# Machine Learning in Finance Lab

## Final Group Project (Classification)


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


Out[2]:

Click here to toggle on/off the raw code.

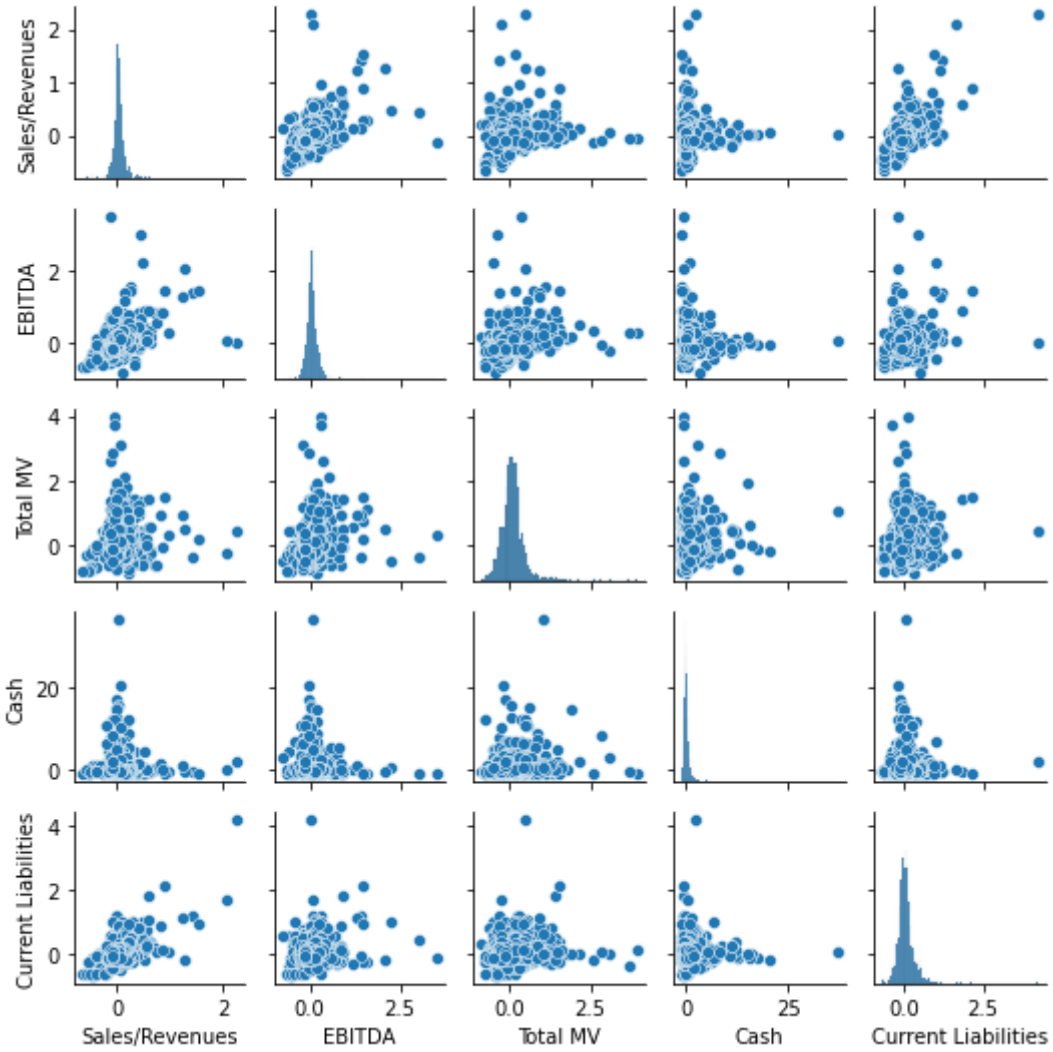
Out[2]:

| | Sales/Revenues | Gross Margin | EBITDA | EBITDA Margin | Net Income Before Extras | Total Debt | Net Debt | LT Debt | |
|---|---|---|---|---|---|---|---|---|---|
| **0** | -0.005496 | 0.030763 | 0.018885 | 0.024515 | 0.146849 | -0.029710 | -0.019296 | -0.042648 | 0 |
| **1** | -0.005496 | 0.030763 | 0.088716 | 0.094733 | 0.146849 | -0.029710 | -0.019296 | -0.042648 | 0 |
| **2** | -0.007045 | 0.023159 | 0.088716 | 0.096440 | 0.108590 | 0.039410 | 0.034268 | 0.009059 | 0 |
| **3** | -0.009396 | 0.028400 | 0.088716 | 0.099046 | 0.146137 | 0.030071 | 0.036938 | -0.016964 | 0 |
| **4** | -0.009009 | 0.027714 | 0.088716 | 0.098611 | 0.123500 | 0.024224 | 0.034445 | -0.034132 | 0 |

5 rows × 28 columns

# 1) Introduction/Exploratory Data Analysis,

Scatter Matrix




Print the Shape Out

```
The number of Columns is 28.
The number of Rows is 1700.
```

Print the nature out

Out[5]:

| | Label | Number | String | Other |
|---|---|---|---|---|
| **0** | Sales/Revenues | 1700 | 0 | 0 |
| **1** | Gross Margin | 1700 | 0 | 0 |
| **2** | EBITDA | 1700 | 0 | 0 |
| **3** | EBITDA Margin | 1700 | 0 | 0 |
| **4** | Net Income Before Extras | 1700 | 0 | 0 |
| **5** | Total Debt | 1700 | 0 | 0 |
| **6** | Net Debt | 1700 | 0 | 0 |
| **7** | LT Debt | 1700 | 0 | 0 |
| **8** | ST Debt | 1700 | 0 | 0 |
| **9** | Cash | 1700 | 0 | 0 |
| **10** | Free Cash Flow | 1700 | 0 | 0 |
| **11** | Total Debt/EBITDA | 1700 | 0 | 0 |
| **12** | Net Debt/EBITDA | 1700 | 0 | 0 |
| **13** | Total MV | 1700 | 0 | 0 |
| **14** | Total Debt/MV | 1700 | 0 | 0 |
| **15** | Net Debt/MV | 1700 | 0 | 0 |
| **16** | CFO/Debt | 1700 | 0 | 0 |
| **17** | CFO | 1700 | 0 | 0 |
| **18** | Interest Coverage | 1700 | 0 | 0 |
| **19** | Total Liquidity | 1700 | 0 | 0 |
| **20** | Current Liquidity | 1700 | 0 | 0 |
| **21** | Current Liabilities | 1700 | 0 | 0 |
| **22** | EPS Before Extras | 1700 | 0 | 0 |
| **23** | PE | 1700 | 0 | 0 |
| **24** | ROA | 1700 | 0 | 0 |
| **25** | ROE | 1700 | 0 | 0 |
| **26** | InvGrd | 1700 | 0 | 0 |
| **27** | Rating | 0 | 1700 | 0 |

