Main subject

COVID19 + ECONOMY

Predict Model

DLP

Data Loss Prevention

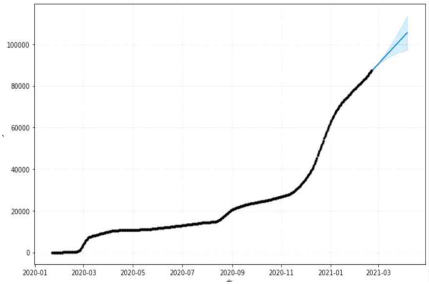
김민기 김영민 송은진

CONTENT

코로나 데이터만 적용

ONLY COVID19

1. 시계열 데이터를 활용한 코로나19 동향 예측



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

02 데이터 설명 및 출처

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

03 데이터 전처리 설명

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**

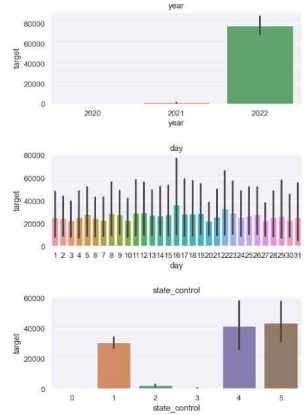
0단계 **= 코로나 시기 이전**

1단계 **= 기본적인 생활방역**

2~ 5 단계 **= 사회적 거리두기 단계에 따라 적용**

04 데이터 시각화

1. 범주형 데이터 시각화



year
➡ 2021년보다 2022년에 확진자 수가 더 많음

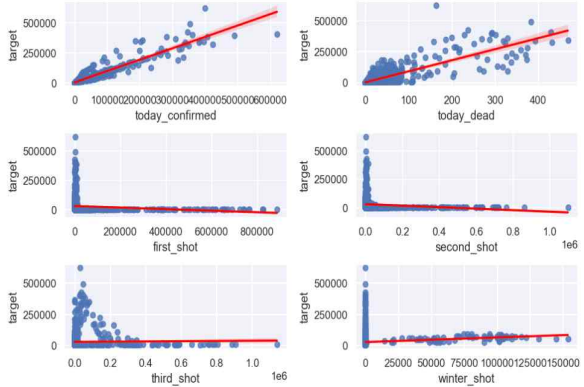
month
➡ 3·4·8월 순으로 확진자 많음

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

05 베이스 라인 모델

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.914
Model:	OLS	Adj. R-squared:	0.913
Method:	Least Squares	F-statistic:	706.2
Date:	Sun, 12 Feb 2023	Prob (F-statistic):	0.00
Time:	21:47:00	Log-Likelihood:	-12136.
No. Observations:	1076	AIC:	2.431e+04
Df Residuals:	1059	BIC:	2.439e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.055e+07	6.76e+06	-4.522	0.000	-4.38e+07	-1.73e+07
today_confirmed	0.8154	0.023	34.958	0.000	0.770	0.861
today_dead	13.6209	21.207	0.642	0.521	-27.992	55.234
first_shot	-0.0061	0.007	-0.907	0.364	-0.019	0.007
second_shot	-0.0067	0.007	-0.907	0.364	-0.021	0.008
third_shot	-0.0156	0.007	-2.130	0.033	-0.030	-0.001
winter_shot	0.0533	0.049	1.088	0.277	-0.043	0.149
accumulate_confirmed	0.0077	0.001	5.487	0.000	0.005	0.010
accumulate_dead	-10.1601	1.608	-6.320	0.000	-13.315	-7.005
accumulate_first_shot	0.0005	0.000	1.596	0.111	-0.000	0.001
accumulate_second_shot	-0.0003	0.000	-1.082	0.280	-0.001	0.000
accumulate_third_shot	0.0021	0.000	7.325	0.000	0.002	0.003
accumulate_winter_shot	-0.0011	0.001	-0.867	0.386	-0.004	0.001
state_control	-959.2099	685.117	-1.400	0.162	-2303.551	385.132
year	1.513e+04	3344.402	4.524	0.000	8566.827	2.17e+04
month	806.3772	301.991	2.670	0.008	213.809	1398.945
weekday	-2655.0105	303.336	-8.753	0.000	-3250.219	-2059.802

Omnibus:	942.021	Durbin-Watson:	2.235
Prob(Omnibus):	0.000	Jarque-Bera (JB):	91124.367
Skew:	3.503	Prob(JB):	0.00
Kurtosis:	47.536	Cond. No.	5.21e+11

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.21e+11. This might indicate that there are strong multicollinearity or other numerical problems.

1. R-squared는 **0.914**로 높은 수치를 보임

2. Cond. No.는 **5.21e + 11**로 높은 수치를 보여 다중공선성이 의심됨

3. 왜도와 첨도도 높은 수치를 보임
   ➡종속변수가 정규분포 모형이 아님

※ **종속변수에 log를 통해 정규분포 모형으로 변환 필요**

※ **독립변수들 간 단위가 다르기 때문에 scale 적용 필요**

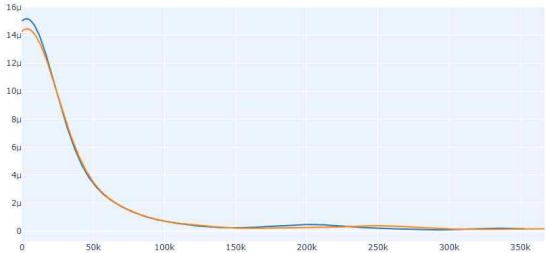
05 베이스 라인 모델

2. Decision Tree 모델

모델 성능 비교

```
[('catboost', 3720.178724871175),
 ('extra_tree', 4312.925906976745),
 ('xgboost', 4626.692139507935),
 ('random_forest', 5161.960418604651),
 ('lightgbm', 5299.751388003597),
 ('baysian_ridge', 6691.002635053039),
 ('ardr_linear', 8499.377243353121),
 ('elasticnet', 8684.72004218218),
 ('lasso', 8904.975660200997),
 ('ridge', 9059.904961949665),
 ('linear', 9080.352717636699),
 ('adaboost', 14005.620317056735),
 ('svr', 26895.364050448792)]
```

