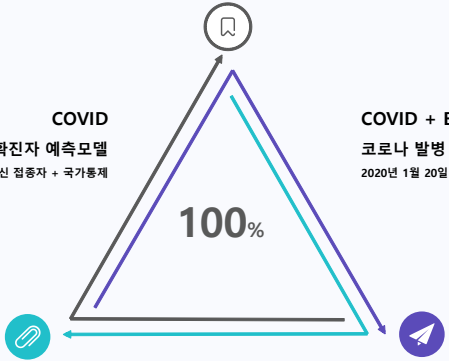
코로나 확진자 예측 모델

DLP(Data Loss Prevention)

김민기 김영민 송은진

목 차



COVID

코로나 데이터 확진자 예측모델

코로나 확진자 + 사망자 + 백신 접종자 + 국가통제

COVID + ECONOMY

코로나 발병 시기부터 적용한 확진자 예측모델

2020년 1월 20일부터 코로나 경제 데이터 적용 예측 모델

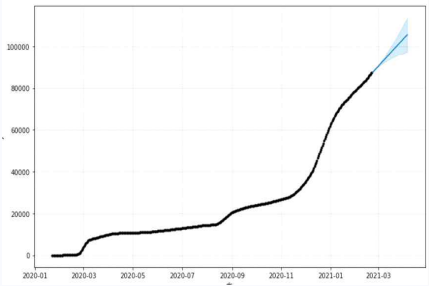
ECONOMY

코로나 ,경제 데이터 확진자 예측모델

2018년부터 코스피 + 코스닥 + 실업률 + 소비자물가지수

COVID

1. 시계열 데이터를 활용한 코로나19 동향 예측(수원대학교 논문)



Coefficient of determination	0.9144467335
MAE	1340.9974356183
MPE	-19.39964443585
MSE	5886837.758508
RMSE	2426.280642981

※ Prophet 알고리즘을 이용해 코로나 국내 발생 ~ 2021년 5월까지 예측
➡ 시계열의 추세를 이용해 미래 예측(발병 초기에는 예측 불가)

2. 모델 소개

1) 코로나 전날 확진자와 관련 데이터를 통해 다음날 확진자 예측
   ➡ 발병 초기부터 코로나 확진자를 예측가능
   ➡ 코로나 확진자 외 추가 데이터 적용
      - 코로나 사망자
      - 백신 접종자
      - 국가 통제

2) 코로나 확진자가 0명 ~ 60만명까지 큰 차이가 있음
   ➡ R-Squered와 MAE(평균절대오차)를 기준으로 모델 성능 평가

1. 데이터 설명

컬 럼 명	설 명	출 처
today_confirmed	일일 확진자	https://ncov.kdca.go.kr/
today_dead	일일 사망자	https://ncov.kdca.go.kr/
first_shot	1차 백신 접종자	https://ncv.kdca.go.kr/
second_shot	2차 백신 접종자	https://ncv.kdca.go.kr/
third_shot	3차 백신 접종자	https://kdx.kr/data/view/30239
winter_shot	동절기 백신 접종자	https://ncv.kdca.go.kr/
state_control	국가 통제(사회적 거리두기)	정부 발표자료 참고

