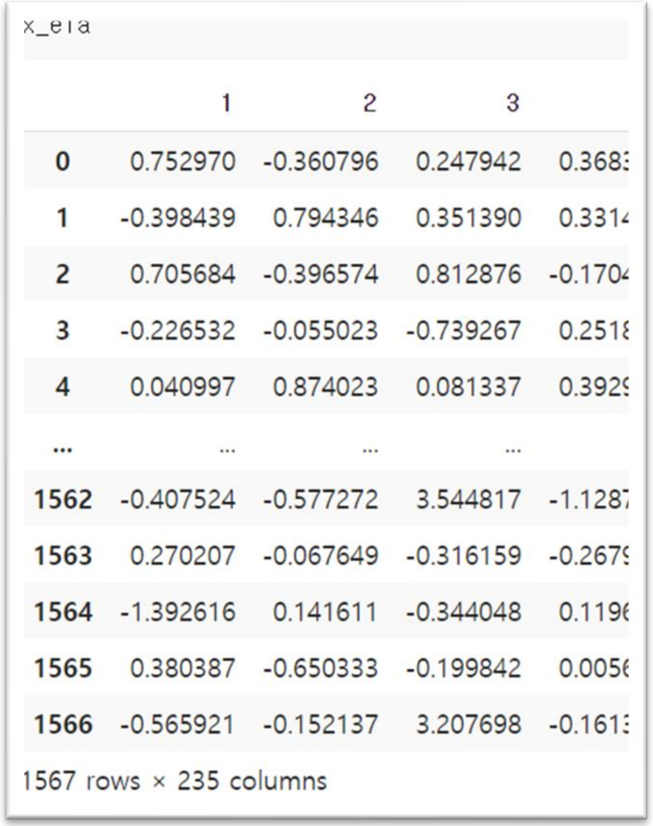
BigData Analysis
SECOM

3조 | 김규연 정연서 정유미 정지윤

HANYANG UNIV. (ERICA)
INDUSTRIAL & MANAGEMENT ENGINEERING

1. Feature Selection

[Elastic Net]

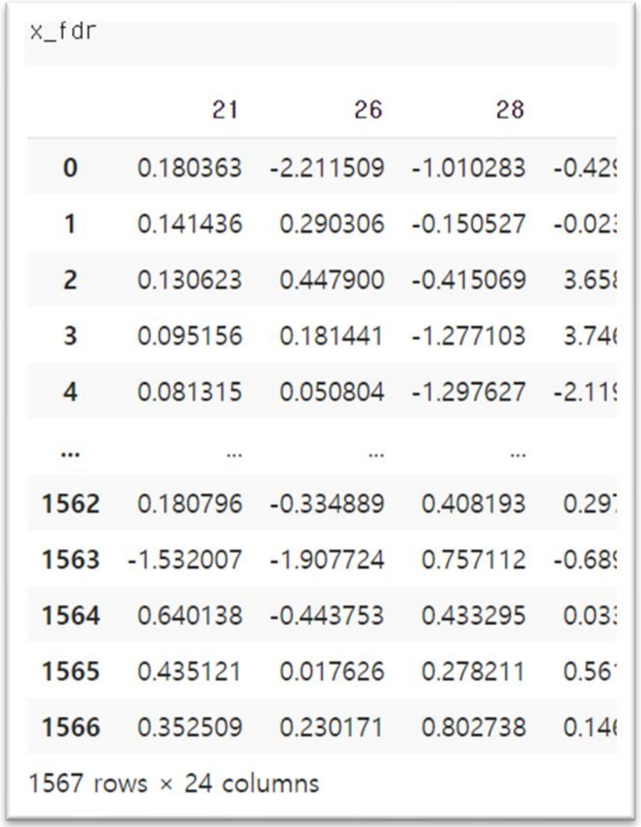


x_ela

	1	2	3	
0	0.752970	-0.360796	0.247942	0.368
1	-0.398439	0.794346	0.351390	0.331
2	0.705684	-0.396574	0.812876	-0.170
3	-0.226532	-0.055023	-0.739267	0.251
4	0.040997	0.874023	0.081337	0.392
...	...	...	...	
1562	-0.407524	-0.577272	3.544817	-1.128
1563	0.270207	-0.067649	-0.316159	-0.267
1564	-1.392616	0.141611	-0.344048	0.119
1565	0.380387	-0.650333	-0.199842	0.005
1566	-0.565921	-0.152137	3.207698	-0.161

1567 rows × 235 columns

[FDR]

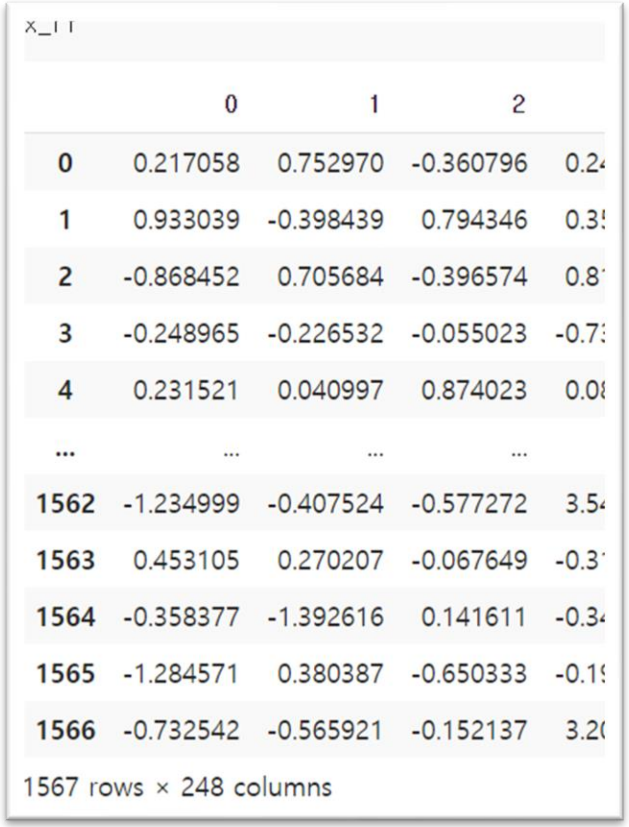


x_fdr

	21	26	28	
0	0.180363	-2.211509	-1.010283	-0.42
1	0.141436	0.290306	-0.150527	-0.02
2	0.130623	0.447900	-0.415069	3.65
3	0.095156	0.181441	-1.277103	3.74
4	0.081315	0.050804	-1.297627	-2.11
...	...	...	...	
1562	0.180796	-0.334889	0.408193	0.29
1563	-1.532007	-1.907724	0.757112	-0.68
1564	0.640138	-0.443753	0.433295	0.03
1565	0.435121	0.017626	0.278211	0.56
1566	0.352509	0.230171	0.802738	0.14

1567 rows × 24 columns

[Random Forest]

