230501 시계열팀 주분 회의; 모델링 역할분담

지금까지 한 것

- 데이터 수집 및 전처리 (기사 감성분석, 자료형 통일, 결측치 보간, 데이터 병합)
- EDA (X변수 상관분석, X&Y 상관분석)
- 추가 X변수 수집 시도 but 반영 X ㅠ...
- 라벨링 Y변수 생성 (1일치 등락률 3% / 3일치 등락률 5%)
- 변수선택 (PCA, FA, 상관계수)
- LSTM 분류 모델 코드 완

Issue

- 백교수님과의 메일을 통한 피드백... '정답은 없으니 파라미터 튜닝을 여러 번 해보시고 시행착오를 통해 적절한 모델을 찾으세요'가 답변의 요지...
- Lstm 돌아가는데 정확도는 한 80% 나오지만 유지(1)로만 거의 때려맞춤
- 0(매수), 1(유지), 2(매도) 중에서 예측 값에 0, 2가 거의 없고... 있어도 오답임
- 클래스 불균형 + 클래스 간 유의미한 차이를 x변수로부터 찾기 힘든 듯
- 변수선택 set 7개 (fa, pca, 상관계수 높으면 제거한 것) lstm으로 다 돌려봤는데, 성능 다 비슷 or 차원축소한 set이 더 낮았음
- x,y 관계로 선제적으로 변수선택한 게 별로였을 가능성도 고려해야 함... (선형관계만 본 건가?)
- 모델링하고 성능 평가할 때 꼭!!!! >>> **F1-score** <<< 로 평가할 것!!!
- y_pred 프린트해보고 1 말고도 0이나 2가 얼마나 있는지 꼭 보기

모델 후보

- LSTM
- CNN

- (Cnn + lstm)
- 패널모형
- SVM, Logistic reg
- XGB(+feature importance), 나이브베이즈

To do

- 변수선택 재시도 (tree 기반 feature importance, KS검정, kernelPCA, principle surface)
- ML 모델링하는 사람들은 windowing한 train set 만들어야 함
- 모델링 -> 성능 좋은 모델 만들기 ~~(제발)~~
- 2주차 주분 ppt 제작 (+여러분은 패키지도...)
- 공모전 제출용 아이디어 제안서
- 공모전 제출용 코드 정리
- 공모전 제출용 ppt (주제 소개, 데이터 소개, 변수선택, 모델링, 기대효과)