**[ 팀 회의 1 ]** *20200123 / 고영빈, 박소희, 박찬종, 송이레, 최대건*

## 내용 요약

1. 공모전 문제 설정
2. 전처리 원칙 설정
3. 향후 작업 방향 논의
   1. 소희) 병렬적으로 작업 돌려야 하지 않는가
   2. 영빈) model validation
4. Dacon 글 업데이트 내용 확인

## 문제 설정

“ 각각의 train data(id 0~827)와 상태 값을 match하여 학습한다. Test data set에 넣어서 각 test data가 가질 수 있는 모든 상태(label 0~197)에 대한 확률을 구한다.”

* 제출해야 하는 파일 : test data별로 각각의 label이 될 확률이 몇인지 나타낸 파일.

## 전처리 원칙 설정

1. Unique값 날리기 : 상태 변화가 없는 column이라고 판단.

* unique한 열 번호, 그 열의 값 구하기
* unique하다고 도출된 각각의 열 번호에 대해, 그 열이 가지는 unique한 값이 part1 ~ part4에 대해 모두 동일하면 그 행은 날리기.

1. Column 간 상관관계, 시계열 추이 구하는 것은 진행하지 않음.

* feature 개수가 너무 많아서 노동집약적이고, 의미를 찾아내기 어렵다고 판단.

## 향후 작업 방향 논의

1. 소희 안건 : 이전 단계 작업이 모두 완료되기를 기다리지 말고, 다른 작업을 함께 진행할 필요가 있다.

* 동의. 이후 각자 시간 되는 사람 여러 패키지 내에 각 모델 동시에 돌리고, 그 결과를 참고하며 전처리 방식도 바꾸어 나갈 필요성은 있다. 최대한 많은 방식으로 분석을 시도해보고, 많은 모델을 돌려봐야 한다.
* ***추후 논의 필요한 부분) 그렇다면 누가, 어떤 패키지를 맡아서, 어떻게 작업할 것인가?***

1. 영빈 안건 : 우리 모델에 대한 validation 어떻게 진행하는가?

* Train data set 최대한 여러 방식으로 나눠서 shuffle validation 진행한다.
* 일단 모델을 돌리지 않았으니 나중에 다시 자세히 이야기하자.

## Dacon 글 업데이트 내용

1. 채점 방식 : log loss
2. Baseline 코드 업데이트