3. Mean\_Enconding의 Overfitting 방지하기 위한 방법은 무엇이 있을까?

또한 강의안에서 제시한 카테고리 변수 enconding들 이외에 다른 변환방법들이 무엇이 있을까 조사하고 정리해보기

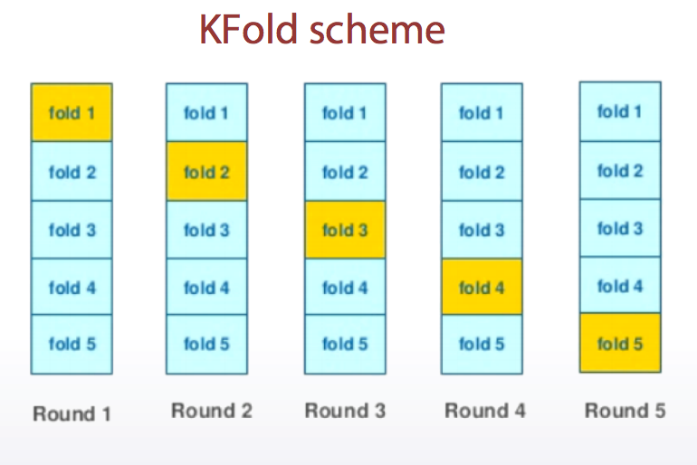
Mean Encoding에 대해서 좀 더 자세한 정보를 획득하였다. [[1]](#endnote-1)

Overfitting이 발생하는 이유는 encoding된 값에 target에 대한 정보가 포함되어있다는 것 이외에도 다르면서 비슷한 이유가 있는데, 바로 label의 대표값을 trainset의 하나의 mean으로만 사용한다는 점이다.

예를 들어서 trainset에서는 국가라는 label에 USA가 50개, Europe이 5개 존재하는데에 반해서, testset에서는 각각 20개, 20개씩 존재한다고 한다면, trainset에서 Europe의 경우 평균을 /5를 통해서 구하는데 이것이 testset에서의 20개를 반영할 것인가에 대한 의문점이 생기는 것이다.

이에 따른 해결책이 각각 존재한다.

1. CV (Cross Validation) Loop

* Trainset 안에서 Cross validation을 통해서 encoding의 값의 다양화를 얻어내는 것이다.
* 예를 들어서 다음과 같은 그림이 있다고 하자.
* [[2]](#endnote-2)
* Round 1을 기준으로 fold 1을

1. <https://www.kaggle.com/vprokopev/mean-likelihood-encodings-a-comprehensive-study> [↑](#endnote-ref-1)
2. <https://zzsza.github.io/data/2018/09/08/feature-engineering/> [↑](#endnote-ref-2)