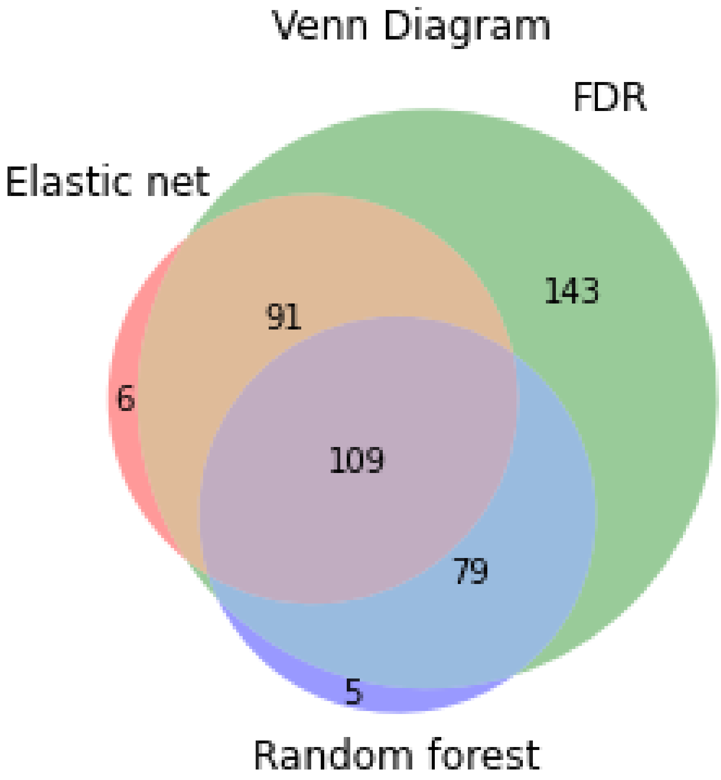


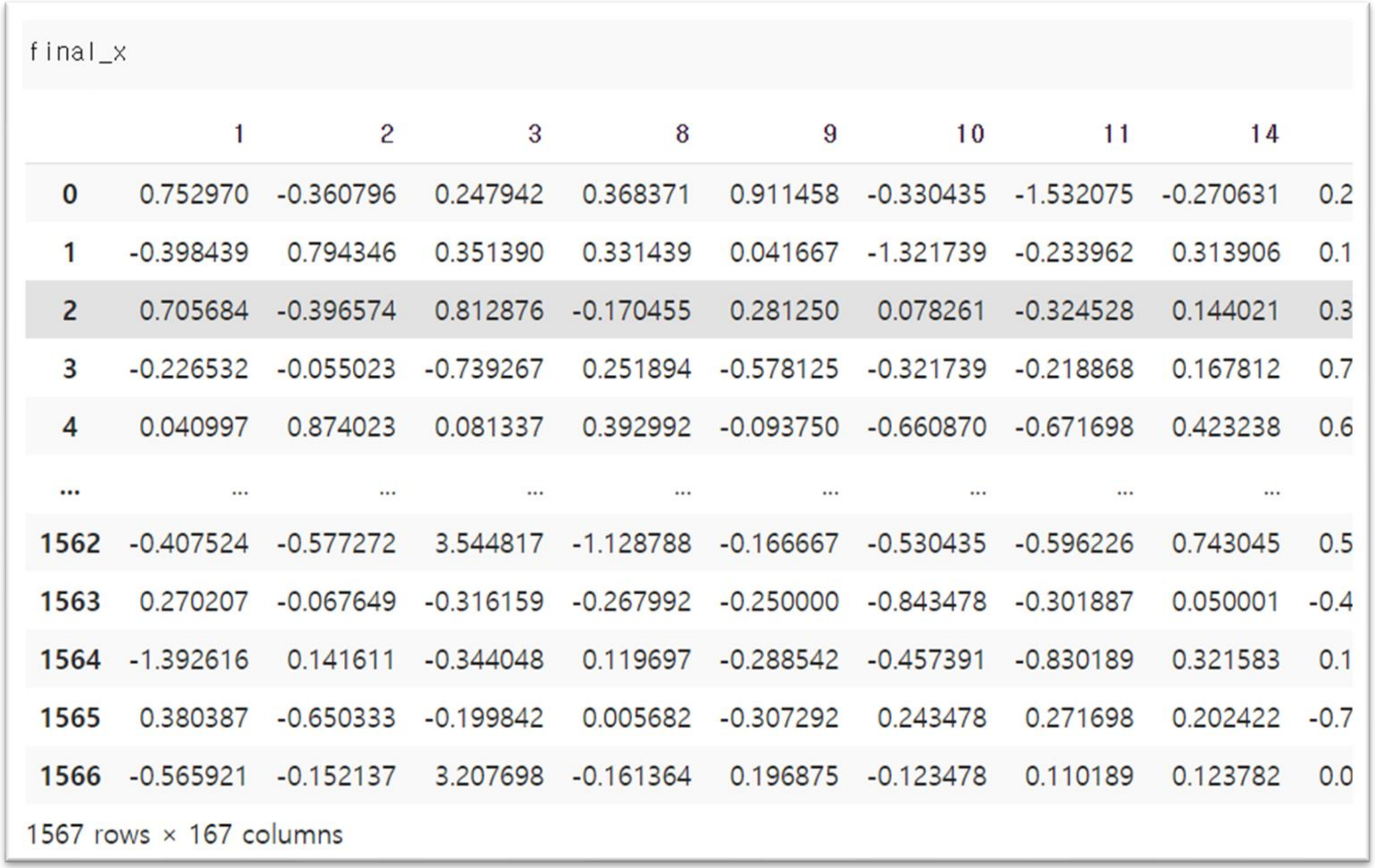
x_rf

	0	1	2	
0	0.217058	0.752970	-0.360796	0.2
1	0.933039	-0.398439	0.794346	0.3
2	-0.868452	0.705684	-0.396574	0.8
3	-0.248965	-0.226532	-0.055023	-0.7
4	0.231521	0.040997	0.874023	0.0
...	...	...	...	
1562	-1.234999	-0.407524	-0.577272	3.5
1563	0.453105	0.270207	-0.067649	-0.3
1564	-0.358377	-1.392616	0.141611	-0.3
1565	-1.284571	0.380387	-0.650333	-0.1
1566	-0.732542	-0.565921	-0.152137	3.2

1567 rows × 248 columns

2. Selected Data

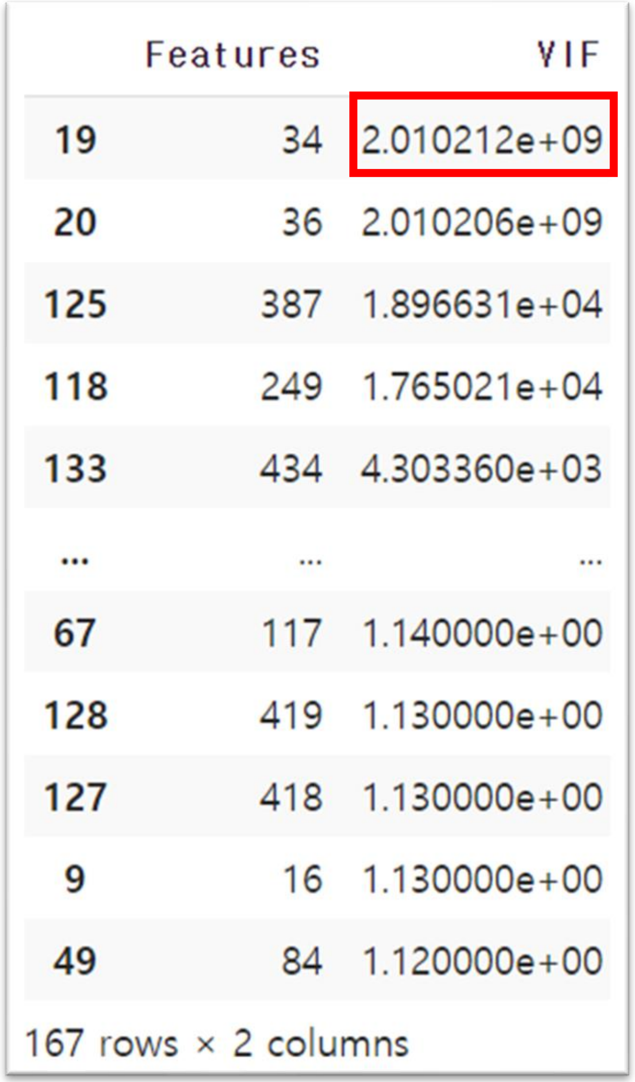



final_x

	1	2	3	8	9	10	11	14	
0	0.752970	-0.360796	0.247942	0.368371	0.911458	-0.330435	-1.532075	-0.270631	0.2
1	-0.398439	0.794346	0.351390	0.331439	0.041667	-1.321739	-0.233962	0.313906	0.1
2	0.705684	-0.396574	0.812876	-0.170455	0.281250	0.078261	-0.324528	0.144021	0.3
3	-0.226532	-0.055023	-0.739267	0.251894	-0.578125	-0.321739	-0.218868	0.167812	0.7
4	0.040997	0.874023	0.081337	0.392992	-0.093750	-0.660870	-0.671698	0.423238	0.6
...	...	...	...	...	...	...	...	...	
1562	-0.407524	-0.577272	3.544817	-1.128788	-0.166667	-0.530435	-0.596226	0.743045	0.5
1563	0.270207	-0.067649	-0.316159	-0.267992	-0.250000	-0.843478	-0.301887	0.050001	-0.4
1564	-1.392616	0.141611	-0.344048	0.119697	-0.288542	-0.457391	-0.830189	0.321583	0.1
1565	0.380387	-0.650333	-0.199842	0.005682	-0.307292	0.243478	0.271698	0.202422	-0.7
1566	-0.565921	-0.152137	3.207698	-0.161364	0.196875	-0.123478	0.110189	0.123782	0.0

