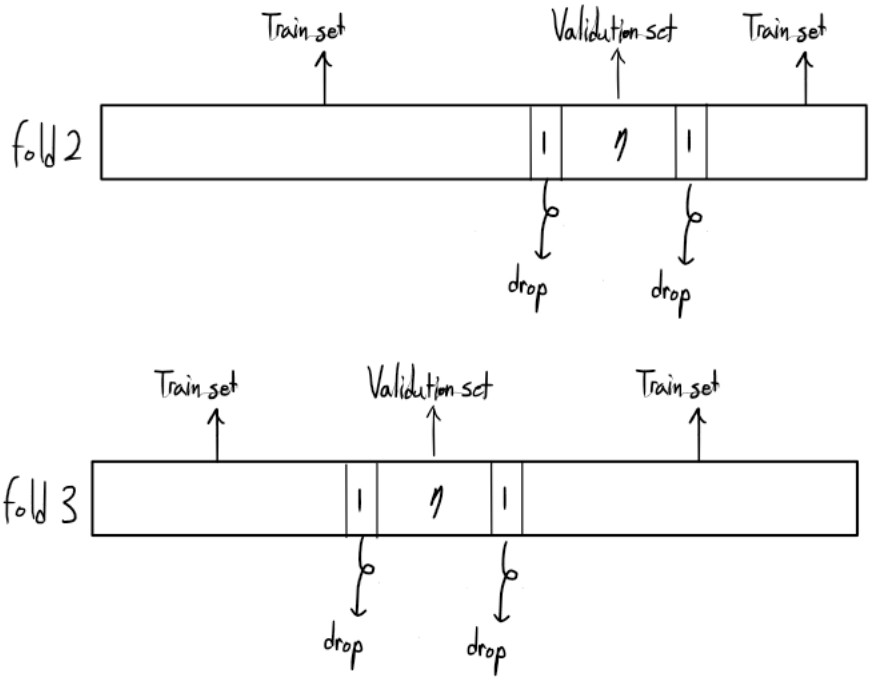
CV strategy

이번 대회는 연간 데이터가 아닌 부분적인 월간 데이터였기 때문에 일반화에 더욱더 신경을 써야 했던 대회였습니다. 따라서, 이번 대회의 핵심은 안정적인 CV전략과 일반화 가능한 변수를 구상하는 것이었습니다.

대회를 시작하고 안정적인 CV전략을 세우는 데만 2~3주가 걸렸던 것 같습니다.

처음에는 15fold에서도 불안정한 모습을 보였습니다. 그래서 50fold까지 생각해보았으나, 훈련시간이 많이 소요되므로 훈련시간과 CV - Public Score correlation 간 적절히 trade-off하는 것이 중요했습니다. 또한 Private set은 7일이므로 이 점 또한 고려해야 했습니다.

7일씩 겹치지 않고 앞, 뒤 1일 간격으로 Validation set을 설정하였고, 건물에 따라 다르지만 위의 전략으로 대부분 11fold로 구성할 수 있었습니다.



또다른 문제는 Target Statistic와 범주형 변수입니다.

데이터셋이 작고 아래에서 언급되었던 문제(power_log1p_stdd_mean이 전체기간의 정보를 담고 있음)로 인해 빠르게 overfitting되었습니다. 또한 특정 범주형 변수(month, university_vacation, school_vacation)와 TS를 함께 사용하는 경우 CV - Public Score correlation이 급격하게 하락하였습니다. 이 원인으로 TS에서 해당 범주형 변수의 정보를 일부 포함하고 있기 때문이라고 생각했습니다. 따라서 **범주형 변수의 정보를 TS에 포함시키면서 그와 동질적인 범주형 변수들을 제거하여 CV – Public Score correlation을 높이는 전략**이 요구되었습니다.

결과적으로 여타 범주형 변수들(month, university_vacation, school_vacation, dhc, weekend)를 제거하고, 1개의 범주형 변수(whc)만을 사용하였습니다. (One-Hot Encoding을 통해 5개로 변환) hour 변수의 경우 Cyclical Encoding을 통해 수치형 변수로 변환하였습니다.

	dhc	weekend	whc
0	Monday	Workday	Workday

1	Tuesday	Weekend1	Weekend1
2	Wednesday	Weekend2	Weekend2
3	Thursday		Holiday
4	Friday		Closed
5	Saturday		-
6	Sunday		-
7	Holiday		-
8	Closed		-

대회 마감 2주 전까지는 TS를 제외하고 필수적인 범주형 변수와 나머지 기본변수들로 실험하였습니다. 기온, 강수량, 습도, 풍속, hour, dhc로만 실험하였고, 만족스러운 결과를 얻은 후에 TS와 범주형 변수들을 추가하며 실험하였습니다. 모든 변수는 CV와 Public Score 모두 최소 1% 이상 개선을 보여야만 변수로 사용하였습니다. 동시에 -1% ~ 1% 내에서 변수를 제거할 수 있다면, 혹은 범주형 변수의 차원을 줄일 수 있다면 그렇게 하였습니다.

		dhc + weekend	whc
One-Hot Encoding 적용시 차원의 수		12	**5**
전체 CV 평균		4.772	**4.759**
submission_1	Public Score	5.295	**5.270**
	Private Score	**6.093**	6.105
submission_2	Public Score	5.344	**5.279**
	Private Score	6.114	**6.109**