Machine Learning in Finance Lab

Final Group Project (Linear Regression)

- Yu-Ching Liao ycliao3@illinois.edu (mailto:ycliao3@illinois.edu)
- Saranpat Praserrtthum sp73@illinois.edu (mailto:sp73@illinois.edu)
- Hyoung Woo Hahm hwham2@illinois.edu (mailto:hwham2@illinois.edu)

Out[2]:

Click here to toggle on/off the raw code.

Basic Import and Definition

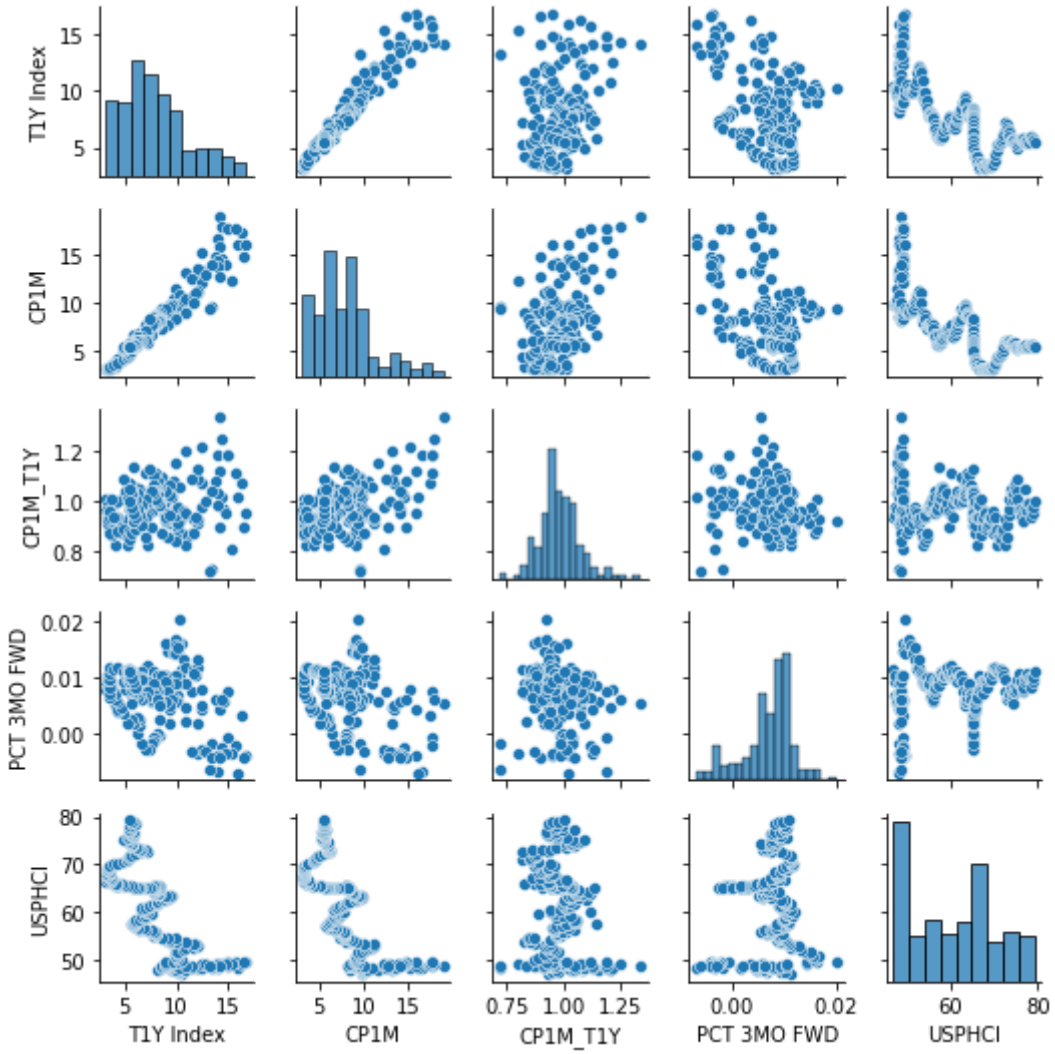
Out[4]:

Date	T1Y Index	T2Y Index	T3Y Index	T5Y Index	T7Y Index	T10Y Index	CP1M	CP3M	CP6M	CP1M_T1Y	CP3M_T1Y
1979-01-31	10.41	9.86	9.50	9.20	9.14	9.10	9.75	9.95	10.01	0.936599	0.955812
1979-02-28	10.24	9.72	9.29	9.13	9.11	9.10	9.74	9.90	9.96	0.951172	0.966797
1979-03-31	10.25	9.79	9.38	9.20	9.15	9.12	9.72	9.85	9.87	0.948293	0.960976
1979-04-30	10.12	9.78	9.43	9.25	9.21	9.18	9.86	9.95	9.98	0.974308	0.983202
1979-05-31	10.12	9.78	9.42	9.24	9.23	9.25	9.77	9.76	9.71	0.965415	0.964427
...	...	...	...	...	...	...	...	...	...	...	...
1997-03-31	5.80	6.22	6.38	6.54	6.65	6.69	5.61	5.71	5.79	0.967241	0.984483
1997-04-30	5.99	6.45	6.61	6.76	6.86	6.89	5.61	5.69	5.78	0.936561	0.949917
1997-05-31	5.87	6.28	6.42	6.57	6.66	6.71	5.60	5.65	5.69	0.954003	0.962521
1997-06-30	5.69	6.09	6.24	6.38	6.46	6.49	5.56	5.57	5.60	0.977153	0.978910
1997-07-31	5.54	5.89	6.00	6.12	6.20	6.22	5.55	5.56	5.59	1.001805	1.003610