※ 국가 통제를 제외한 모든 컬럼의 누적 컬럼도 적용

※ 일자는 년, 월, 일, 요일 컬럼으로 생성

2. 데이터 전처리

1) 코로나 데이터

필요 컬럼 추출 ➡ **일일 확진자 + 일일 사망자 + 누적 확진자 + 누적 사망자**

2) 백신 데이터

필요 컬럼 추출 ➡ **1·2차, 동절기 백신 접종자**

정부 데이터 소실로 인한 거래소 데이터 적용 ➡ **3차 백신 접종자**

3) 국가 통제

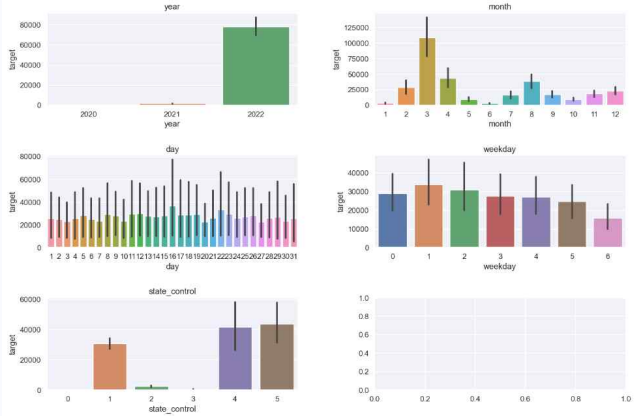
정부 발표 자료 적용 ➡ **사회적 거리두기 적용**

0단계 **= 코로나 시기 이전**

1단계 **= 기본적인 생활방역**

2~ 5 단계 **= 사회적 거리두기 단계에 따라 적용**

1. 범주형 데이터 시각화



year
➡ 2021년보다 2022년에 확진자 수가 더 많음

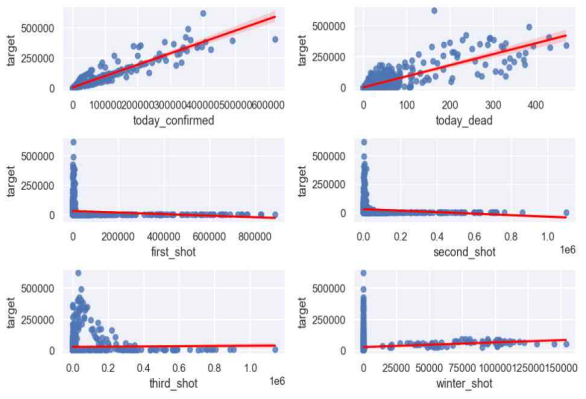
month
➡ 3·4·8월 순으로 확진자 많음

day ➡ 일자에 따른 변화는 없음

weekday
➡ 일요일 데이터는 낮게 측정
(0: 월요일 ~ 6 : 일요일)

state_control
➡ 1, 4, 5단계에서 확진자 높음

2. 코로나 데이터 시각화



일일 사망자
➡ 확진자 증가에 따라 증가

1 · 2차 백신 접종자
➡ 1 · 2차 접종자 증가에 따라 약하게 감소

3차, 동절기 백신 접종자
➡ 3차, 동절기 접종자 증가에 따라 약하게 증가

1. OLS 모델

R-squared	0.914
F-statistic	706.2
AIC	2.431e+04
BIC	2.431e+04
Skew	3.503
Kurtosis	47.536
Cond. No.	5.21e+11

R-squared
➡ 0.914로 높은 수치를 보임

AIC, BIC
➡ 비 정상적으로 높은 수치

Skew
➡ 3.503로 잔차가 대칭이 아님

Kurtosis
➡ 47.536으로 높은 수치

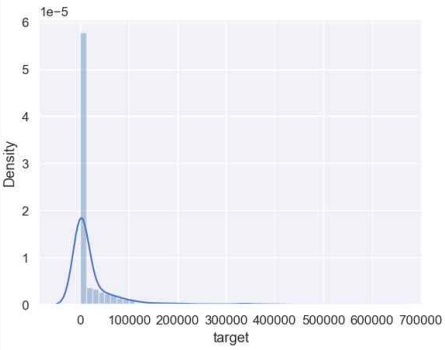
Cond. No.
➡ 비정상적으로 높은 수치
➡ 다중공선성(VIF, p-value 확인 필요)

※ 종속변수를 log를 통해 정규분포 모형으로 변환 필요

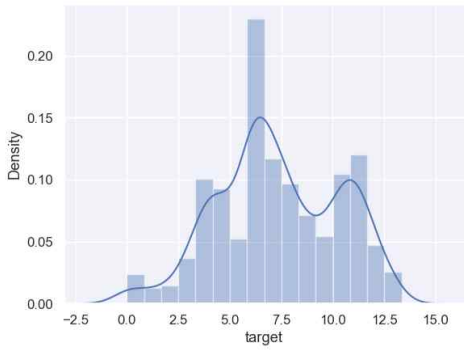
※ 독립변수 간 단위가 다르기 때문에 scale 필요

COVID 종속변수 시각화

일반적 종속변수 시각화



로그 종속변수 시각화



※ log를 통한 종속변수 정규분포 모형 적용

COVID 모델 생성(Log, Scale)