Summary of Statistics

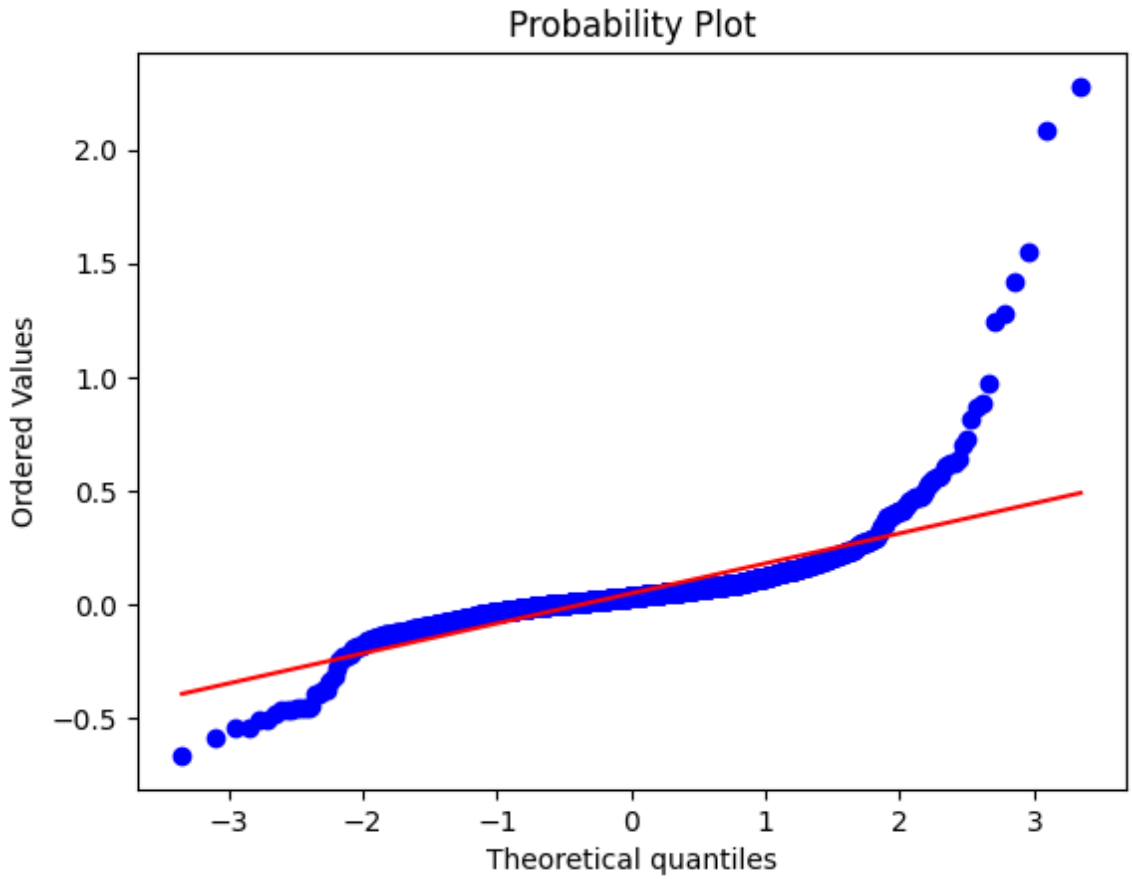
```
All the ratings
['A1' 'A2' 'A3' 'Aa2' 'Aa3' 'Aaa' 'B1' 'B2' 'B3' 'Ba1' 'Ba2' 'Ba3' 'Baa1'
 'Baa2' 'Baa3' 'Caa1']
Investment Recommended
μ = 0.05716446828049729 Var = 0.04507965336673378 σ = 0.21231969613470575

Investment Not Recommended
μ = 0.1047225816416465 Var = 0.08958951418890121 σ = 0.29931507511133015

Boundaries for 4 Equal Percentiles for Investment Recommended
 [-0.782254303, -0.0192508605, 0.044912281, 0.1074112785, 3.542424887]

Boundaries for 10 Equal Percentiles for Investment Recommended
 [-0.782254303, -0.09602339900000001, -0.03783060260000006, -0.0027125775
999999996, 0.021396558399999998, 0.044912281, 0.0693806898, 0.094985605199
99999, 0.1263567034, 0.20178858760000004, 3.542424887]

Boundaries for 4 Equal Percentiles for Investment Not Recommended
 [-0.582112521, -0.036688059, 0.065095003, 0.200948894, 2.084133001]

Boundaries for 10 Equal Percentiles for Investment Not Recommended
 [-0.582112521, -0.1707249226, -0.07302318859999998, -0.01554577360000001
6, 0.03127736020000001, 0.065095003, 0.12283974699999999, 0.175052519, 0.2
321507122, 0.3397043488, 2.084133001]

Unique Label Values
 ['Total Debt', 'Net Debt', 'Interest Coverage', 'Current Liabilities', 'S
ales/Revenues', 'PE', 'EBITDA', 'ROA', 'EBITDA Margin', 'CFO', 'Total MV',
'Total Liquidity', 'Net Debt/MV', 'Total Debt/MV', 'Total Debt/EBITDA', 'N
et Income Before Extras', 'LT Debt', 'ST Debt', 'Current Liquidity', 'RO
E', 'CFO/Debt', 'InvGrd', 'Free Cash Flow', 'Rating', 'Gross Margin', 'Cas
h', 'EPS Before Extras', 'Net Debt/EBITDA']
```

Out[6]:

| Types | Total Debt | Net Debt | Interest Coverage | Current Liabilities | Sales/Revenues | PE | EBITDA | ROA | EBITDA Margin | CFO |
|---|---|---|---|---|---|---|---|---|---|---|
| **Counts** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

1 rows × 28 columns

QQplot

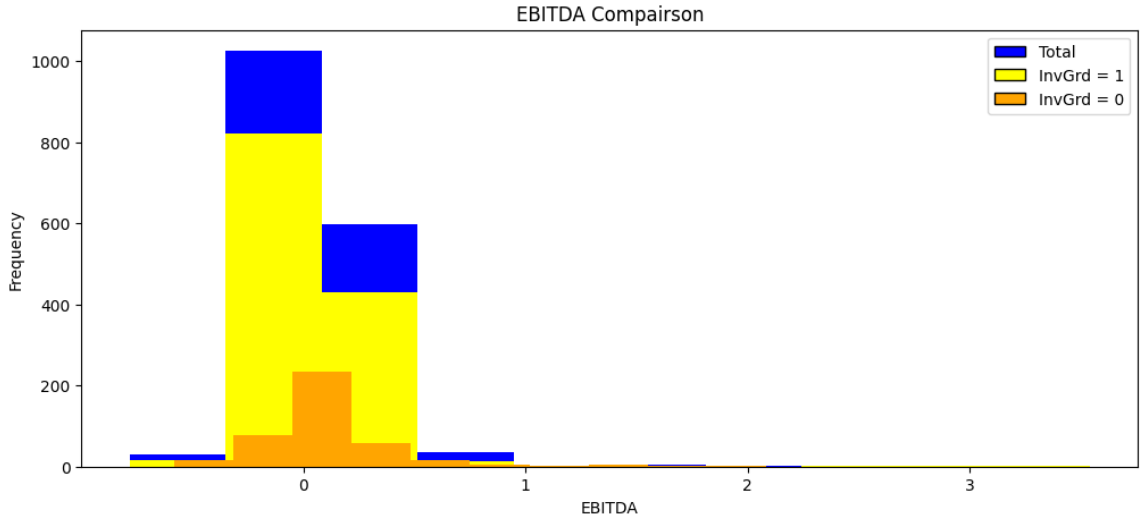


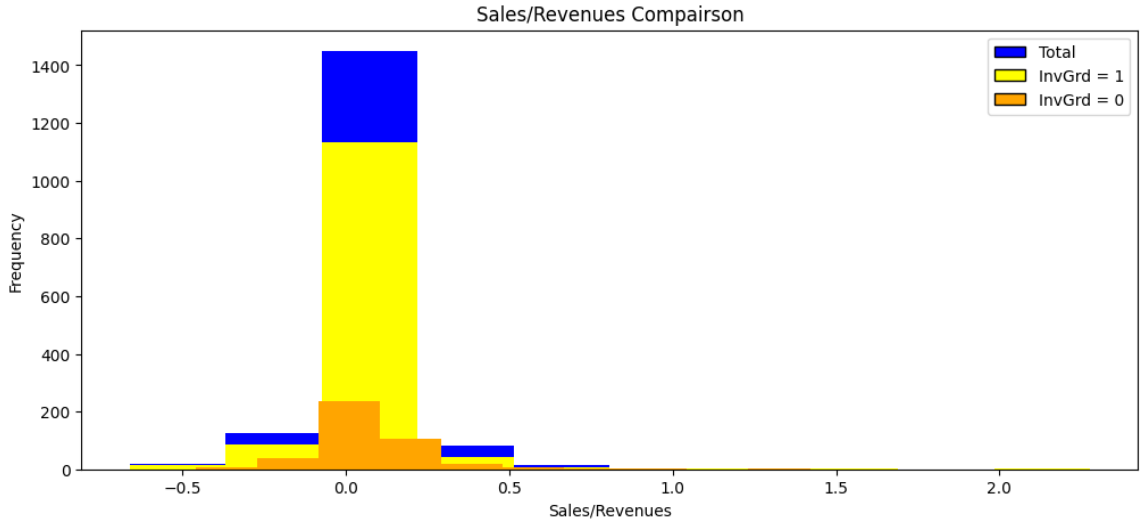

```
P-Value: 0.0
Reject H0: Client_Trade_Percentage is Normally distributed.
```