223 rows × 16 columns

1) Introduction/Exploratory Data Analysis,

Scatter Matrix



Print the Shape Out

```
The number of Columns is 16.
The number of Rows is 223.
```

Print the nature out

Out[7]:

	Label	Number	String	Other
0	T1Y Index	223	0	0
1	T2Y Index	223	0	0
2	T3Y Index	223	0	0
3	T5Y Index	223	0	0
4	T7Y Index	223	0	0
5	T10Y Index	223	0	0
6	CP1M	223	0	0
7	CP3M	223	0	0
8	CP6M	223	0	0
9	CP1M_T1Y	223	0	0
10	CP3M_T1Y	223	0	0

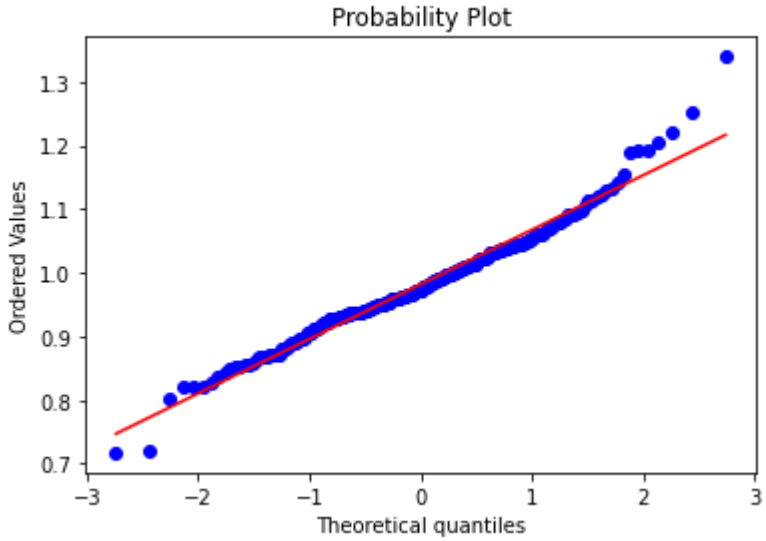
Summary of Statistics

```
μ = 60.59466367713005 Var = 90.07907242051922 σ = 9.490999548020179

Boundaries for 4 Equal Percentiles
 [47.08, 50.370000000000005, 61.09, 67.005, 79.21]

Boundaries for 10 Equal Percentiles
 [47.08, 48.598000000000006, 49.266, 53.12, 56.724000000000004, 61.09, 64.
96600000000001, 65.55799999999999, 69.302, 74.33200000000001, 79.21]

Unique Label Values
 ['T7Y Index', 'T5Y Index', 'PCT 6MO FWD', 'T1Y Index', 'CP6M', 'PCT 3MO F
WD', 'CP3M', 'T2Y Index', 'CP1M_T1Y', 'PCT 9MO FWD', 'CP3M_T1Y', 'T10Y Ind
ex', 'USPHCI', 'CP1M', 'CP6M_T1Y', 'T3Y Index']
```

Out[8]:

Types	T7Y Index	T5Y Index	PCT 6MO FWD	T1Y Index	CP6M	PCT 3MO FWD	CP3M	T2Y Index	CP1M_T1Y	PCT 9MO FWD	CP3M_T1Y
Counts	1	1	1	1	1	1	1	1	1	1	1