1. OLS 모델

R-squared	0.938
F-statistic	1001
AIC	2403
BIC	2487
Skew	0.072
Kurtosis	5.422
Cond. No.	95.5

R-squared
➡ 0.938로 높은 수치를 보임

AIC, BIC
➡ 2403, 2487

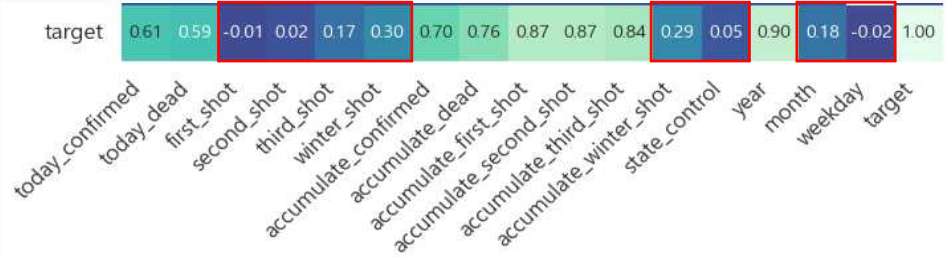
Skew
➡ 0.072로 잔차가 대칭으로 볼 수 있음

Kurtosis
➡ 5.422로 정상 범위 내

Cond. No.
➡ 95.5
➡ 기존 모델보다는 개선, 여전히 다중공선성 의심

※ 최적화 필요
➡ p-value, VIF 높은 컬럼, 상관관계 높은 컬럼 제거

COVID Target에 대한 상관계수



※ 0.3 이하의 상관관계를 보이는 컬럼 제거

- ➡ 코로나 백신 데이터(1, 2, 3, 동절기 백신 접종자, 누적 동절기 백신 접종자)
- ➡ 국가 통제
- ➡ 월, 요일

COVID 모델 생성(Target에 대한 상관계수)

1. OLS 모델

R-squared	0.875
F-statistic	935
AIC	3139
BIC	3184
Skew	-0.938
Kurtosis	6.883
Cond. No.	80.5

R-squared
➡ 0.875로 기존 모델보다 낮은 수치를 보임

AIC, BIC
➡ 3139, 3184로 기존 모델보다 증가

Skew
➡ -0.938로 잔차가 대칭으로 볼 수 있음

Kurtosis
➡ 6.883로 정상 범위 내

Cond. No.
➡ 80.5
➡ 기존 모델보다는 개선, 여전히 다중공선성 의심

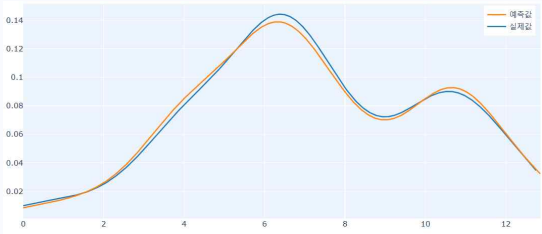
※ 상관관계가 낮은 값을 제외하니 모델 성능에 부정적 영향을 미침

2. Decision Tree 모델

모델 성능 비교

```
[('extra_tree', 0.2280096226109563),
 ('catboost', 0.2321790813560986),
 ('lightgbm', 0.2330132422049959 2),
 ('random_forest', 0.2393907033664739),
 ('xgboost', 0.265731464937665),
 ('adaboost', 0.2809249024591011),
 ('svr', 0.563346942944187),
 ('baysian_ridge', 0.6815636548319872),
 ('ardr_linear', 0.6815692585941188),
 ('ridge', 0.6816867696123331),
 ('linear', 0.6831936181094244),
 ('elasticnet', 0.9713373325836555),
 ('lasso', 1.1795778595117836)]
```

