**데이터 소개 및 분석목적**

분석목적은 21.11.29 부터 동년 12.3일까지의 주식 종가를 11/28일 까지의 데이터를 사용하여 예측하는 것입니다.

평가지표는 mae를 정규화 시킨 nmae로 주식의 종가와 같이 단위가 다른 데이터들의 단위를 맞춰주어 평가를 용이하게 해줍니다.

다음으로 데이터는 csv파일로 2가지, 파이썬 모듈을 통한 데이터로 총 3가지가 주어졌습니다. stock\_list는 2021년 6월 기준 kospi-200내 주식과 kosdaq-150내 주식들로 총 350개의 종목명과 종목코드, 상장시장으로 구성되어 있습니다.

Sample submission은 제출양식으로 350종목의 요일자별 가격으로 구성되어있습니다.

Sample은 파이썬 내 finance data reader package 를 이용하여 원하는 주식시장의 일자 별 시가, 종가,고가,저가 등 다양한 데이터를 불러올 수 있습니다. 저희는 1.4일부터 11.28 약 1년간의 데이터를 가져왔습니다.

**EDA 및 전처리**

한국시장에서는 주가가 상한가와 하한가로 일자 별 등락폭이 정해져 있기때문에 주가의 전일대비 큰 변동에 대해서는 따로 이상치로 처리하지 않았습니다. 21년도에 상장된 위 3가지 주식은 데이터가 적으므로 제외하였으며, 데이콘 내에서 처리한 것과 마찬가지로 기간 내 기존 주식에 병합되거나 상장 폐지된 아래 3가지 주식도 제거하고 진행하였습니다.

먼저 일자별로 나뉜 종가를 주별로 합쳐 시각화를 해보았습니다. 그래프에서 보이듯 좀 더 완만 해졌으며 이는 주식 특성상 당일의 이슈와 같이 정량적 분석에서 고려하기 힘든 시장에 영향을 주는 요인들에 대해 덜 예민하게 반응하므로, 일 단위 예측과 함께 주 단위 예측도 병행하면 좋을 것 같다고 판단하였습니다.

다음으로는 finance data reader을 이용해 한국 시장 지표인 코스피, 코스닥과 함께, 미장이 한국시장에 어떠한 영향을 미치는지 확인하기 위해 동 기간 나스닥과 다우지수를 가져왔습니다. 1년치 시장의 종가를 히트맵으로 시각화 한 결과 높은 상관관계를 보이는 코스닥과는 다르게, 코스피는 미장 지수에 대해 낮은 상관관계를 가지는 것을 확인하였습니다. 주식 특성상 변동성이 많기 때문에 1년에서 10년치로 확장해본 결과, 코스피 또한 높은 상관관계를 볼 수 있었습니다. 이를 통해 단기간의 분석에서는 상장시장이 코스피인 종목은 미장 데이터를 사용하지 않고, 코스닥은 나스닥 지표를 새로운 변수로 추가하였습니다.

다음으로는 주식의 경향성을 확인하기위해 특정 종목의 종가를 가져와 시각화하였습니다. 평균, 분산등이 일정하지 않아 정상성을 만족하지 않으며 대체적인 하락 추세가 있지만, 상수항이 없는 임의 보행 모형이며 주기성은 크게 보이지 않는다고 판단하였습니다.

추가로 20개의 주식을 임의 추출하여 종가를 시각화 하였을 때, 위와 마찬가지로 뚜렷한 계절성과 주기성이 거의 없음을 확인하였습니다. 또한 주가 간 독립성이 없는 모습 또한 확인할 수 있어 주식들 간에도 상관관계가 존재한다고 판단하였습니다.

이는 주식마다 가진 특정 테마들의 영향이 있어 같은 분야의 주식 간에는 가격의 추세도 유사하지 않을까 라고 예상하였으며, 따라서 이 주식들에 대해 테마별로 정리해보았습니다.

먼저 세계 표준 기준 중 하나이자 네이버금융에서 테마지표로 활용되는 공공데이터인 WICS의 섹터 분류지표를 크롤링하였으며, 이를 기존 데이터에 결합하여 대분류와 소분류를 추가하였습니다.

