BUSINESS FINAL PRESENTATION

- 16102275 Park Hyun Woo
- 16102391 Aziz Dadabaev
- 17102063 Lee YooSeok

CONTEXT

1. Summarize the data analysis process

    - The most important points in each steps.

2. Conclusion

    - Final Selected Model

    - Needs to improve

SUMMARY DATA ANALYSIS PROCESS

1-1. PURPOSE OF THE PROJECT

- Analyze the Stock Data by industry with news data

The stock industry is filled with lots of data, however being able to analyze and utilize those data is very important in order to take successful risks

- Understanding and Observing Market Trends

Using data mining techniques we can extract publicly available data to properly identify and study the stock industries of interest

- Predict how much attention each industry will receive the next day through news data from the past to today.

1-2. INPUT DATA SELECTION

- Keyword data from news data of each date and sector

We used the frequency of word appearances in news by sector for each date And by using Mutual info classification to discover each sector’s meaningful words.

- Financial Data

Exchange rate, KBond rate, Dollar index

- ETF Volume Data for each Sectors And USA Sector Volume for each Sectors

1-3. EXPLANATION OUTPUT TARGET

- Check how much each sector will receives the attention on next day based on using previous day’s economy ,USA stock and News data in our training model
- Output Target Sectors selection

  selected sectors that are not related to each others

  SemiConductor, Bio, Car, Gas, Game

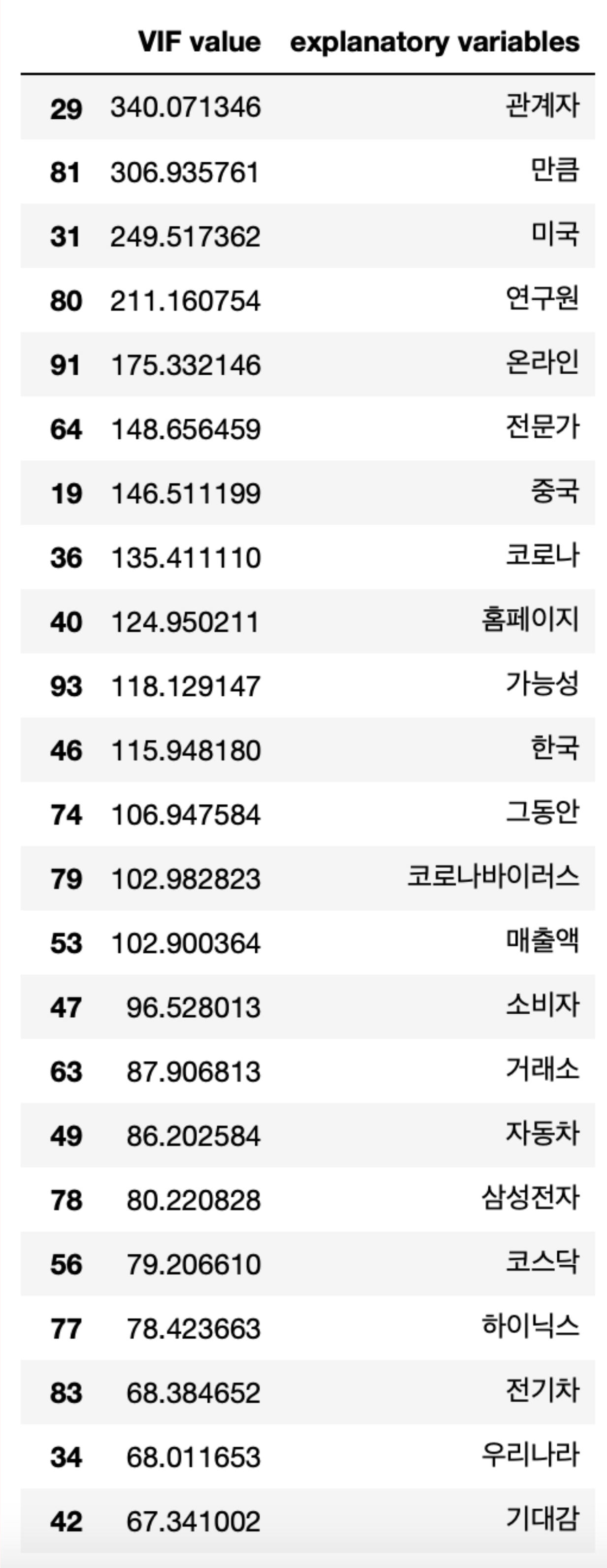
2-1. FINAL SELECTED MODEL

- Base estimator is Linear Regression to predict volume of each sectors in next date
  - Input : word counting , Economy and USA stock market datas
  - Output : for each sectors volumes

- Ensemble (Bagging)
  - Because of small data numbers, we need more training and more accuracy of predicting by adding training models

BY USING FIRSTLY VIF VALUE, REMOVE WORDS

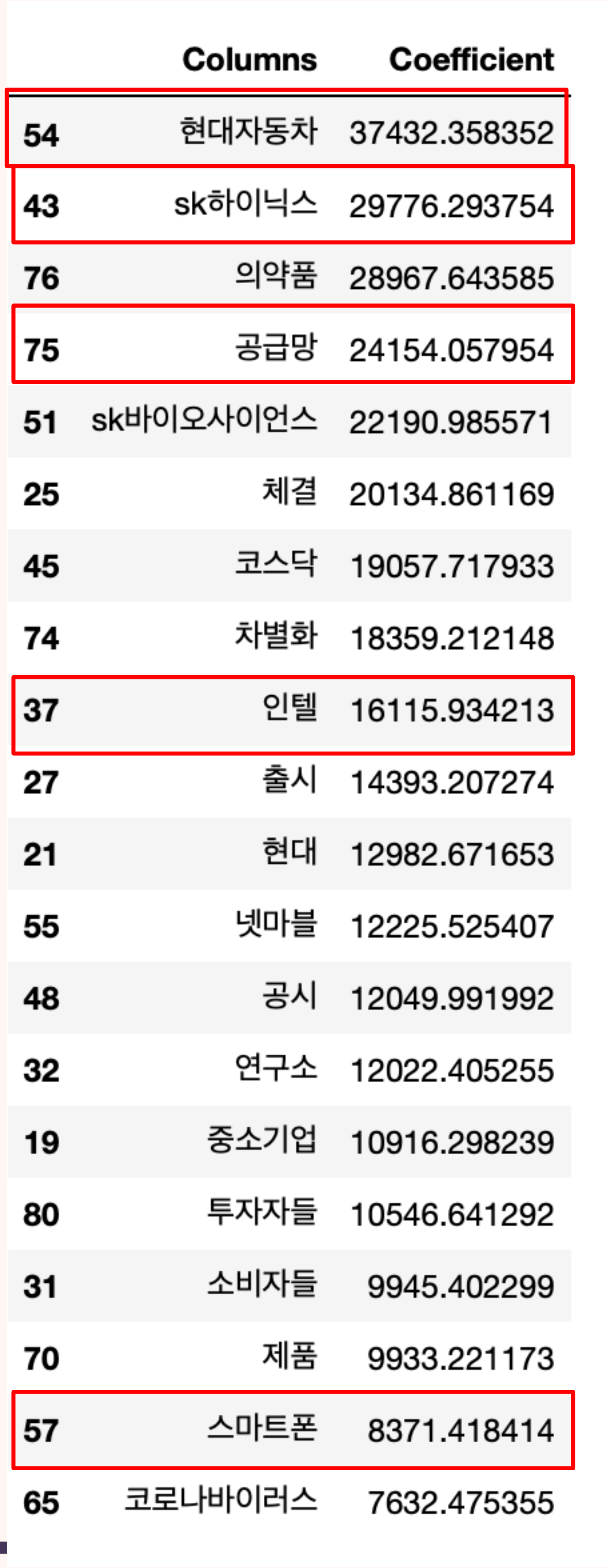
except for these words removed those words ["코로나","코로나바이러스","자동차","삼성전자","코스닥","하이닉스","전기차","기대감"]



	VIF value	explanatory variables
29	340.071346	관계자
81	306.935761	만큼
31	249.517362	미국
80	211.160754	연구원
91	175.332146	온라인
64	148.656459	전문가
19	146.511199	중국
36	135.411110	코로나
40	124.950211	홈페이지
93	118.129147	가능성
46	115.948180	한국
74	106.947584	그동안
79	102.982823	코로나바이러스
53	102.900364	매출액
47	96.528013	소비자
63	87.906813	거래소
49	86.202584	자동차
78	80.220828	삼성전자
56	79.206610	코스닥
77	78.423663	하이닉스
83	68.384652	전기차
34	68.011653	우리나라
42	67.341002	기대감

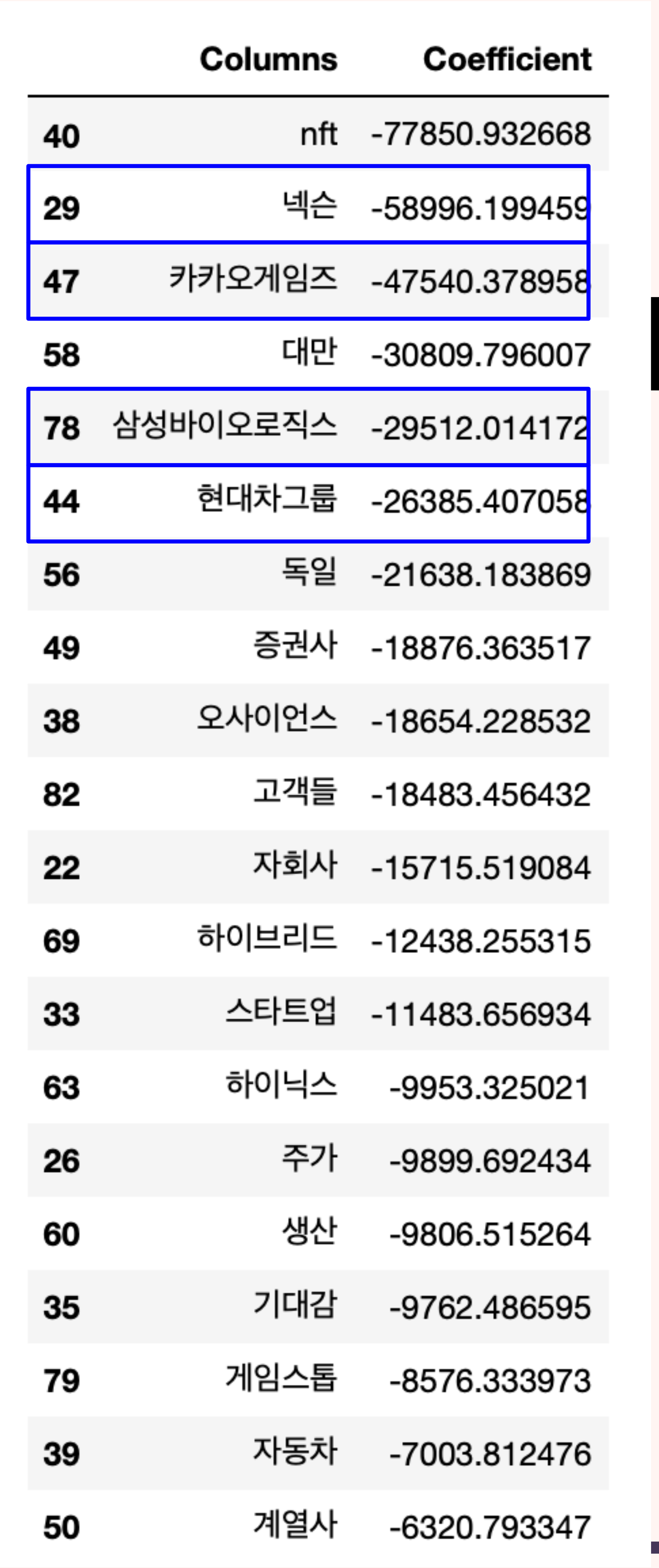
RESULT OF DATA ANALYSIS

BY ENSEMBLE MODEL'S COEFFICIENT OF SEMICONDUCTOR SECTOR



	Columns	Coefficient
54	현대자동차	37432.358352
43	sk하이닉스	29776.293754
76	의약품	28967.643585
75	공급망	24154.057954
51	sk바이오사이언스	22190.985571
25	체결	20134.861169
45	코스닥	19057.717933
74	차별화	18359.212148
37	인텔	16115.934213
27	출시	14393.207274
21	현대	12982.671653
55	넷마블	12225.525407
48	공시	12049.991992
32	연구소	12022.405255
19	중소기업	10916.298239
80	투자자들	10546.641292
31	소비자들	9945.402299
70	제품	9933.221173
57	스마트폰	8371.418414
65	코로나바이러스	7632.475355