모델 그래프

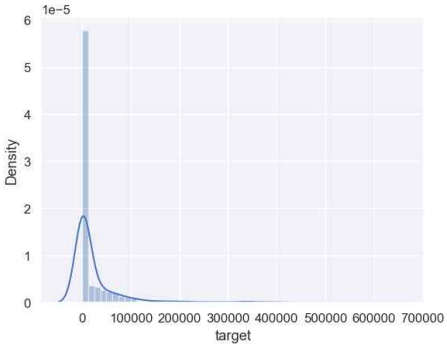


※ Catboost로 모델을 만들었을때 MAE가 3720으로 가장 낮은 수치를 보였음

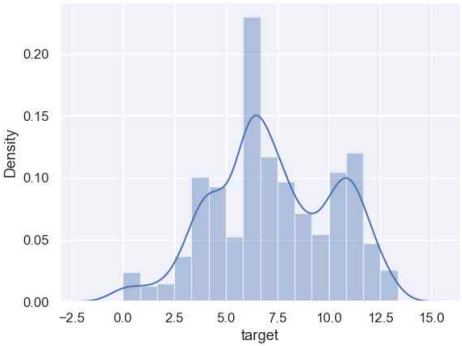
※ 결정계수는 0.975로 높은 수치를 보이나 다중공선성의 의심돼 데이터 처리 후 훈련 필요

06 종속변수 시각화

일반적 종속변수 시각화



로그 종속변수 시각화



※ log를 통한 종속변수 정규분포 모형 적용

07 Log, Scale

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.938
Model:	OLS	Adj. R-squared:	0.937
Method:	Least Squares	F-statistic:	1001.
Date:	Sun, 12 Feb 2023	Prob (F-statistic):	0.00
Time:	21:47:18	Log-Likelihood:	-1184.3
No. Observations:	1076	AIC:	2403.
Df Residuals:	1059	BIC:	2487.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3218	0.022	327.579	0.000	7.278	7.366
today_confirmed	0.5364	0.058	9.257	0.000	0.423	0.650
today_dead	0.1115	0.049	2.269	0.023	0.015	0.208
first_shot	0.0251	0.031	0.830	0.407	-0.036	0.088
second_shot	-0.0734	0.034	-2.147	0.032	-0.140	-0.006
third_shot	-0.0711	0.031	-2.281	0.023	-0.132	-0.010
winter_shot	-0.0056	0.038	-0.149	0.882	-0.079	0.068
accumulate_confirmed	-1.6015	0.511	-3.132	0.002	-2.605	-0.598
accumulate_dead	1.9595	0.666	2.940	0.003	0.651	3.266
accumulate_first_shot	-1.4797	0.230	-6.443	0.000	-1.930	-1.029
accumulate_second_shot	1.1952	0.242	4.947	0.000	0.721	1.669
accumulate_third_shot	0.3171	0.158	2.007	0.045	0.007	0.627
accumulate_winter_shot	-0.1181	0.038	-3.131	0.002	-0.192	-0.044
state_control	0.6749	0.041	16.475	0.000	0.595	0.755
year	1.9967	0.103	19.336	0.000	1.794	2.199
month	0.9519	0.039	24.423	0.000	0.875	1.028
weekday	-0.0978	0.023	-4.244	0.000	-0.143	-0.053

Omnibus:	66.350	Durbin-Watson:	0.340
Prob(Omnibus):	0.000	Jarque-Bera (JB):	283.953
Skew:	0.072	Prob(JB):	4.82e-62
Kurtosis:	5.422	Cond. No.	95.5

1. R-squared는 **0.938**로 높은 수치를 보임

2. Cond. No.는 **95.5**로 베이스라인 모델보다는 낮아졌지만 여전히 다중공선성은 의심됨

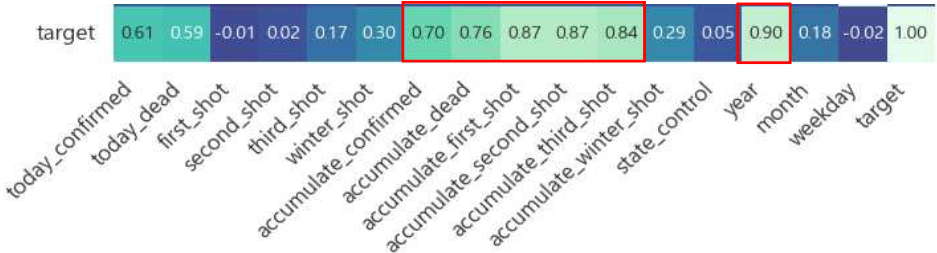
3. VIF 계수 ➡ 10이상의 컬럼 존재

4. 최적화 필요(p-value, VIF 높은 컬럼 제거)

	vif_factor	feature
0	888.2	const
1	523.4	today_confirmed
2	116.8	today_dead
3	105.6	first_shot
4	50.0	second_shot
5	21.3	third_shot
6	6.7	winter_shot
7	4.8	accumulate_confirmed
8	3.4	accumulate_dead
9	3.0	accumulate_first_shot
10	2.9	accumulate_second_shot
11	2.8	accumulate_third_shot
12	2.3	accumulate_winter_shot
13	2.0	state_control
14	1.9	year
15	1.1	month
16	1.0	weekday