Print Summary of Data

```
       Sales/Revenues  Gross Margin       EBITDA  EBITDA Margin  \
count     1700.000000   1700.000000  1700.000000    1700.000000
mean         0.050378      0.026007     0.068718       0.021074
std          0.161910      0.273768     0.237365       0.189025
min         -0.661715     -0.794722    -0.782254      -0.805153
25%         -0.005693     -0.020028    -0.022640      -0.042771
50%          0.034000      0.003403     0.049482       0.011134
75%          0.083004      0.025595     0.124533       0.060566
max          2.277229      3.202713     3.542425       4.141182

       Net Income Before Extras   Total Debt     Net Debt      LT Debt  \
count               1700.000000  1700.000000  1700.000000  1700.000000
mean                  0.123026     0.822405    -0.419810     1.255168
std                  14.475689    13.317075    28.385702    16.224453
min                -289.000000    -0.903014  -493.305578    -0.921515
25%                  -0.158478    -0.076316    -0.120725    -0.094767
50%                   0.056627     0.005886    -0.003060    -0.002078
75%                   0.222219     0.136449     0.160251     0.174735
max                 478.280075   281.604237   865.194595   289.388178

           ST Debt         Cash  ...          CFO  Interest Coverage  \
count  1700.000000  1700.000000  ...  1700.000000        1700.000000
mean      3.142797     0.466620  ...    -0.189317           0.298785
std      51.986550     1.859494  ...     5.668669           5.265291
min      -0.997692    -0.990982  ...  -161.609425          -0.991976
25%      -0.337959    -0.195117  ...    -0.115159          -0.096996
50%       0.043092     0.075820  ...     0.046983           0.043216
75%       0.649475     0.483113  ...     0.216432           0.177340
max    2038.000000    36.980037  ...    13.005788         182.131887

       Total Liquidity  Current Liquidity  Current Liabilities  \
count      1700.000000        1700.000000          1700.000000
mean         -0.855714           0.436002             0.072802
std          22.926862           1.904282             0.266471
min        -502.000000          -0.994141            -0.684678
25%          -0.857013          -0.227327            -0.072734
50%          -0.229098           0.040446             0.041785
75%           0.512778           0.416067             0.161215
max         280.138728          34.372455             4.194381

       EPS Before Extras           PE          ROA          ROE       InvG
rd
count        1700.000000  1700.000000  1700.000000  1700.000000  1700.0000
00
mean            0.032196     0.497705     0.019394    -0.217604     0.7570
59
std             6.151994    12.102502    14.594193    15.389000     0.4289
86
min           -96.250000   -59.795133  -305.462167  -373.837267     0.0000
00
25%            -0.152894    -0.293521    -0.208483    -0.233955     1.0000
00
50%             0.066027    -0.040405    -0.009403    -0.020392     1.0000
00
75%             0.236046     0.168897     0.156136     0.201596     1.0000
00
max           187.000000   381.243282   474.847172   343.145356     1.0000
00

[8 rows x 27 columns]
```

Plot Data

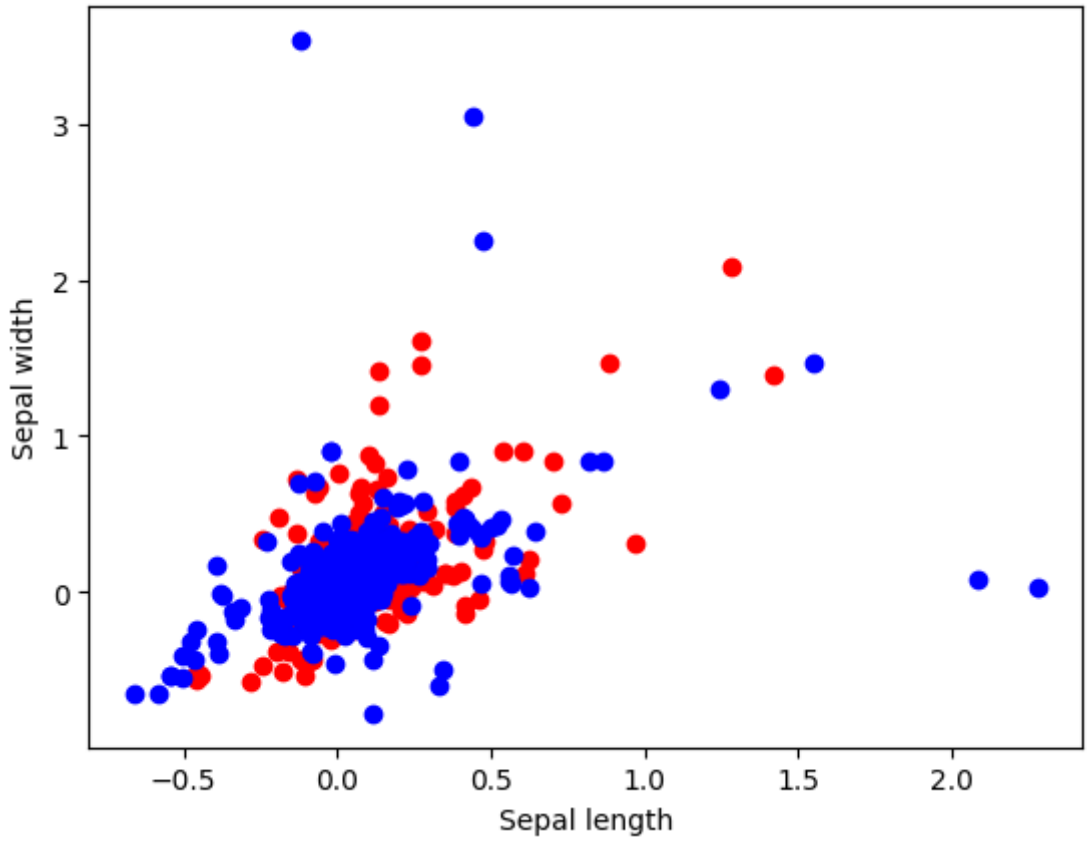



Cross Plotting Pairs of Attributes (Scatter Plot)

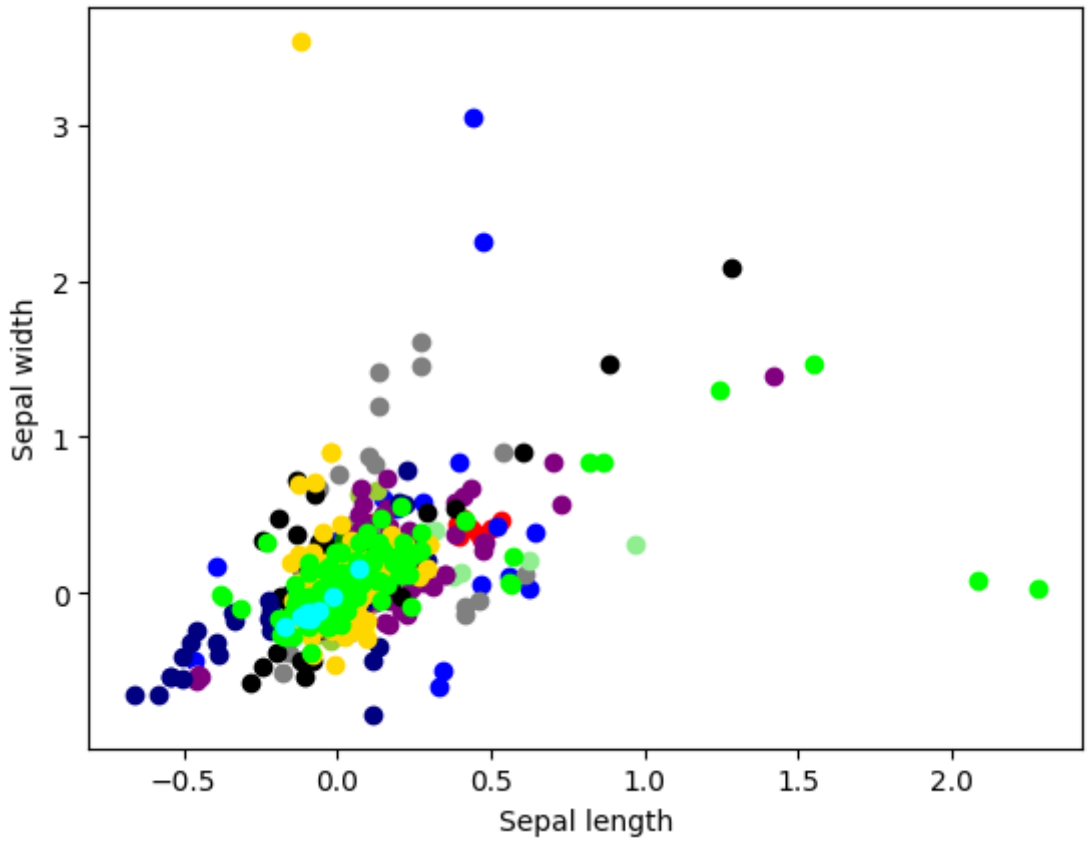
Out[11]:

```
Text(0, 0.5, 'Sepal width')
```