Have positive effect

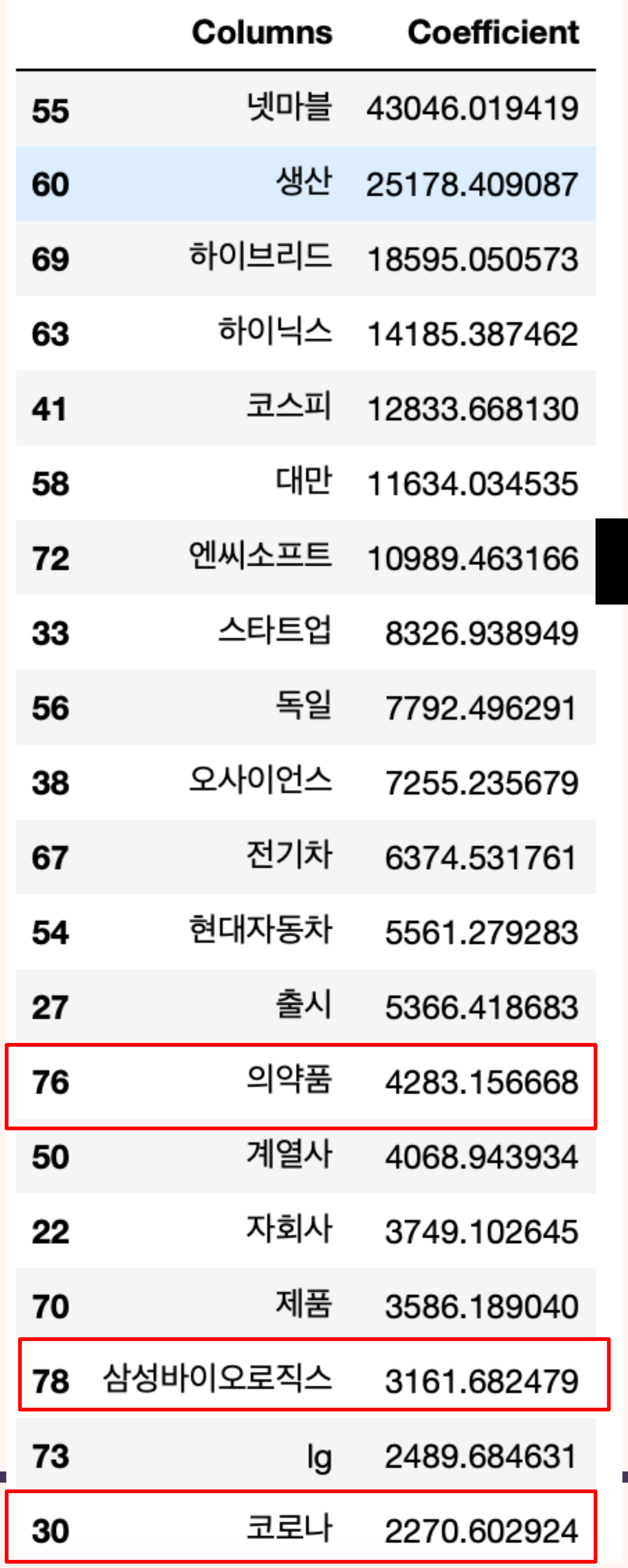


	Columns	Coefficient
40	nft	-77850.932668
29	넥슨	-58996.199459
47	카카오게임즈	-47540.378958
58	대만	-30809.796007
78	삼성바이오로직스	-29512.014172
44	현대차그룹	-26385.407058
56	독일	-21638.183869
49	증권사	-18876.363517
38	오사이언스	-18654.228532
82	고객들	-18483.456432
22	자회사	-15715.519084
69	하이브리드	-12438.255315
33	스타트업	-11483.656934
63	하이닉스	-9953.325021
26	주가	-9899.692434
60	생산	-9806.515264
35	기대감	-9762.486595
79	게임스톱	-8576.333973
39	자동차	-7003.812476
50	계열사	-6320.793347

Negative effect

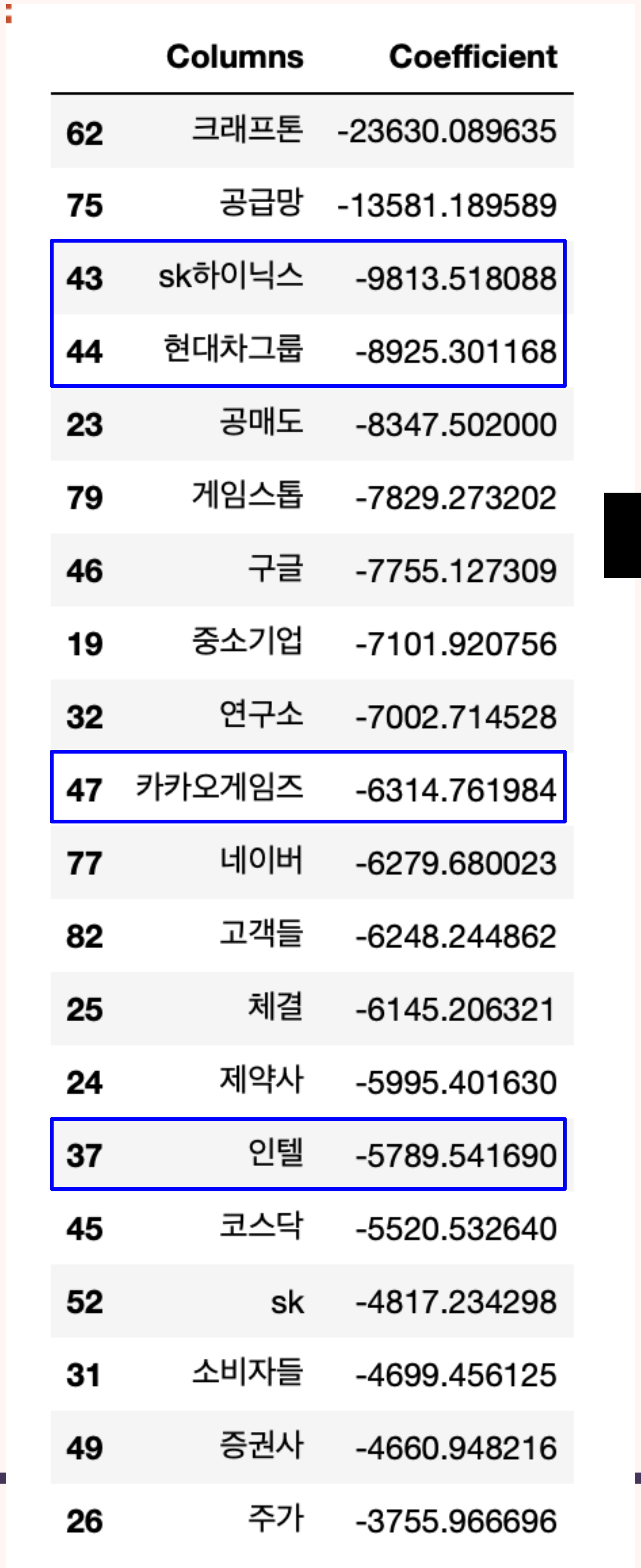
For Other Sectors

BY ENSEMBLE MODEL'S COEFFICIENT OF BIO SECTOR



	Columns	Coefficient
55	넷마블	43046.019419
60	생산	25178.409087
69	하이브리드	18595.050573
63	하이닉스	14185.387462
41	코스피	12833.668130
58	대만	11634.034535
72	엔씨소프트	10989.463166
33	스타트업	8326.938949
56	독일	7792.496291
38	오사이언스	7255.235679
67	전기차	6374.531761
54	현대자동차	5561.279283
27	출시	5366.418683
76	의약품	4283.156668
50	계열사	4068.943934
22	자회사	3749.102645
70	제품	3586.189040
78	삼성바이오로직스	3161.682479
73	lg	2489.684631
30	코로나	2270.602924