QQ PLot

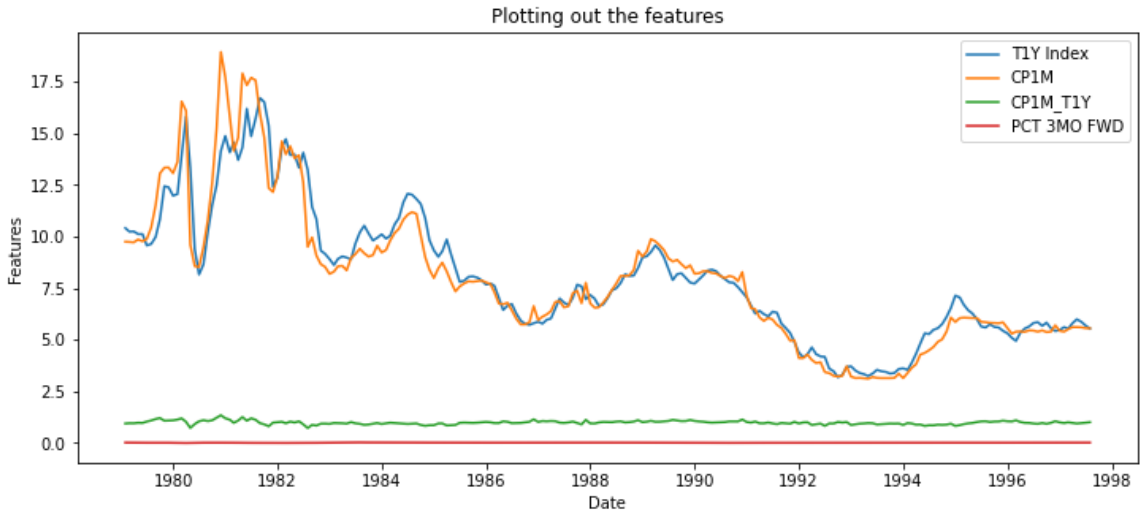


P-Value: 8.424094272178417e-05
Reject H0: Client_Trade_Percentage is Normally distributed.

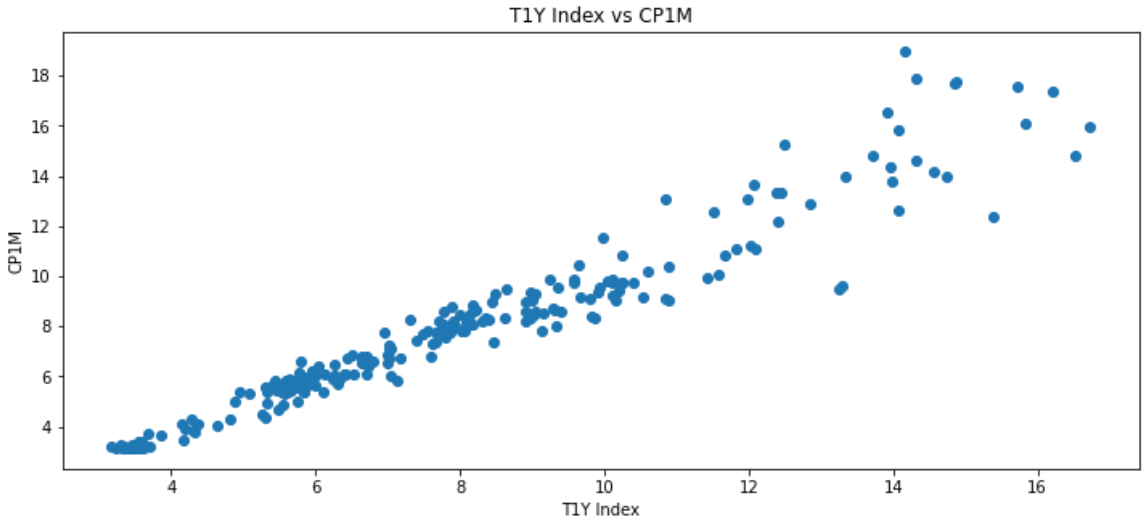
Print Summary of Data

```
        T1Y Index   T2Y Index   T3Y Index   T5Y Index   T7Y Index  T10Y
Index  \
count  223.000000  223.000000  223.000000  223.000000  223.000000  223.0
00000
mean     8.030717    8.410673    8.563587    8.808655    8.979776    9.0
73498
std      3.158575    2.954431    2.820405    2.647742    2.542686    2.4
47525
min      3.180000    3.840000    4.170000    4.710000    5.050000    5.3
30000
25%      5.735000    6.180000    6.410000    6.695000    6.965000    7.1
75000
50%      7.670000    8.000000    8.130000    8.330000    8.520000    8.6
10000
75%      9.840000   10.075000   10.375000   10.525000   10.640000   10.6
85000
max     16.720000   16.460000   16.220000   15.930000   15.650000   15.3
20000
```

Plot Data



Cross Plotting Pairs of Attributes (Scatter Plot)