Out[12]:

```
Text(0, 0.5, 'Sepal width')
```



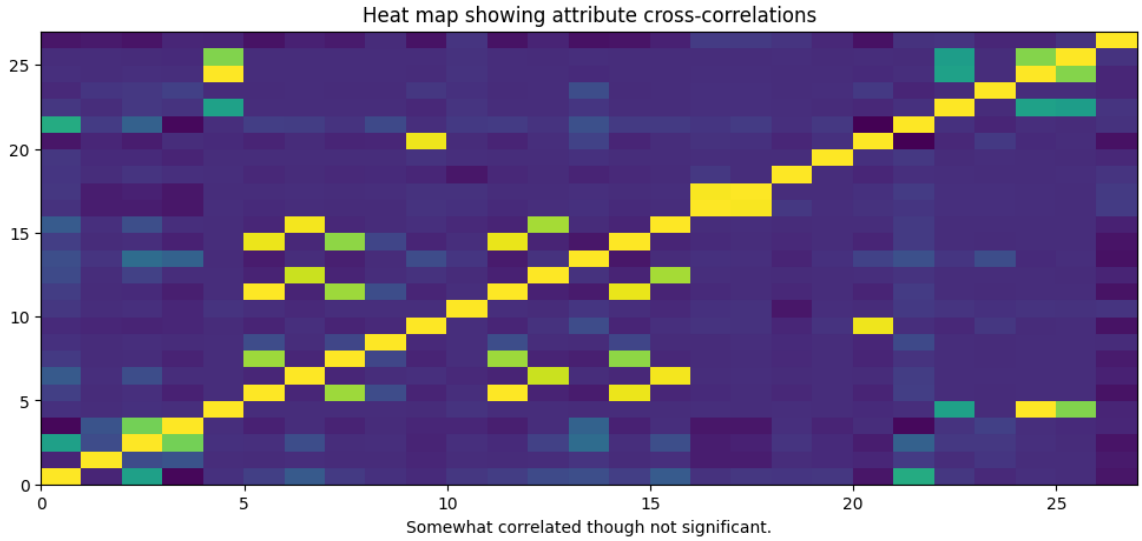
Correlation

Out[13]:

| | Sales/Revenues | Gross Margin | EBITDA | EBITDA Margin | Net Income Before Extras | Total Debt | Net D |
|---|---|---|---|---|---|---|---|
| **Sales/Revenues** | 1.000000 | -0.026318 | 0.500178 | -0.124079 | 0.012024 | 0.068938 | 0.1757 |
| **Gross Margin** | -0.026318 | 1.000000 | 0.114092 | 0.147886 | -0.001061 | -0.002665 | 0.0042 |
| **EBITDA** | 0.500178 | 0.114092 | 1.000000 | 0.757142 | 0.012565 | 0.008949 | 0.1192 |
| **EBITDA Margin** | -0.124079 | 0.147886 | 0.757142 | 1.000000 | 0.003331 | -0.039804 | 0.0003 |
| **Net Income Before Extras** | 0.012024 | -0.001061 | 0.012565 | 0.003331 | 1.000000 | -0.001065 | 0.0008 |
| **Total Debt** | 0.068938 | -0.002665 | 0.008949 | -0.039804 | -0.001065 | 1.000000 | -0.0222 |
| **Net Debt** | 0.175741 | 0.004266 | 0.119251 | 0.000336 | 0.000867 | -0.022209 | 1.0000 |
| **LT Debt** | 0.048960 | -0.003149 | -0.000665 | -0.037009 | -0.001162 | 0.833567 | -0.0213 |
| **ST Debt** | 0.014987 | -0.005417 | 0.004844 | -0.006310 | -0.000221 | 0.118240 | 0.0011 |
| **Cash** | -0.008088 | -0.024540 | -0.030773 | -0.023997 | -0.006703 | -0.030002 | 0.0071 |
| **Free Cash Flow** | 0.035716 | 0.001920 | 0.009102 | -0.021301 | 0.023523 | 0.002539 | -0.0132 |
| **Total Debt/EBITDA** | 0.056092 | -0.005690 | -0.009527 | -0.051062 | -0.001490 | 0.999328 | -0.0314 |
| **Net Debt/EBITDA** | 0.110201 | -0.000236 | 0.074284 | -0.000145 | 0.000312 | -0.035136 | 0.9074 |
| **Total MV** | 0.123111 | 0.024361 | 0.256941 | 0.209314 | -0.005511 | -0.056449 | 0.0001 |
| **Total Debt/MV** | 0.062128 | -0.003954 | 0.002857 | -0.041728 | -0.001252 | 0.964306 | -0.0192 |
| **Net Debt/MV** | 0.176797 | 0.009945 | 0.123084 | 0.004226 | 0.000701 | -0.022305 | 0.9787 |
| **CFO/Debt** | 0.016833 | -0.055521 | -0.054541 | -0.076554 | 0.002084 | -0.008508 | 0.0015 |
| **CFO** | 0.034069 | -0.055569 | -0.041064 | -0.075326 | 0.001832 | 0.000924 | 0.0027 |
| **Interest Coverage** | 0.032716 | -0.002079 | 0.028118 | 0.008147 | 0.002233 | -0.016078 | -0.0035 |
| **Total Liquidity** | 0.035747 | -0.011190 | -0.008799 | -0.033862 | -0.001938 | -0.000801 | 0.0004 |
| **Current Liquidity** | -0.081346 | -0.022793 | -0.054680 | -0.000036 | -0.008131 | -0.031600 | -0.0038 |
| **Current Liabilities** | 0.553807 | 0.051386 | 0.207526 | -0.119068 | 0.004226 | 0.062364 | 0.0591 |
| **EPS Before Extras** | 0.034722 | 0.000857 | 0.042191 | 0.021589 | 0.506547 | -0.001424 | 0.0010 |
| **PE** | -0.014842 | 0.029146 | 0.040732 | 0.071426 | -0.003166 | -0.003652 | -0.0003 |
| **ROA** | 0.007251 | -0.001583 | 0.007913 | 0.002072 | 0.997349 | -0.001406 | -0.0007 |
| **ROE** | -0.000206 | 0.000182 | -0.005701 | -0.008328 | 0.782491 | 0.000301 | -0.0019 |
| **InvGrd** | -0.080836 | -0.066103 | -0.085951 | -0.024112 | -0.027919 | -0.090372 | -0.0421 |