모델 그래프



※ **Extratree로 모델을 만들었을때 MAE가 다른 모델에 비해 낮으나 6834.699으로 매우 높은 수치를 보였음**

※ **결정계수는 0.986으로 높은 수치를 보임**

※ **Decision Tree 모델은 높게 나오지만 OLS 모델과 MAE는 낮기 때문에 다른 방법 적용**

1. 최적화 방법

1) 전체적으로 VIF 값이 높은 값을 순차적으로 제거

➡ today_dead 제거 ➡ 116.8
➡ first_shot 제거 ➡ 114.7
➡ second_shot 제거 ➡ 90.3
➡ third_shot 제거 ➡ 83.9
➡ winter shot 제거 ➡ 77.4
➡ accumulate_confirmed 제거 ➡ 77.3

※ 이후 컬럼을 추가로 제외하면 성능이 저하됨

2. OLS 모델

R-squared	0.937
F-statistic	1575
AIC	2413
BIC	2468
Skew	0.063
Kurtosis	5.302
Cod. No.	27.7

R-squared
➡ 0.937로 높은 수치를 보임

AIC, BIC
➡ 2413, 2468

Skew
➡ 0.063로 잔차가 대칭으로 볼 수 있음

Kurtosis
➡ 5.302로 정상 범위 내

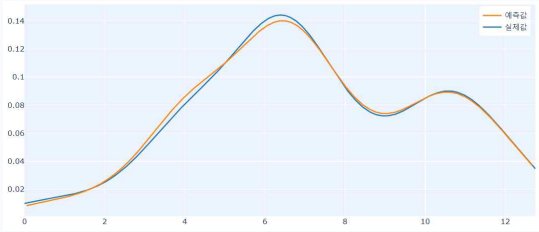
Cond. No.
➡ 27.7
➡ 다중공선성을 해결한 것으로 보임

3. Decision Tree 모델

모델 성능 비교

```
[('catboost', 0.16820957526089245),
 ('lightgbm', 0.17869817004513694),
 ('extra_tree', 0.18364362072366766),
 ('random_forest', 0.1965103744038466),
 ('xgboost', 0.19701355745818994),
 ('adaboost', 0.29733163107008603),
 ('svr', 0.40809977696797206),
 ('ardr_linear', 0.5516242408385891),
 ('baysian_ridge', 0.5528870464255821),
 ('ridge', 0.5529367407660578),
 ('linear', 0.5533343803600916),
 ('elasticnet', 0.9771709675559382),
 ('lasso', 1.179577782315734)]
```

모델 그래프



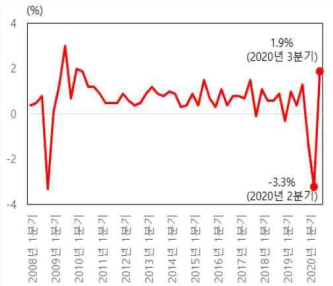
※ Catboost로 모델을 만들었을때 MAE가 다른 모델보다 2599.182로 가장 낮음
※ 결정계수는 0.9938로 높은 수치를 보임
※ 컬럼을 추가로 제거했을 때 모델 성능이 저하됨

4. 예측 결과

날짜	실제값	예측값	오차	오차(백분율)
2023-02-08	17927	17783	-144	-0.80
2023-02-09	14662	15070	408	2.78
2023-02-10	13504	14071	567	4.20
2023-02-11	12805	12475	-330	-2.57
2023-02-12	12051	11134	-917	-7.61
2023-02-13	5174	5351	177	3.42
2023-02-14	14371	13975	-396	-2.76

ECONOMY

1. 코로나19 사태가 국내 경제에 미치는 영향과 향후 과제(경제 논문)



출처: 한국은행, 저자 작성

<그림 2> 국내 실질GDP 추이

출처: 한국은행, 저자 작성

<그림 3> 국내 GDP, 민간소비, 수출 추이

<그림2>

코로나 발병 이후 국내 실질 GDP 변화 추이로 매우 큰 하락세를 보임

<그림3>

코로나 발병 이후 국내 GDP, 민간소비, 수출 변화 추이도 매우 큰 하락세를 보임

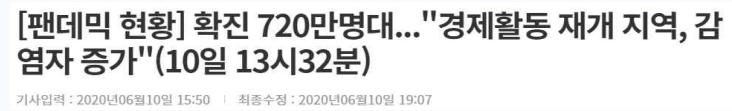
2. 모델 소개

1) 2018 ~ 2022년 코로나19 + 경제 데이터

- 코스피
- 코스닥
- 소비자 물가 지수
- 환 율
- 실업률

2) 코로나19와 경제의 관계

- 코로나19 초기 경제 타격을 받음
- 초기 코로나19가 지나고 경제활동이 증가함에 따라 코로나 확진자 증가


[팬데믹 현황] 확진 720만명대..."경제활동 재개 지역, 감염자 증가"(10일 13시32분)

