경제와 코로나 데이터를 통한 코로나19 동향 예측

DLP(Data Loss Prevention) 김민기

김영민

송은진

초 록

2020년부터 국내 코로나19 확진자 증가에 따라 백신, 사회적 거리두기 등을 통해 확진자 수가 점차 감소하는 듯 보였으나 특정 사건들에 의해 6차례 대유행을 겪었다. 현재 2023년에는 앞서 시행했던 마스크 착용, 사회적 거리두기 등을 해제하는 상황이다. 선행 연구에서는 Prophet 알고리즘을 통해 시계열 데이터의 추세로 미래를 예측하는 모델을 만들었다. 이 모델은 발병 초기 상황을 예측하지 못하는 단점을 가지고 있어 정부에서 발표한 코로나 확진자, 백신 접종자 등의 코로나 데이터와 코스피, 코스닥, 환율과 같은 경제 데이터를 통해 코로나 다음날 확진자를 회귀분석 모델로 예측하고 사이킷런의 결정계수, 평균절대오차, 평균 제곱근 편차 등으로 모델을 설명한다. 이러한 과정에서 사회적 거리두기 등을 통한 국가 통제로 사회 전반에 영향을 끼쳤는데 그중에 가장 국민이 체감할 수 있는 부분으로 경제가 있다. 코로나19를 통한 예측을 통해 미래의 질병에 대한 방역 대책에 따른 변화를 예측할 수 있도록 하고 방역에 대한 중요성을 강조한다.

목 차

Ⅰ. 서 론

1.1. 개발동기

대한민국은 2020년부터 국내 첫 감염사례를 시작으로 방역 수칙과 국가 통제를 속에서도 지속적으로 확진자가 증가했었다. 이에 따른 국가의 방역 대책을 실시했지만 초반 사회적 거리두기에 대한 반발과 백신접종에 대한 불신, 코로나 변종 바이러스로 위기가 있었다. 하지만 백신접종, 사회적 거리두기 등 국민의 노력을 통해 2023년 현재 코로나 확진자는 과거에 비해 많은 숫자를 보이지만 마스크 해제 등으로 상대적으로 약한 방역 수칙으로 생활할 수 있게 됐다.

미래에 또 다른 전염성 유행병이 발병하면 코로나19의 경험을 통해 국가적 방역 대책이 시행될 것이고 방역 대책이나 다양한 우발상황에 따른 확진자의 변화가 코로나19와 유사하게 진행될 것이라고 생각한다. 이러한 생각을 바탕으로 한국정보통신학회논문지 “시계열 데이터를 활용한 코로나19 동향 예측” 이라는 코로나 데이터를 시계열 데이터를 통한 확진자 예측 모델을 접했고 사이킷런의 모델들을 통해 설명력이 높은 예측 모델을 만들어 미래의 전염성 질병에 대응할 수 있도록 하고 현시점 느슨해진 방역으로 인한 안일하게 생각할 수 있는 코로나19를 강조한다.

1.2. 코로나19 정보

1.2.1 코로나19

중국에서 보고되기 시작한 코로나19는 2020년 1월 20일부터 우리나라에서도 지속적으로 보고되고 해외 국가 역시 같은 상황이다.

fig1. Covid_Graph

사람과 사람 사이 전파는 호흡기 분비물을 통해서 발생하며 사람의 기침이나 말을 할 때 음식물을 섭취할 때 나오는 호흡기 분비물에 포함된 바이러스가 나오고 타인의 점막에 닿으면 감염될 수 있다. 잠복기는 1~14일이고 코로나19 진단 기준은 검사기준에 따라 코로나19 감염에 대하여 양성이 뜬 자이며, 검사기준은 코로나19 유전자(PCR) 검출, 바이러스 분리된 경우이다. 주요 증상은 무증상자부터 발열, 기침, 인후통, 후각 및 미각 소실, 어지러움, 오한, 두통, 근육통, 호흡곤란 등의 다양한 증상이 나타난다.