Have positive effect

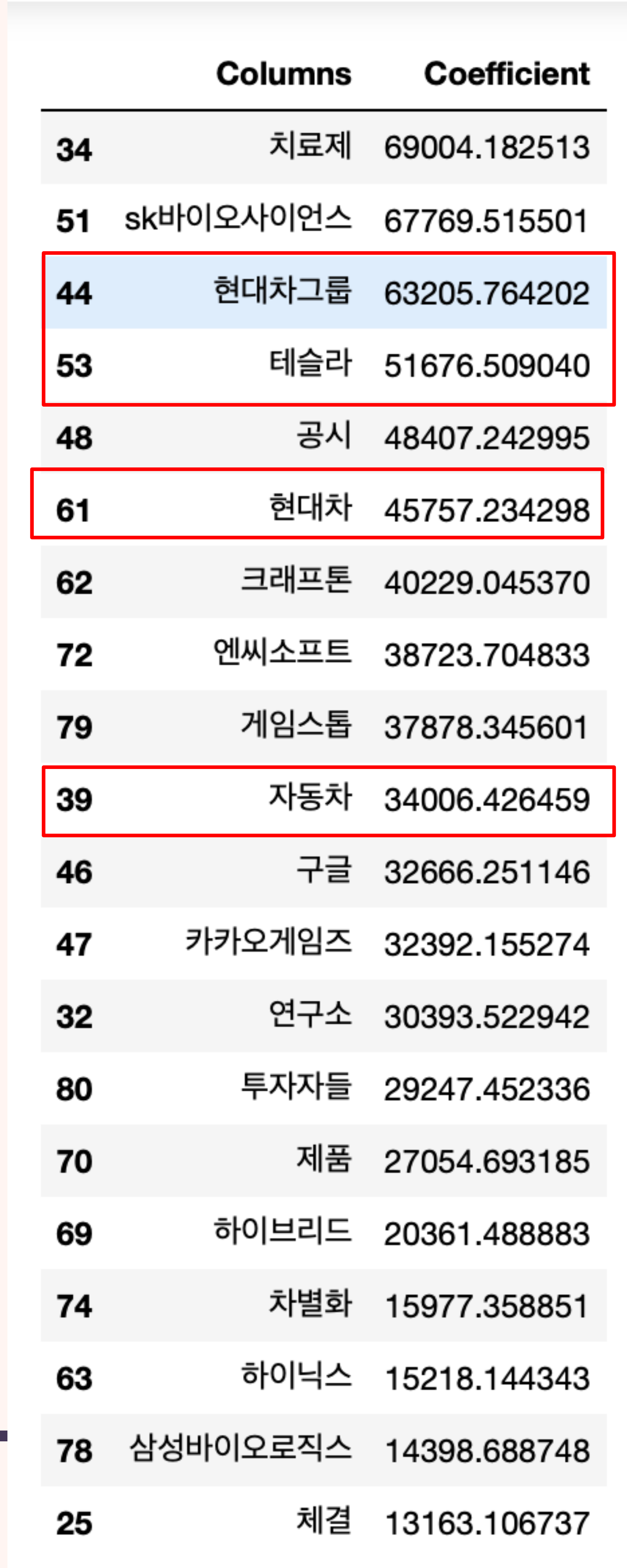


	Columns	Coefficient
62	크래프톤	-23630.089635
75	공급망	-13581.189589
43	sk하이닉스	-9813.518088
44	현대차그룹	-8925.301168
23	공매도	-8347.502000
79	게임스톱	-7829.273202
46	구글	-7755.127309
19	중소기업	-7101.920756
32	연구소	-7002.714528
47	카카오게임즈	-6314.761984
77	네이버	-6279.680023
82	고객들	-6248.244862
25	체결	-6145.206321
24	제약사	-5995.401630
37	인텔	-5789.541690
45	코스닥	-5520.532640
52	sk	-4817.234298
31	소비자들	-4699.456125
49	증권사	-4660.948216
26	주가	-3755.966696

Negative effect

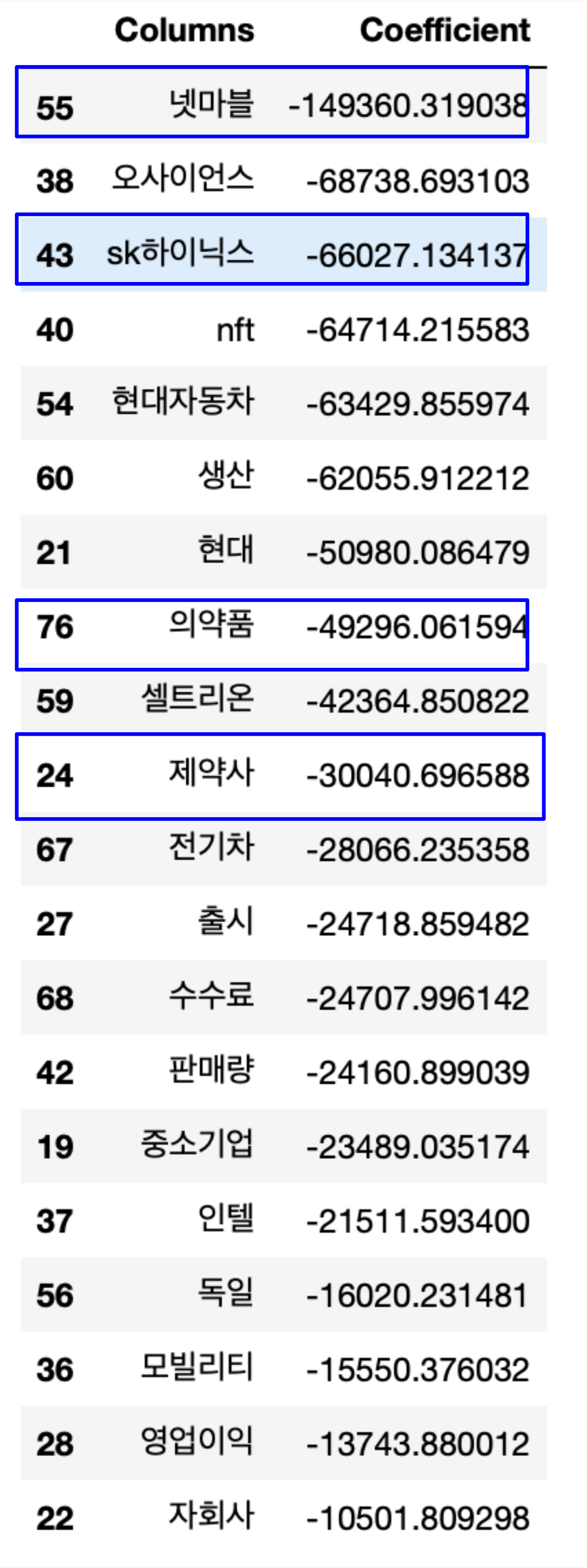
For Other Sectors

BY ENSEMBLE MODEL'S COEFFICIENT OF CAR SECTOR



	Columns	Coefficient
34	치료제	69004.182513
51	sk바이오사이언스	67769.515501
44	현대차그룹	63205.764202
53	테슬라	51676.509040
48	공시	48407.242995
61	현대차	45757.234298
62	크래프톤	40229.045370
72	엔씨소프트	38723.704833
79	게임스톱	37878.345601
39	자동차	34006.426459
46	구글	32666.251146
47	카카오게임즈	32392.155274
32	연구소	30393.522942
80	투자자들	29247.452336
70	제품	27054.693185
69	하이브리드	20361.488883
74	차별화	15977.358851
63	하이닉스	15218.144343
78	삼성바이오로직스	14398.688748
25	체결	13163.106737

Have positive effect

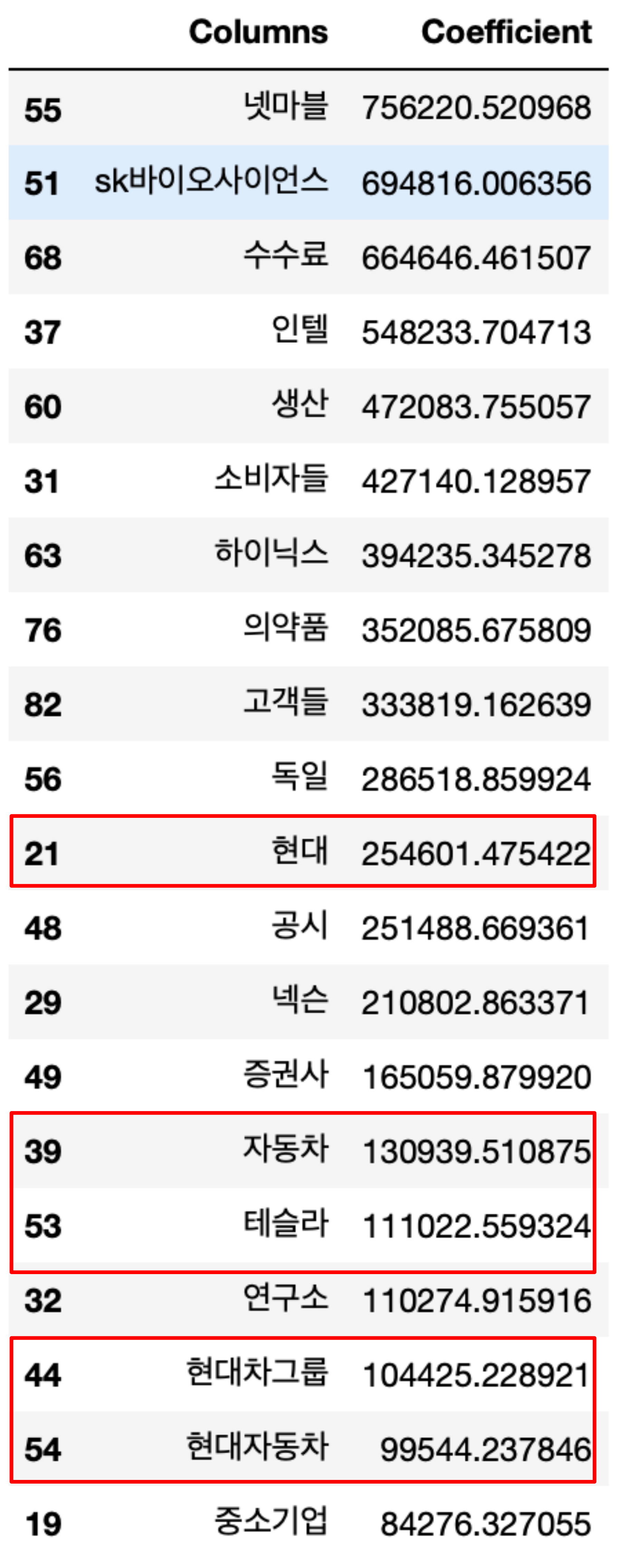


	Columns	Coefficient
55	넷마블	-149360.319038
38	오사이언스	-68738.693103
43	sk하이닉스	-66027.134137
40	nft	-64714.215583
54	현대자동차	-63429.855974
60	생산	-62055.912212
21	현대	-50980.086479
76	의약품	-49296.061594
59	셀트리온	-42364.850822
24	제약사	-30040.696588
67	전기차	-28066.235358
27	출시	-24718.859482
68	수수료	-24707.996142
42	판매량	-24160.899039
19	중소기업	-23489.035174
37	인텔	-21511.593400
56	독일	-16020.231481
36	모빌리티	-15550.376032
28	영업이익	-13743.880012
22	자회사	-10501.809298