08 Target과 상관관계

1. target에 대한 상관계수



※ 0.7 이상의 상관관계를 보이는 컬럼 제거

➡ 누적에 대한 데이터, 연도

08 Target과 상관관계

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.509
Model:	OLS	Adj. R-squared:	0.505
Method:	Least Squares	F-statistic:	110.6
Date:	Sun, 12 Feb 2023	Prob (F-statistic):	5.05e-157
Time:	21:47:35	Log-Likelihood:	-2296.8
No. Observations:	1076	AIC:	4616.
Df Residuals:	1065	BIC:	4670.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3218	0.063	116.817	0.000	7.199	7.445
today_confirmed	1.3952	0.132	10.552	0.000	1.136	1.655
today_dead	0.4682	0.133	3.520	0.000	0.207	0.729
first_shot	0.0892	0.075	1.182	0.238	-0.059	0.237
second_shot	0.1601	0.081	1.972	0.049	0.001	0.319
third_shot	0.4294	0.068	6.336	0.000	0.296	0.562
winter_shot	0.4550	0.101	4.504	0.000	0.257	0.653
accumulate_winter_shot	0.1850	0.101	1.825	0.068	-0.014	0.384
state_control	0.0676	0.078	0.863	0.388	-0.086	0.221
month	0.5095	0.073	6.952	0.000	0.365	0.652
weekday	-0.0751	0.064	-1.175	0.240	-0.200	0.050

Omnibus:	15.853	Durbin-Watson:	0.122
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16.177
Skew:	-0.298	Prob(JB):	0.000307
Kurtosis:	3.081	Cond. No.	4.41

1. R-squared는 **0.509**로 매우 성능이 저하됨

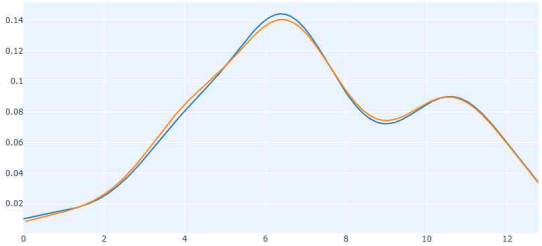
2. Cond. No.는 **4.41**로 다중공선성은 없음

2. Decision Tree 모델

모델 성능 비교

```
[('catboost', 0.1785179569827269 3),
 ('lightgbm', 0.18336289803837857),
 ('extra_tree', 0.1899913479707818 8),
 ('random_forest', 0.19500545474404934),
 ('xgboost', 0.20065671069662167),
 ('adaboost', 0.3010182761897087),
 ('svr', 0.9090193259426057),
 ('baysian_ridge', 1.622313673408226),
 ('ardr_linear', 1.6230142054665724),
 ('ridge', 1.626197848135191 7),
 ('linear', 1.6264744127715456),
 ('elasticnet', 1.925950585363432),
 ('lasso', 2.079753671121483)]
```

모델 그래프



※ Catboost로 모델을 만들었을때 MAE가 2847.383으로 가장 낮은 수치를 보였음
※ 결정계수는 0.993로 높은 수치를 보임
※ Decision Tree 모델은 높게 나오지만 OLS 모델은 낮기 때문에 다른 방법 적용

09 최종 모델

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.937
Model:	OLS	Adj. R-squared:	0.936
Method:	Least Squares	F-statistic:	1575.
Date:	Sun, 12 Feb 2023	Prob (F-statistic):	0.00
Time:	21:49:28	Log-Likelihood:	-1195.4
No. Observations:	1076	AIC:	2413.
Df Residuals:	1065	BIC:	2468.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3218	0.023	325.130	0.000	7.278	7.366
today_confirmed	0.5575	0.031	18.108	0.000	0.497	0.618
accumulate_dead	-0.0258	0.114	-0.226	0.822	-0.250	0.198
accumulate_first_shot	-1.4843	0.185	-8.011	0.000	-1.848	-1.121
accumulate_second_shot	1.2052	0.198	6.088	0.000	0.817	1.594
accumulate_third_shot	0.6025	0.122	4.921	0.000	0.362	0.843
accumulate_winter_shot	-0.1630	0.028	-5.759	0.000	-0.219	-0.107
state_control	0.6452	0.039	16.404	0.000	0.568	0.722
year	2.2129	0.077	28.671	0.000	2.061	2.364
month	0.9635	0.038	25.586	0.000	0.890	1.037
weekday	-0.0937	0.023	-4.159	0.000	-0.138	-0.049

Omnibus:	62.713	Durbin-Watson:	0.314
Prob(Omnibus):	0.000	Jarque-Bera (JB):	238.219
Skew:	0.063	Prob(JB):	1.87e-52
Kurtosis:	5.302	Cond. No.	27.7