1567 rows × 167 columns

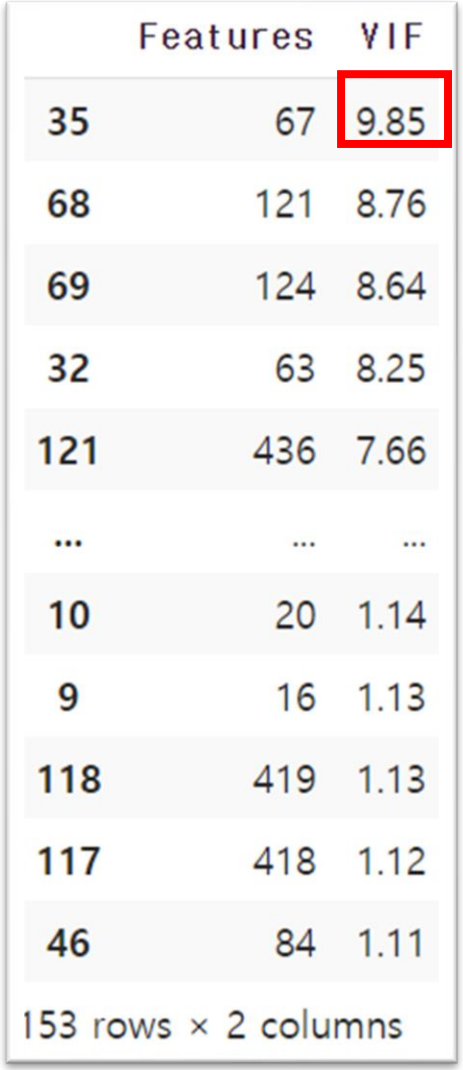
3. 상관계수_VIF



	Features	VIF
19	34	2.010212e+09
20	36	2.010206e+09
125	387	1.896631e+04
118	249	1.765021e+04
133	434	4.303360e+03
...	...	...
67	117	1.140000e+00
128	419	1.130000e+00
127	418	1.130000e+00
9	16	1.130000e+00
49	84	1.120000e+00

167 rows × 2 columns

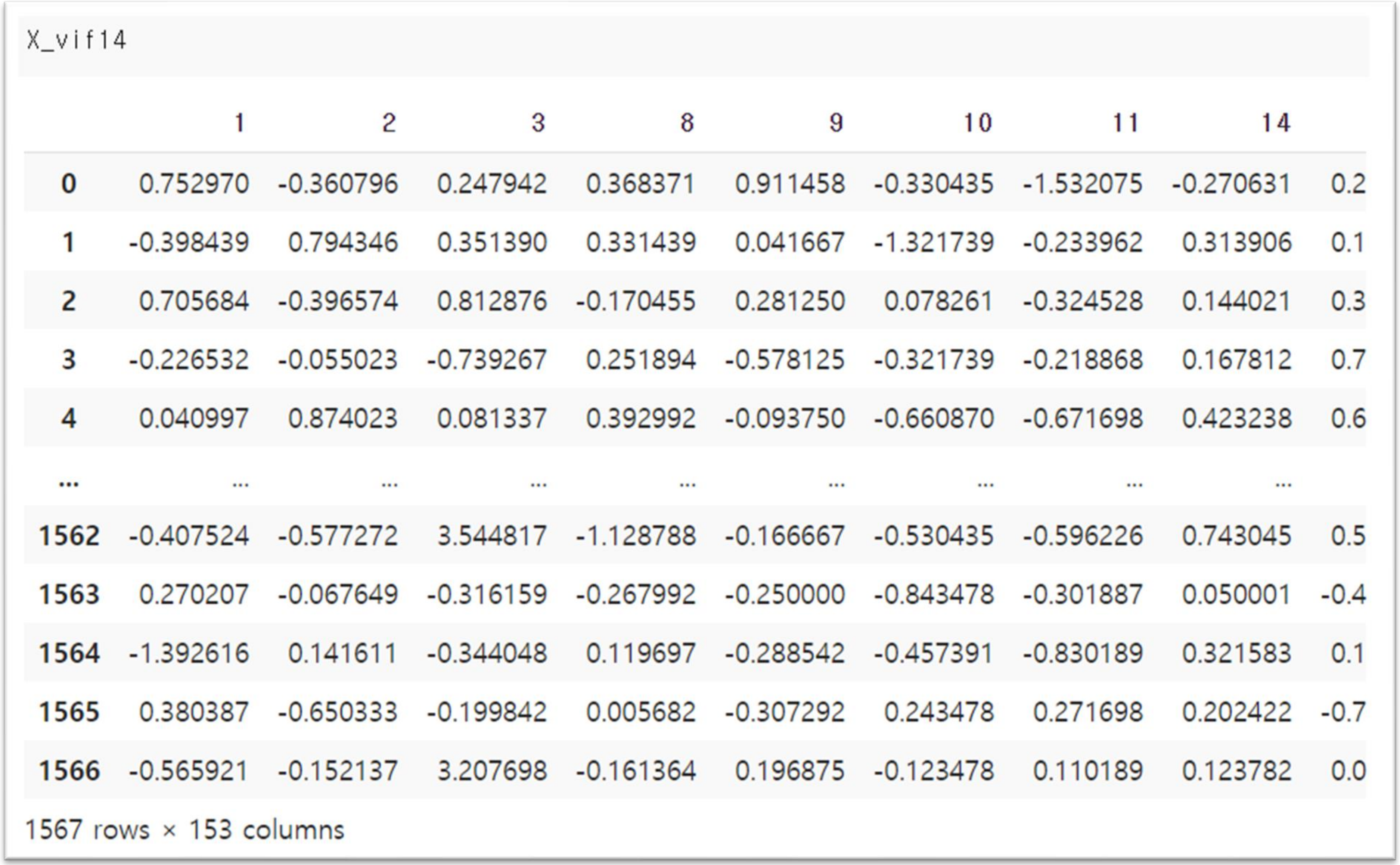
Eliminates 14 features



	Features	VIF
35	67	9.85
68	121	8.76
69	124	8.64
32	63	8.25
121	436	7.66
...	...	...
10	20	1.14
9	16	1.13
118	419	1.13
117	418	1.12
46	84	1.11