Negative effect

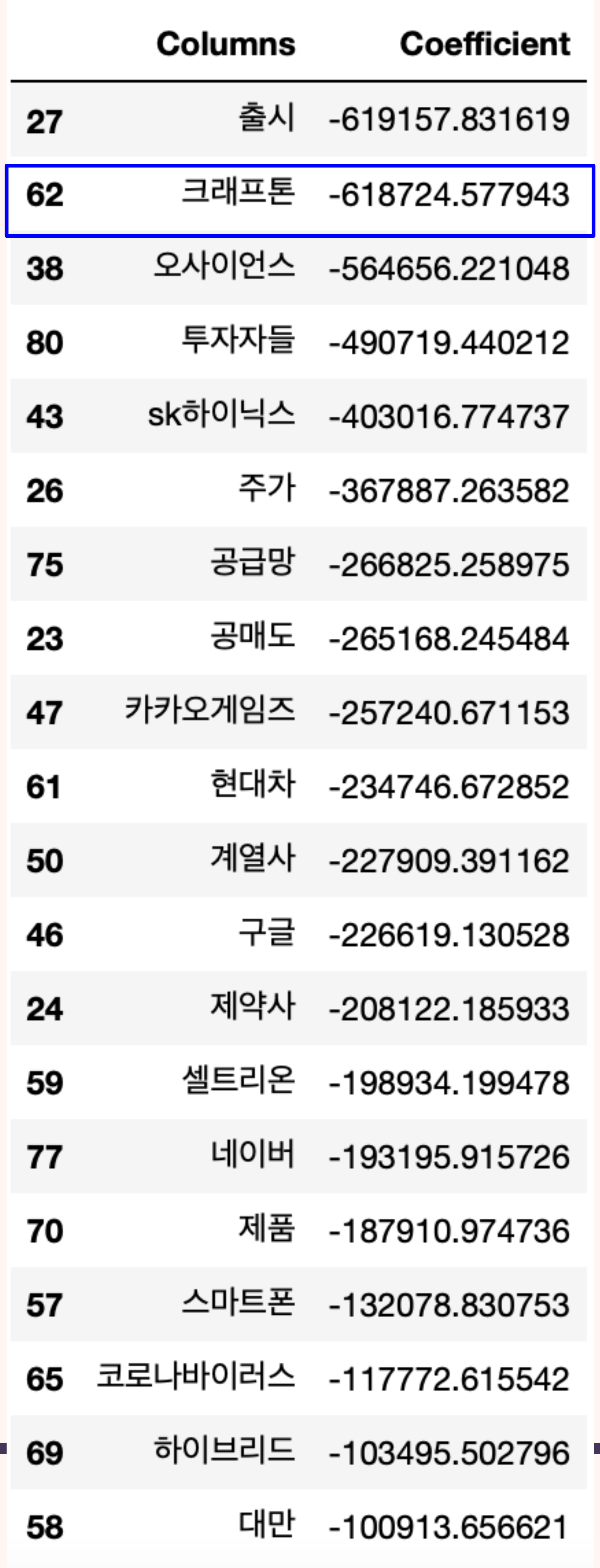
For Other Sectors

BY ENSEMBLE MODEL'S COEFFICIENT OF GAS SECTOR



	Columns	Coefficient
55	넷마블	756220.520968
51	sk바이오사이언스	694816.006356
68	수수료	664646.461507
37	인텔	548233.704713
60	생산	472083.755057
31	소비자들	427140.128957
63	하이닉스	394235.345278
76	의약품	352085.675809
82	고객들	333819.162639
56	독일	286518.859924
21	현대	254601.475422
48	공시	251488.669361
29	넥슨	210802.863371
49	증권사	165059.879920
39	자동차	130939.510875
53	테슬라	111022.559324
32	연구소	110274.915916
44	현대차그룹	104425.228921
54	현대자동차	99544.237846
19	중소기업	84276.327055

Have positive effect with car sector

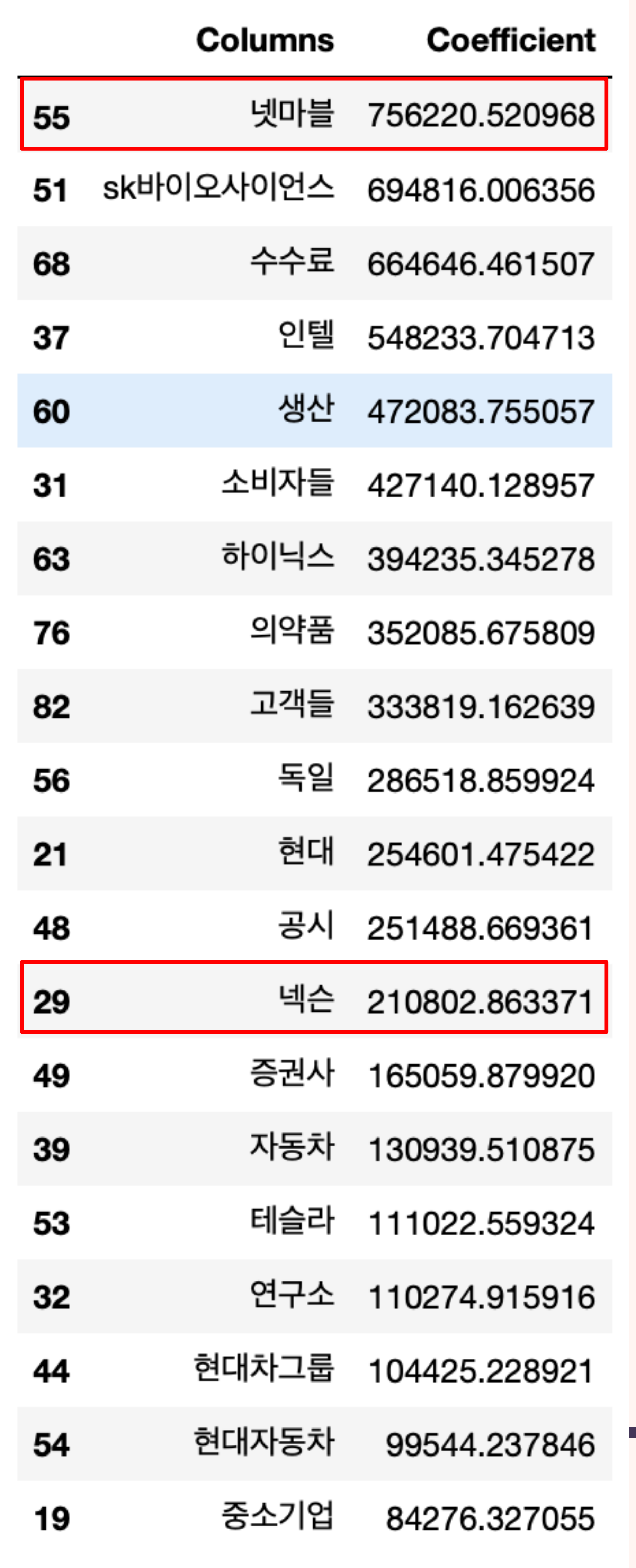


	Columns	Coefficient
27	출시	-619157.831619
62	크래프톤	-618724.577943
38	오사이언스	-564656.221048
80	투자자들	-490719.440212
43	sk하이닉스	-403016.774737
26	주가	-367887.263582
75	공급망	-266825.258975
23	공매도	-265168.245484
47	카카오게임즈	-257240.671153
61	현대차	-234746.672852
50	계열사	-227909.391162
46	구글	-226619.130528
24	제약사	-208122.185933
59	셀트리온	-198934.199478
77	네이버	-193195.915726
70	제품	-187910.974736
57	스마트폰	-132078.830753
65	코로나바이러스	-117772.615542
69	하이브리드	-103495.502796
58	대만	-100913.656621

Negative effect

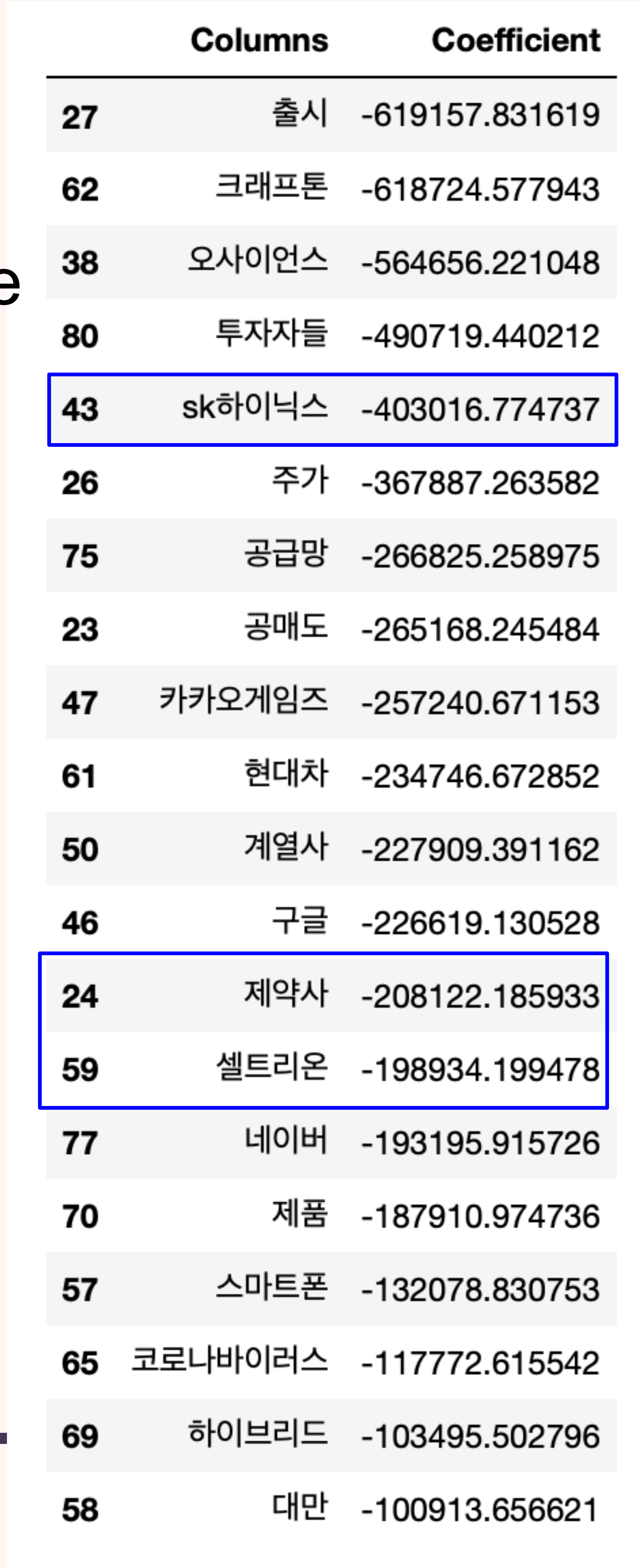
For Other Sectors

BY ENSEMBLE MODEL'S COEFFICIENT OF GAME SECTOR



	Columns	Coefficient
55	넷마블	756220.520968
51	sk바이오사이언스	694816.006356
68	수수료	664646.461507
37	인텔	548233.704713
60	생산	472083.755057
31	소비자들	427140.128957
63	하이닉스	394235.345278
76	의약품	352085.675809
82	고객들	333819.162639
56	독일	286518.859924
21	현대	254601.475422
48	공시	251488.669361
29	넥슨	210802.863371
49	증권사	165059.879920
39	자동차	130939.510875
53	테슬라	111022.559324
32	연구소	110274.915916
44	현대차그룹	104425.228921
54	현대자동차	99544.237846
19	중소기업	84276.327055