1.2.2 코로나19 백신

2021년 2월 26일을 시작으로 국내에서 백신 접종이 시작됐으며 전 국민을 대상으로 실시했다. 초기 예방접종 순서는 백신 도입 및 공급, 접종 상황, 백신별 임상 결과 등을 고려하여 우선 접종했으며 코로나19 의료기관 종사자, 중증 및 사망 예방을 위한 노령자, 지역사회 전파 차단 등 연령과 직업에 따라 접종했다.

백신접종을 통해 인체의 면역 체계를 훈련시켜 면역을 획득시키기 위한 목적이고 이를 통해 예방접종은 코로나19 감염의 위험을 줄여주고 중증 환자 발생이나 사망을 예방한다.

국내 백신접종은 1, 2차 부스터샷, 동절기 예방접종 등으로 사회적 활동을 위해 3차 접종까지는 권고됐다.

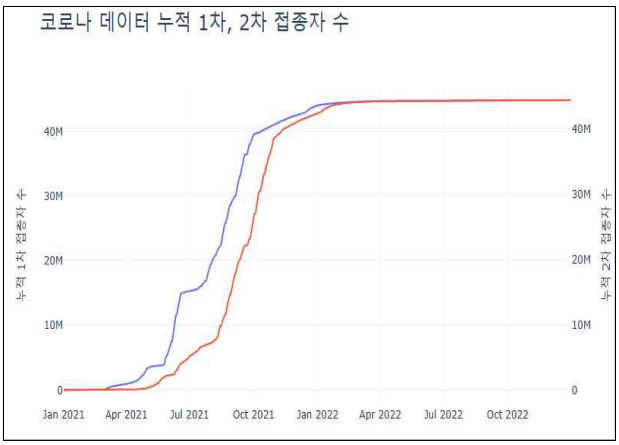

fig2. Covid_Vaccine_Graph

Ⅱ. 기존연구

2.1. Prophet 알고리즘을 통한 예측

1년 3개월의 시계열 데이터를 통해 Prophet의 알고리즘을 이용해 코로나19 확진자를 예측했다. 해당 논문에서 사용된 Prophet 모델은 주기적이지 않은 변화인 트렌드를 나타내는 Trand와 weekly/yearly 등 주기적으로 나타내는 패턴들을 포함하는 Seasonality, 특정 값이 비정상적으로 증가하거나 감소하면 휴일과 같이 불규칙한 이벤트로 설정하는 Holiday의 구성요소로 구성된다. 이러한 모델로 0.914446의 결정계수를 갖는다.

2.2. 코로나19 사태가 국내경제에 미치는 영향과 향후 과제

코로나19 사태로 국내경제가 크게 둔화되고 있으며 지역경제의 위기를 설명하고 있다. 코로나 초기에는 이동제한조치, 국경봉쇄 조치에 따라 소비감소, 노동시장 위축, 세계 무역 규모 축소 등 경제에 직접적인 영향을 미치는데 그에 반해 온라인 등을 통한 비대면 산업은 빠른 성장을 했다. 또한 코로나 확진자, 사회적 거리두기 변화, 긴급재난지원금 등에 따라서 국내경제와 지역 경제의 변화가 있음을 말하고 있다.

Ⅲ. 분석 알고리즘

시계열의 추세를 통해 예측하는 선행 연구와 다르게 코로나와 경제 데이터로 발병 초기부터 다음날 코로나 확진자를 예측하는 회귀분석 모델을 만든다. Decision Tree 모델로 예측하고 평가 방법은 결정계수와 MAE를 기준으로 한다. Tree 모델의 최적화를 위해 직관적인 수치를 보여주는 선형 모델인 OLS 모델을 사용한다.

3.1. Decision Tree

독립변수의 관측값을 모델에 입력해 목표변수를 분류하거나 예측하는 지도학습 기반의 방법론이다. 직관적인 해석이 용이하고 정규성, 독립성, 등분산성, 선형성에 자유로워 비모수적 모델이다. 또한 변수 간 상호작용을 고려하여 선형/비선형 관계 탐색이 가능하다.