153 rows × 2 columns

4. Final Data

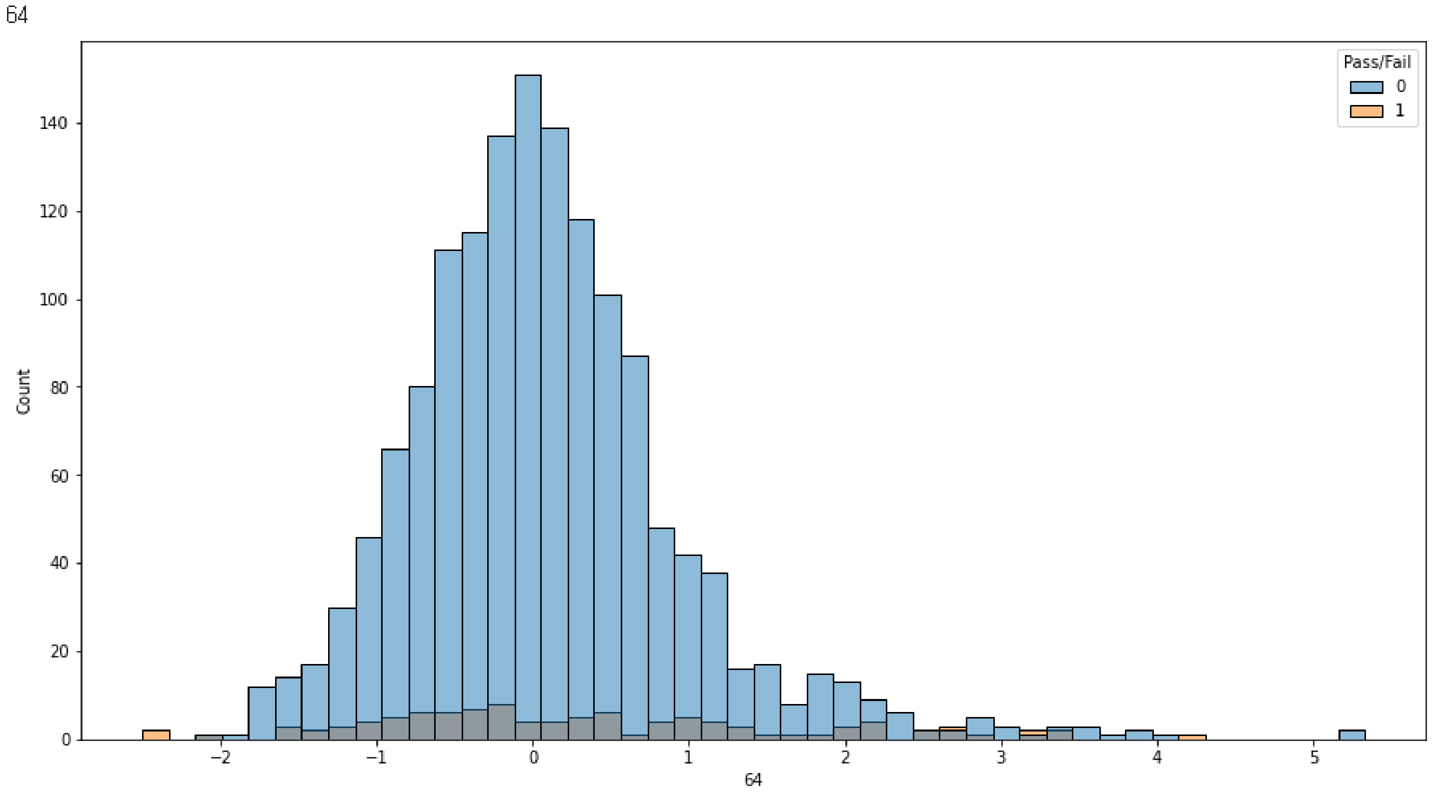


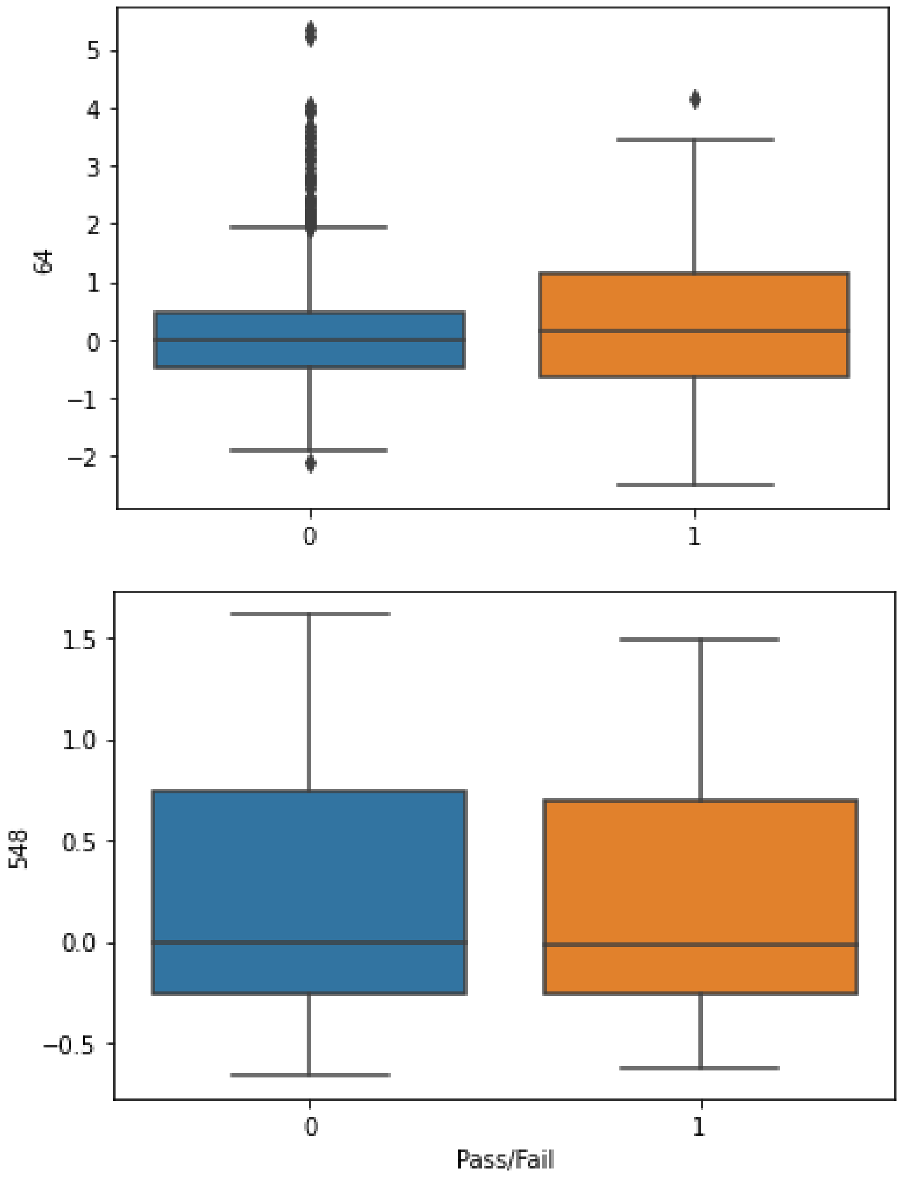
X_vif14

	1	2	3	8	9	10	11	14	
0	0.752970	-0.360796	0.247942	0.368371	0.911458	-0.330435	-1.532075	-0.270631	0.2
1	-0.398439	0.794346	0.351390	0.331439	0.041667	-1.321739	-0.233962	0.313906	0.1
2	0.705684	-0.396574	0.812876	-0.170455	0.281250	0.078261	-0.324528	0.144021	0.3
3	-0.226532	-0.055023	-0.739267	0.251894	-0.578125	-0.321739	-0.218868	0.167812	0.7
4	0.040997	0.874023	0.081337	0.392992	-0.093750	-0.660870	-0.671698	0.423238	0.6
...	...	...	...	...	...	...	...	...	
1562	-0.407524	-0.577272	3.544817	-1.128788	-0.166667	-0.530435	-0.596226	0.743045	0.5
1563	0.270207	-0.067649	-0.316159	-0.267992	-0.250000	-0.843478	-0.301887	0.050001	-0.4
1564	-1.392616	0.141611	-0.344048	0.119697	-0.288542	-0.457391	-0.830189	0.321583	0.1
1565	0.380387	-0.650333	-0.199842	0.005682	-0.307292	0.243478	0.271698	0.202422	-0.7
1566	-0.565921	-0.152137	3.207698	-0.161364	0.196875	-0.123478	0.110189	0.123782	0.0