Have positive effect

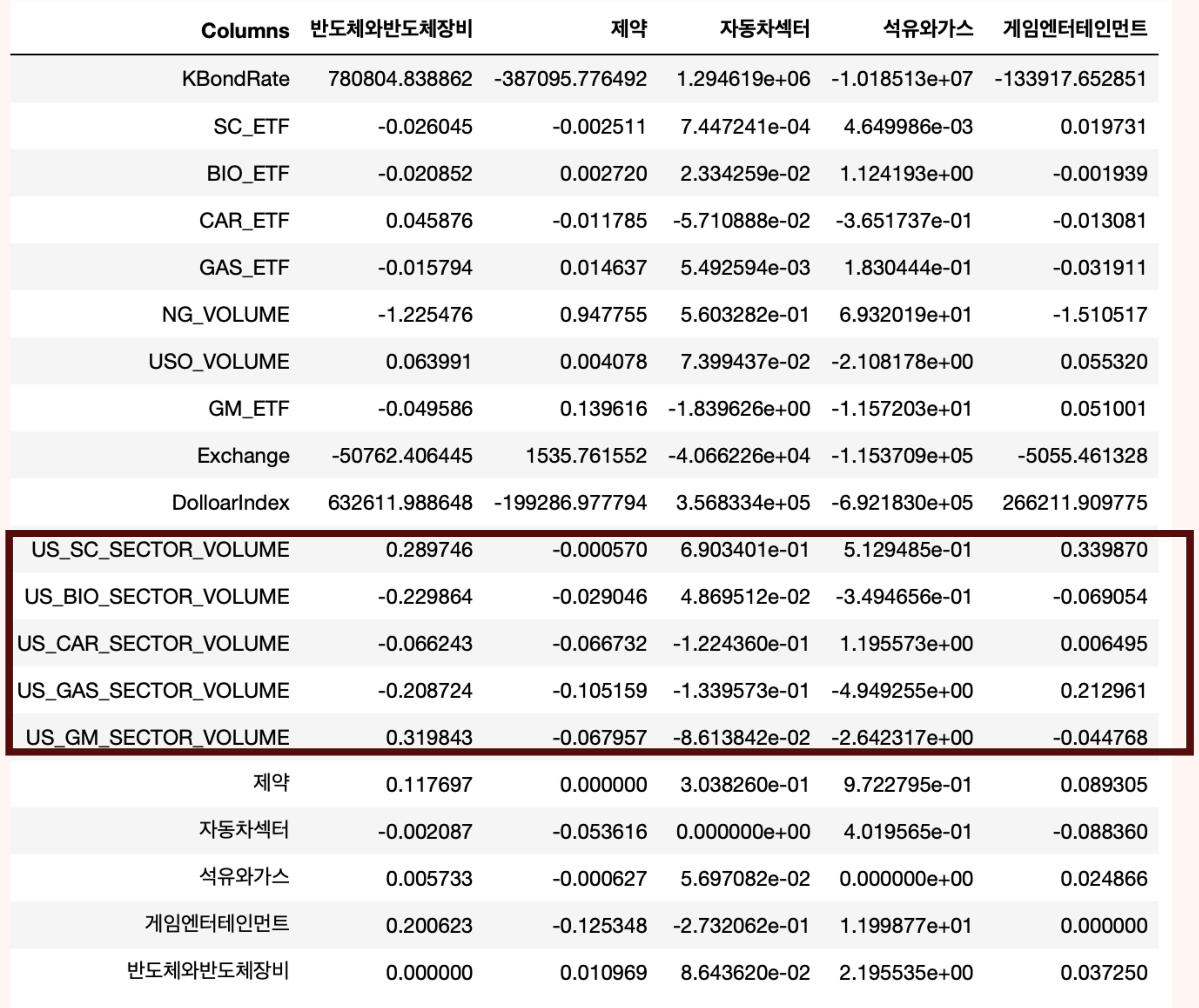


	Columns	Coefficient
27	출시	-619157.831619
62	크래프톤	-618724.577943
38	오사이언스	-564656.221048
80	투자자들	-490719.440212
43	sk하이닉스	-403016.774737
26	주가	-367887.263582
75	공급망	-266825.258975
23	공매도	-265168.245484
47	카카오게임즈	-257240.671153
61	현대차	-234746.672852
50	계열사	-227909.391162
46	구글	-226619.130528
24	제약사	-208122.185933
59	셀트리온	-198934.199478
77	네이버	-193195.915726
70	제품	-187910.974736
57	스마트폰	-132078.830753
65	코로나바이러스	-117772.615542
69	하이브리드	-103495.502796
58	대만	-100913.656621

Negative effect

For Other Sectors

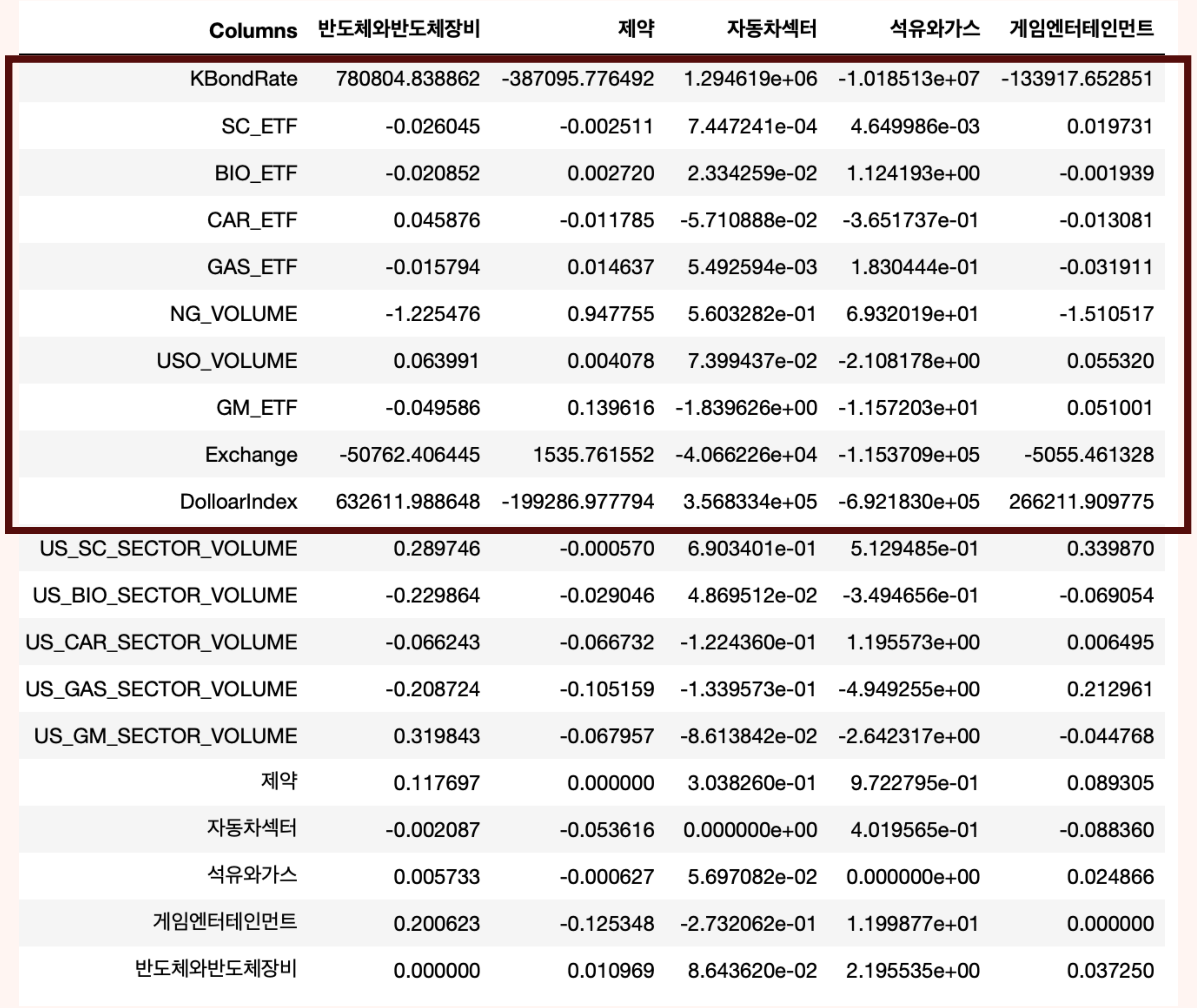
BY ENSEMBLE MODEL'S COEFFICIENT ABOUT ECONOMY AND USA DATA



Columns	반도체와반도체장비	제약	자동차섹터	석유와가스	게임엔터테인먼트
KBondRate	780804.838862	-387095.776492	1.294619e+06	-1.018513e+07	-133917.652851
SC_ETF	-0.026045	-0.002511	7.447241e-04	4.649986e-03	0.019731
BIO_ETF	-0.020852	0.002720	2.334259e-02	1.124193e+00	-0.001939
CAR_ETF	0.045876	-0.011785	-5.710888e-02	-3.651737e-01	-0.013081
GAS_ETF	-0.015794	0.014637	5.492594e-03	1.830444e-01	-0.031911
NG_VOLUME	-1.225476	0.947755	5.603282e-01	6.932019e+01	-1.510517
USO_VOLUME	0.063991	0.004078	7.399437e-02	-2.108178e+00	0.055320
GM_ETF	-0.049586	0.139616	-1.839626e+00	-1.157203e+01	0.051001
Exchange	-50762.406445	1535.761552	-4.066226e+04	-1.153709e+05	-5055.461328
DolloarIndex	632611.988648	-199286.977794	3.568334e+05	-6.921830e+05	266211.909775
US_SC_SECTOR_VOLUME	0.289746	-0.000570	6.903401e-01	5.129485e-01	0.339870
US_BIO_SECTOR_VOLUME	-0.229864	-0.029046	4.869512e-02	-3.494656e-01	-0.069054
US_CAR_SECTOR_VOLUME	-0.066243	-0.066732	-1.224360e-01	1.195573e+00	0.006495
US_GAS_SECTOR_VOLUME	-0.208724	-0.105159	-1.339573e-01	-4.949255e+00	0.212961
US_GM_SECTOR_VOLUME	0.319843	-0.067957	-8.613842e-02	-2.642317e+00	-0.044768
제약	0.117697	0.000000	3.038260e-01	9.722795e-01	0.089305
자동차섹터	-0.002087	-0.053616	0.000000e+00	4.019565e-01	-0.088360
석유와가스	0.005733	-0.000627	5.697082e-02	0.000000e+00	0.024866
게임엔터테인먼트	0.200623	-0.125348	-2.732062e-01	1.199877e+01	0.000000
반도체와반도체장비	0.000000	0.010969	8.643620e-02	2.195535e+00	0.037250

- **When we see coefficient data of USA's each sector's transaction volumes and our output targets, coefficient value of same sector (if usa = semi and Korea = "반도체와장비) don't have usually same direction (positive coefficient).**

