고급 통계적 머신러닝 최종과제

KOSPI50 일일 종가 예측모형 개발

2020150473 김지후

1. 연구목적

주식 투자에 대한 수요가 증가하고 있다. 한국갤럽이 지난 12~14일 전국 18세 이상 1천명에게 현재 펀드를 제외한 주식 투자를 하는지 물은 결과 29%가 '하고 있다'고 답하였다. 투자자 비율은 지난해 8월 21%에서 8%포인트 증가하였다. 따라서 정확하고 신뢰할 수 있는 주가 예측의 필요성이 상승하고 있으며, 증권사에서도 AI를 활용한 주가예측을 부가서비스로 제시하고 있다. 주가는 다양한 요인에 영향을 받기 때문에 예측이 쉽지 않은 분야이다. 시간의 흐름을 잘 반영하고 외부 변수를 활용하여 예측의 정확도를 높이는 것이 중요하다.

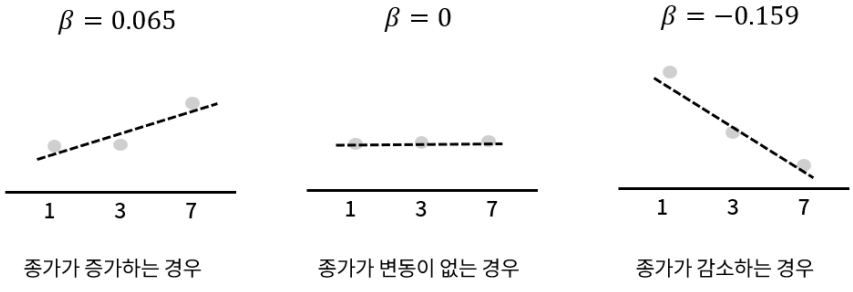
주가의 정확한 가격 예측이 아닌 주가의 상승과 하락을 예측하는 것을 모형의 목표이다. 코스피 200 구성종목 가운데 시가총액이 큰 상위 50개 종목으로 구성된 KOSPI50의 일일 종가를 예측하는 것이 목적이다.

2. 데이터

코스피 50 종목의 2020년 1월1일부터 2022년 11월30일까지의 데이터를 사용해 총 데이터는 33172개이다. 목적변수는 전날 종가 대비 상승 여부이다. 총 데이터 중에서 주가가 상승한 데이터는 15437개, 하락 또는 변화없는 데이터는 17834개로 심각한 데이터 불균형은 없었다.

특성변수는 과거의 주가정보를 사용하였다. 예측 하루 전날의 코스피 지수와 종목별 종가, 전날의 시가, 전날의 최고가와 최저가, 전날의 거래량을 변수로 사용해 가장 가까운 과거의 주가에 대한 정보를 반영하였다. 또 3일 전과 7일 전의 주가 변수를 사용해 1일, 3일, 7일의 시간의 흐름을 반영하고자 하였다.

1일, 3일, 7일전의 종가 및 거래량은 시간의 흐름에 따른 변동성은 나타내지 못한다. 변동성을 나타내기 위해 회귀계수의 개념을 이용했다.[1] 1일, 3일, 7일간의 종가가 하락하는지, 증가하는지, 변함없는지를 나타낼 수 있는 변동 기울기를 구할 수 있다.



[1] 백지혜. "시계열 패턴을 활용한 고객분류모형 개발에 대한 연구." 국내석사학위논문 숭실대학교 대학원, 2014. 서울

일일 시가와 종가의 차이 그리고 고가와 저가의 차이로 변화량을 나타냈다. 또한 달러당 원화 환율, 엔화당 원화환율, 비트코인 원화가격을 사용하였다. 총 30개의 특성변수를 사용했다.

1) 과거 주식 관련 변수

1일전 종가, 1일전 시가, 1일전 최저가, 1일전 최고가, 3일전 종가, 3일전 시가, 3일전 최저가, 3일전 최고가, 7일전 종가, 7일전 시가, 7일전 최저가, 7일전 최고가, 1일전 종가와 시가의 차이, 3일전 종가와 시가 차이, 7일전 종가와 시가 차이, 1일전 최고가와 최저가 차이, 3일전 최고가와 최저가 차이, 종가 변동기울기, 거래량 변동기울기, 1일전 코스피 지수

2) 그 외 변수

달러당 원화 환율, 엔화당 원화 환율, 비트코인 원화가격

3. 분석 방법론

8개의 초모수를 조정하지 않은 기본 모형의 학습데이터에서 교차검증 결과 성능은 다음과 같다.

모형	성능
LogisticRegression	0.5341
DecisionTreeClassifier	0.6011
RandomForestClassifier	0.6150
XGBClassifier	0.6160
XGBClassifier(booster='gblinear')	0.5357
LGBMClassifier	0.6569
LGBMClassifier(boosting_type='goss')	0.6081
CatBoostClassifier	0.6578

이때 0.61 이상의 높은 성능을 보여준 RandomForest와 XGBoost, LightGBM, CatBoost을 선택하였다. 75:25의 비율로 학습데이터와 시험데이터를 분리하여 학습하고 성능을 확인하였다. 학습데이터의 수는 24879개, 시험데이터의 수는 8293개이다. GridSearchCV를 사용해 각 초모수의 범위를 좁히고 RandomizedSearchCV를 사용해 최종 최적의 초모수를 찾았다. 다음은 분류모형 별 초모수와 성능이다.