단, 데이터 수가 적을 때 불안정하고 과적합 발생률이 높다.

3.1.1. Random Forest Regressor

Decision Tree의 단점인 overfitting 문제를 완화해주는 발전된 형태의 트리 모델로 랜덤으로 생성된 무수히 많은 트리를 이용하여 예측한다. Decision Tree 수가 많을수록 다수결에 의한 예측 결과의 품질이 높아진다.

3.1.2. Extra Tree Regressor

Random Forest Regressor와 매우 비슷하게 작동하고 기본적으로 100개의 Decision Tree를 훈련하고 Decision Tree가 제공하는 대부분의 매개 변수를 지원한다. 전체 특성 중 일부 특성을 랜덤하게 선택해 노드를 분할하는데 사용하고 노드를 분할할 때 무작위로 분할한다. 하나의 Decision Tree에서 특성을 무작위로 분할한다면 성능이 낮아지겠지만 많은 트리를 앙상블하기 때문에 과적합을 막고 검증 세트의 점수를 높이는 효과가 있다.

3.1.3. Cat Boost Regressor

대부분이 범주형변수로 이루어진 데이터세트에서 예측 성능이 우수하다. 범주형 변수를 One-Hot-Encoding, Label Encoding 등 Encoding 작업을 하지 않고도 모델의 input으로 사용할 수 있다. 다른 모델이 비해 overfitting이 적다.

3.1.4. Gradient Boosting Regressor

Gradient(또는 잔차)를 이용하여 이전 모형의 약점을 보완하는 새로운 모형을 순차적으로 적합한 뒤 이들을 선형 결합하여 얻어진 모형을 생성하는 지도학습 알고리즘이다. 약점은 이전 모형이 실제값을 정확하게 예측하지 못한 정도이다. 이러한 약점을 순차적으로 구한 뒤 결합한다. 단, 과적합 가능성이 있다.

3.1.5. XGBoost Regressor

앞서 설명한 Gradient Tree Boosting 알고리즘에 과적합 방지를 위한 기법이 추가된 지도학습 알고리즘이다. 과적합 방지가 되어 있어 예측 성능이 좋지만 작은 데이터에 대해서는 과적합 가능성이 있다.

3.1.6. LightGBM Regressor

leaf-wise로 Tree가 수평적으로 확장되는 알고리즘과 다르게 수직적으로 확장한다. 최대 손실 값을 가지는 리프 노드를 지속적으로 분할하면서 트리의 깊이가 깊어지고 비대칭적인 트리로 예측 오류 손실을 최소화할 수 있다. 학습 시간이 짧고 대용량 데이터 처리가 가능하다. 하지만 작은 데이터에 대해서는 과적합 가능성이 있다.

3.1.7. AdaBoost Regressor

초기 모형을 약한 모형으로 설정하며 매 스텝마다 가중치를 이용하여 이전 모형의 약점을 보완하는 새로운 모형을 순차적으로 적합한 뒤 최종적으로 이들을 선형 결합하여 얻어진 모형을 생성시키는 알고리즘이다. 과적합의 영향을 덜 받고 유연하지만 이상치에 민감하다.

3.2. 앙상블 기법

여러 개의 분류기를 생성하고 그 예측값을 결합함으로써보다 정확한 예측을 도출하는 기법을 의미한다. 강력한 하나의 모델을 사용하는 대신 약한 모델 여러 개를 조합하여 더 정확한 예측에 도움을 주는 방식이다.

3.2.1. 보팅(Voting)

여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정하는 방식으로 서로 다른 알고리즘을 여러 개 결합하여 사용한다.