1567 rows × 153 columns

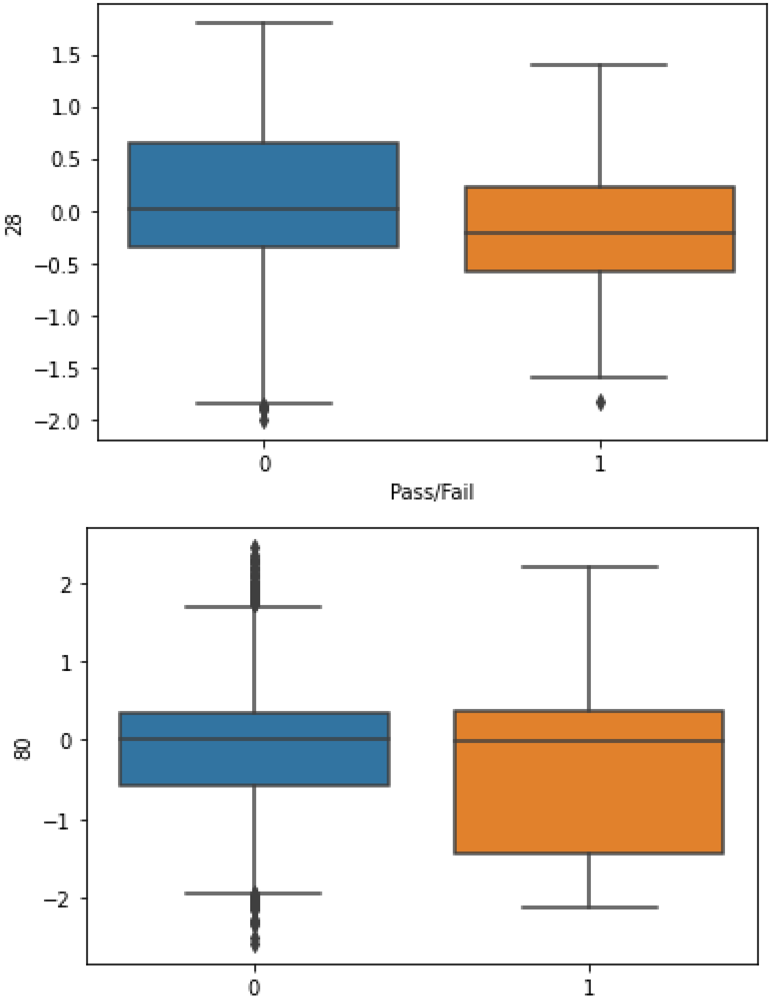
5. Feature Histogram & Boxplot



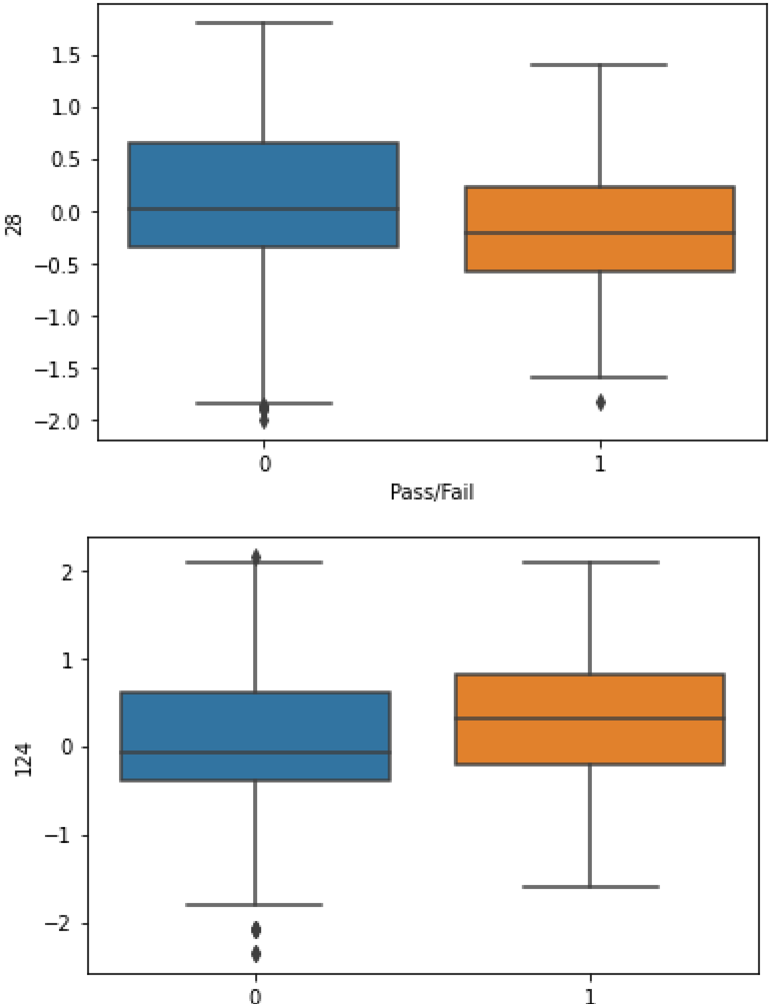


5. Feature Histogram & Boxplot

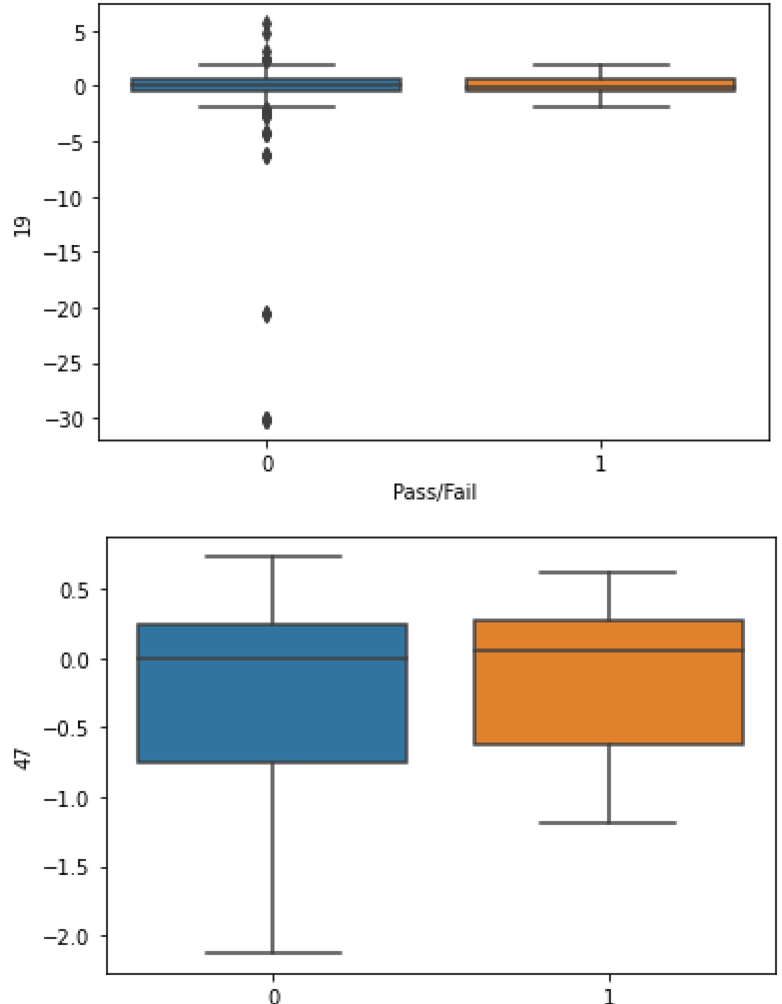
[Elastic Net]