기사입력 : 2020년06월10일 15:50 | 최종수정 : 2020년06월10일 19:07

1. 데이터 설명

컬 럼 명	설 명	출 처
kospi	코스피	https://finance.yahoo.com
kospi_volume	코스피 거래량	https://finance.yahoo.com
kosdaq	코스닥	https://finance.yahoo.com
kosdaq_volume	코스닥 거래량	https://finance.yahoo.com
exchange_rate	환율	https://finance.yahoo.com
jobless	실업률	kosis.kr
price_index	소비자 물가 지수	kosis.kr

※ 일자는 년, 월, 일, 요일 컬럼으로 생성

2. 데이터 전처리

1. 국가 경제 데이터

➡ 필요 컬럼 추출 ➡ **코스피 + 코스피 거래량 + 코스닥 + 코스닥 거래량 + 환율**

2. 국민 경제 데이터

➡ 월별 데이터를 일자별로 일괄 적용 ➡ **실업률 + 물가지수**

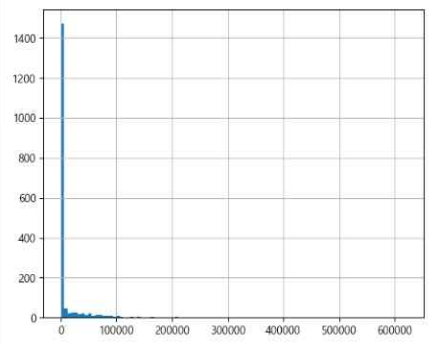
3. 데이터 적용시기

➡ 코로나 이전 경제 데이터 적용 ➡ **2018년 1월 1일**

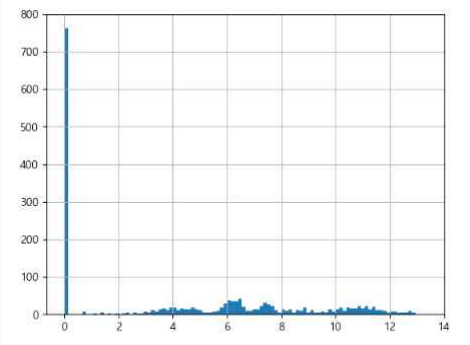
➡ 코로나 발병 이전 코로나 데이터는 0으로 처리

ECONOMY 종속변수 시각화

일반적 종속변수 시각화



로그 종속변수 시각화



※ 종속변수에 0이 많아 log 적용

➡ log를 했으나 여전히 0이 많게 나옴

ECONOMY 모델 생성(베이스라인 모델)

1. OLS 모델

R-squared	0.915
F-statistic	839.7
AIC	4.034e+04
BIC	4.048e+04
Skew	4.710
Kurtosis	78.959
Cond. No.	2.67e+11