- 하드 보팅(Hard Voting) : 다수의 분류기가 예측한 결과값을 최종결과로 선정
- 소프트 보팅(Soft Voting) : 모든 분류기가 예측한 레이블 값의 결정 확률 평균을 구한 뒤 가장 확률이 높은 레이블 값을 최종 결과로 선정

3.2.2. 배깅(Bagging)

데이터 샘플링(Bootstrap)을 통해 모델을 학습시키고 결과를 집계하는 방법이다.

- Bootstrap : 데이터가 조금씩은 편향되도록 샘플링하는 기법으로 분산이 높은 모델의 과적합 위험을 줄이는 효과가 있다.

모두 같은 유형의 알고리즘 기반의 분류기를 사용하고 데이터 분할 시 중복을 허용한다.

3.2.2. 부스팅(Boosting)

여러 개의 분류기가 순차적으로 학습을 수행하고 이전 분류기의 예측이 틀린 데이터에 대해서 올바르게 예측할 수 있도록 다음 분류기에 가중치(weight)를 부여하면서 학습과 예측을 진행한다. 일반적으로 부스팅 방식이 배깅에 비해 성능이 좋지만 속도가 느리고 과적합이 발생할 가능성이 더 높다.

3.3. OLS(Ordinary Least Squares) 모델

OLS는 최소제곱법으로 잔차의 제곱합을 최소화하는 가중치 벡터를 구하는 방식이다. OLS 모델을 통해 모델의 성능을 확인할 수 있다.(첨부자료 참고)

3.4. 평가 지표

3.4.1. 결정계수

상관계수와는 대조적으로 변수 간 서로 영향을 주는 정도이고 인과관계 정도를 나타낸 수치이다. 추정한 모형이 주어진 값이나 자료에 얼마나 적합한 정도의 값으로 0과 1 사이의 값이며 종속변수와 독립변수 사이에 상관관계가 높을수록 1에 가까워진다.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- SST(Total Sum of Squares) : 관측값에서 관측값의 평균을 뺀 결과의 총합
- SSE(Explained Sum of Squares) : 추정값에서 관측값의 평균을 뺀 결과의 총합
- SSR(Residual Sum of Squares) : 관측값에서 추정값을 뺀 값, 즉 잔차의 총합

3.4.2. MAE(Mean Absolute Error)

평균 절대 오차로 모델의 예측값과 실제값의 차를 모두 합한 후 절댓값을 취하고 평균을 낸다. 직관적으로 알 수 있는 지표지만 절댓값을 가지기 때문에 under performance인지 over performance인지 알 수 없다. 평균 백분율 오차를 통해 알 수 있다.

3.4.3. MPE(Mean Percentage Error)

평균 백분율 오차로 양수이면 under performance이고 음수이면 over performance이다.

3.4.4. MSE(Mean Squared Error)

평균 제곱차로 실제값과 예측값을 빼고 제곱해 평균화한다. 예측값과 실제값 차이의 면적의 합이고 지표 자체가 직관적이고 단순하다. 특이값이 존재하면 수치가 많이 늘어나 값이 작다면 에러율이 적다는 것을 의미한다.

3.4.5. RMSE(Root Mean Squared Error)

평균 제곱근 편차로 예측한 값과 실제 환경에서의 값의 차를 다룰 때 주로 사용하고 정밀도를 나타낸다. 또한 잔차들을 하나의 값으로 종합할 때 사용하며 MSE에 루트를 취해 산출한다.

값은 에러값이 클수록 가중치가 크게 반영되며 값이 작을수록 정밀도가 높다고 측정하며 이상치에 민감하며 해석이 쉬워진다.

3.4.6. RMSLE(Root Mean Squared Log Error)

RMSE에 로그를 적용해준 지표로 아웃라이어에 강건하고 상대적 Error를 측정해준다. 예측값이 실제값보다 클 때보다 예측값이 실제값보다 작을 때 더 큰 페널티를 부여한다.

3.4.7. 다중공선성(Multicollinearity)

목표변수와 2개 이상의 설명변수 간 선형관계를 분석하는 다중 회귀모델에서 설명변수 간의 강한 상관관계로 인해 회귀모델의 회귀계수에 대한 신뢰성이 떨어지는 현상이다.