1) RandomForest

```
Best params: {'min_samples_split': 7, 'max_features': 0.7,
'max_depth': 8}
```

```
Training score: 0.622
Test score: 0.619
```

2) XGBoost

```
Best params: {'subsample': 0.8, 'min_child_weight': 5, 'max_depth':
6, 'learning_rate': 0.4, 'colsample_bytree': 0.5}
Training score: 0.628
Test score: 0.633
```

3) LightGBM

```
Best params: {'subsample': 0.2, 'n_estimators': 300,
'min_child_weight': 6, 'max_depth': 7, 'learning_rate': 0.06,
'colsample_bytree': 0.7}
Training score: 0.660
Test score: 0.669
```

4) CatBoost

```
Best params: {'subsample': 0.7, 'min_child_samples': 1,
'max_depth': 8, 'learning_rate': 0.07, 'colsample_bylevel': 1}
Training score: 0.639
Test score: 0.651
```

4개의 모형 모두 과대적합의 문제가 나타나지 않았고 LightGBM이 test 성능 0.669로 가장 좋았다. 성능이 월등히 개선되지는 않았지만 모두 기본모형보다 성능이 향상이 되었다.

4. 최종 모형

초모수 조절한 최종 4가지 예측모형의 예측치를 특성변수의 입력하여 최종예측치를 출력하는 stacking 메타모형을 사용해 성능향상을 이루고자 하였다. 비교적 간단하고 로버스트한 로지스틱 회귀모형을 메타모형으로 사용하였다. 메타모형으로 적용한 결과 성능은 0.67이다.

1) Base model

```
LGBMClassifier(subsample=0.2, n_estimators= 300, min_child_weight=
6, max_depth=7, learning_rate= 0.06, colsample_bytree=0.7)
```

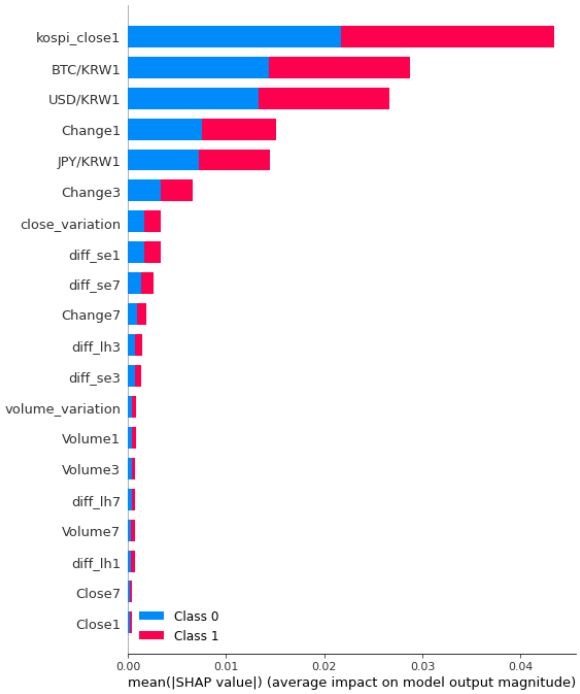
```
CatBoostClassifier(subsample=0.7, n_estimators= 8000, min_child_sam
ples= 1, max_depth=8, learning_rate= 0.07, colsample_bylevel=1)

XGBClassifier(subsample=0.8, n_estimators= 300, min_child_weight= 5
, max_depth=6,learning_rate= 0.4, colsample_bytree=0.5)

RandomForestClassifier(n_estimators= 300, min_samples_split= 7, max
_depth=8, max_features=0.7)
```