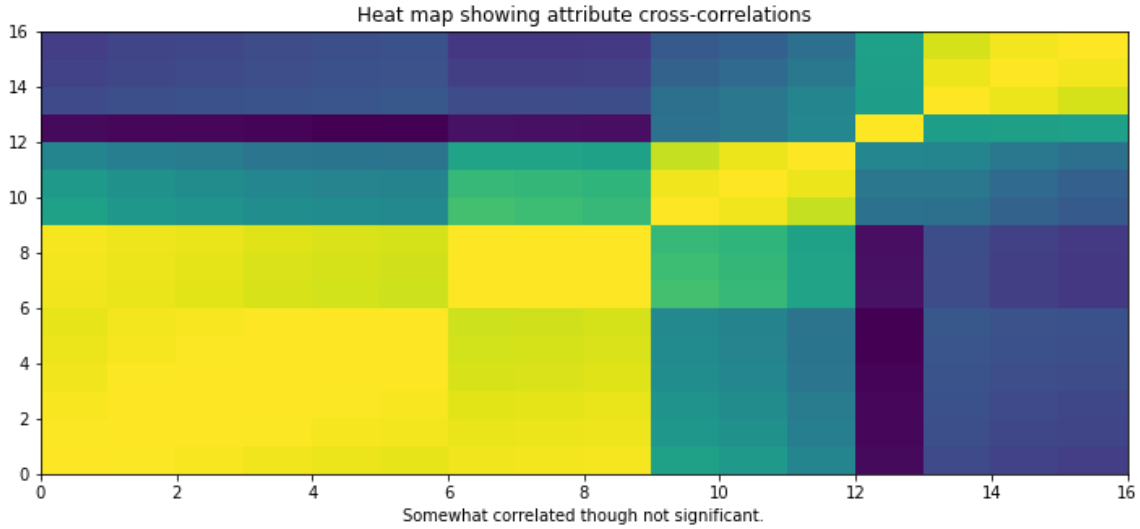


Correlation

Out[13]:

	T1Y Index	T2Y Index	T3Y Index	T5Y Index	T7Y Index	T10Y Index	CP1M	CP
T1Y Index	1.000000	0.992299	0.981237	0.961512	0.946299	0.934787	0.962917	0.9678
T2Y Index	0.992299	1.000000	0.997306	0.986983	0.977260	0.968840	0.938417	0.9451
T3Y Index	0.981237	0.997306	1.000000	0.995546	0.989145	0.982837	0.919866	0.9272
T5Y Index	0.961512	0.986983	0.995546	1.000000	0.998315	0.995331	0.890890	0.8990
T7Y Index	0.946299	0.977260	0.989145	0.998315	1.000000	0.999073	0.872348	0.8809
T10Y Index	0.934787	0.968840	0.982837	0.995331	0.999073	1.000000	0.859418	0.8682
CP1M	0.962917	0.938417	0.919866	0.890890	0.872348	0.859418	1.000000	0.9984
CP3M	0.967800	0.945139	0.927224	0.899064	0.880997	0.868233	0.998414	1.0000
CP6M	0.973094	0.954145	0.937839	0.911446	0.894304	0.881913	0.993353	0.9979
CP1M_T1Y	0.213583	0.147634	0.113604	0.066948	0.049383	0.038051	0.453449	0.4315
CP3M_T1Y	0.158550	0.094849	0.062140	0.017599	0.001674	-0.008190	0.398043	0.3884
CP6M_T1Y	0.006001	-0.046372	-0.072444	-0.108187	-0.119328	-0.125453	0.233306	0.2353
USPHCI	-0.771879	-0.786831	-0.790018	-0.802284	-0.811539	-0.818440	-0.734319	-0.7410
PCT 3MO FWD	-0.407624	-0.382981	-0.368031	-0.351309	-0.336880	-0.327772	-0.404970	-0.4022
PCT 6MO FWD	-0.460467	-0.428199	-0.409257	-0.386366	-0.368737	-0.357288	-0.481658	-0.4780
PCT 9MO FWD	-0.488882	-0.448940	-0.427909	-0.400488	-0.380166	-0.367086	-0.525706	-0.5206

Correlation Visualization



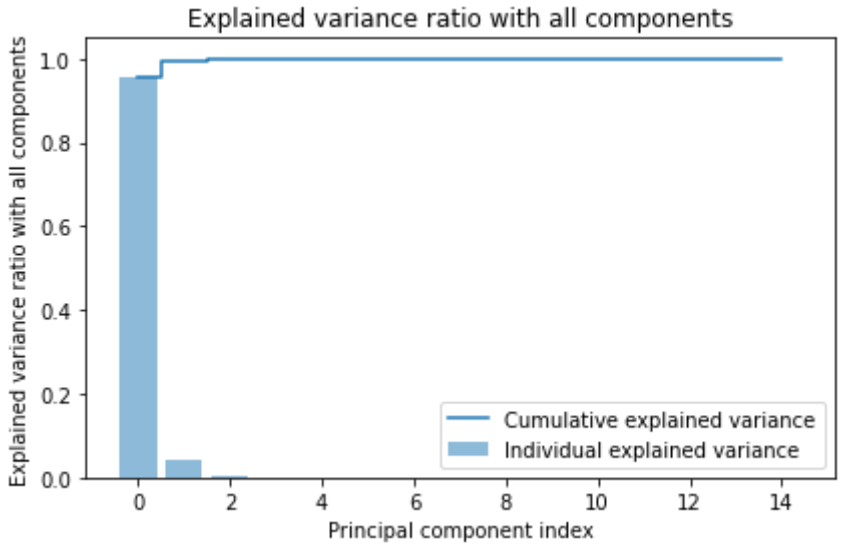
2) Preprocessing, feature extraction, feature selection,