Ⅳ. 실험결과 및 분석

4.1. 코로나19 데이터 회귀분석

4.1.1. 최소제곱법(OLS)

국내 코로나가 처음 발생한 2020년 1월 20일부터 2022년 12월 31일까지의 코로나 확진자, 사망자, 백신접종, 국가 통제(사회적 거리두기), 년, 월, 요일 컬럼을 통해 다음날의 확진자 수를 예측하는 모델을 생성했다. 백신 컬럼은 1 · 2 · 3차와 동절기 접종자를 첨부했고 국가 통제는 사회적 거리두기에 맞춰 0~5까지의 단계로 적용했다.

최초 베이스라인 모델의 경우 target에 0이 많아 정규분포 모형이 아니기 때문에 target에 log를 통해 정규분포 모형을 만들었다.

이후 모델에서는 Cond. No.가 100에 가까워 다중공선성이 의심돼 VIF가 10을 넘는 컬럼들을 제외했다.

Coefficient of determination	0.937
F-statistic	1575
Prob (F-statistic)	0.00
AIC	2413
BIC	2468

Table.1. Result OLS model value

R-squared는 0.937의 예측정도를 나타낸다. 1에 가까울수록 예측이 잘되었다는 것을 의미해 위 모델은 잘 예측되었음을 뜻한다. 초기 모델에서는 0.914로 최종모델보다 낮게 나왔지만 VIF가 높은 값을 제외해 다중공선성의 위험을 줄이고 R-squared를 올렸다.

4.1.2. Decision Tree

성능 지표 중 MAE를 통해 실제값과 예측값의 오차를 비교해서 MAE가 낮은 모델을 적용해서 모델을 생성했다. 모델 중 회귀분석에서 등분산성, 독립성, 선형성 등 수치에서 자유로운 Tree 모델이 MAE가 낮게 나와 적용했다. 기존 최소제곱법을 통해 적용한 모델에 MAE를 확인해 낮은 모델을 적용해 훈련했으며 성능 지표는 R-squered와 MAE를 기준으로 잡았다.

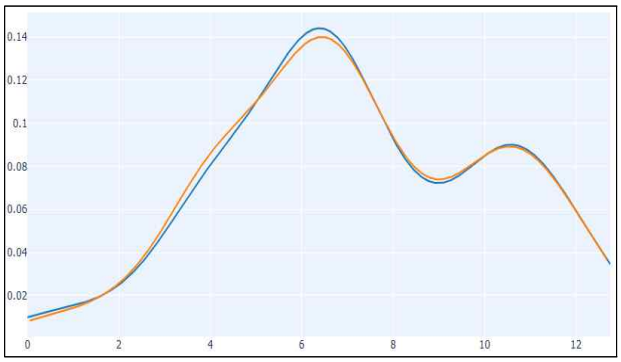

fig3. Covid_Catboost_Model_Graph

주황색 선은 예측값을 의미하고 파란색 선은 실제값을 의미한다. 평가지표인 **R-squared는 0.994, MAE는 2599**로 높은 성능을 보인다.

날 짜	실 제 값	예 측 값	차 이
2023-02-08	17927	17783	-144
2023-02-09	14662	15070	408
2023-02-10	13504	14071	567
2023-02-11	12805	12475	-330
2023-02-12	12051	11134	-917
2023-02-13	5174	5351	177
2023-02-14	14371	13975	-396

Table.2. predict value

4.2. 코로나19, 경제 데이터 회귀분석

앞서 실시한 분석 데이터인 코로나19와 기업의 성장을 확인할 수 있는 kospi, kosdaq 컬럼과 exchange_rate(환율), 소비자 물가지수, 실업률 컬럼을 통해 모델을 생성했다. 초기 모델 작성 시 코로나19 이전의 경제 동향을 확인하기 위해 2018년부터 데이터를 적용했으나 코로나 발생 이전까지 코로나19의 모든 컬럼이 0으로 입력해 모델 성능에 영향을 주어 결정계수가 무의미하게 높게 나온다. 그러므로 코로나19 국내 발생 일자에 따라 2020년 1월 20일부터 2022년 12월 31일까지 범위의 데이터를 사용했다.