BY ENSEMBLE MODEL'S COEFFICIENT ABOUT ECONOMY AND USA DATA

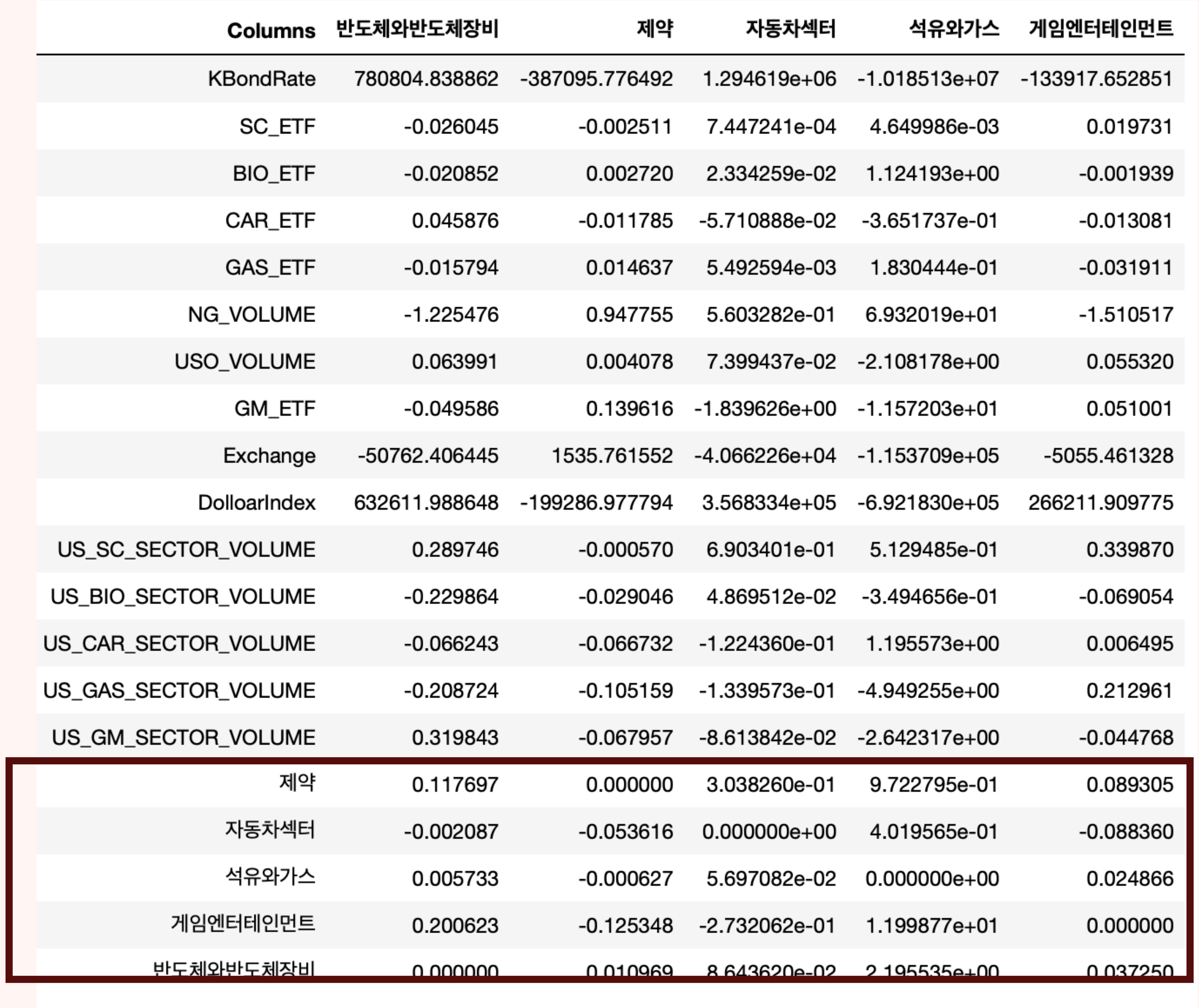


Columns	반도체와반도체장비	제약	자동차섹터	석유와가스	게임엔터테인먼트
KBondRate	780804.838862	-387095.776492	1.294619e+06	-1.018513e+07	-133917.652851
SC_ETF	-0.026045	-0.002511	7.447241e-04	4.649986e-03	0.019731
BIO_ETF	-0.020852	0.002720	2.334259e-02	1.124193e+00	-0.001939
CAR_ETF	0.045876	-0.011785	-5.710888e-02	-3.651737e-01	-0.013081
GAS_ETF	-0.015794	0.014637	5.492594e-03	1.830444e-01	-0.031911
NG_VOLUME	-1.225476	0.947755	5.603282e-01	6.932019e+01	-1.510517
USO_VOLUME	0.063991	0.004078	7.399437e-02	-2.108178e+00	0.055320
GM_ETF	-0.049586	0.139616	-1.839626e+00	-1.157203e+01	0.051001
Exchange	-50762.406445	1535.761552	-4.066226e+04	-1.153709e+05	-5055.461328
DolloarIndex	632611.988648	-199286.977794	3.568334e+05	-6.921830e+05	266211.909775
US_SC_SECTOR_VOLUME	0.289746	-0.000570	6.903401e-01	5.129485e-01	0.339870
US_BIO_SECTOR_VOLUME	-0.229864	-0.029046	4.869512e-02	-3.494656e-01	-0.069054
US_CAR_SECTOR_VOLUME	-0.066243	-0.066732	-1.224360e-01	1.195573e+00	0.006495
US_GAS_SECTOR_VOLUME	-0.208724	-0.105159	-1.339573e-01	-4.949255e+00	0.212961
US_GM_SECTOR_VOLUME	0.319843	-0.067957	-8.613842e-02	-2.642317e+00	-0.044768
제약	0.117697	0.000000	3.038260e-01	9.722795e-01	0.089305
자동차섹터	-0.002087	-0.053616	0.000000e+00	4.019565e-01	-0.088360
석유와가스	0.005733	-0.000627	5.697082e-02	0.000000e+00	0.024866
게임엔터테인먼트	0.200623	-0.125348	-2.732062e-01	1.199877e+01	0.000000
반도체와반도체장비	0.000000	0.010969	8.643620e-02	2.195535e+00	0.037250

- When we see coefficient data of USA's each sector's ETF transaction volumes and our output targets, coefficient value of same sector ETF and output target is positive with bio , gas , game sector's ETF.

- And KbondRate, DollarIndex and Exchange reflect economy situation. Commonly KbondRate & Dollar index increase mean inflation increase and this mean investment sentiment decrease. By depending on these coefficient value , except "Semi sector and car sector, other sectors have negative coefficient with those variables.

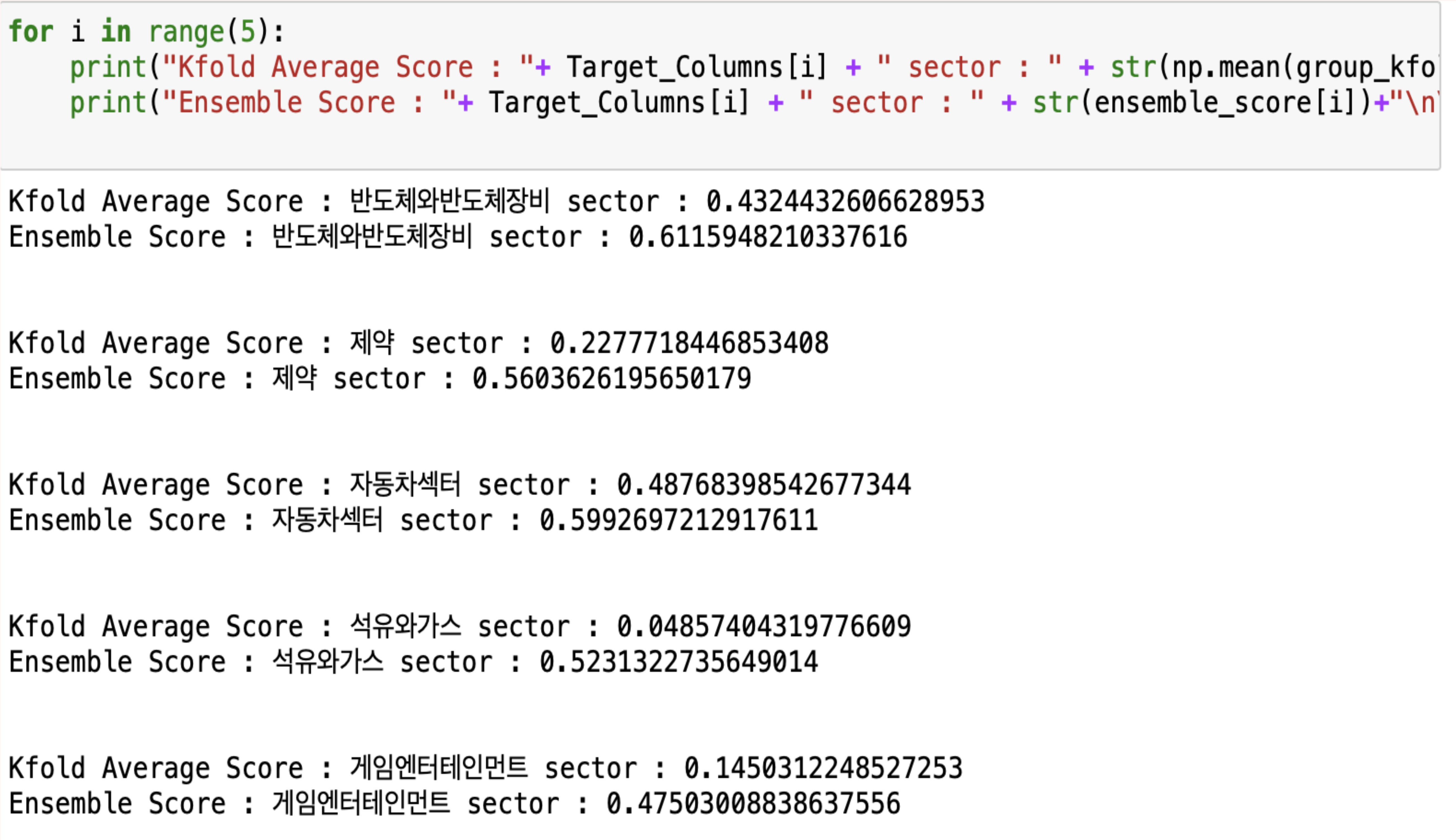
BY ENSEMBLE MODEL'S COEFFICIENT ABOUT ECONOMY AND USA DATA



Columns	반도체와반도체장비	제약	자동차섹터	석유와가스	게임엔터테인먼트
KBondRate	780804.838862	-387095.776492	1.294619e+06	-1.018513e+07	-133917.652851
SC_ETF	-0.026045	-0.002511	7.447241e-04	4.649986e-03	0.019731
BIO_ETF	-0.020852	0.002720	2.334259e-02	1.124193e+00	-0.001939
CAR_ETF	0.045876	-0.011785	-5.710888e-02	-3.651737e-01	-0.013081
GAS_ETF	-0.015794	0.014637	5.492594e-03	1.830444e-01	-0.031911
NG_VOLUME	-1.225476	0.947755	5.603282e-01	6.932019e+01	-1.510517
USO_VOLUME	0.063991	0.004078	7.399437e-02	-2.108178e+00	0.055320
GM_ETF	-0.049586	0.139616	-1.839626e+00	-1.157203e+01	0.051001
Exchange	-50762.406445	1535.761552	-4.066226e+04	-1.153709e+05	-5055.461328
DolloarIndex	632611.988648	-199286.977794	3.568334e+05	-6.921830e+05	266211.909775
US_SC_SECTOR_VOLUME	0.289746	-0.000570	6.903401e-01	5.129485e-01	0.339870
US_BIO_SECTOR_VOLUME	-0.229864	-0.029046	4.869512e-02	-3.494656e-01	-0.069054
US_CAR_SECTOR_VOLUME	-0.066243	-0.066732	-1.224360e-01	1.195573e+00	0.006495
US_GAS_SECTOR_VOLUME	-0.208724	-0.105159	-1.339573e-01	-4.949255e+00	0.212961
US_GM_SECTOR_VOLUME	0.319843	-0.067957	-8.613842e-02	-2.642317e+00	-0.044768
제약	0.117697	0.000000	3.038260e-01	9.722795e-01	0.089305
자동차섹터	-0.002087	-0.053616	0.000000e+00	4.019565e-01	-0.088360
석유와가스	0.005733	-0.000627	5.697082e-02	0.000000e+00	0.024866
게임엔터테인먼트	0.200623	-0.125348	-2.732062e-01	1.199877e+01	0.000000
반도체와반도체장비	0.000000	0.010969	8.643620e-02	2.195535e+00	0.037250