1. R-squared는 **0.937**로 최적화 전보다 0.02 성능 저하됨

2. Cond. No.는 **27.7**로 다중공선성 가능성을 낮춤

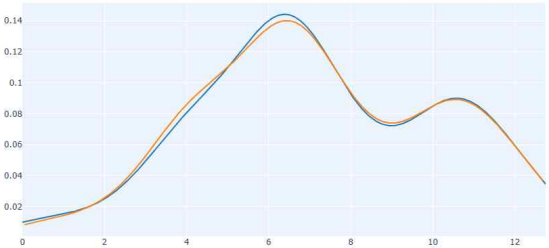
3. MAE(평균절대오차)가 가장 낮은 모델을 했기 때문에 최종 모델로 선정

2.Decision Tree 모델

모델 성능 비교

```
[('catboost', 0.16820957526089245),
 ('lightgbm', 0.17869817004513694),
 ('extra_tree', 0.1821480267269l383),
 ('random_forest', 0.195977245403601),
 ('xgboost', 0.19701355745818994),
 ('adaboost', 0.291992409887707),
 ('svr', 0.40809977696797206),
 ('ardr_linear', 0.5516242408385891),
 ('baysian_ridge', 0.5528870464255821),
 ('ridge', 0.552936740766O578),
 ('linear', 0.5533343803600916),
 ('elasticnet', 0.9771709675559382),
 ('lasso', 1.1795777823151734)]
```

모델 그래프



※ Catboost로 모델을 만들었을때 MAE가 2599.182으로 가장 낮은 수치를 보였음
※ 결정계수는 0.994로 높은 수치를 보임
※ 위 모델을 적용하면 다음날 코로나 확진자를 최대 2600명 정도의 오차 내에서 확인할 수 있음

10 최종 모델로 예측

※ TEST 데이터와 예측값 비교

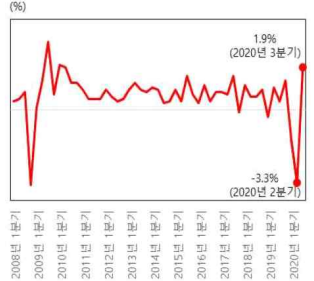
날짜	실제값	예측값	오차	오차(백분율)
2023-02-08	17927	17783	-144	-0.80
2023-02-09	14662	15070	408	2.78
2023-02-10	13504	14071	567	4.20
2023-02-11	12805	12475	-330	-2.57
2023-02-12	12051	11134	-917	-7.61
2023-02-13	5174	5351	177	3.42
2023-02-14	14371	13975	-396	-2.76

DLP

2018~2022
코로나+경제

COVID19 + ECONOMY

1. 코로나19 사태가 국내 경제에 미치는 영향과 향후 과제



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

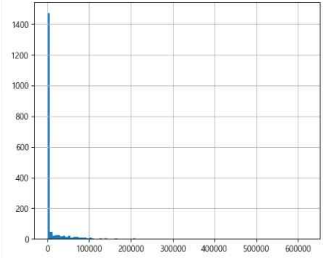
02 데이터 컬럼 설명

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

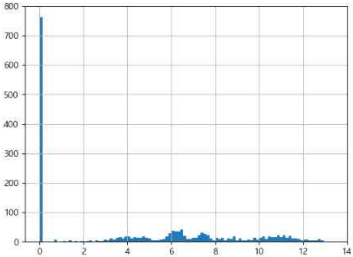
※ 기존 코로나 데이터에 추가

1. 코로나19 사태가 국내 경제에 미치는 영향과 향후 과제

일반적 종속변수 시각화



로그 종속변수 시각화



※ 종속변수에 0이 많아 log를 적용

➡ log를 했으나 여전히 0이 많게 나옴

04 베이스라인모델

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	859.7
Date:	Mon, 13 Feb 2023	Prob (F-statistic):	0.00
Time:	20:52:10	Log-Likelihood:	-20148.
No. Observations:	1824	AIC:	4.034e+04
Df Residuals:	1800	BIC:	4.048e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.402e+06	2.74e+06	-2.697	0.007	-1.28e+07	-2.02e+06
exchange_rate	-62.0382	21.503	-2.885	0.004	-104.212	-19.864
kospi	-2.0830	4.830	-0.431	0.666	-11.555	7.389
kospi_volume	0.0005	0.002	0.326	0.745	-0.003	0.004
kosdaq	8.0274	10.537	0.762	0.446	-12.639	28.694
kosdaq_volume	-0.0003	0.002	-0.199	0.842	-0.003	0.003
today_confirmed	0.8047	0.019	42.828	0.000	0.768	0.842
today_dead	-4.6861	17.838	-0.263	0.793	-39.671	30.299
first_shot	0.0015	0.005	0.292	0.771	-0.009	0.012
second_shot	-0.0017	0.006	-0.295	0.768	-0.013	0.010
third_shot	-0.0160	0.006	-2.650	0.008	-0.028	-0.004
wintershot	0.1297	0.039	3.347	0.001	0.054	0.206
accumulate_confirmed	0.0094	0.001	7.420	0.000	0.007	0.012
accumulate_dead	-12.0264	1.409	-8.534	0.000	-14.790	-9.262
accumulate_first_shot	0.0002	0.000	0.944	0.346	-0.000	0.001
accumulate_second_shot	-3.204e-05	0.000	-0.121	0.904	-0.001	0.000
accumulate_third_shot	0.0023	0.000	9.574	0.000	0.002	0.003
accumulate_winter_shot	-0.0027	0.001	-2.260	0.024	-0.005	-0.000
jobless	-1558.4432	1221.012	-1.276	0.202	-3953.194	836.308
price_index	3898.6617	1154.157	3.378	0.001	1635.033	6162.290
state_control	-692.4969	477.629	-1.450	0.147	-1629.263	244.269
year	3512.3517	1378.598	2.548	0.011	808.531	6216.172
month	-67.1769	202.591	-0.332	0.740	-464.514	330.161
day	19.2808	41.513	0.464	0.642	-62.136	100.699