4.2.1. 최소제곱법(OLS)

기존 코로나19와 경제 데이터를 새로 병합하여 사용했다. 코로나19 데이터와 동일한 초기 모델은 target값에 0이 많아 정규분포 모형이 아니다. 그래서 log를 통해 정규분포 모형으로 만들어 단위가 독립변수들은 scale하고 모델에 적용했다.

Coefficient of determination	0.940
F-statistic	714.1
Prob (F-statistic)	0.00
AIC	2384
BIC	2504

Table.3. log_scale OLS model value

코로나19만 적용된 모델보다는 성능이 향상된 것으로 나왔으나 Cond. No.가 148로 다중공선성이 의심된다. target값과 다른 독립변수들의 상관관계가 높은 컬럼을 제외한 모델을 만들고 p-value가 높은 컬럼을 제외해 최적화를 시도했다.

Coefficient of determination	0.938
F-statistic	1149
Prob (F-statistic)	0.00
AIC	2395
BIC	2470

Table.4. p-value OLS model value

R-squared는 0.938로 scale과 log를 적용한 값보다는 0.002정도 떨어졌으나 Cond. No.는 46.6으로 다중공선성 가능성을 낮췄다. 회귀분석은 독립변수 간 독립성이 있어야 해 다중공선성 낮은 모델을 최종모델로 선정했다.

4.2.2. Decision Tree

코로나 데이터만 사용한 모델과 동일하게 성능 지표 중 MAE를 통해 실제값과 예측값의 오차를 비교해서 MAE가 낮은 모델을 적용해서 모델을 생성했다. 상관계수를 확인해 0.3 이하 컬럼을 제외하고 모델을 생성했으나 MAE가 6000 이상만 나와 상관계수를 통한 최적화가 아닌 p-value에 따라 제외해 모델을 최적화했다.

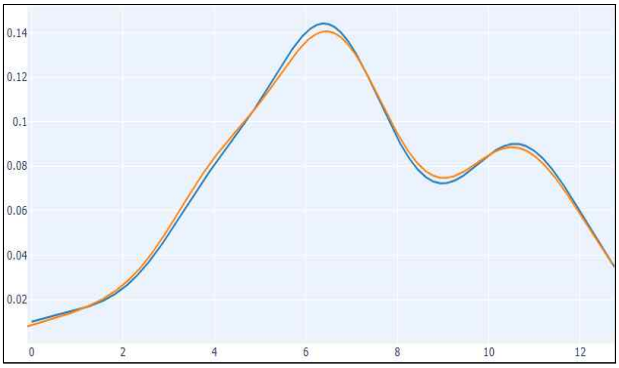

fig4. p-value_Catboost_Model_Graph

평가지표인 R-squared는 0.994, MAE는 2378.657로 높은 성능을 보인다.

Coefficient of determination	0.99342
MAE	2860.594

Table.5. Hyperparameter Catboost model value

Coefficient of determination	0.99311
MAE	3526.652

Table.5. Hyperparameter LightGBM model value

Optuna를 통해 Catboost와 lightgbm 모델의 하이퍼파라미터 튜닝을 통해 모델 최적화를 시도했으나 MAE가 증가해서 기존 모델을 적용했다.

Coefficient of determination	0.99407
MAE	2697.312

Table.6. Voting model value

마찬가지로 Catboost, Lightgbm, Extratree, Gradientboost, xgboost 모델을 다양하게 조합해 Voting으로 모델을 생성했으나 MAE값이 여전히 높게 나와 p-value를 통한 모델 최적화를 최종모델로 선정했다.