27 rows × 27 columns

Correlation Visualization



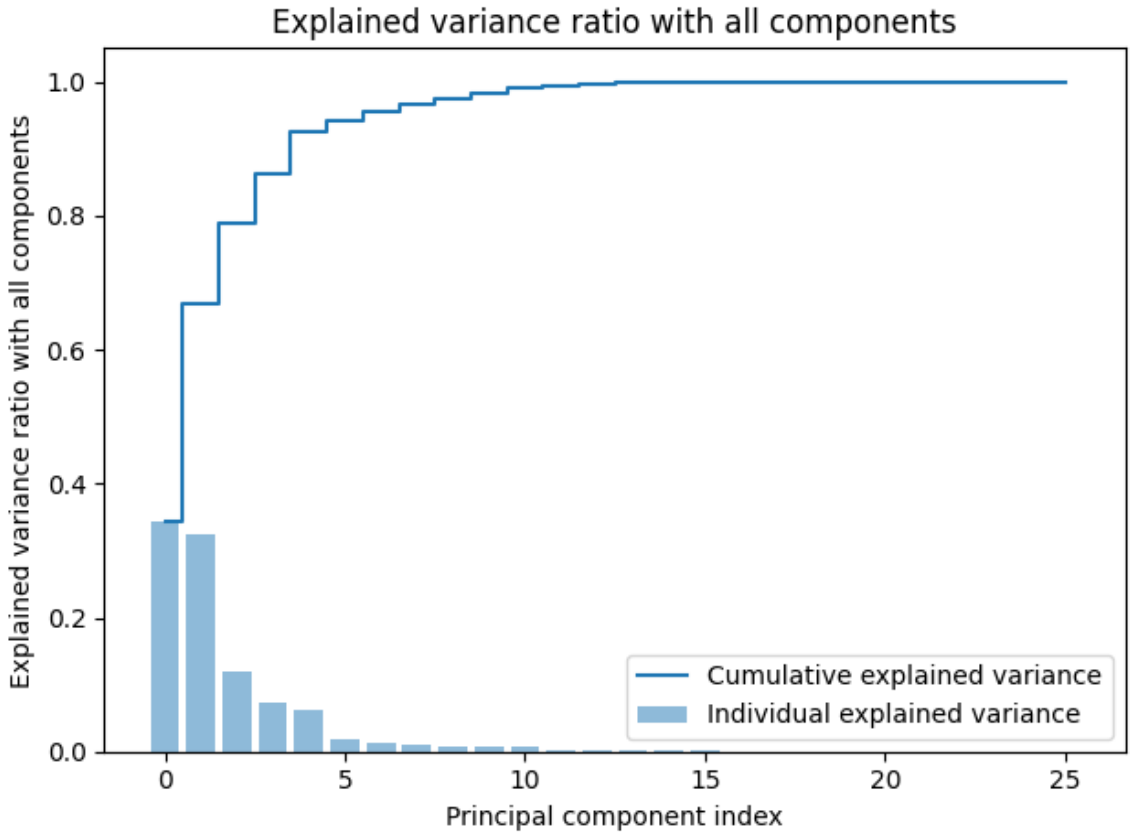

# 2) Preprocessing, feature extraction, feature selection

Drop Missing Value

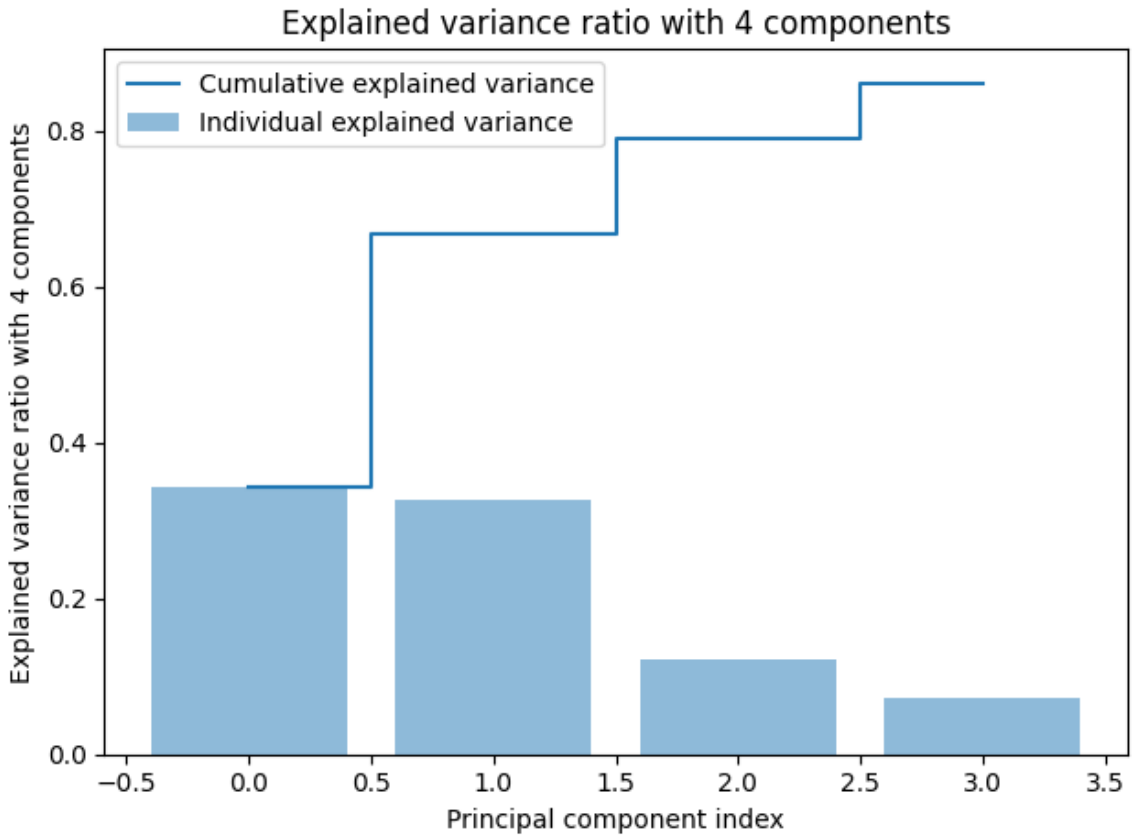
Preprocessing the Data

Variance Ratio

```
Explained Variance Ratio with all components:
 [3.43047199e-01 3.25324928e-01 1.20717674e-01 7.24409782e-02
 6.28795206e-02 1.75279579e-02 1.34937683e-02 9.49562533e-03
 8.50716279e-03 8.32593599e-03 7.69529261e-03 3.29035410e-03
 3.20490490e-03 1.70619518e-03 1.34038953e-03 8.30711631e-04
 6.59285613e-05 3.64523500e-05 2.28387916e-05 1.42339841e-05
 1.20727109e-05 8.64794299e-06 5.25124749e-06 3.85369188e-06
 1.97773212e-06 1.44521331e-07]

Culmulative Variance Ratio with all components:
 [0.3430472  0.66837213 0.7890898  0.86153078 0.9244103  0.94193826
 0.95543203 0.96492765 0.97343481 0.98176075 0.98945604 0.9927464
 0.9959513  0.9976575  0.99899789 0.9998286  0.99989453 0.99993098
 0.99995382 0.99996805 0.99998012 0.99998877 0.99999402 0.99999788
 0.99999986 1.        ]
```



```
Explained Variance Ratio with 4 components:
 [0.3430472  0.32532493 0.12071767 0.07244098]

Culmulative Variance Ratio with 4 components:
 [0.3430472  0.66837213 0.7890898  0.86153078]
```




# 3 + 4) Model fitting and evaluation + Hyperparameter tuning

## Model: Binary Classification (Investment grade)

Model : SVM + Gridsearch (Hyperparameter tuning) w PCA

```
SVC best parameter : {'C': 0.1, 'kernel': 'linear'}
SVC best estimator : SVC(C=0.1, kernel='linear')
SVC best score:  0.7594117647058823
```

Model : Decision Tree + Gridsearch (Hyperparameter tuning) w PCA

```
DT best parameter : {'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_s
plit': 2}
DT best estimator : DecisionTreeClassifier(max_depth=4)
DT best score:  0.7476470588235294
```

## Model: Multiclass Classification (Moody)

One rest model

Model : Logistic Regresssion on Scaled data + One rest w/ PCA

```
Multi-Class LR best parameter : {'C': 0.1, 'penalty': 'l2'}
Multi-Class LR best estimator : LogisticRegression(C=0.1, multi_class='ov
r')
Multi-Class LR best score:  0.16882352941176473
```

Model : Decision Tree + OneRest w PCA

```
Multi-Class DT best parameter : {'estimator__max_depth': 10, 'estimator__m
in_samples_leaf': 3, 'estimator__min_samples_split': 2}
Multi-Class DT best estimator : OneVsRestClassifier(estimator=DecisionTree
Classifier(max_depth=10,
                                                    min_samples_leaf=3))
Multi-Class DT best score:  0.18294117647058825
```

Model : SVC + One v One w/ PCA

```
Multi-Class SVC best parameter : {'estimator__C': 0.1, 'estimator__gamma':
1, 'estimator__kernel': 'rbf'}
Multi-Class SVC best estimator : OneVsOneClassifier(estimator=SVC(C=0.1, g
amma=1))
Multi-Class SVC best score:  0.20294117647058824
```

# 5) Ensembling

## Model: Binary Classification (Investment grade)

Model : Random Forest + Gridsearch (Hyperparameter tuning) w/ PCA

```
RF best parameter : {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_
split': 5, 'n_estimators': 100}
RF best estimator : RandomForestClassifier(max_depth=10, min_samples_split
=5, random_state=1)
RF best score:  0.741764705882353
```

Model : Random Forest + Gridsearch (Hyperparameter tuning) w PCA

```
RF + PCA best parameter : {'max_depth': 5, 'min_samples_leaf': 1, 'min_sam
ples_split': 5, 'n_estimators': 50}
RF + PCA best estimator : RandomForestClassifier(max_depth=5, min_samples_
split=5, n_estimators=50,
                       random_state=1)
RF + PCA best score:  0.7452941176470589
```

## Model: Multiclass Classification (Moody)

Model : Random Forest + One Rest w/ PCA

```
Multi-Class RF best parameter : {'estimator__max_depth': 5, 'estimator__n_
estimators': 100}
Multi-Class RF best estimator : OneVsRestClassifier(estimator=RandomForest
Classifier(max_depth=5,
                                                    random_state=1))
Multi-Class RF best score:  0.20941176470588233
```

Model : Random Forest + One Rest w PCA

```
Multi-Class RF + PCA best parameter : {'estimator__max_depth': 2, 'estimat
or__n_estimators': 50}
Multi-Class RF + PCA best estimator : OneVsRestClassifier(estimator=Random
ForestClassifier(max_depth=2,
                                                    n_estimators=50,
                                                    random_state=1))
Multi-Class RF + PCA best score:  0.18176470588235294
```

# 6) Conclusions

**Binary Classification**

When it comes to binary classification, the Support Vector Machine (SVM) method has shown to be highly effective, achieving an accuracy rate of 76% using 5-fold cross-validation. Remarkably, this even outperformed ensemble models such as Random Forest. The second-best model was the decision tree. Based on these results, we can conclude that non-ensemble models are advantageous in binary classification might cause by low risk of overfitting.

**Multi-Class Classification**

For multi-class classification, I opted to use the One v One + SVM model due to its faster computation time compared to the One v Rest method. However, the Random Forest + One v Rest model outperformed it by 21% in terms of accuracy. In this scenario, an ensemble model proved to be the superior choice.

