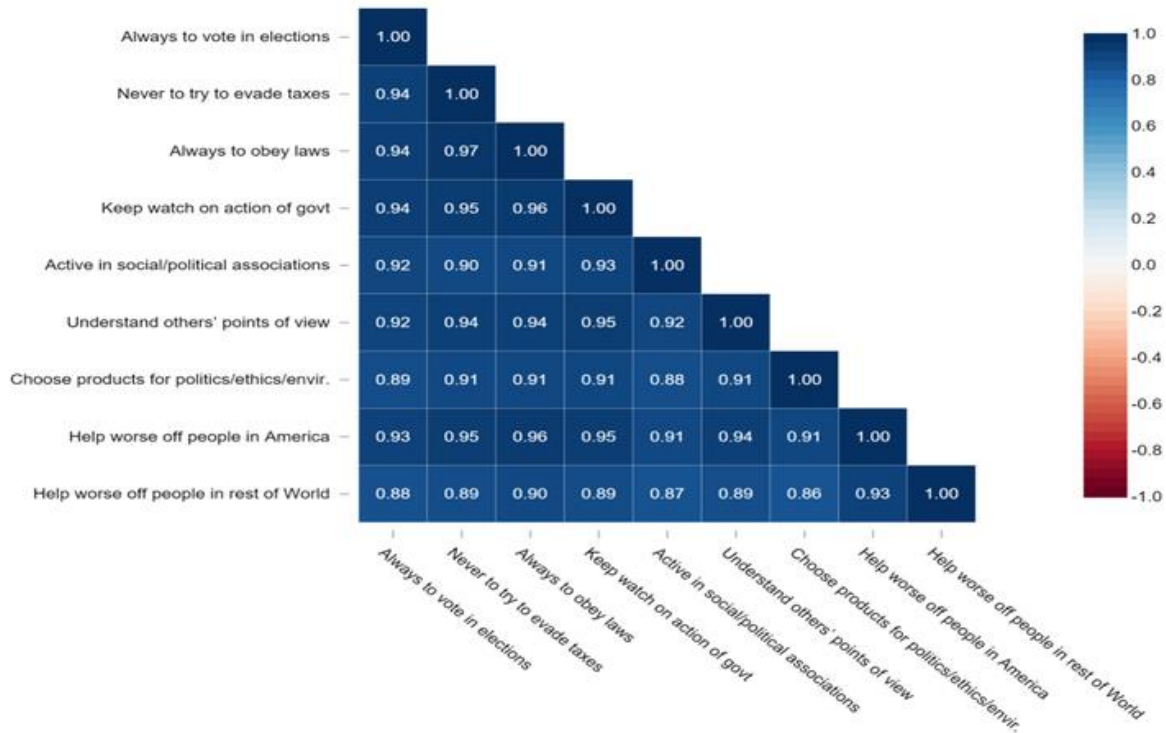
4. Correlation removing

WHY

모델 오버피팅 방지

회귀의 경우 필수(다중공선성은 가정 위반)

0.9넘으면 피쳐 제거 추천



+ Feature split

예

- ✓ 이름과 성을 따로 분리
- ✓ 'Toy Story (1995)'라면 'Toy Story' 와 '1995'를 분리.
- ✓ 주간의 이름(1주, 2주, 등) 주말인지 아닌지, 공휴일인지 아닌지.

+ 도메인 지식

예

금융데이터라면 갑자기 주식 price가 -뜨면 무조건 잘못된 데이터라고 판단가능

과제

본인 깃헙에
이름_FE_과제.ipynb를 올려주세요!
내용은 캐글/데이컨에서 아무 데이터셋이나 찾아서
제가 알려드린 걸 적용시켜서 모델 집어넣기전까지 만들어보세요

진짜 마지막으로 제일 중요한 것..

YBIGTA 일원이 되신 거 축하합니다
제 이름은 유동준입니다
만나서 반가워요 ㅎㅎ

열심히 들어주셔서 감사합니다
실습하러 ㄱㄱ