2) Base model

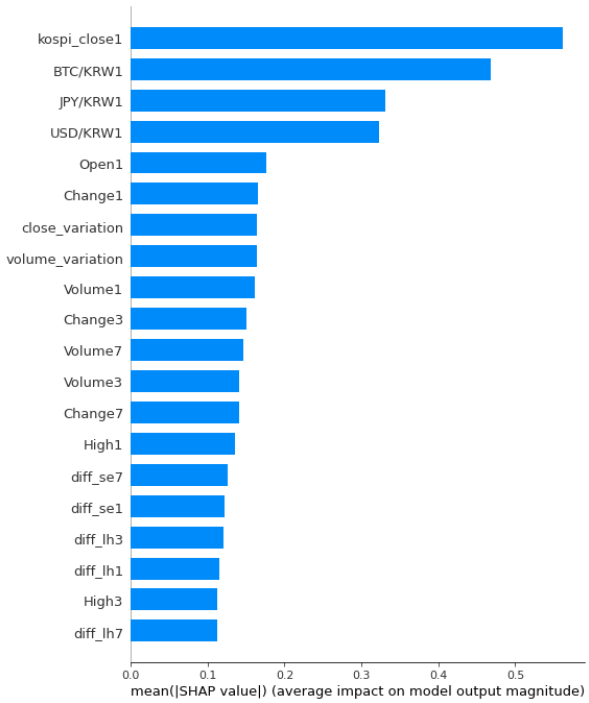
```
LogisticRegression()
```

5. 모형의 해석

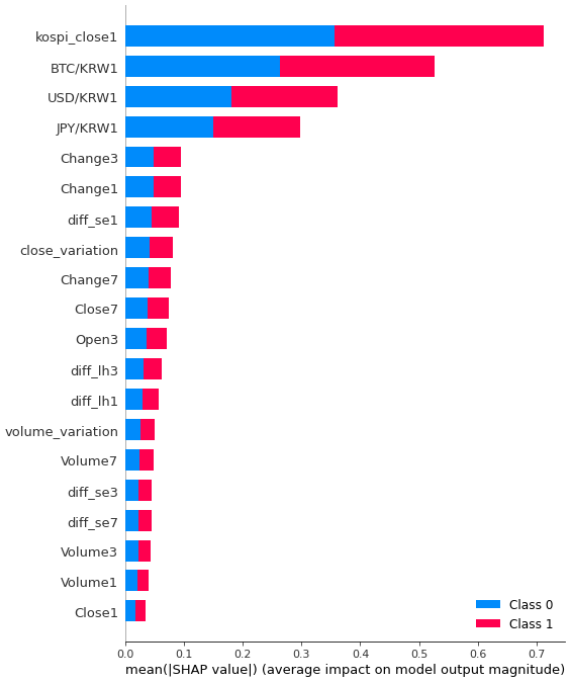
4가지 모형의 변수중요도는 다음과 같다. 예측 날짜의 1일전의 코스피 지수(kospi_close1)이 가장 중요한 것으로 나타났고, 이어서 주식에 영향을 미치는 환경에 대한 변수인 환율(USD/KRW1, JPY/KRW1), 비트코인(BTC/KRW1) 가격이 공통적으로 높게 나타났다. 또한 1일, 3일, 7일 간의 종가의 변동성(close_variation)과 1일전 거래량 등의 과거의 주식 정보가 예측에 주요한 역할을 했다.



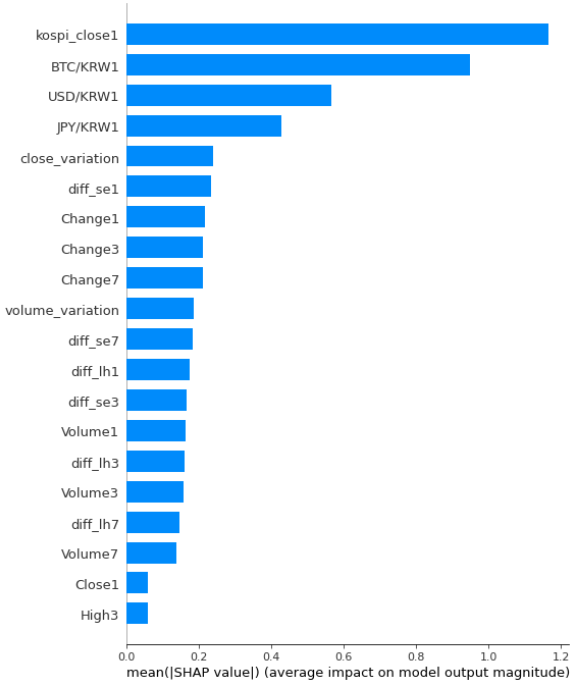
[RandomForest 변수 중요도]



[XGBoost 변수 중요도]



[LightGBM 변수 중요도]



[Catboost 변수 중요도]

6. 결론

Bagging과 Boosting 기반의 앙상블 머신러닝 모델을 활용한 메타모형을 사용하여 코스피 50 일 일 종가 예측 모델을 개발하였다. 최종 성능은 0.67이다. GridSearchCV와 RandomizedSearchCV를 활용해 최적의 초모수를 찾고 메타모형을 적용해 성능향상을 이루어 낼 수 있었다. 앞으로 더 성능향상을 위한 시도가 필요해 보인다.

다양한 과거 주식과 관련된 변수와 그 이외의 변수를 사용하여 주식가격에 영향을 미치는 요인을 고려하였다. 본 개발에서는 1일, 3일, 7일 전의 과거 주식에 대한 정보를 활용하였는데 다른 과거 시점을 추가로 변수로 반영할 수 있을 것이다. 변수 중요도를 나타낸 결과 주식에 영향을 미치는 주변 상황을 반영할 수 있는 환율이나 비트코인 가격이 중요하였다. 추가적으로 주변 상황을 반영하는 변수를 추가한다면 더 좋은 성능을 보일 것이라고 예상한다.

주가 예측에 대한 수요가 증가하고 있는 상황에 머신러닝을 활용한 주가 예측 모델을 개발하였다는 의의가 있다.