# 7) Appendix

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

# 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

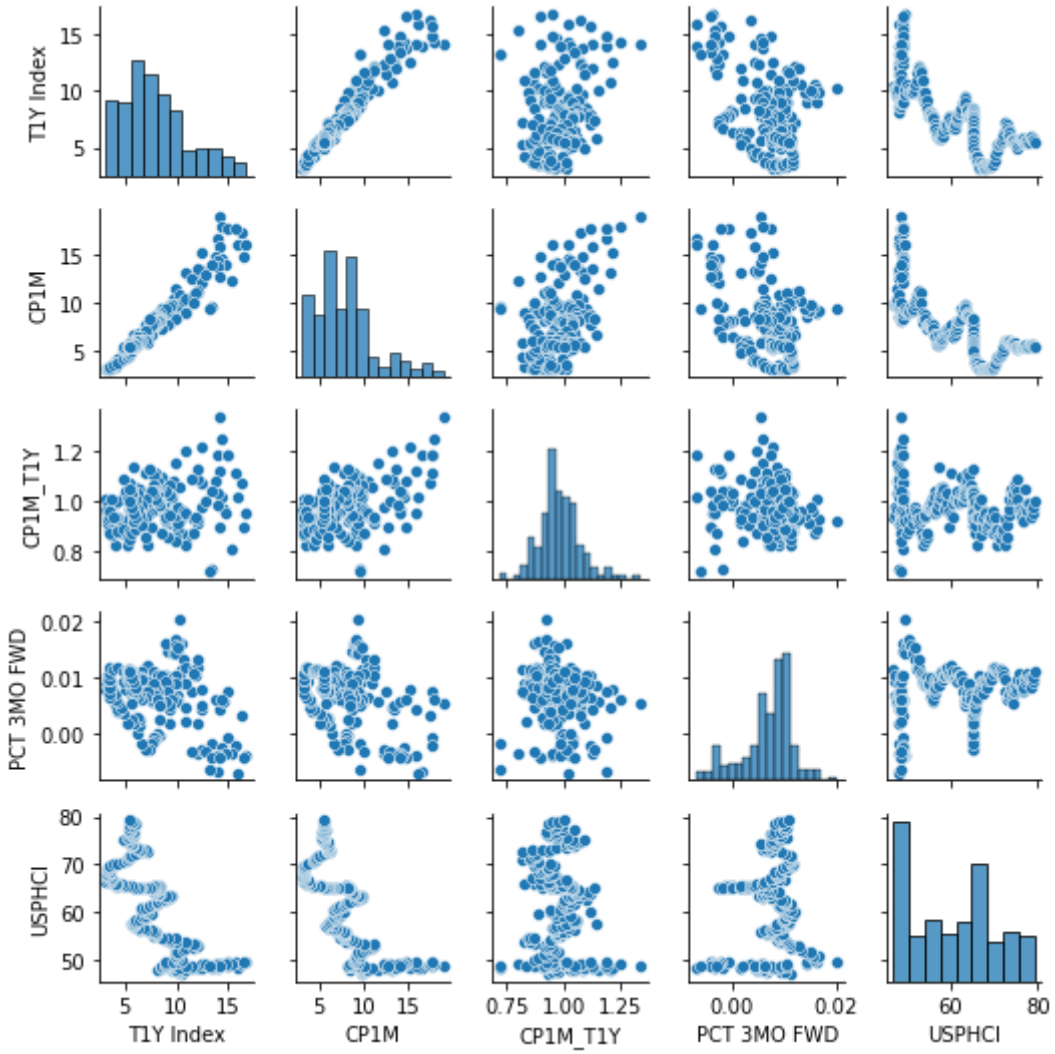
Out[4]:

| 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 |

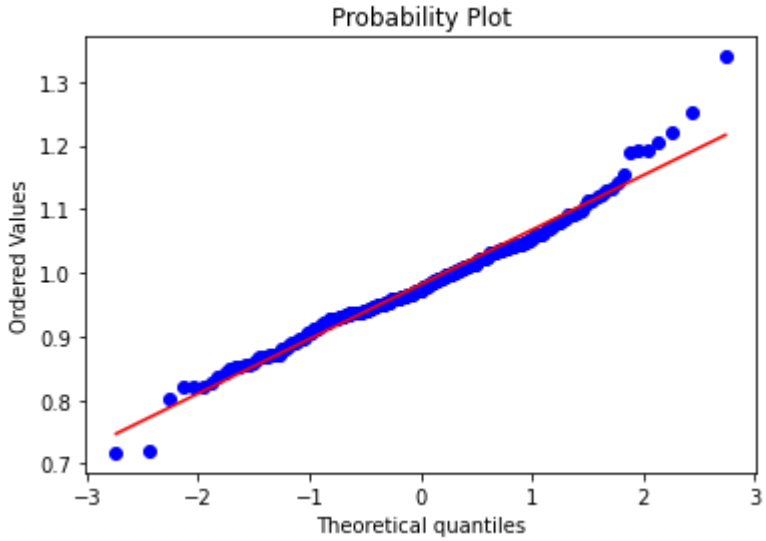
# Summary of Statistics

```
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

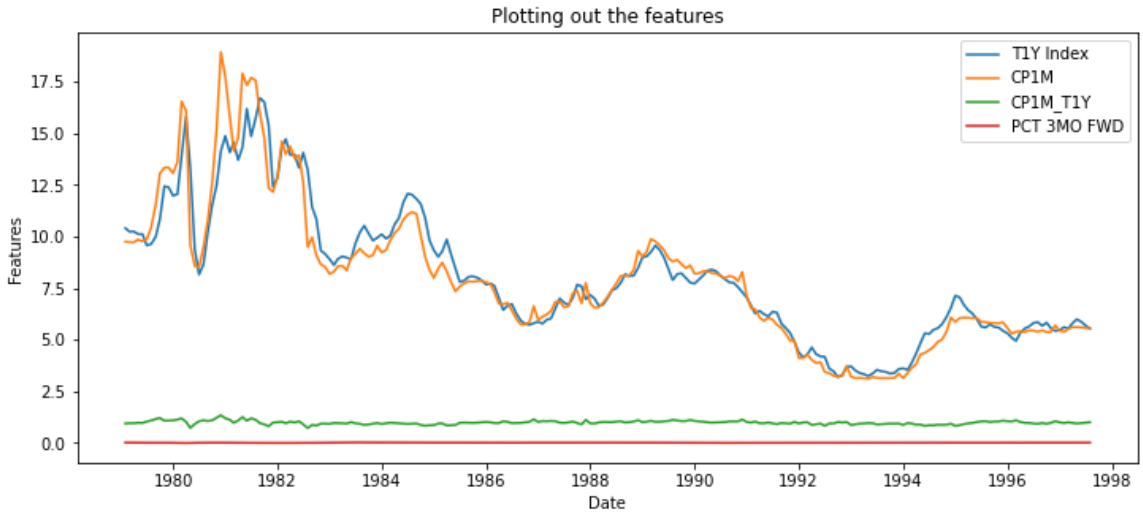



P-Value: 8.424094272178417e-05
Reject H0: Client_Trade_Percentage is Normally distributed.

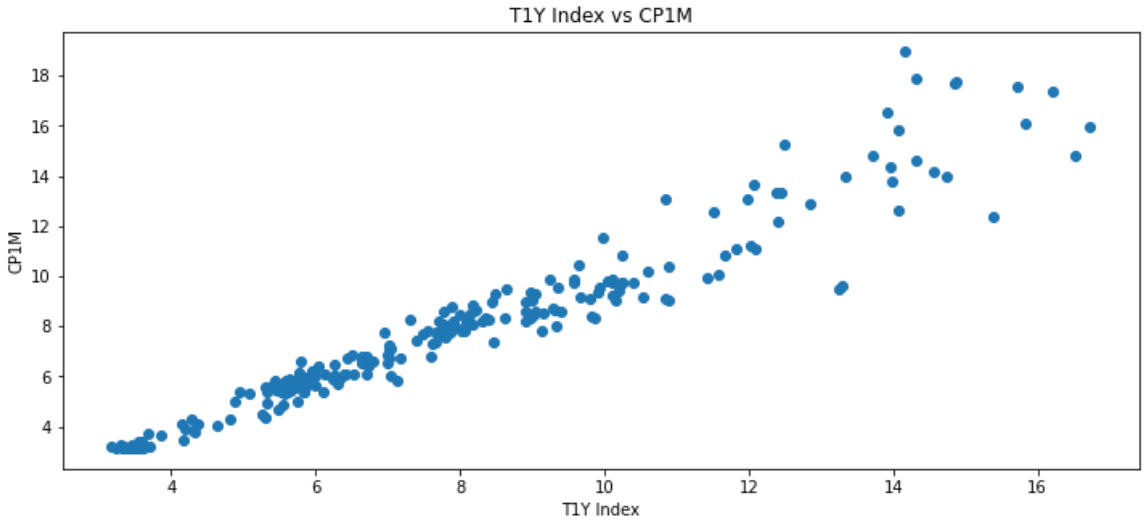
# Print Summary of Data

| | T1Y Index | T2Y Index | T3Y Index | T5Y Index | T7Y Index | T10Y Index |
|---|---|---|---|---|---|---|
| count | 223.000000 | 223.000000 | 223.000000 | 223.000000 | 223.000000 | 223.000000 |
| mean | 8.030717 | 8.410673 | 8.563587 | 8.808655 | 8.979776 | 9.073498 |
| std | 3.158575 | 2.954431 | 2.820405 | 2.647742 | 2.542686 | 2.447525 |
| min | 3.180000 | 3.840000 | 4.170000 | 4.710000 | 5.050000 | 5.330000 |
| 25% | 5.735000 | 6.180000 | 6.410000 | 6.695000 | 6.965000 | 7.175000 |
| 50% | 7.670000 | 8.000000 | 8.130000 | 8.330000 | 8.520000 | 8.610000 |
| 75% | 9.840000 | 10.075000 | 10.375000 | 10.525000 | 10.640000 | 10.685000 |
| max | 16.720000 | 16.460000 | 16.220000 | 15.930000 | 15.650000 | 15.320000 |

# Plot Data




# Cross Plotting Pairs of Attributes (Scatter Plot)

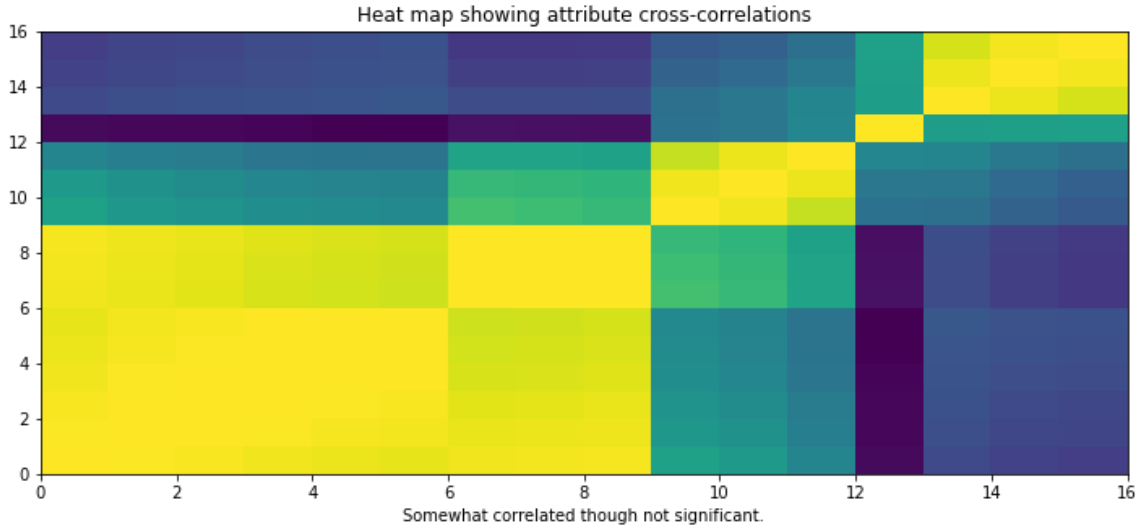


# Correlation

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



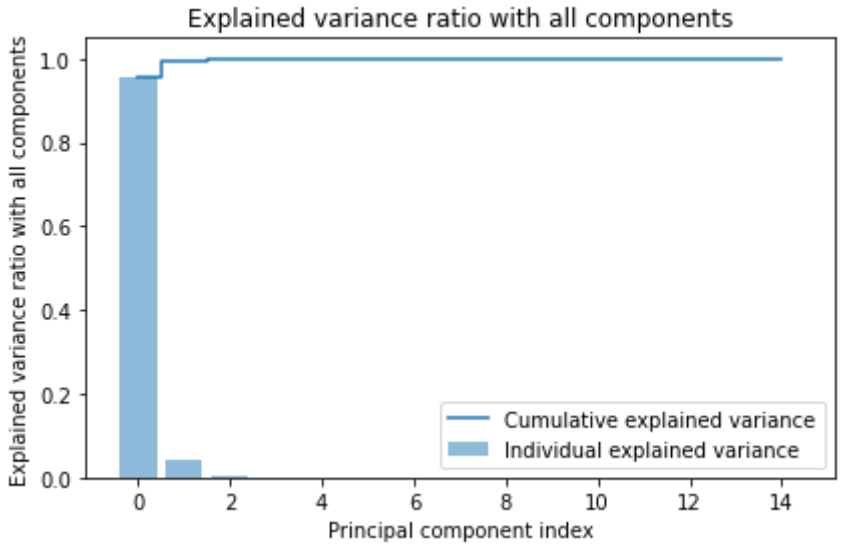
# 2) Preprocessing, feature extraction, feature selection,

## Drop Missing Value

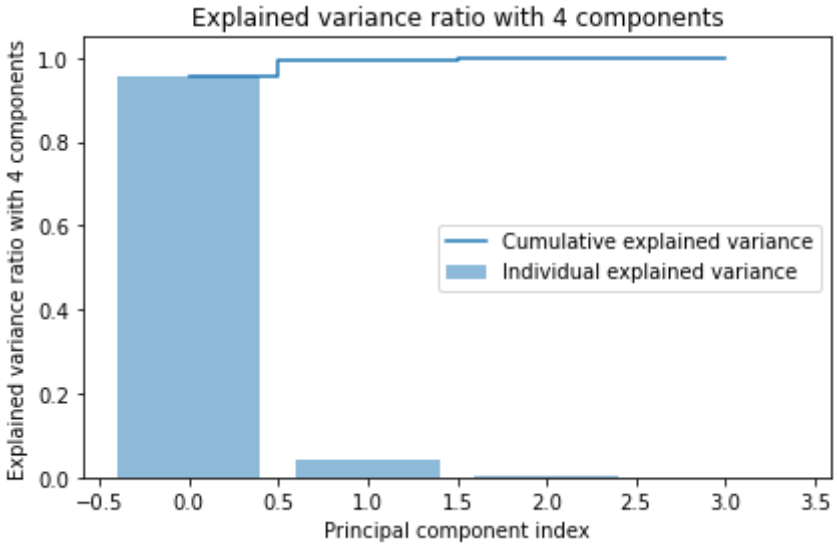
## Preprocessing the Data

## 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.        ]
```