R-squared
➡ 0.915로 높은 수치를 보임

AIC, BIC
➡ 4.034e + 11, 4.048e + 11 매우 높음

Skew
➡ 4.710로 log를 적용해도 높게 나옴

Kurtosis
➡ 78.959로 log를 적용해도 정상범위 외

Cond. No.
➡ 2.67e + 11
➡ 다중공성선이 높음

※ 코로나 국내 발병 이전 코로나 관련 컬럼 0으로 처리

➡종속 변수를 log 해도 정규분포 모형을 나타내지 않음

➡ 2020년 1월 20일 코로나 국내 발병 시기부터 적용 필요

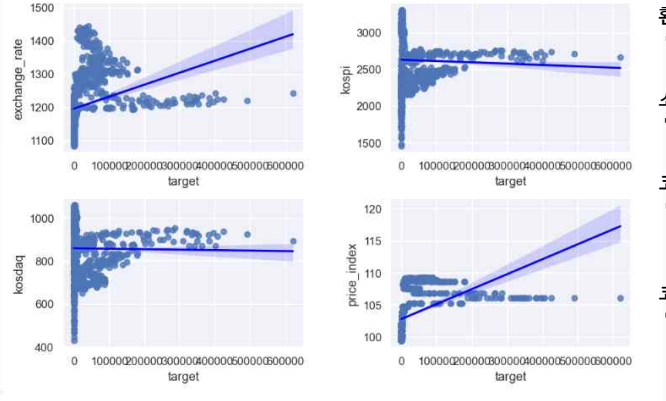
COVID + ECONOMY

1. 모델 소개

1) 2020.1.20 ~ 2022년 코로나19 + 경제 데이터
- ➡ 코로나19 국내 발병 시기부터 데이터 적용
- ➡ 코로나19 + 경제 데이터 적용 모델 생성
- ➡ 모델 생성 후 최적화

2) 회귀분석
- ➡ 선형성, 독립성, 정규성, 등분산성 확인
  - 위배되는 컬럼 다수 존재
  - 비교적 자유로운 Decision Tree 모델 적용
  - 직관적으로 볼 수 있는 선형모델로 모델 최적화 후 Tree모델 적용

1. 경제 데이터 시각화



환율
➡ 확진자 증가에 따라 증가

소비자 물가지수
➡ 확진자 증가에 따라 증가

코스피
➡ 확진자 증가에 따라 약하게 감소

코스닥
➡ 확진자 증가에 따라 약하게 감소

1. OLS 모델

R-squared	0.916
F-statistic	499.5
AIC	2.430e+04
BIC	2.442e+04
Skew	3.568
Kurtosis	49.067
Cond. No.	1.18e+12

R-squared
➡ 0.916으로 높은 수치를 보임

AIC, BIC
➡ 비 정상적으로 높은 수치

Skew
➡ 3.568로 잔차가 대칭이 아님

Kurtosis
➡ 49.067으로 높은 수치

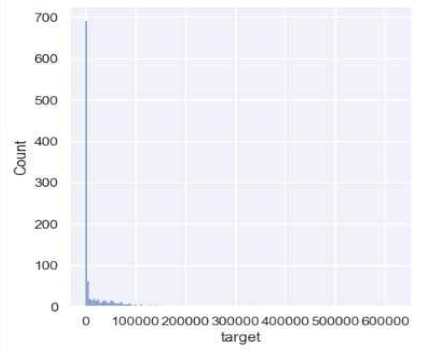
Cond. No.
➡ 비정상적으로 높은 수치
➡ 다중공선성(VIF, p-value 확인 필요)

※ 종속변수를 log를 통해 정규분포 모형으로 변환 필요

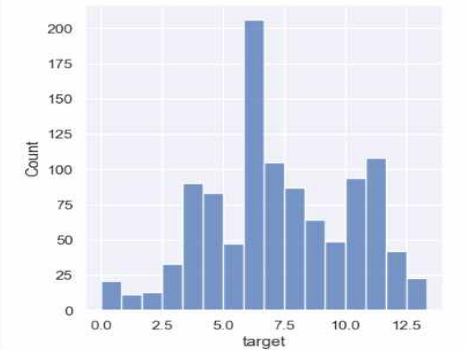
※ 독립변수 간 단위가 다르기 때문에 scale 필요

COVID + ECONOMY 종속변수 시각화

일반적 종속변수 시각화



로그 종속변수 시각화



※ log를 통한 종속변수 정규분포 모형 적용

COVID + ECONOMY 모델 생성(Log, Scale)

1. OLS 모델

R-squared	0.940
F-statistic	714.1
AIC	2384
BIC	2504
Skew	0.082
Kurtosis	5.629
Cond. No.	148