Omnibus:	1935.672	Durbin-Watson:	2.213
Prob(Omnibus):	0.000	Jarque-Bera (JB):	445243.348
Skew:	4.710	Prob(JB):	0.00
Kurtosis:	78.959	Cond. No.	2.67e+11

1. R-squared는 **0.915**로 높은 수치를 보임

2. Cond. No.는 **2.67e+11**로 다중공선성이 의심됨

3. 코로나 이전인 2018년부터 데이터 적용 (코로나 관련 컬럼은 0 적용)

   ➡모든 데이터에 0이 많아서 제대로된 훈련이 이루어지지않음

DLP

2020~2022
코로나+경제

COVID19 + ECONOMY

01 데이터 전처리 설명

1. 국가 경제 데이터

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

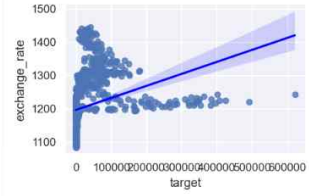
2. 국민 경제 데이터

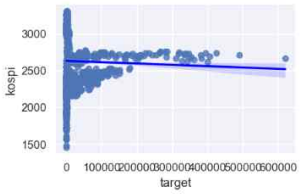
월별 데이터를 일자별로 일괄 적용 ➡ **실업률 + 물가지수**

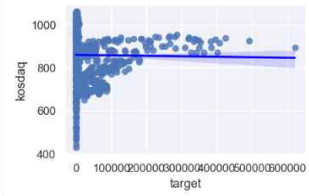
3. 데이터 적용시기

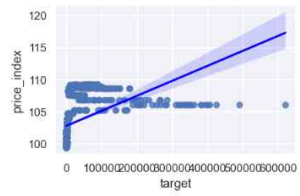
코로나 발병 시기부터 적용 ➡ **2020년 1월 20일**

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

03 베이스 라인 모델

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.916
Model:	OLS	Adj. R-squared:	0.914
Method:	Least Squares	F-statistic:	499.5
Date:	Tue, 14 Feb 2023	Prob (F-statistic):	0.00
Time:	21:21:12	Log-Likelihood:	-12125.
No. Observations:	1076	AIC:	2.430e+04
Df Residuals:	1052	BIC:	2.442e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.793e+07	1.52e+07	-1.178	0.239	-4.78e+07	1.19e+07
exchange_rate	-110.1674	35.360	-3.116	0.002	-179.572	-40.762
kospi	-17.4011	10.003	-1.740	0.082	-37.030	2.227
kospi_volume	0.0016	0.002	0.715	0.475	-0.003	0.006
kosdaq	18.7593	22.714	0.826	0.409	-25.810	63.328
kosdaq_volume	-0.0039	0.003	-1.315	0.189	-0.010	0.002
today_confirmed	0.7600	0.025	31.659	0.000	0.712	0.808
today_dead	14.5411	25.312	0.574	0.566	-35.127	64.209
first_shot	-0.0076	0.007	-1.137	0.256	-0.021	0.006
second_shot	-0.0065	0.007	-0.886	0.376	-0.021	0.008
third_shot	-0.0071	0.008	-0.885	0.376	-0.023	0.009
winter_shot	0.0826	0.050	1.660	0.097	-0.015	0.180
accumulate_confirmed	0.0115	0.002	6.147	0.000	0.008	0.015
accumulate_dead	-15.0108	2.059	-7.290	0.000	-19.051	-10.970
accumulate_first_shot	0.0008	0.000	2.252	0.025	0.000	0.002
accumulate_second_shot	-0.0008	0.000	-1.987	0.047	-0.002	-9.39e-06
accumulate_third_shot	0.0023	0.000	5.752	0.000	0.002	0.003
accumulate_winter_shot	-0.0027	0.002	-1.648	0.100	-0.006	0.001
jobless	-570.7358	2119.652	-0.269	0.788	-4729.962	3588.490
price_index	9469.1357	2878.287	3.162	0.002	3812.743	1.37e+04
state_control	-1586.0349	722.880	-2.194	0.028	-3004.495	-167.574
year	8542.8822	7615.566	1.122	0.262	-6400.742	2.35e+04
month	75.6371	489.171	0.155	0.877	-884.224	1035.498
weekday	-2586.9280	303.318	-8.528	0.000	-3282.106	-2091.750

Omnibus:	955.192	Durbin-Watson:	2.197
Prob(Omnibus):	0.000	Jarque-Bera (JB):	97426.101
Skew:	3.569	Prob(JB):	0.00
Kurtosis:	49.057	Cond. No.	1.19e+12

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.19e+12. This might indicate that there are strong multicollinearity or other numerical problems.