- **By using coefficient between each sectors volume(our target datas) , there are some meaning.**
- **"Bio Sector" volume increase affect all target sector's volume increase.**
- **"Car sector" volume increase affect only "GAS" sector volume increase but others decrease.**
- **"Game Sector" volume increase affect Gas and Semi sector volume increase but others decrease.**
- **"Semi Sector" volume increase affect all target sector's volume increase**

SCORE OF GROUP KFOLD AND ENSEMBLE



```
for i in range(5):
    print("Kfold Average Score : "+ Target_Columns[i] + " sector : " + str(np.mean(group_kfo
    print("Ensemble Score : "+ Target_Columns[i] + " sector : " + str(ensemble_score[i])+"\n
```

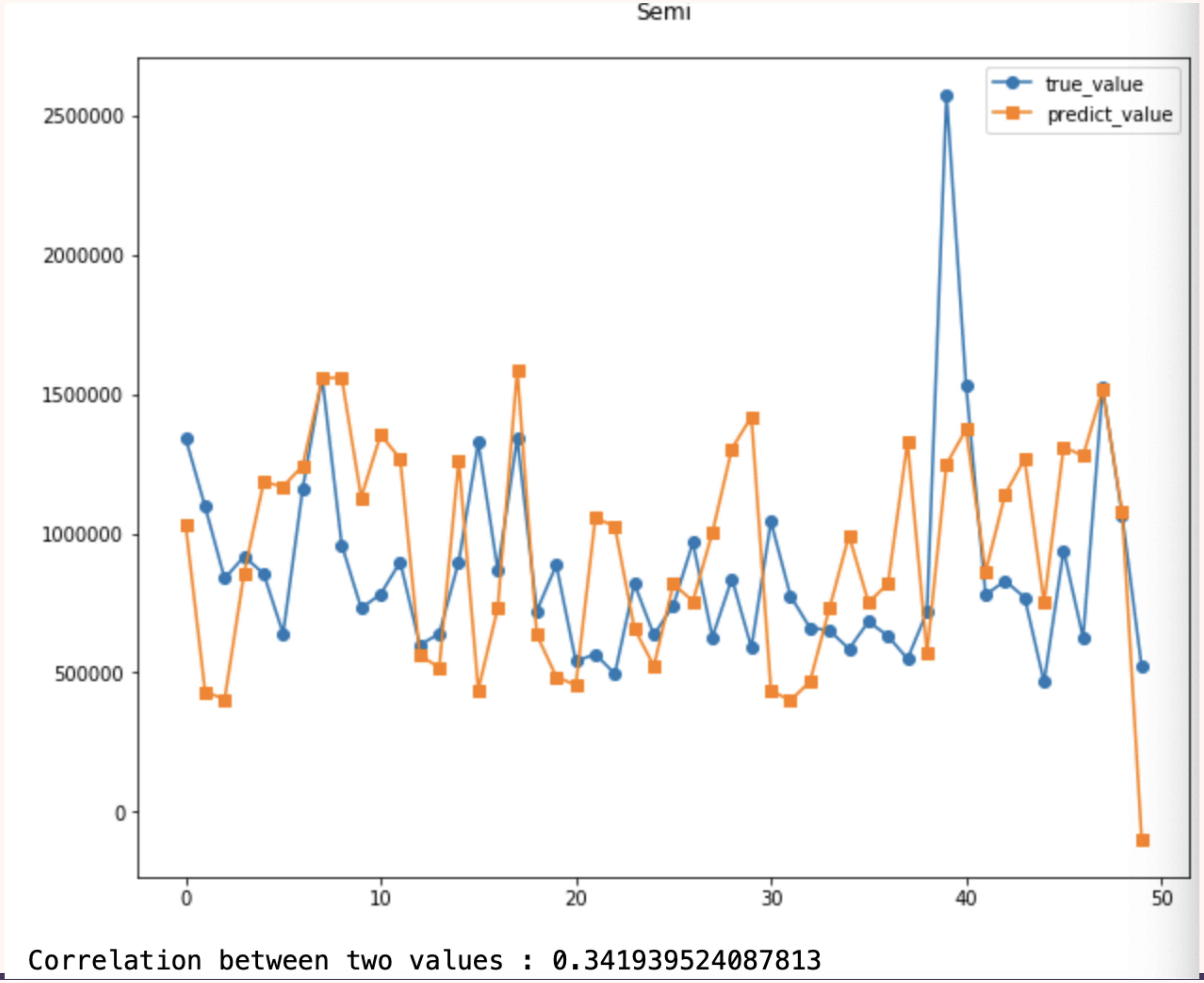
```
Kfold Average Score : 반도체와반도체장비 sector : 0.4324432606628953
Ensemble Score : 반도체와반도체장비 sector : 0.6115948210337616


Kfold Average Score : 제약 sector : 0.2277718446853408
Ensemble Score : 제약 sector : 0.560362619565O179


Kfold Average Score : 자동차섹터 sector : 0.4876839854267734
Ensemble Score : 자동차섹터 sector : 0.5992697212917611


Kfold Average Score : 석유와가스 sector : 0.04857404319776609
Ensemble Score : 석유와가스 sector : 0.5231322735649014


Kfold Average Score : 게임엔터테인먼트 sector : 0.1450312248527253
Ensemble Score : 게임엔터테인먼트 sector : 0.47503008838637556
```

We make each sector dataset group columns.

We group by Target Values by using 1st, 2nd and third quantile values. So 4 groups are made.

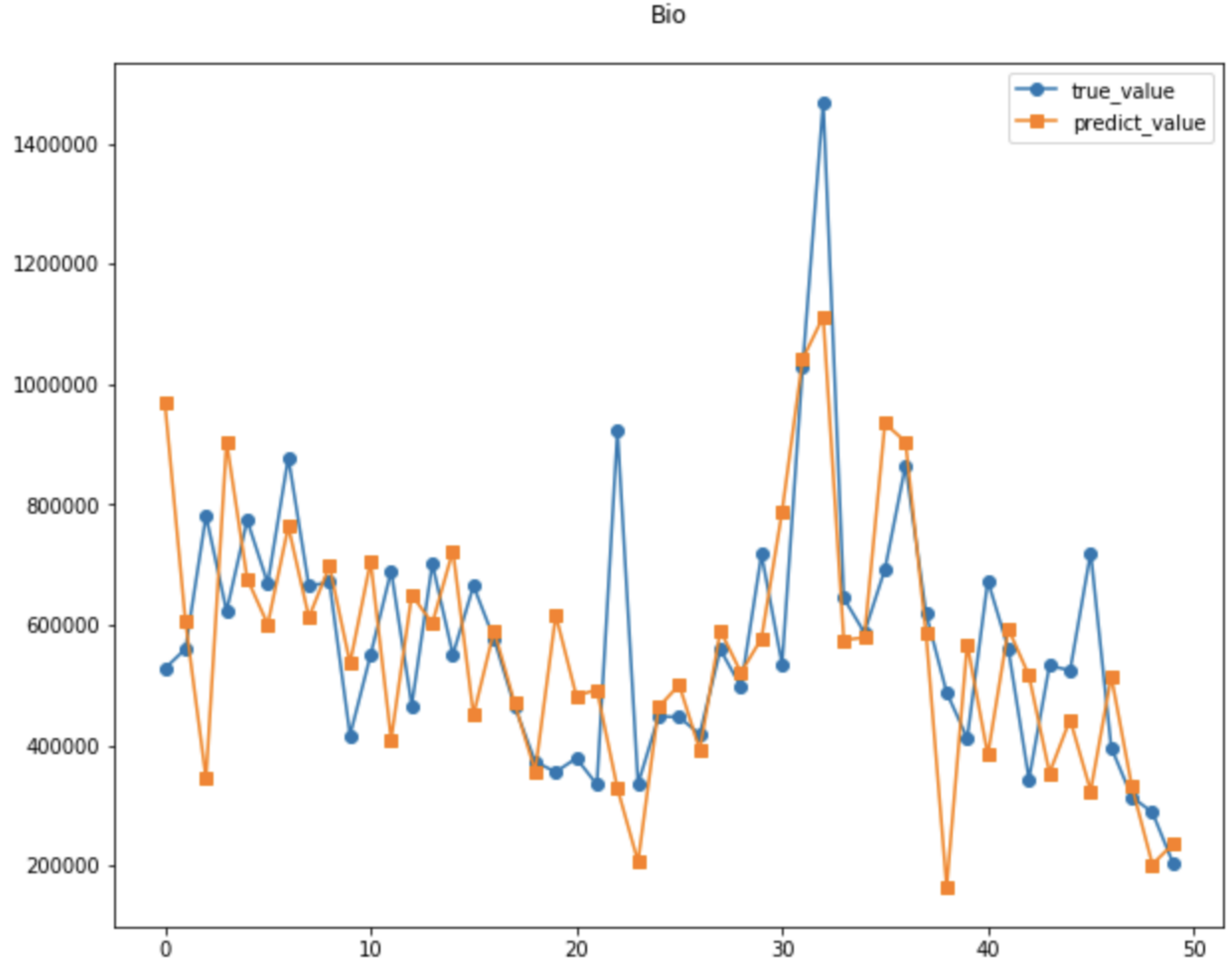
The reason why we use group k fold is higher volume prediction and low volume prediction are important. So we know how well predict volume that have high or low or median.