R-squared
➡ 0.940로 높은 수치를 보임

AIC, BIC
➡ 2384, 2504

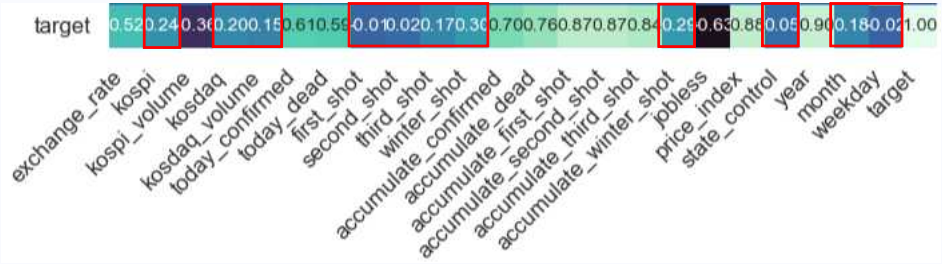
Skew
➡ 0.082로 잔차가 대칭으로 볼 수 있음

Kurtosis
➡ 5.629로 정상 범위 내

Cond. No.
➡ 148
➡ 기존 모델보다는 개선, 여전히 다중공선성 의심

※ 최적화 필요
➡ p-value, VIF 높은 컬럼, 상관관계 높은 컬럼 제거

COVID + ECONOMY Target에 대한 상관계수



※ 0.3 이하의 상관관계를 보이는 컬럼 제거

- ➡ 경제 데이터(코스피, 코스닥, 코스닥 거래량)
- ➡ 코로나 백신 데이터(1, 2, 3, 동절기 백신 접종자, 누적 동절기 백신 접종자)
- ➡ 국가 통제
- ➡ 월, 요일

1. OLS 모델

R-squared	0.902
F-statistic	811.7
AIC	2891
BIC	2955
Skew	-1.722
Kurtosis	9.979
Cond. No.	111

R-squared
➡ 0.902로 기존 모델보다 낮은 수치를 보임

AIC, BIC
➡ 2891, 2955로 기존 모델보다 증가

Skew
➡ -1.722로 잔차가 대칭으로 볼 수 있음

Kurtosis
➡ 9.979로 정상 범위에서 약간 벗어남

Cond. No.
➡ 111
➡ 기존 모델보다는 개선, 여전히 다중공선성 의심

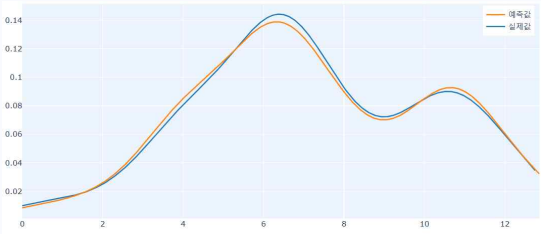
※ 상관관계가 낮은 값을 제외하니 모델 성능에 부정적 영향을 미침

2. Decision Tree 모델

모델 성능 비교

```
[('extra_tree', 0.228077947793048),
 ('random_forest', 0.23851823461640237),
 ('catboost', 0.2424444182792967),
 ('gradient_boost', 0.24302364436502502),
 ('lightgbm', 0.24303569628100083),
 ('xgboost', 0.272979322623911),
 ('adaboost', 0.27693448851879854),
 ('svr', 0.4665068445102184),
 ('linear', 0.6061913631117989),
 ('ridge', 0.6120414296577636),
 ('baysian_ridge', 0.6145177081788076),
 ('ardr_linear', 0.617596500314496),
 ('elasticnet', 0.9363810140869158),
 ('lasso', 1.1696019850424115)]
```

모델 그래프



※ **Extratree로 모델을 만들었을때 MAE가 다른 모델에 비해 낮으나 6440.491으로 매우 높은 수치를 보였음**

※ **결정계수는 0.989으로 높은 수치를 보임**

※ **Decision Tree 모델은 높게 나오지만 OLS 모델과 MAE는 낮기 때문에 다른 방법 적용**

1. 최적화 방법

1) 전체적으로 p-value 값이 높은 값을 순차적으로 제거

- ➡ accumulate_confirmed 제거 ➡ 0.977
- ➡ kosdaq 제거 ➡ 0.978
- ➡ today_dead 제거 ➡ 0.89
- ➡ third_shot 제거 ➡ 0.858
- ➡ winter shot 제거 ➡ 0.804
- ➡ kospi_volume 제거 ➡ 0.645
- ➡ first_shot 제거 ➡ 0.402
- ➡ kosdaq_volume 제거 ➡ 0.290

