머신러닝 스터디 Week 1

Target Encoding

대표적인 범주형 변수 encoding 방법

One-hot encoding	Label Encoding
- 특정 범주를 갖고 있으면 1, 그렇지 않으면 0으로 표현하는 Encoding 방법 - Cardinality가 큰 경우 데이터의 dimension이 지나치게 커지는 현상 발생(Curse of dimension)	- 각 범주를 0, 1, 2, .. 같이 정수로 encoding하는 방법 - One-hot encoding과 달리 cardinality가 커져도 dimension이 커지지 않음 - 그러나 0, 1, 2와 같이 코딩할 시 숫자로 인식하여 특정 범주의 값을 더 크게 인식하는 문제점 발생

Target Encoding

: any kind of encoding that replaces a feature's categories with some number derived from the target

1) Mean Encoding

: 각 sex에서 target의 mean을 구하는 방법

	Sex	Sex_mean
0	male	0.188908
1	female	0.742038
2	female	0.742038
3	female	0.742038
4	male	0.188908

문제점

- Overfitting 발생 가능: 본래 train set에는 target에 대한 정보가 들어가면 안되는데, target의 mean을 변수로 넣을 경우 target에 대한 정보를 포함하게 되므로 overfitting이 발생할 수 있다.
- 데이터에 대한 대표성 문제: 만약 클래스 비율이 1:1:1 비율로 일정하다면, mean encoding 한 결과는 target별로 같다. 또한 특정 클래스의 개수가 매우 적은 경우 encoding을 한 의미가 없어지며, 개별 데이터를 대표할 수 없는 문제가 발생한다.

Target Encoding

2) Smoothing

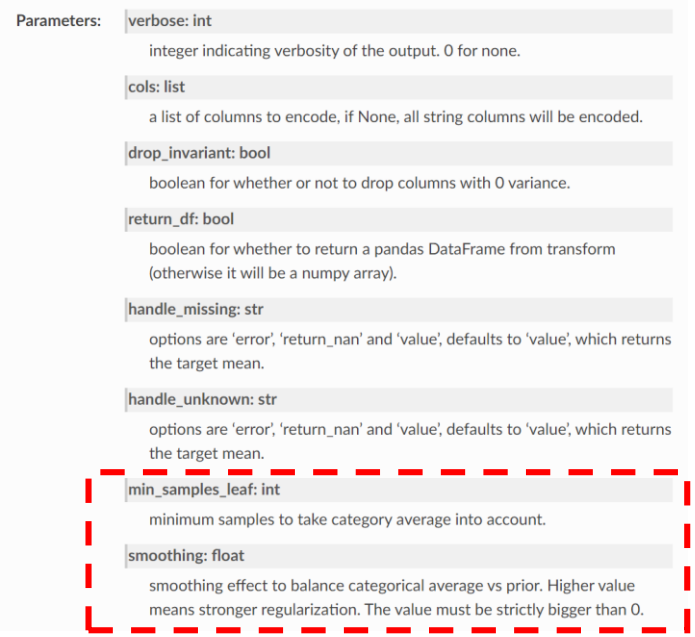
앞에서 제시한 Mean Encoding 시 대표성(representative)문제를 해결하는 방법

In-category 평균과 overall category 평균을 적절히 섞는다

Result = (전체 target 평균) * (1-smoothing) + (각 카테고리에 대한 target 평균) * smoothing

- Smoothing이 0에 가까울수록 overall mean에 가깝도록 조정한다.

→ Python에서는 target_encoder로 구현 가능



Parameters:

verbose: int
integer indicating verbosity of the output. 0 for none.

cols: list
a list of columns to encode, if None, all string columns will be encoded.

drop_invariant: bool
boolean for whether or not to drop columns with 0 variance.

return_df: bool
boolean for whether to return a pandas DataFrame from transform (otherwise it will be a numpy array).

handle_missing: str
options are 'error', 'return_nan' and 'value', defaults to 'value', which returns the target mean.

handle_unknown: str
options are 'error', 'return_nan' and 'value', defaults to 'value', which returns the target mean.

min_samples_leaf: int
minimum samples to take category average into account.

smoothing: float
smoothing effect to balance categorical average vs prior. Higher value means stronger regularization. The value must be strictly bigger than 0.

$$\lambda(n) = \frac{1}{1+e^{\frac{-(n-k)}{f}}}$$

Stacking Ensemble with CV

- 일반적은 Stacking ensemble 기법을 사용할 경우, overfitting 위험이 있기 때문에 Cross validation과 결합하여 사용한다
- Stacking과 같이 각 개별 모델이 예측한 데이터를 다시 학습하고 예측하는 방식을 **메타 모델**이라 부른다