SEMICONDUCTOR SECTOR'S GRAPH



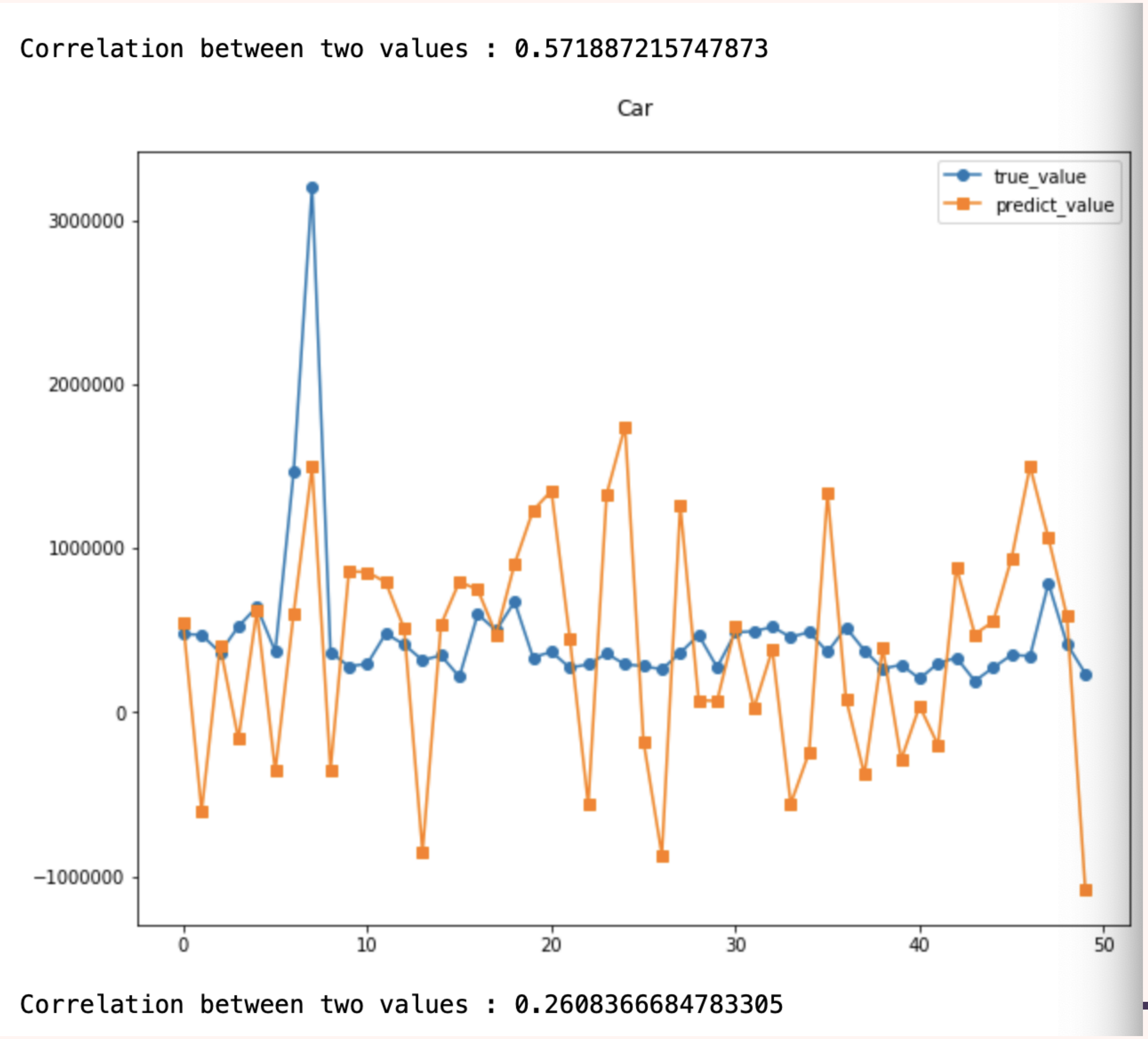
```
Correlation between two values : 0.341939524087813
```

BIO SECTOR'S GRAPH



```
Correlation between two values : 0.571887215747873
```

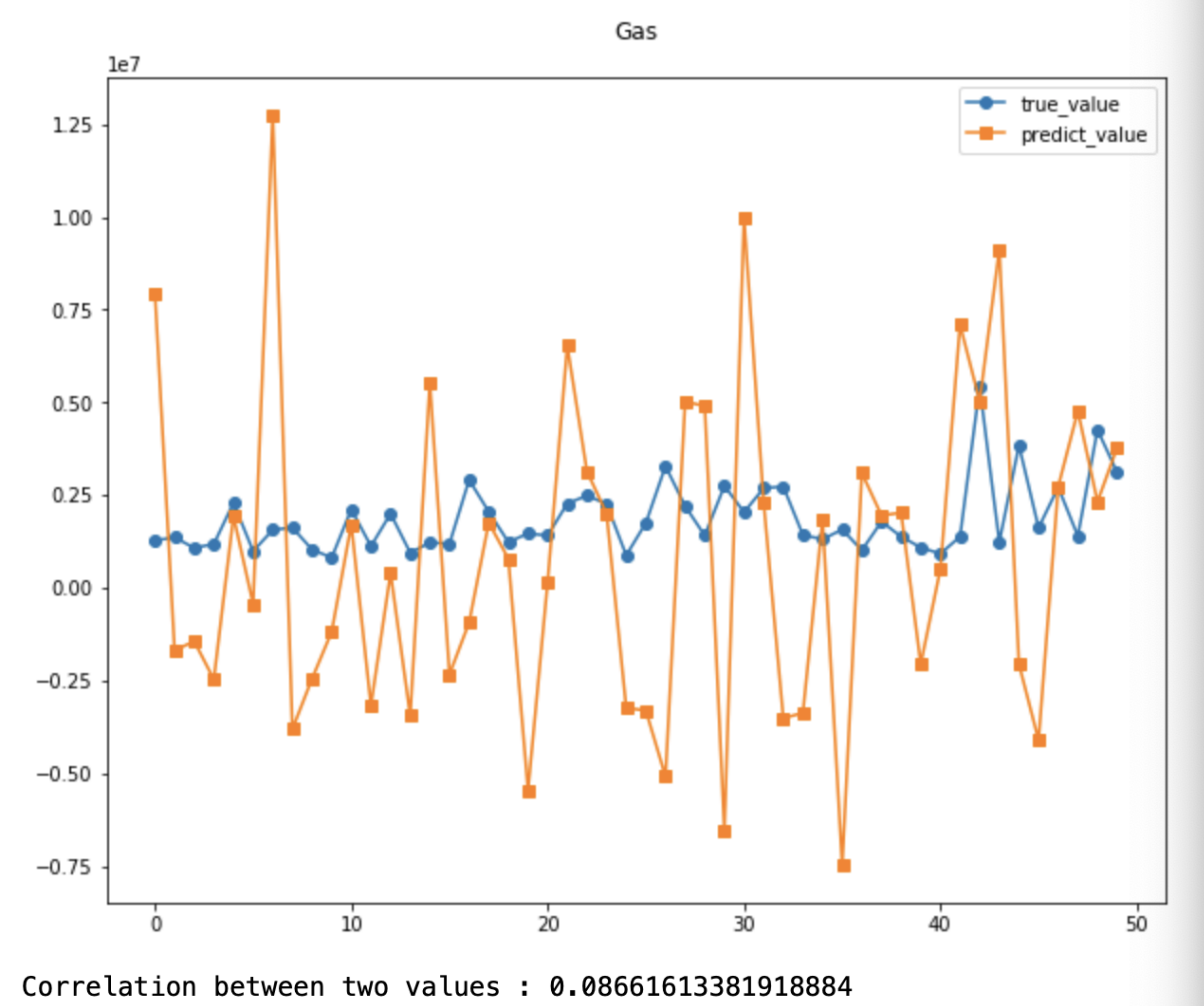
CAR SECTOR'S GRAPH



```
Correlation between two values : 0.571887215747873
```

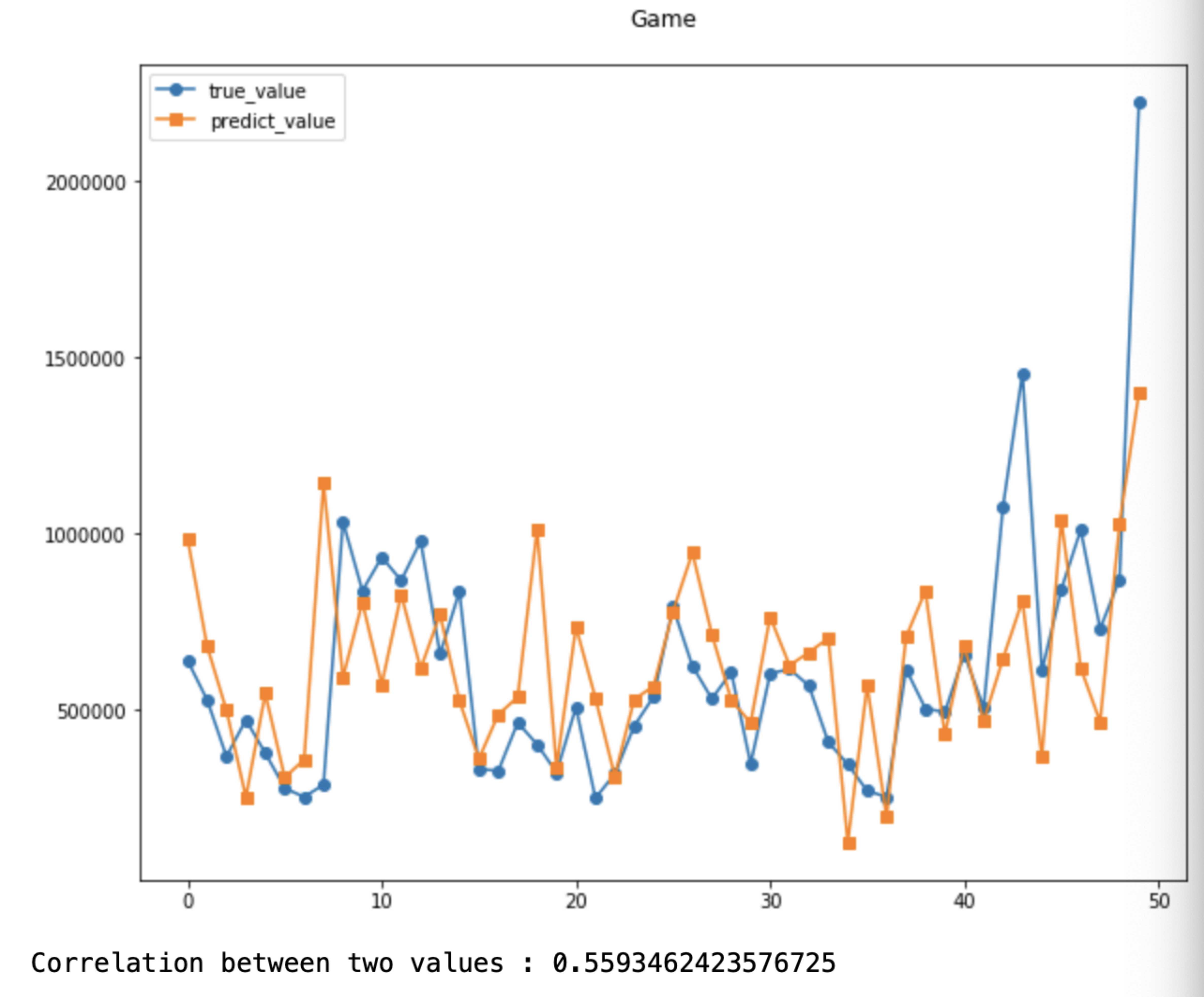
```
Correlation between two values : 0.2608366684783305
```

GAS SECTOR'S GRAPH



```
Correlation between two values : 0.08661613381918884
```

GAME SECTOR'S GRAPH



```
Correlation between two values : 0.5593462423576725
```

RESULT OF PROJECT

- **Our Analysis don’t predict Gas and Bio sector’s Target Values.**
- **News data from a day ago and U.S. stock data are used as input data, so if we increase performance, you can predict the target's interest the next day.**
- **Find Useful word set for each sectors.**

RESULT OF PROJECT

Highly informative news words

In Semiconductor sector -> 인텔 , sk하이닉스, 스마트폰, 넥슨, 삼성바이오로직스 ,현대자동차

in Bio sector -> 의약품 ,삼성바이오로직스 ,코로나

in Car sector are 현대차 ,자동차 , 하이브리드, 현대차그룹, 테슬라

in Gas sector -> 현대, 자동차, 테슬라, 현대차그룹, 현대자동차

in Game sector -> 넷마블, 넥슨

LIMITATION OF ANALYSIS

- **Step of Processing word data**
  - **Lack of data**
  - **Preprocessing of word data is weak. we need more study about it not only use mutual info classification , coefficient and VIF but to use also others**
- **Whole data set samples more needed and other Economy data is more needed.**