230411 시계열팀 주분 회의

To do: 일정 정리, 종목 확정, 수집 데이터 기간 결정, 의문점 공유

수집 데이터 기간?

토론방이 2017.06.07부터 데이터가 있음 (다른 데이터 기간은 아직 파악 x)

17.06.07 ~ 23.03.31 데이터를 사용하면 총 5년 9개월치 데이터

주가 모델링하기에 충분한 양인가???

종목 확정? 테마 당 1개

반도체: SK 하이닉스

자동차: 현대차, 기아

운송: CJ대한통운

식품: 농심, 오뚜기, CJ 제일제당

에너지: 현대에너지솔루션

은행: 신한지주

화장품: LG 생활건강, 아모레퍼시픽

항공: 대한항공, 아시아나항공

일반상점: 이마트, 롯데쇼핑, 신세계

주분 일정? = 2주만 빡세게 몰빵하자!!!!!!

4/10 (월): 공모전 시작

4/21 (금): 중간고사 기간 끝 (그런데 저는 4/23에 시험이 하나 있어요.....)

~ 4/24 (월): 데이터 수집 완료, 공모전 팀 구성

~ 4/29 (토): 결측치 처리, 각 변수별 EDA, 파생변수 생성, 주가와의 상관분석, 기사 감성분석

~ 4/30 (일): EDA 결과 공유, 하락장 정의 및 라벨링

~ 5/4 (목): 변수선택, 모델링 (좋은 방법론으로 모델 결정하기)

5/5 (금): 주분 2주차 발표

~ 5/7 (일): 확정된 모델로 2023.04월 주가 예측 및 하락장 포인트 예측, 실제 주가와 비교 + 예측 주가 및 실제 주가와 예측한 하락장 포인트가 일치하는지 확인

~ 5/10 (수): (빠를수록 좋음) 보고서 작성 및 제출

~ 5/15 (월): 따봉 홍보

의문점?

- 일단 생각한 건 23년 3월까지의 데이터만 사용하고 4월 한 달에 대해 예측한 다음, 실제 4월 주가와 비교... 그렇다면 4월 데이터도 추가로 수집해야 하긴 함
- 일 단위 예측이라면, X 값으로 오늘의 데이터들을 입력했을 때, Y값으로 내일의 주가가 나와야 함 (왜냐면 X로 오늘 데이터를 입력하고 Y로 오늘 값을 예측하는 건 무의미, 이미 알고 있는 주가를 예측할 필요는 없음)
- 미래 한 달(4월)의 추이를 예측한다고 하면, 3월까지의 x값 만으로 4월 전체를 예측하는 게 과연 유의미한가? 과거의 x값을 사용해서 t 기간에 대해서 예측을 할 때, 어느 정도 기간까지의 예측이 의미있는지도 알아봐야 함
- 그렇다고 딱 내일의 값만 예측하기에는 예측할 수 있는 범위가 너무 짧지 않나....
- ML 모델은 해당 row의 X변수 값을 바탕으로 y의 값을 예측하는 것인데, 과거 변화 추이의 영향을 담으려면 현재 row 말고도 과거 row도 고려? 아니면 x변수 자체에 t-1 ~ t-7시점과 같이 여러 시점의 과거 값을 반영? 주가의 과거 값뿐만 아니라, x 변수의 과거 값도 고려해야 하는 거 아닌가... 그렇다면 이전 학습 결과에 영향을 받는 RNN 기반 모델을 써야 하나.....
- 아니면 주가나 하락장을 예측하지 말고 그냥 오늘까지의 데이터를 기반으로 내일 상승할지 하락할지를 예측? 그런데 주식은 선반영이라며.....

- 그렇다면 선반영 말고 후반영을 잡을 수 있는 피처 위주로 진행? 근데 후반영이라는 거 자체가 있기는 한거야...??

- 강건한 모델이라면 해당 테마에 대해서는 동일한 모델을 다 적용할 수 있어야 하나? 즉, 동일한 테마에 속한 종목인 경우 개별지표가 주가 관련 데이터(주가, 시가총액, 투자자별 지분율) 빼고 모두 동일한지 = 개별지표 중에서 기사, 네이버토론방, 검색어 관련 데이터를 테마별로 통일할 것인지 or 종목마다 다른 데이터를 사용할 것인지

- 위에서 테마별로 기사, 토론방, 검색어 데이터를 통일한다면, 그게 유의미한지는 동일한 테마 내의 다른 종목을 예측해 봤을 때도 결과가 잘 나오는지를 확인해 봐야 함 (그러니까 현차에 대해 fitting한 모델에서 주가 관련 데이터만 바꿔서 기아로 돌렸을 때도 모델이 괜찮아야 함)