Ⅴ. 결 론

본 모델은 코로나19 데이터만 적용한 것이 아닌 경제 데이터를 적용해 코로나19 초기부터 다음날 코로나19 확진자를 예측한다. 모델에 대한 최적화 과정을 최소제곱법을 통한 Tree모델 성능 향상, 하이퍼파라미터, 앙상블 기법을 적용해 최적의 성능을 찾았고 평균절대오차(MAE)를 기준으로 모델을 평가했다.

코로나19와 같은 미래의 호흡기 질병으로 인한 전염성 유행병을 초기부터 확진자와 경제 데이터를 통해 예측하면서 방역에 대한 중요성을 강조할 수 있을 것으로 판단된다.

향후 연구로 year 컬럼을 1~6차 대유행, 변종 바이러스에 대한 내용을 반영하는 등 세분화하고 국가 통제를 초기 대응과 면역이 생긴 후의 대응을 반영해 분석하면 좀 더 정확한 분석이 가능하다고 판단된다.

그래프 설명

fig1. Covid_Graph : 코로나 누적 확진자, 사망자 수

fig2. Covid_Vaccine_Graph : 코로나 백신 1,2차 누적 접종자 수

fig3. Covid_Catboost_Model_Graph : 코로나 데이터 적용한 Catboost모델 최적화 결과

fig4. p-value_Catboost_Model_Graph : 최종 데이터 적용한 Catboost모델 최적화 결과

표 설명

Table.1. Result OLS model value : 코로나 데이터만 적용한 OLS모델 결과

Table.2. predict value : 코로나 데이터를 Catboost 모델로 예측한 실제, 예측값

Table.3. log_scale OLS model value : 코로나, 경제 데이터에 log, scale를 적용한 OLS 모델

Table.4. p-value OLS model value : 표3의 데이터를 p-value를 통해 최적화

Table.5. Hyperparameter Catboost model value : 최종 데이터 하이퍼파라미터 튜닝 적용 모델 결과

Table.6. Voting model value : Extratree, Catboost, Gradienboost 모델 voting 결과

참고문헌

김재호, 김장영 (2021). 시계열 데이터를 활용한 코로나19 동향 예측, 한국정보통신학회논문지, Vol.25

이수은 (2020). 자동 분류 성능 개선을 위한 다중 스태킹 앙상블 기법, 서울시립대학교 석사 논문.

조대형, 김정주 (2020). 코로나-19 사태가 국내경제에 미치는 영향과 향후 과제, 순천향대학교 석사 논문.

[첨부자료]

3.3. OLS(Ordinary Least Squares) 모델

3.3.1. R-squared

결정계수, 전체 데이터 중 해당 회귀 모델이 설명할 수 있는 데이터의 비율이다. 회기식의 설명력을 나타내고 성능을 나타낸다. 1에 가까울수록 성능이 좋다.

3.3.2. Adj. R-squared

도움이 되는지에 따라 조정된 결정계수

3.3.3. F-statistic

도출된 회귀식이 적절한지 볼 수 있다. 낮을수록 적절한 것으로 변수의 t값을 제곱한 값이다. 회귀계수가 큰 경우 분산의 차이도 커지므로 F-값도 크나 나오며 회귀계수가 크면 두 변수 간에 유의미하고 강한 인과관계가 있다는 것을 의미한다.

3.3.4. Prob(F-statistic)

F-통계량에 대한 p-value이며 F값이 0에서 얼마나 가까운지 확률적으로 측정한 값이다.

3.3.5. AIC(Akaike information criterion)

새롭게 얻을 데이터를 얼마나 잘 예측할 수 있는지를 바탕으로 모형의 적합도를 결정하는 지표이다. 모형의 파라미터 개수까지 고려하여 모형을 평가함으로써 과대적합이 없는 모형을 선택하는 지표이다. 표본의 개수와 모델의 복잡성으로 모델을 평가한다.