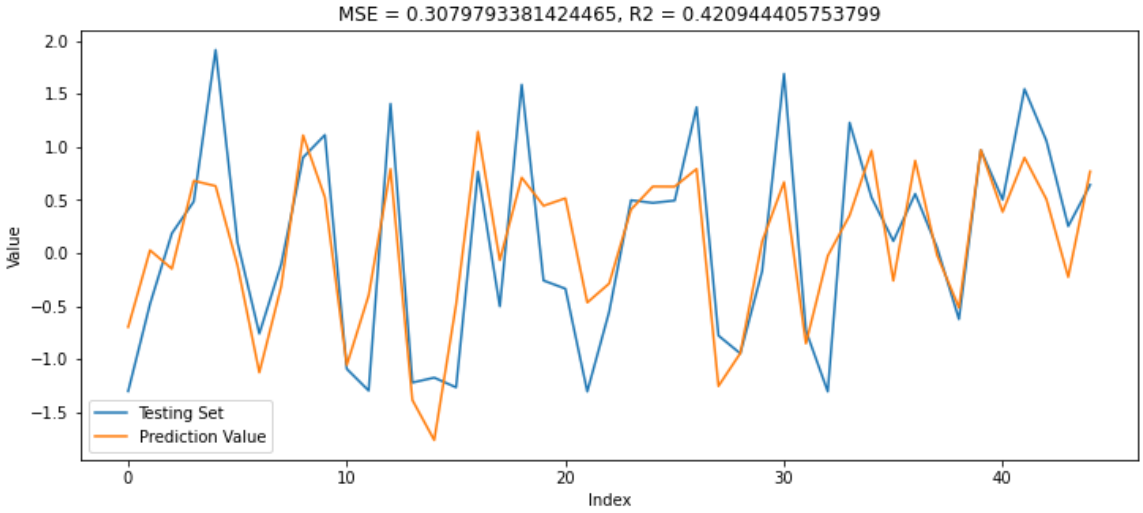
```
Explained Variance Ratio with 4 components:
 [9.55657070e-01 3.99135802e-02 3.35994699e-03 7.55389913e-04]

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

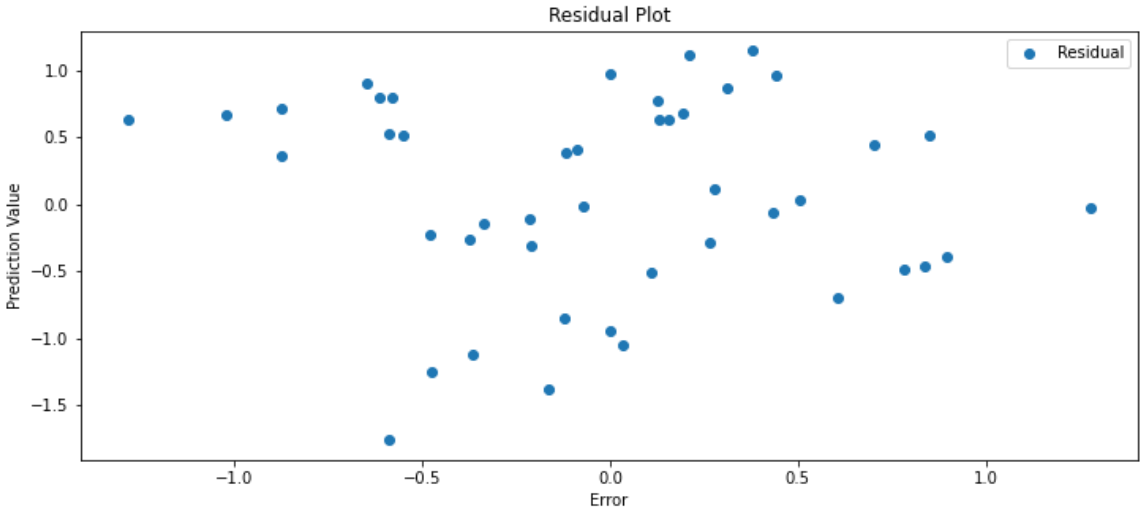



# Model fitting and evaluation, (you should fit at least 3 different machine learning models) & Hyperparameter tunning

# Simple Linear Regresion with PCA

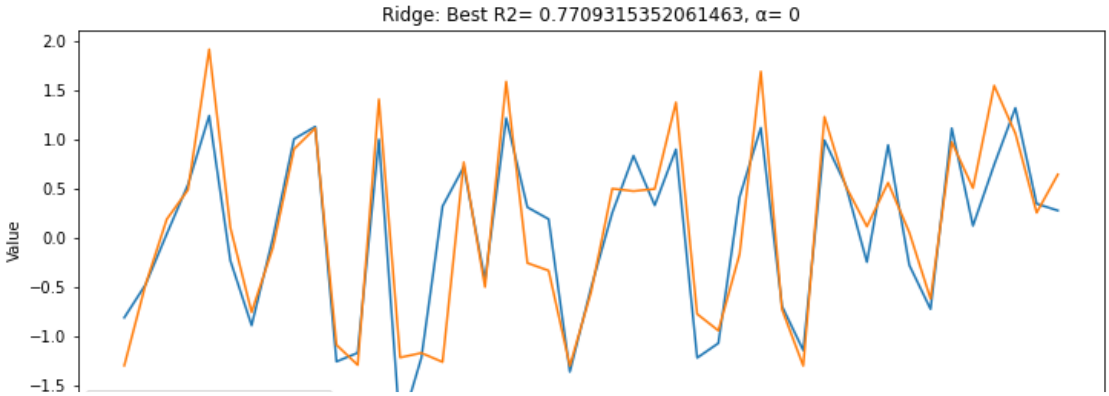
```
Coefficients:  [[-0.24109023 -0.09217819 -0.18879978 -0.18903653]]
Intercept:  [-0.004937]
```



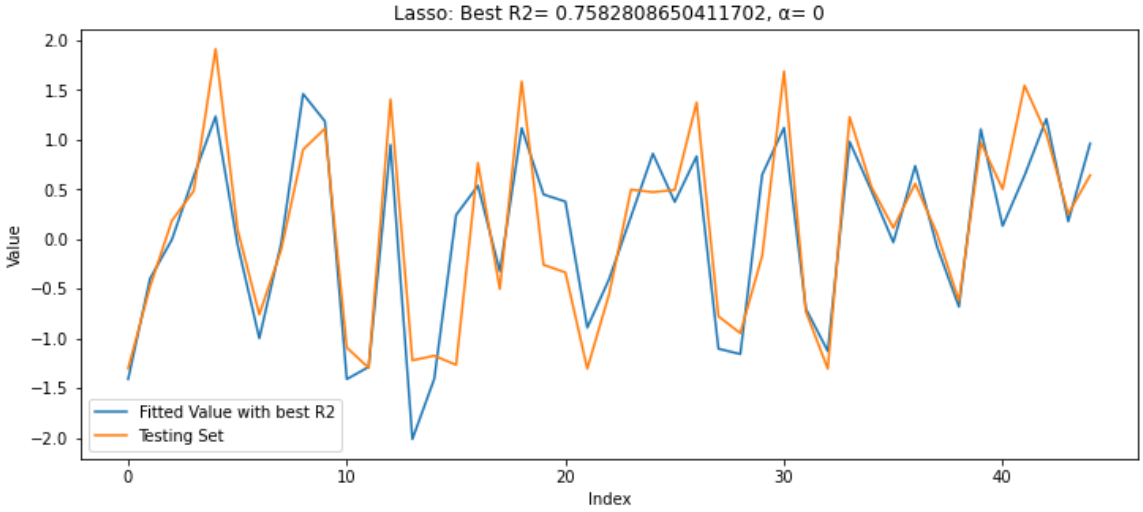
# Ridge Regression with Hyperparameter Tunning

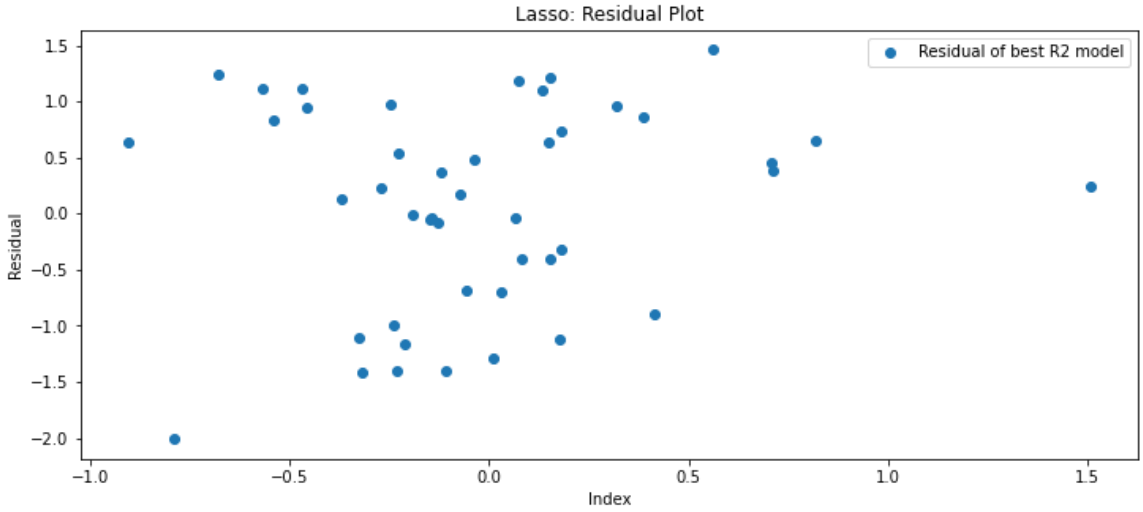
```
α=  0 , R2=  0.7709315352061463 , MSE =  0.17110414461321527
α=  0.1 , R2=  0.7606253656128817 , MSE =  0.1664053287775124
α=  1.0 , R2=  0.6291381510053733 , MSE =  0.2084804119213952
α=  10.0 , R2=  0.5060668309520844 , MSE =  0.26135570870022196
α=  100.0 , R2=  0.37994266062846904 , MSE =  0.2918654056718467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