Drop Missing Value

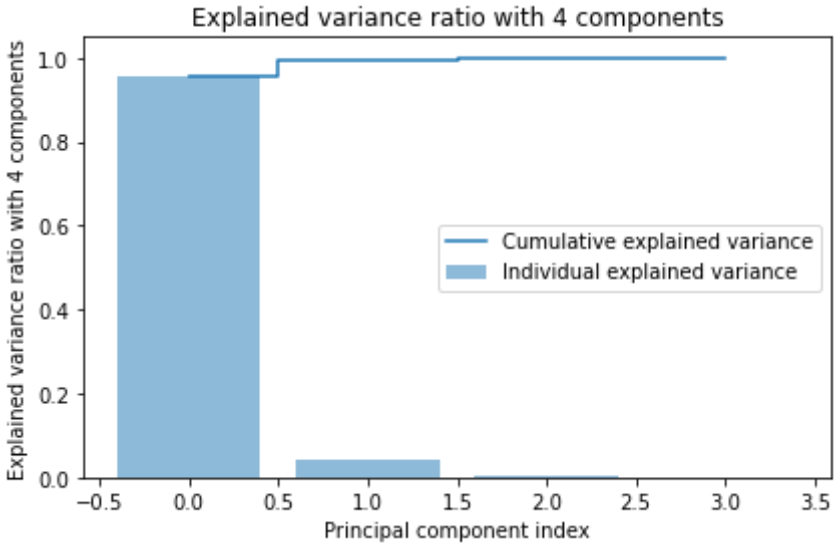
Preprocessing the Data

See Variance Ratio

```
Explained Variance Ratio with all components:
 [9.55657070e-01 3.99135802e-02 3.35994699e-03 7.55389913e-04
 2.10423219e-04 3.81406915e-05 2.17681942e-05 1.74860869e-05
 1.41572298e-05 7.98034462e-06 2.33594094e-06 1.56298102e-06
 9.34359752e-08 5.51891546e-08 1.00610705e-08]

Culmulative Variance Ratio with all components:
 [0.95565707 0.99557065 0.9989306  0.99968599 0.99989641 0.99993455
 0.99995632 0.9999738  0.99998796 0.99999594 0.99999828 0.99999984
 0.99999993 0.99999999 1.        ]
```



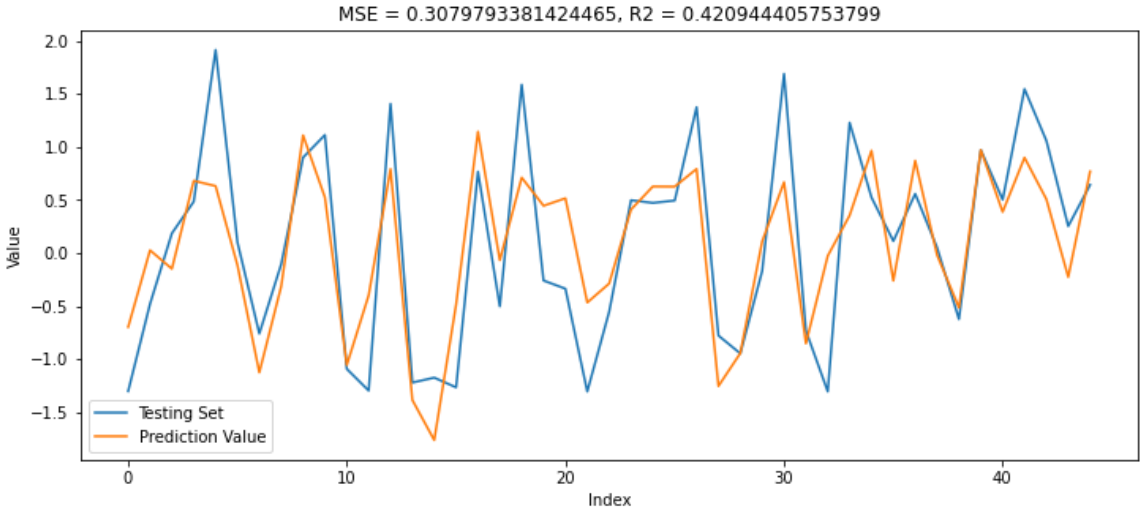
```
Explained Variance Ratio with 4 components:
 [9.55657070e-01 3.99135802e-02 3.35994699e-03 7.55389913e-04]

Culmulative Variance Ratio with 4 components:
 [0.95565707 0.99557065 0.9989306  0.99968599]
```