1. R-squared는 **0.916**로 높은 수치를 보임

2. Cond. No.는 **1.18e + 12**로 높은 수치를 보여 다중공선성이 의심됨

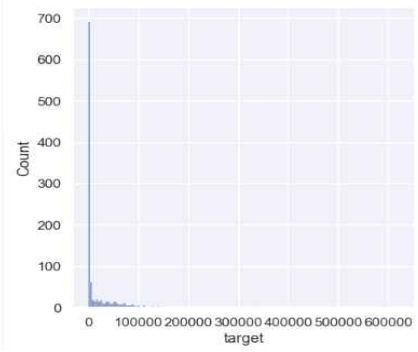
3. 왜도와 첨도도 높은 수치를 보임
   ➡종속변수가 정규분포 모형이 아님

※ **종속변수에 log를 통해 정규분포 모형으로 변환 필요**

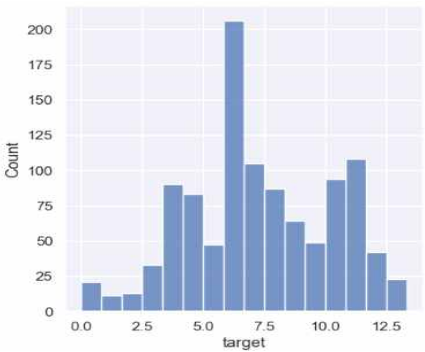
※ **독립변수들 간 단위가 다르기 때문에 scale 적용 필요**

04 종속변수 시각화

일반적 종속변수 시각화



로그 종속변수 시각화



※ log를 통한 종속변수 정규분포 모형 적용

05 Log, Scale

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.940
Model:	OLS	Adj. R-squared:	0.938
Method:	Least Squares	F-statistic:	714.1
Date:	Mon, 13 Feb 2023	Prob (F-statistic):	0.00
Time:	12:16:51	Log-Likelihood:	-1160.0
No. Observations:	1076	AIC:	2384.
Df Residuals:	1052	BIC:	2504.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3218	0.022	331.461	0.000	7.278	7.365
exchange_rate	-0.4354	0.109	-3.977	0.000	-0.650	-0.221
kospi	-0.1614	0.159	-1.012	0.312	-0.474	0.152
kospi_volume	0.0142	0.033	0.434	0.664	-0.050	0.079
kosdaq	-0.0037	0.115	-0.032	0.974	-0.229	0.221
kosdaq_volume	0.0235	0.023	1.027	0.304	-0.021	0.068
today_confirmed	0.5293	0.061	8.689	0.000	0.410	0.649
today_dead	-0.0069	0.058	-0.118	0.906	-0.121	0.108
first_shot	0.0254	0.031	0.809	0.419	-0.036	0.087
second_shot	-0.0863	0.034	-2.541	0.011	-0.153	-0.020
third_shot	-0.0055	0.034	-0.161	0.872	-0.072	0.061
winter_shot	0.0091	0.038	0.240	0.810	-0.065	0.084
accumulate_confirmed	-0.0195	0.680	-0.029	0.977	-1.353	1.314
accumulate_dead	0.8497	0.849	1.000	0.317	-0.817	2.516
accumulate_first_shot	-1.4547	0.289	-5.025	0.000	-2.023	-0.887
accumulate_second_shot	1.5119	0.257	5.886	0.000	0.929	2.095
accumulate_third_shot	0.2610	0.192	1.363	0.173	-0.115	0.637
accumulate_winter_shot	-0.2114	0.049	-4.339	0.000	-0.307	-0.116
jobless	-0.2004	0.064	-3.125	0.002	-0.326	-0.075
price_index	-0.7940	0.343	-2.290	0.022	-1.457	-0.112
state_control	0.6906	0.043	16.051	0.000	0.606	0.775
year	2.4653	0.234	10.532	0.000	2.006	2.925
month	0.9012	0.063	14.339	0.000	0.778	1.025
weekday	-0.0948	0.023	-4.135	0.000	-0.140	-0.050

Omnibus:	72.342	Durbin-Watson:	0.331
Prob(Omnibus):	0.000	Jarque-Bera (JB):	311.106
Skew:	0.082	Prob(JB):	2.79e-68
Kurtosis:	5.629	Cond. No.	148.

1. R-squared는 **0.940**로 높은 수치를 보임

2. Cond. No.는 **148**로 베이스라인 모델보다는 낮아졌지만 여전히 다중공선성은 의심됨

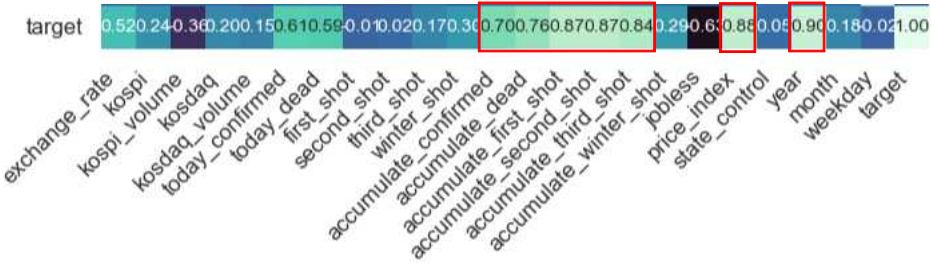
3. VIF 계수 ➡ 10이상의 컬럼 존재

4. 최적화 필요(p-value, VIF 높은 컬럼 제거)

	vif_factor	feature
0	1478.2	const
1	947.0	exchange_rate
2	240.8	kospi
3	181.1	kospi_volume
4	171.7	kosdaq
5	112.3	kosdaq_volume
6	75.2	today_confirmed
7	52.1	today_dead
8	27.0	first_shot
9	24.6	second_shot
10	8.4	third_shot
11	8.1	winter_shot
12	7.6	accumulate_confirmed
13	7.0	accumulate_dead
14	4.9	accumulate_first_shot
15	3.8	accumulate_second_shot
16	3.0	accumulate_third_shot
17	2.4	accumulate_winter_shot
18	2.4	jobless
19	2.2	price_index
20	2.0	state_control
21	1.1	year
22	1.1	month
23	1.0	weekday