3.3.6. BIC(Bayesian information criterion)

AIC와 마찬가지로 새롭게 얻을 데이터를 잘 예측할 수 있는지를 바탕으로 적합도 결정하는 지표이다. AIC와 다른 점은 표본크기 n에 따라 달라지며 n이 클수록 파라미터의 개수 k의 페널티가 커진다. AIC 보다 모델 평가 성능이 더 좋다.

3.3.7. Omnibus

디아고스티노 검정이며 비대칭도와 첨도를 결합한 정규성 테스트로 값이 클수록 정규 분포를 따른다는 의미이다.

3.3.8. pred(Omnibus)

디아고스티노 검정이 유의한지 판단(0.05 이하일 경우 유의하다고 판단)

3.3.9 Skew(왜도)

평균 주위의 잔차들의 대칭하는지를 보이는 것이며 0에 가까울수록 대칭이다.

3.3.10 Kurtosis(첨도)

잔차들의 분포 모양이며 3에 가까울수록 정규분포이다.(음수이면 평평한 형태, 양수는 뾰족한 형태이다.)

[첨부자료]

3.3.11 Durbin-Watson

더빈-왓슨 정규성 검정으로 잔차의 독립성 여부를 판단한다. 1.5 ~ 2.5 사이일 때 잔차는 독립적이라 판단하며 0이나 4에 가까울수록 잔차들은 자기상관을 가지고 있다고 판단한다. 단, 설명변수에 종속변수의 과거값이 들어가면 유의미한 값을 볼 수 없다.

3.3.12 Jarque-Bera(JB)

자크베라 정규성 검정으로 값이 클수록 정규분포의 데이터를 사용했다는 것을 의미한다.

3.3.13 Cond. No.

다중공선성 검정으로 독립변수 간 상관관계가 있는지를 보는 것이며 10이상이면 다중공선성이 있다고 판단한다.

[첨부자료]

OLS Regression Results						
Dep. Variable	target		R-squared		결정계수	
Model	OLS		Adj. R-squared		조정된 결정계수	
Method	Least Squares		F-statistic		F 통계량	
Date:	날짜		Prob(F-statistic)		F 통계량에 대한 p-value	
Time	시간		Log-Likelihood			
No. Observations	총 표본의 수		AIC		표본의 개수와 모델 복잡성	
Df Residuals	잔차의 자유도		BIC		AIC와 유사함	
Df Model	독립변수의 개수					
Covariance Type	nonrobust					
	coef	std err	t	P>|t|	[0.025	0.975
	회귀계수	계수 추정치의 표준오차	t-test	p-value	회귀계수의 신뢰구간	
Intercept	-0.0003	0.001	-0.257	0.797	-0.003	0.002
kospi	0.9895	0.005	193.450	0.000	0.979	1.000
kosdaq	0.0052	0.004	1.151	0.250	-0.004	0.014
kosdaq_volume	-0.0028	0.002	-1.723	0.085	-0.006	0.000
today_confirmed	0.0073	0.003	2.790	0.005	0.002	0.012
third_shot	0.0014	0.002	0.823	0.411	-0.002	0.005
wintershot	0.0044	0.002	1.980	0.048	4.25e-05	0.009
accumulate_confirmed	-0.0656	0.030	-2.174	0.030	-0.125	-0.006
accumulate_dead	0.0917	0.039	2.358	0.018	0.015	0.168
accumulate_first_shot	-0.0101	0.004	-2.319	0.021	-0.019	-0.002
accumulate_third_shot	-0.0259	0.010	-2.615	0.009	-0.045	-0.006
accumulate_winter_shot	-0.0034	0.002	-1.478	0.140	-0.008	0.001
Omnibus	디아고스티노 검정		Durbin-Watson		더빈왓슨 정규성 검정	
Prob(Omnibus)	Omnibus의 p-value		Jarque-Bera (JB)		자크베라 정규성 검정	
Skew	왜도		Prob(JB)		Jarque-Bera의 p-value	
Kurtosis	첨도		Cond. No.		다중공선성 검정	

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.89e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.(다중공선성이 있다는 의미)