위 대분류 중 it테마를 가져와 시각화 했을때 전체적으로 비슷한 경향성을 보이지만 몇몇은 다른 움직임을 보이는 걸 확인할 수 있습니다. 또한 ppt에는 누락되어 있으나, 소분류에서도 이런 움직임을 볼 수 있었습니다. 이를 통해 전체적으로는 비슷한 경향성을 보이지만, 외부지표에 국한하지 않고 같은 대분류 내에서도 각기 다른 움직임을 가지는 것들끼리 군집화를 통해 묶어보면 좋겠다고 판단하였습니다.

시계열데이터를 군집화 하기위해 tslearn패키지를 사용하여 클러스터링을 진행하였습니다. Scree plot을 확인한 결과 30개 이상의 종목을 보유한 대분류는 클러스터의 개수가 3개이면 적당하다고 판단하였으며 이를 통해 클러스터링을 진행하였습니다. 최종적으로 It종목의 3군집별로 임의의5개 주식을 추출하여 시각화 한 결과, 비슷한 흐름이 보이는 종목들끼리 잘 군집화 된 것을 확인할 수 있습니다.

**랜덤포레스트**

다음 모델은 랜덤포레스트입니다. 먼저 간단히 모델소개를 하겠습니다. 복원추출인 부트스트랩을 반복하는 배깅에 모델의 변수인 feature를 임의로 선정하여 결정트리를 생성하는 앙상블 기법 입니다.

앞서 진행한 두 모델과 다르게 일별 종가를 예측하지 않고 예측할 11월 마지막 주의 평균 종가를 예측하여 월화수목금에 동일하게 넣었습니다. 이는 앞서 전처리에서 얘기했듯이 이슈와 같은 고려되지 않은 노이즈의 영향을 최소화하기 위함입니다. 새로운 변수로는 예측일 기준 앞으로 5일간 평균 종가인 after5days, 예측일 이전 5일의 평균수익률인 last5days, 예측일 이전 5일의 평균 종가인 5avg를 추가했으며, 이외에 종가, 나스닥, 코스피, 코스닥, 클러스터들을 더미변수로 처리한 변수들을 추가하였습니다. 총 73402개의 데이터와 26개의 변수로 구성하였습니다.

Grid search로 교차검증과 함께 최적파라미터를 선정하였습니다. 다른 파라미터들은 기본값으로 두고, 과적합 방지를 위해 최대변수개수를 줄이고 생성트리개수를 늘리는 방향으로 접근하였습니다. Cv=5로 훈련데이터를 총 5번의 평가가 이루어지게 나누었습니다. 하이퍼파라미터의 영향력을 파악하기위해 기본모델과 임의로 max\_features를 24로 설정한 모델과 비교하였습니다. Validation 과 최종 test점수인 private간 차이가 크지 않으며 트리개수를 제일 크게 한 모델에서 가장 좋은 점수인 3.6이 나왔습니다.

최적화 모델로 돌린 평가점수는 오히려 위 두 모델에 비해 val 과 private 둘 다 높은 점수를 받았습니다. 이는 현재 훈련데이터에 주식들이 내림차순으로 일자별 정렬되어 cv를 나눌때 훈련데이터에 포함되지 않은 데이터가 생겨서 발생되는 문제라고 판단하였습니다. 따라서 validation set을 임의로 추출하는 방식으로 변경하여 진행한 결과 위와같이 점수가 많이 개선된것을 확인 할 수 있었습니다. 하지만 위 default 24모델과 비교했을때 val점수와 private 점수의 차이가 커진 것으로 보아 과적합이 발생된것으로 판단하였습니다. 최적화 모델과 24모델간의 변수중요도를 확인해보았을때, 파라미터의 개수에 따라 변수중요도가 달라짐을 확인할 수 있습니다. 모델을 개선하기위해 더미변수를 제외하고 중요도가 제일낮은 kospi와 kosdaq을 제거하기 위해 상장시장별 모델로 분리하여 다시 모델링을 진행하였습니다. 코스피모델에서는 코스닥과 앞서 확인했듯이 미장지표를 제외하였습니다. 기본모델과 최적화모델을 비교한 결과 기본모델에서 더 좋은 점수를 보여줬으며 코스닥 모델또한 같은 모습을 보여줬습니다. 모델별 종목수에 따른 평균 모델점수는 기본모델에서 3.86으로 앞의 임의모델보다 0.26점 높습니다. Kospi 모델의 변수중요도를 시각화하여 확인해보면 역시나 피처 개수에따라 변수중요도가 달라진것을 확인할 수 있습니다. Validation set에 대해 예측값과 실제값의 경향성을 시각화 했을때 큰 차이가 없음을 시각화로 확인하였습니다.