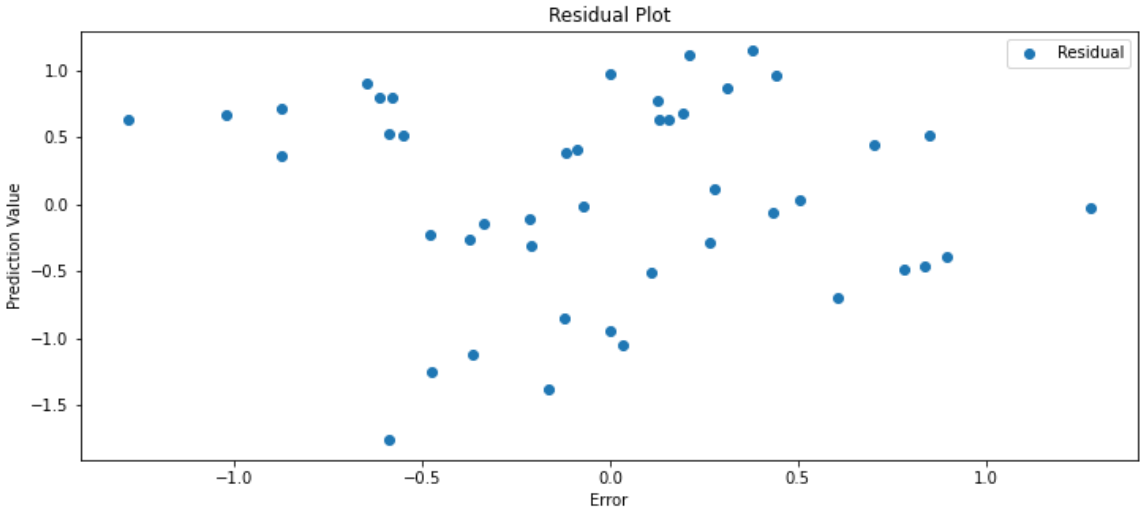


Model fitting and evaluation, (you should fit at least 3 different machine learning models) & Hyperparameter tunning

Simple Linear Regresion with PCA

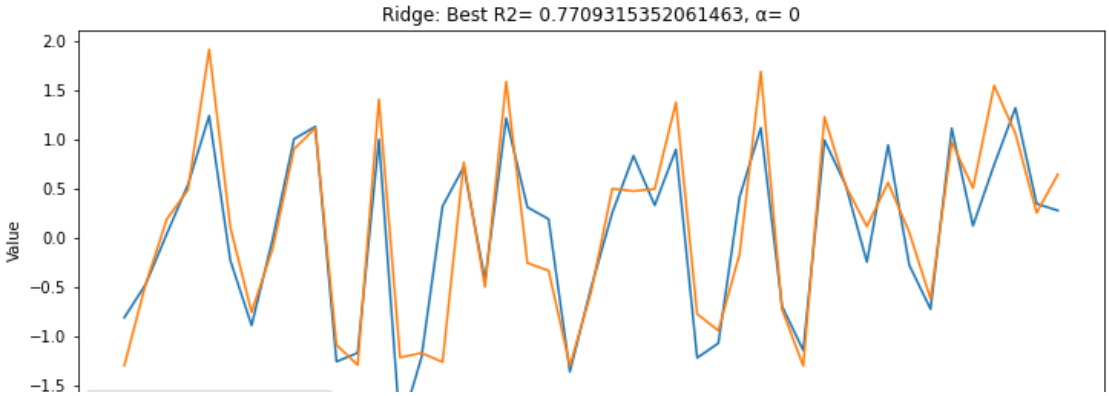
```
Coefficients:  [[-0.24109023 -0.09217819 -0.18879978 -0.18903653]]
Intercept:  [-0.004937]
```





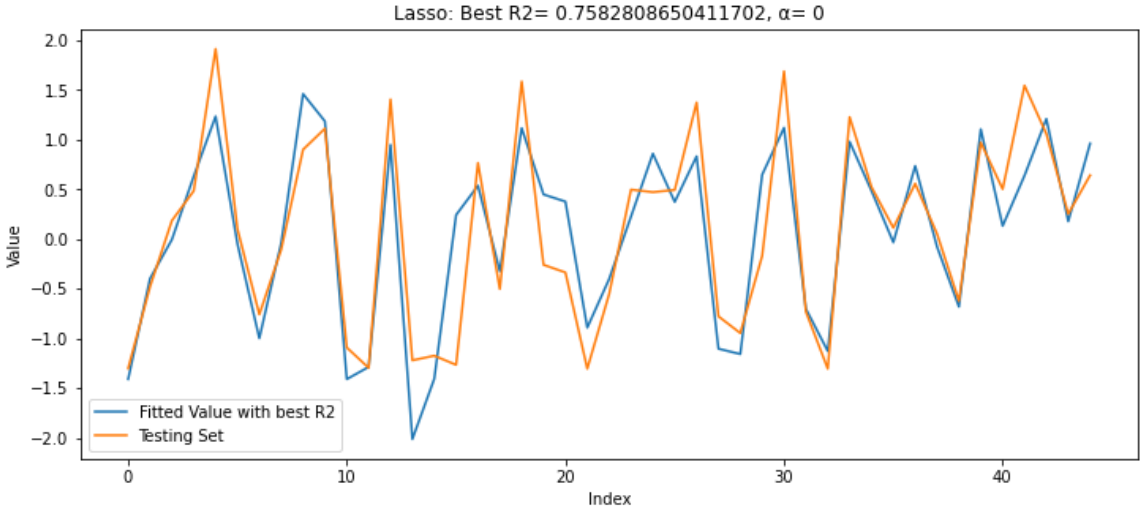
Ridge Regression with Hyperparameter Tunning