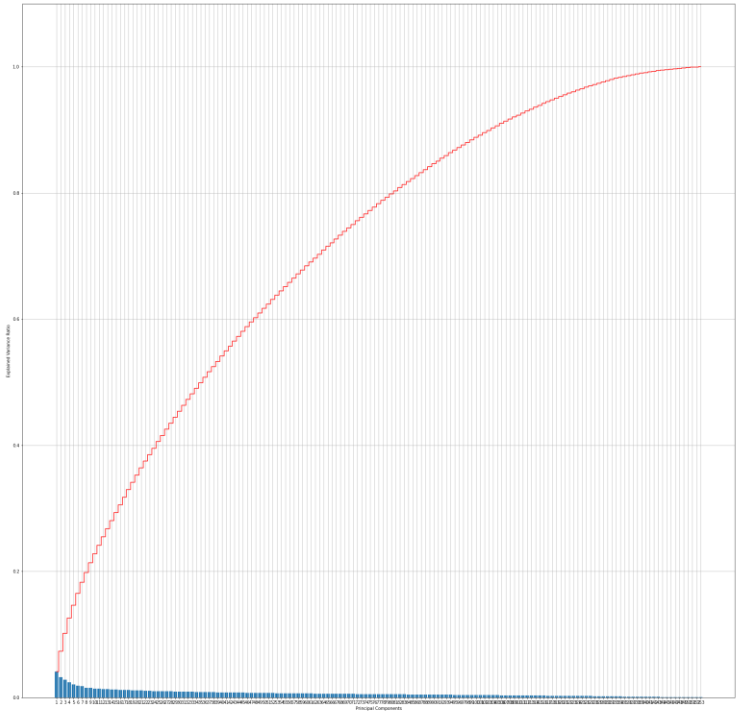
[FDR]

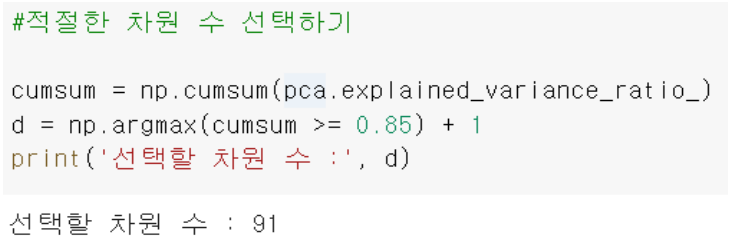


[Random Forest]



6. PCA

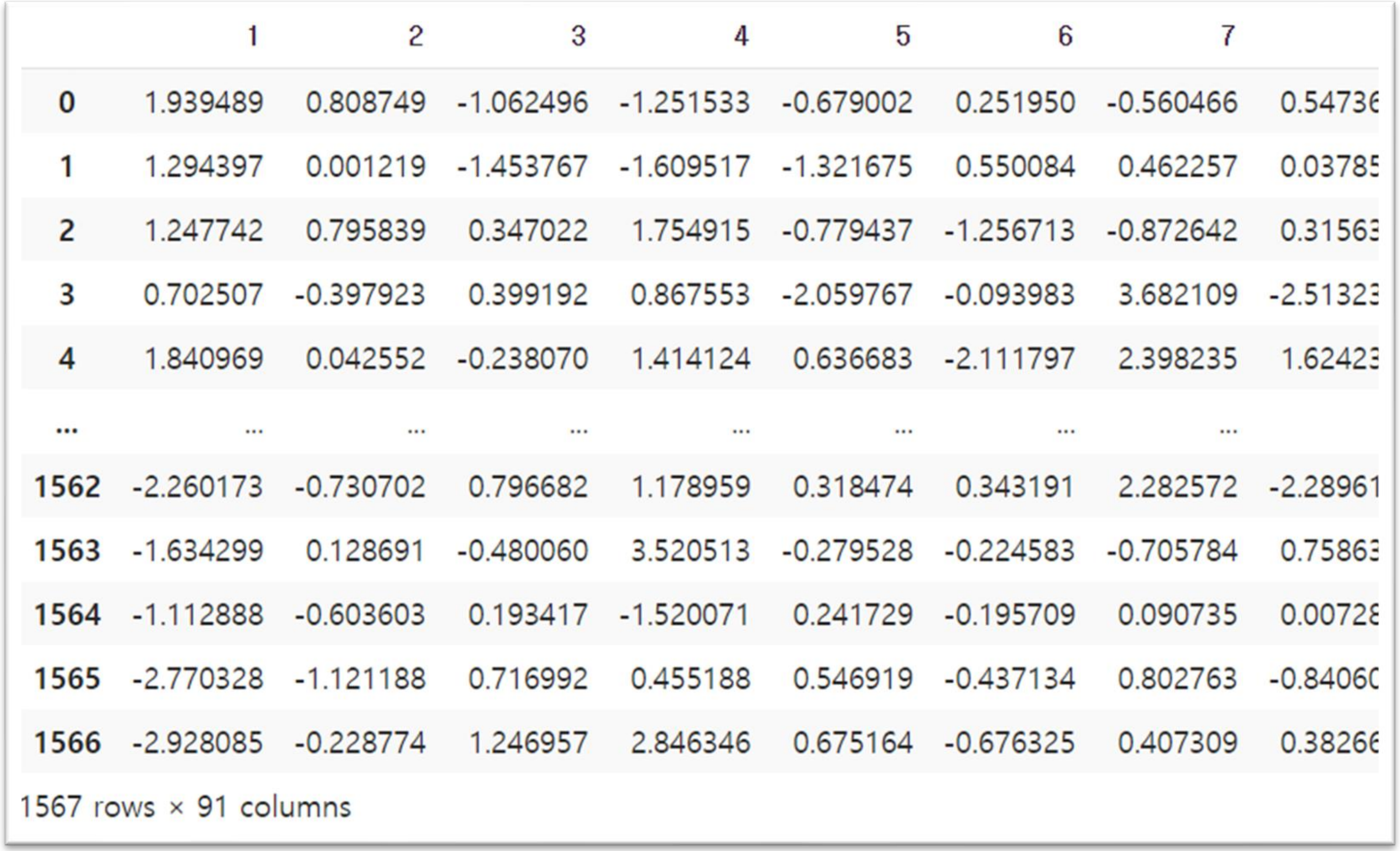




```
#적절한 차원 수 선택하기

cumsum = np.cumsum(pca.explained_variance_ratio_)
d = np.argmax(cumsum >= 0.85) + 1
print('선택할 차원 수 :', d)
```

```
선택할 차원 수 : 91
```

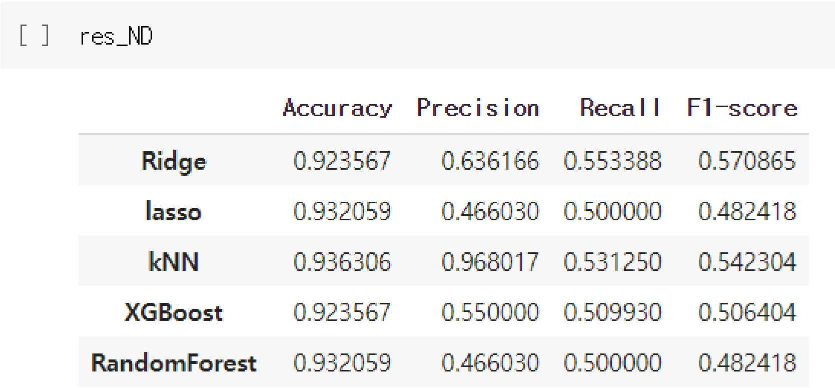


	1	2	3	4	5	6	7	
0	1.939489	0.808749	-1.062496	-1.251533	-0.679002	0.251950	-0.560466	0.54736
1	1.294397	0.001219	-1.453767	-1.609517	-1.321675	0.550084	0.462257	0.03785
2	1.247742	0.795839	0.347022	1.754915	-0.779437	-1.256713	-0.872642	0.31563
3	0.702507	-0.397923	0.399192	0.867553	-2.059767	-0.093983	3.682109	-2.51323
4	1.840969	0.042552	-0.238070	1.414124	0.636683	-2.111797	2.398235	1.62423
...	...	...	...	...	...	...	...	
1562	-2.260173	-0.730702	0.796682	1.178959	0.318474	0.343191	2.282572	-2.28961
1563	-1.634299	0.128691	-0.480060	3.520513	-0.279528	-0.224583	-0.705784	0.75863
1564	-1.112888	-0.603603	0.193417	-1.520071	0.241729	-0.195709	0.090735	0.00728
1565	-2.770328	-1.121188	0.716992	0.455188	0.546919	-0.437134	0.802763	-0.84060
1566	-2.928085	-0.228774	1.246957	2.846346	0.675164	-0.676325	0.407309	0.38266