06 Target과 상관관계

1. target에 대한 상관계수



※ 0.7 이상의 상관관계를 보이는 컬럼 제거

➡ 누적에 대한 데이터, 소비자 물가 지수, 연도

06 Target과 상관관계

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.832
Method:	Least Squares	F-statistic:	334.0
Date:	Mon, 13 Feb 2023	Prob (F-statistic):	0.00
Time:	12:18:07	Log-Likelihood:	-1711.8
No. Observations:	1076	AIC:	3458.
Df Residuals:	1059	BIC:	3542.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3218	0.036	200.634	0.000	7.250	7.393
exchange_rate	1.9122	0.083	22.935	0.000	1.749	2.076
kospi	1.3272	0.150	9.314	0.000	1.103	1.692
kospi_volume	-0.0200	0.051	-0.395	0.693	-0.119	0.079
kosdaq	0.3157	0.164	1.923	0.055	-0.006	0.638
kosdaq_volume	0.0479	0.038	1.272	0.204	-0.026	0.122
today_confirmed	0.8006	0.082	9.760	0.000	0.640	0.962
today_dead	0.2062	0.081	2.550	0.011	0.048	0.365
first_shot	-0.2295	0.047	-4.924	0.000	-0.321	-0.138
second_shot	-0.3305	0.049	-6.703	0.000	-0.427	-0.234
third_shot	0.1263	0.043	2.915	0.004	0.041	0.211
winter_shot	-0.1026	0.062	-1.657	0.098	-0.224	0.019
accumulate_winter_shot	0.4486	0.054	7.051	0.000	0.324	0.573
jobless	-0.4286	0.083	-5.156	0.000	-0.592	-0.266
state_control	0.5052	0.054	9.434	0.000	0.400	0.610
month	0.0268	0.063	0.427	0.669	-0.096	0.150
weekday	-0.1634	0.038	-4.351	0.000	-0.237	-0.090

Omnibus:	24.620	Durbin-Watson:	0.257
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27.154
Skew:	-0.325	Prob(JB):	1.27e-06
Kurtosis:	3.427	Cond. No.	12.2

1. R-squared는 **0.835**로 매우 성능 저하됨

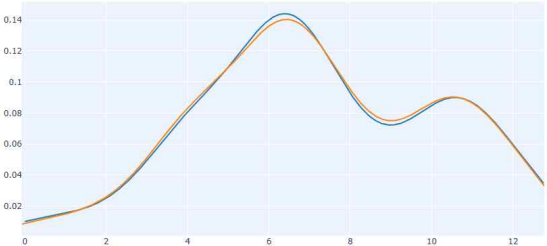
2. Cond. No.는 **12.2**로 다중공선성은 없음

2. Decision Tree 모델

모델 성능 비교

```
[('catboost', 0.17925757077407378),
 ('extra_tree', 0.1819862212651334),
 ('lightgbm', 0.1845392471320968),
 ('random_forest', 0.19305501878162376),
 ('xgboost', 0.1960203437323948),
 ('adaboost', 0.2796721871923199),
 ('svr', 0.4369049521923215),
 ('baysian_ridge', 0.9636240678772424),
 ('ardr_linear', 0.9646519146790634),
 ('ridge', 0.9651072765733589),
 ('linear', 0.9659296470753249),
 ('elasticnet', 1.5071673512435817),
 ('lasso', 1.7781091182153475)]
```

모델 그래프



※ Catboost로 모델을 만들었을때 MAE가 3436.838으로 가장 낮은 수치를 보였음
※ 결정계수는 0.994로 높은 수치를 보임
※ Decision Tree 모델은 높게 나오지만 OLS 모델은 낮기 때문에 다른 방법 적용

07 최종 모델

※ 전체적으로 p-value값이 높은 값을 순차적으로 제거

1. Kosdaq 제거 -> p-value 값 0.974
2. Today_dead 제거 -> p-value 값 0.9
3. Accumulate confirmed 제거 -> p-value 값 0.95
4. Third shot 제거 -> p-value 값 0.858
5. Winter shot 제거 -> p-value 값 0.804
6. Kospi volume 제거 -> p-value 값 0.645
7. First shot 제거 -> p-value 값 0.402
8. Kosdaq volume -> p-value 값 0.2