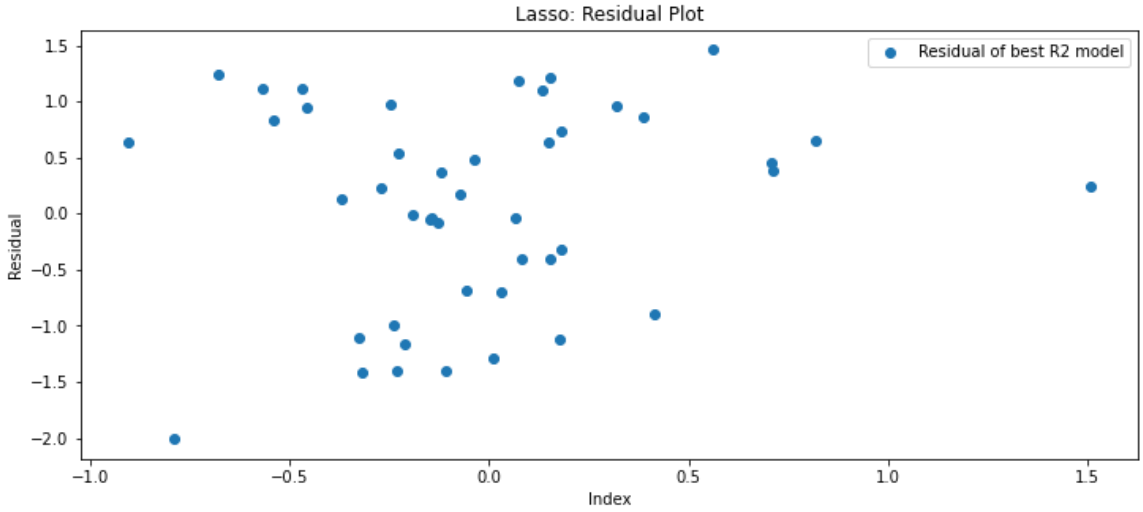
```
α=  0 , R2=  0.7709315352061463 , MSE =  0.17110414461321527
α=  0.1 , R2=  0.7606253656128817 , MSE =  0.1664053287775124
α=  1.0 , R2=  0.6291381510053733 , MSE =  0.2084804119213952
α=  10.0 , R2=  0.5060668309520844 , MSE =  0.26135570870022196
α=  100.0 , R2=  0.37994266062846904 , MSE =  0.2918654056718467
α=  1000.0 , R2=  -1.2422537634035846 , MSE =  0.41568765881273323
α=  10000.0 , R2=  -77.32440611960723 , MSE =  0.7578752378476383
```