1567 rows × 91 columns

7. Model Performance

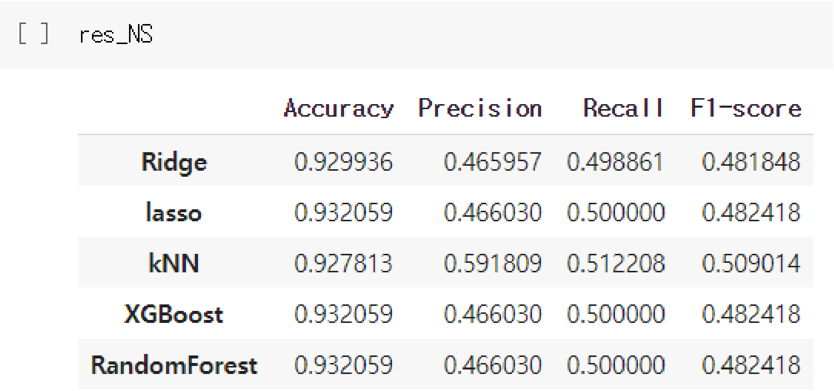
[Normal Data]



```
[ ]  res_ND
```

	Accuracy	Precision	Recall	F1-score
Ridge	0.923567	0.636166	0.553388	0.570865
lasso	0.932059	0.466030	0.500000	0.482418
kNN	0.936306	0.968017	0.531250	0.542304
XGBoost	0.923567	0.550000	0.509930	0.506404
RandomForest	0.932059	0.466030	0.500000	0.482418

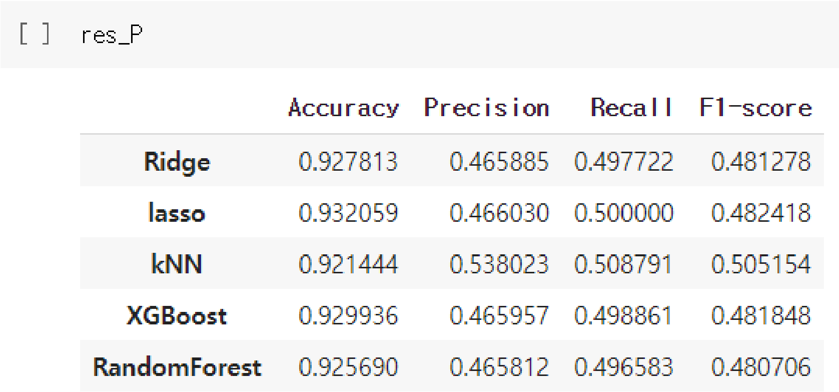
[Selected Feature Data]



```
[ ]  res_NS
```

	Accuracy	Precision	Recall	F1-score
Ridge	0.929936	0.465957	0.498861	0.481848
lasso	0.932059	0.466030	0.500000	0.482418
kNN	0.927813	0.591809	0.512208	0.509014
XGBoost	0.932059	0.466030	0.500000	0.482418
RandomForest	0.932059	0.466030	0.500000	0.482418

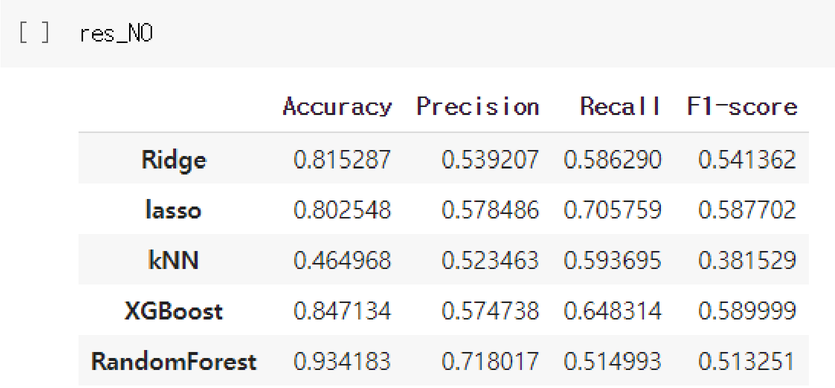
[PCA Data]



```
[ ]  res_P
```

	Accuracy	Precision	Recall	F1-score
Ridge	0.927813	0.465885	0.497722	0.481278
lasso	0.932059	0.466030	0.500000	0.482418
kNN	0.921444	0.538023	0.508791	0.505154
XGBoost	0.929936	0.465957	0.498861	0.481848
RandomForest	0.925690	0.465812	0.496583	0.480706

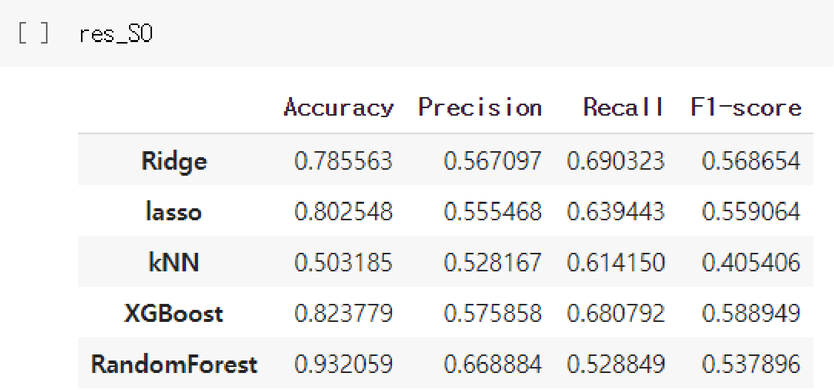
[ND Oversampling]



```
[ ]  res_NO
```

	Accuracy	Precision	Recall	F1-score
Ridge	0.815287	0.539207	0.586290	0.541362
lasso	0.802548	0.578486	0.705759	0.587702
kNN	0.464968	0.523463	0.593695	0.381529
XGBoost	0.847134	0.574738	0.648314	0.589999
RandomForest	0.934183	0.718017	0.514993	0.513251

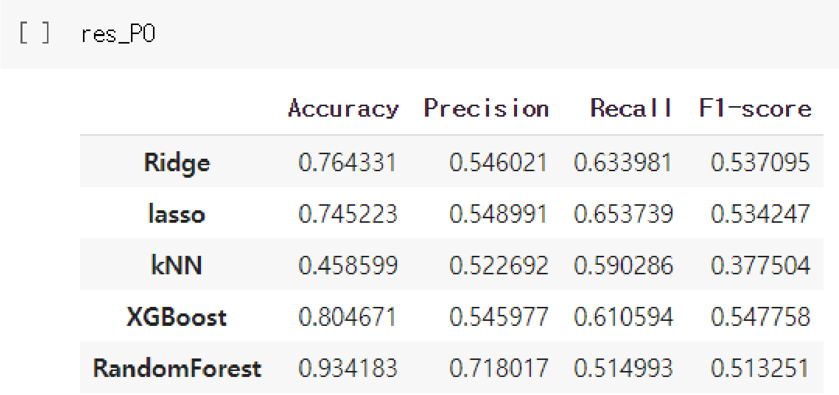
[SFD Oversampling]



```
[ ]  res_SO
```

	Accuracy	Precision	Recall	F1-score
Ridge	0.785563	0.567097	0.690323	0.568654
lasso	0.802548	0.555468	0.639443	0.559064
kNN	0.503185	0.528167	0.614150	0.405406
XGBoost	0.823779	0.575858	0.680792	0.588949
RandomForest	0.932059	0.668884	0.528849	0.537896

[PCAD Oversampling]