※ VIF 계수값이 큰 exchange rate 제거

07 최종 모델

1. OLS 모델

OLS Regression Results

Dep. Variable:	target	R-squared:	0.938
Model:	OLS	Adj. R-squared:	0.937
Method:	Least Squares	F-statistic:	1149.
Date:	Wed, 15 Feb 2023	Prob (F-statistic):	0.00
Time:	09:23:00	Log-Likelihood:	-1182.6
No. Observations:	1076	AIC:	2395.
Df Residuals:	1061	BIC:	2470.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.3218	0.022	328.397	0.000	7.278	7.366
kospi	0.2145	0.066	3.251	0.001	0.085	0.344
today_confirmed	0.5292	0.032	16.444	0.000	0.466	0.592
second_shot	-0.0694	0.033	-2.018	0.044	-0.131	-0.002
accumulate_dead	0.6500	0.201	3.240	0.001	0.256	1.044
accumulate_first_shot	-1.6428	0.248	-6.611	0.000	-2.130	-1.155
accumulate_second_shot	1.6558	0.255	6.496	0.000	1.155	2.157
accumulate_third_shot	0.5705	0.138	4.123	0.000	0.299	0.842
accumulate_winter_shot	-0.1403	0.031	-4.575	0.000	-0.201	-0.080
jobless	-0.1705	0.061	-2.799	0.005	-0.290	-0.051
price_index	-0.9048	0.298	-3.032	0.002	-1.490	-0.319
state_control	0.6941	0.040	17.188	0.000	0.615	0.773
year	2.1364	0.211	10.108	0.000	1.722	2.551
month	0.8551	0.057	14.973	0.000	0.743	0.967
weekday	-0.0992	0.023	-4.404	0.000	-0.143	-0.055

Omnibus:	63.928	Durbin-Watson:	0.319
Prob(Omnibus):	0.000	Jarque-Bera (JB):	251.627
Skew:	0.011	Prob(JB):	2.20e-55
Kurtosis:	5.369	Cond. No.	46.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

1. R-squared는 **0.938로 높은 성능을 보임**

2. Cond. No.는 **46.6으로** 기준 범위 내

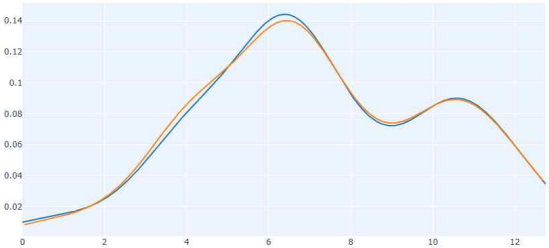
3. VIF 계수에 따라 추가로 제외할 수록 성능이 저하 됨

2. Decision Tree 모델

모델 성능 비교

```
[('catboost', 0.17089729706765694),
 ('extra_tree', 0.17767708353084868),
 ('lightgbm', 0.17800013724699643),
 ('random_forest', 0.1895967551779154),
 ('xgboost', 0.1979414693127811),
 ('adaboost', 0.290117402389130 8),
 ('svr', 0.3830516423205685),
 ('baysian_ridge', 0.5508814786345014),
 ('ridge', 0.5509262099545211),
 ('ardr_linear', 0.550956941896184 7),
 ('linear', 0.5510974946798106),
 ('elasticnet', 0.9373640583971884),
 ('lasso', 1.169602616736278)]
```

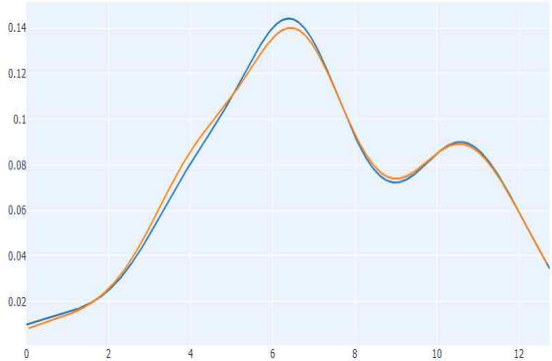
모델 그래프



※ Catboost로 모델을 만들었을때 MAE가 2378.657으로 가장 낮은 수치를 보였음
※결정계수는 0.994로 높은 수치를 보임
※ 위 모델을 적용하면 다음날 코로나 확진자를 최대 2379명 정도의 오차 내에서 확인할 수 있음

08 모델 최적화

1. 하이퍼 파라미터 튜닝



1. RandomizedSearchCV를 이용한 하이퍼 파라미터 튜닝

2. R-squared는 **0.9938**로 높은 수치를 보임

3. MAE는 **2763.704**로 높은 수치를 보이나 p-value를 통한 최적화 보다는 성능이 저하

2. 앙상블 기법(Voting)

Extratree, catboost, gradienboost

1. ExtraTreeRegressor, CatBoostRegressor, GradientBoostingRegressor을 Voting

2. R-squared는 **0.9939**
   MAE는 **2892.869**

Extratree, catboost

1. ExtraTreeRegressor, CatBoostRegressor, 을 Voting

2. R-squared는 **0.9941**
   MAE는 **2824.023**

08 모델 최적화

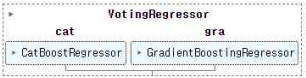
2. 앙상블 기법(Voting)

Extratree, gradienboost

1. ExtraTreeRegressor, GradientBoostingRegressor을 Voting

2. R-squared는 **0.9935**
   MAE는 **3191.253**

catboost, gradienboost



1. CatBoostRegressor, GradientBoostingRegressor을 Voting

2. R-squared는 **0.9936**
   MAE는 **2723.597**

※ 하이퍼 파라미터튜닝을 먼저 적용해보고 **Boosting** 모델을 **Voting**해 적용했으나 **MAE**는 증가해 **p-value**를 통한 최적화 모델을 최종 모델로 선택

09 Final Model

※ 최종 모델 성능

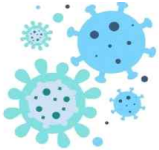
OLS 모델	
R-squared	0.938
Cond. No.	46.6
Decision Tree	
R-squared	0.994
MAE	2378.657

※ 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

10 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

DLP

THANK YOU

Final project