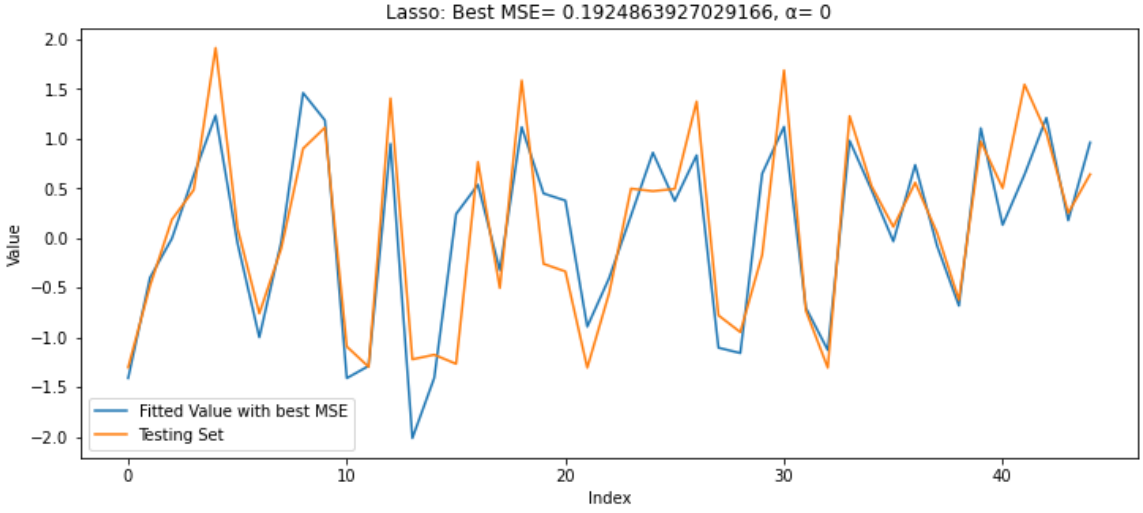


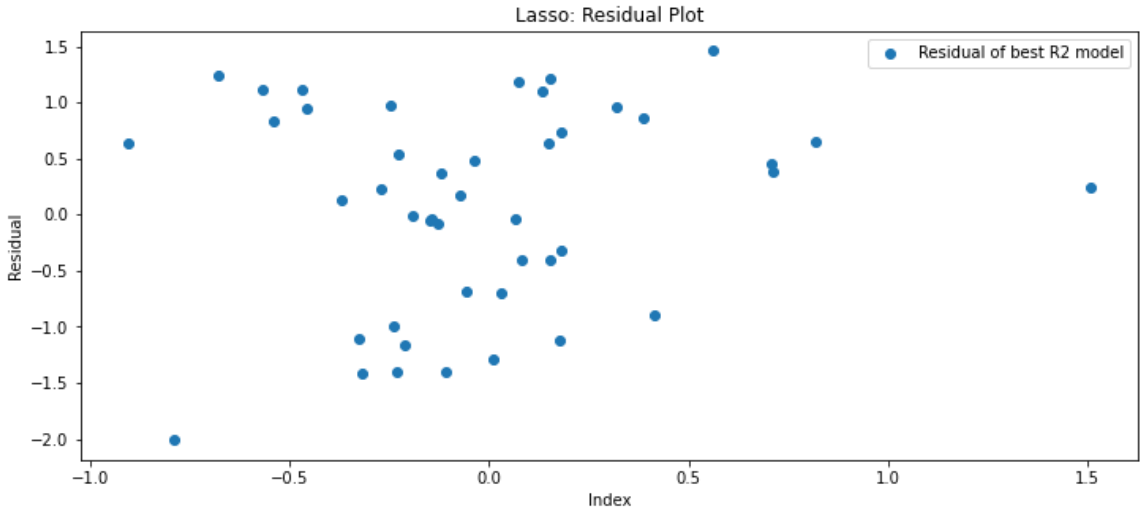
Lasso Regression with Hyperparameter tunning

```
α=  0 , R2=  0.7582808650411702 , MSE =  0.1924863927029166
α=  0.1 , R2=  0.3288151922141559 , MSE =  0.27584050295119744
α=  1.0 , R2=  -1.867853816152284e+34 , MSE =  0.8993388990664136
α=  10.0 , R2=  -1.867853816152284e+34 , MSE =  0.8993388990664136
α=  100.0 , R2=  -1.867853816152284e+34 , MSE =  0.8993388990664136
α=  1000.0 , R2=  -1.867853816152284e+34 , MSE =  0.8993388990664136
α=  10000.0 , R2=  -1.867853816152284e+34 , MSE =  0.8993388990664136
```

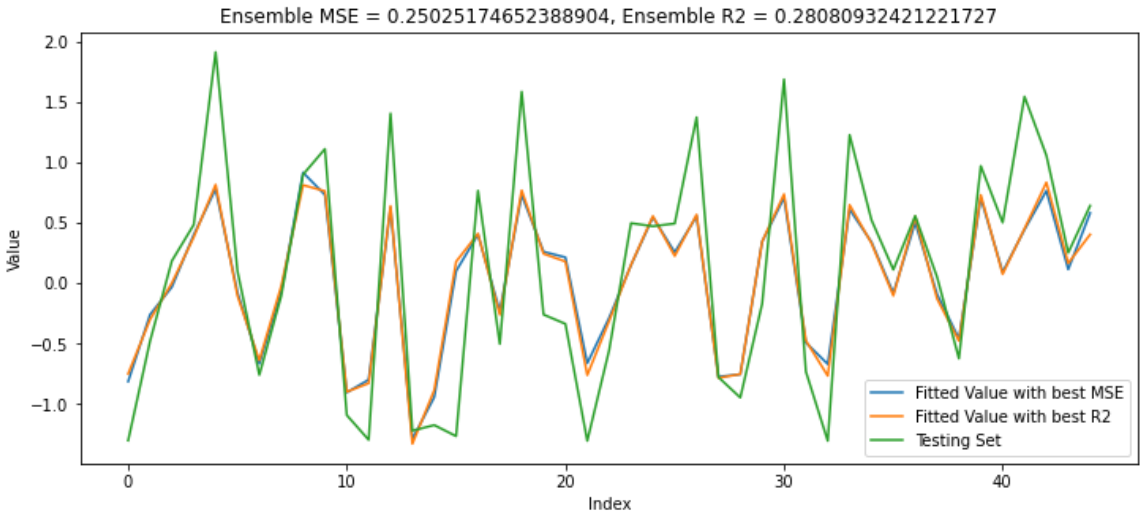








5) Ensembling



Conclusion

By applying Ridge and Lasso regularization, we can enhance the performance of the fit. However, this is not held in the case of ensemble. It is highly possible that, since Ridge and Lasso regularization assume a linear relationship between the features and the target variable, if there are complex nonlinear relationships in the data, then these regularization techniques may not be effective. In such cases, it may be better to use nonlinear models such as decision trees, random forests, or neural networks.

Appendix

Like to github:
https://github.com/Saranpatp/IE517_F2023_HW/blob/main/IE517MLF_Group_project/LinearRegression.ipynb
(https://github.com/Saranpatp/IE517_F2023_HW/blob/main/IE517MLF_Group_project/LinearRegression.ipynb