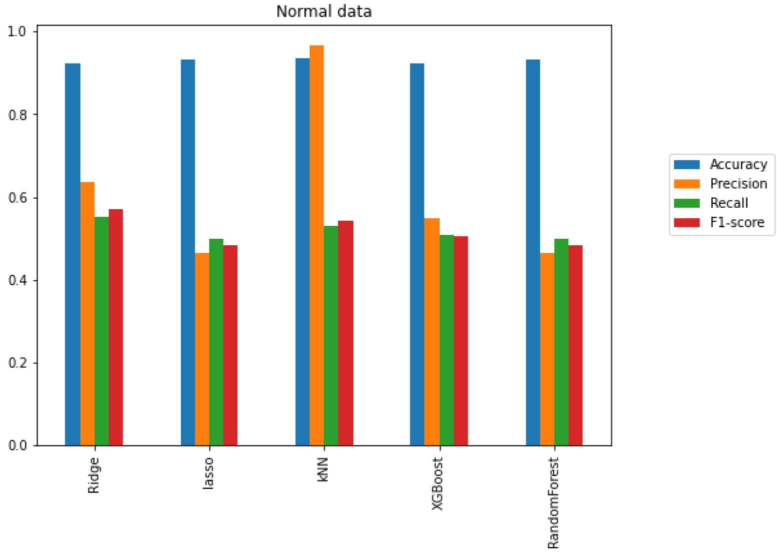


```
[ ]  res_PO
```

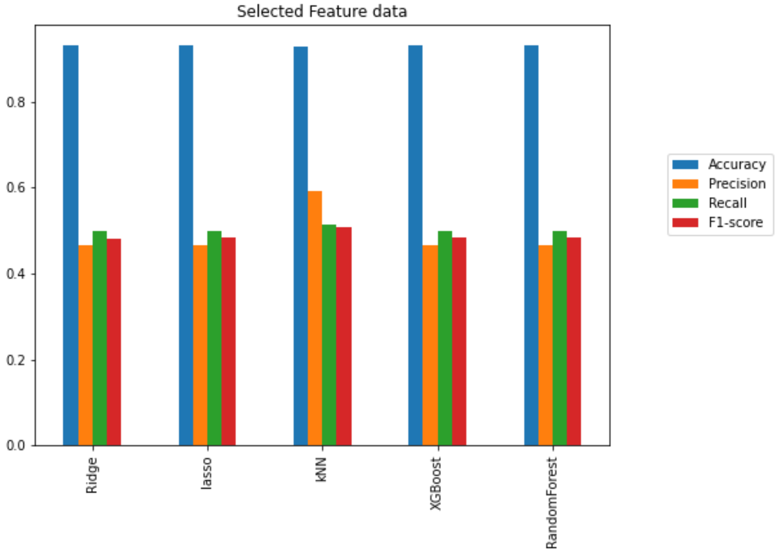
	Accuracy	Precision	Recall	F1-score
Ridge	0.764331	0.546021	0.633981	0.537095
lasso	0.745223	0.548991	0.653739	0.534247
kNN	0.458599	0.522692	0.590286	0.377504
XGBoost	0.804671	0.545977	0.610594	0.547758
RandomForest	0.934183	0.718017	0.514993	0.513251

7. Model Performance

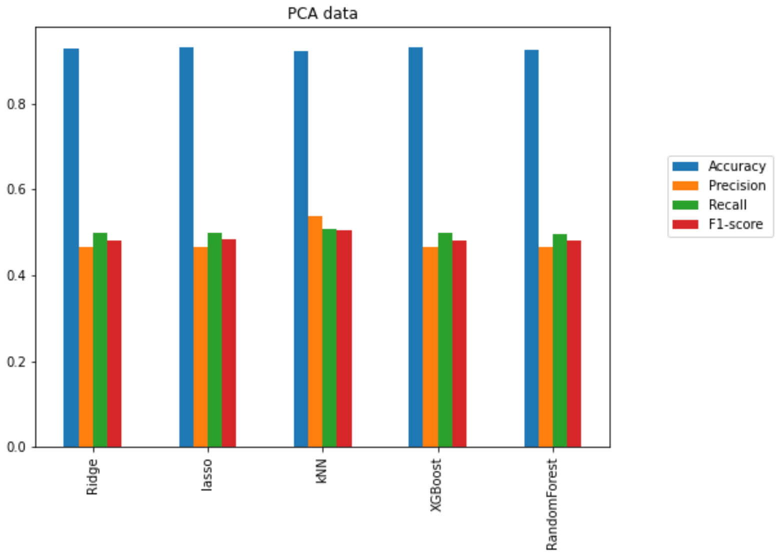
[Normal Data]



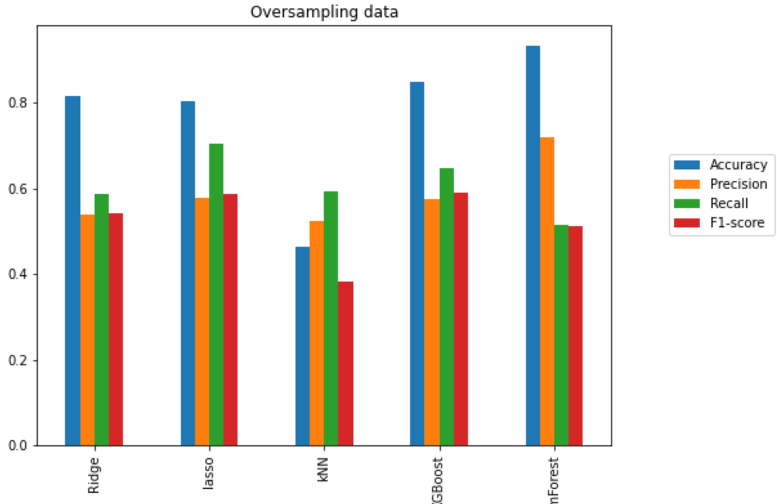
[Selected Feature Data]



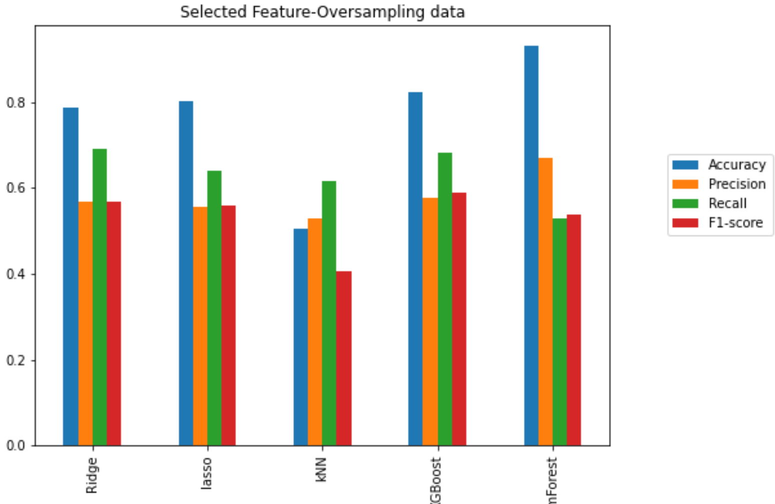
[PCA Data]



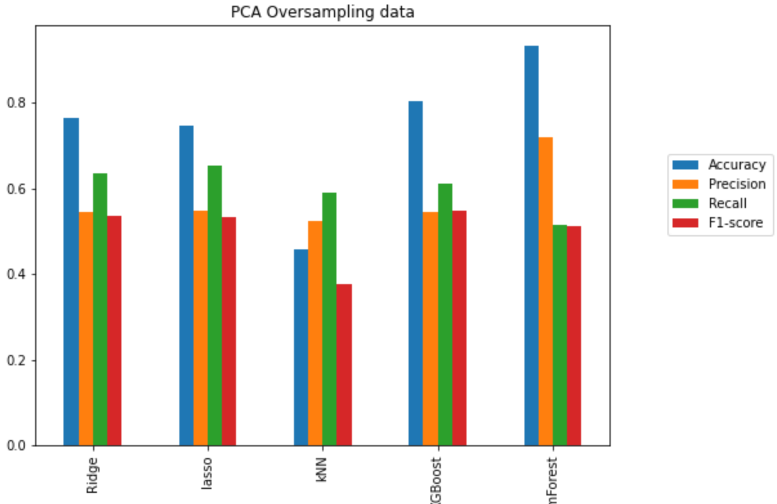
[ND Oversampling]



[SFD Oversampling]



[PCAD Oversampling]



Thank You

3조 | 김규연 정연서 정유미 정지윤

HANYANG UNIV. (ERICA)

INDUSTRIAL & MANAGEMENT ENGINEERING