2) VIF가 높은 exchange_rate 제거

※ 이후 컬럼을 추가로 제외하면 성능이 저하됨

2. OLS 모델

R-squared	0.938
F-statistic	1149
AIC	2395
BIC	2470
Skew	0.011
Kurtosis	5.369
Cond. No.	46.6

R-squared
➡ 0.938로 높은 수치를 보임

AIC, BIC
➡ 2395, 2470

Skew
➡ 0.011으로 잔차가 대칭으로 볼 수 있음

Kurtosis
➡ 5.369로 정상 범위 내

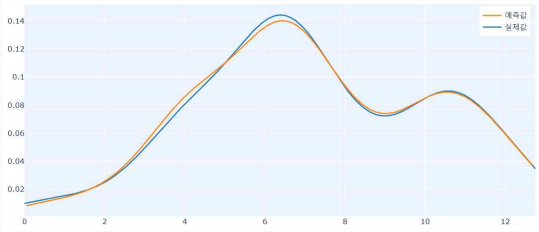
Cond. No.
➡ 46.6
➡ 다중공선성을 해결한 것으로 보임

3. Decision Tree 모델

모델 성능 비교

```
[('catboost', 0.171828025415439),
 ('lightgbm', 0.1780001724699643),
 ('extra_tree', 0.17945371827320072),
 ('gradient_boost', 0.18645421790336975),
 ('random_forest', 0.19200475482965992),
 ('xgboost', 0.1979414693127811),
 ('adaboost', 0.29796947216793185),
 ('svr', 0.3830516423205685),
 ('baysian_ridge', 0.5508814786345014),
 ('ridge', 0.5509262099545211),
 ('ardr_linear', 0.5509569418961847),
 ('linear', 0.5510974946798106),
 ('elasticnet', 0.9373640583971884),
 ('lasso', 1.169602616736278)]
```

모델 그래프



※ Catboost로 모델을 만들었을때 MAE가 다른 모델보다 2378.657로 가장 낮음
※ 결정계수는 0.9937로 높은 수치를 보임
※ 컬럼을 추가로 제거했을 때 모델 성능이 저하됨

4. 추가 최적화 시도

1) 하이퍼파라미터 튜닝

➡ optuna를 통한 최적의 하이퍼파라미터 확인
- catboost 결과 : R-squared(0.99358), MAE(3063.053)
- lightgbm결과 : R-squared(0.99318), MAE(3569.972)

※ 성능개선 안됨

2) 앙상블 기법

➡ catboost, extratree, lightgbm, gradient, xgboost를 다양하게 voting
- catboost, lightgbm은 하이퍼파라미터 적용해서도 적용
- voting 결과 MAE가 증가

※ 성능개선 안됨

COVID + ECONOMY 최종모델

※ 최종 모델 성능

OLS 모델	
R-squared	0.938
Cond. No.	46.6
Decision Tree	
R-squared	0.994
MAE	2378.657

※ 2.8 ~ 2.14 실제 예측값

날짜	실제값	예측값	오차	오차(백분율)
2023-02-08	17927	17879	-48	-0.27
2023-02-09	14662	14797	135	0.92
2023-02-10	13504	13691	187	1.38
2023-02-11	12805	12694	-111	-0.87
2023-02-12	12051	11737	-314	-2.60
2023-02-13	5174	5232	58	1.13
2023-02-14	14371	14238	-133	-0.93

결 론

※ 미래의 전염성 유행병에 대응이 가능

코로나 데이터

확진자

누적 백신접종

국가 통제

경제 데이터

코스피

실업률

소비자 물가지수

날 짜

연 도

월

요 일

결 론

※ 발전 방향

1) year, state_control 컬럼 개선

➡ year 컬럼을 1~6차 대유행, 변종 바이러스에 대한 내용 반영

➡ 초기 대응과 면역이 생긴 후의 대응을 반영

2) 데이터 추가

➡ 병원 내원 기록이나 감기 환자 수 등에 대한 데이터가 추가

감 사 합 니 다