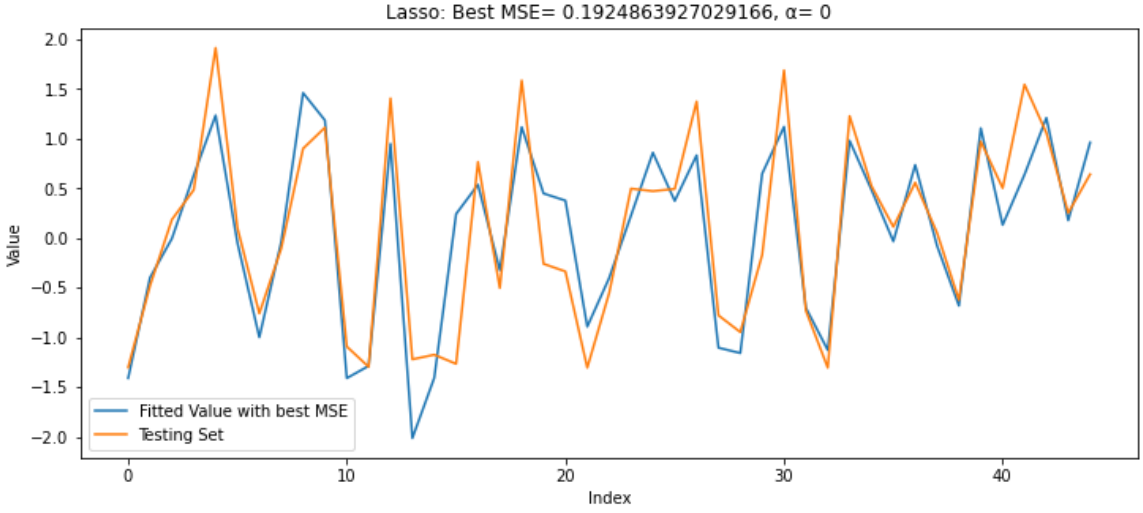


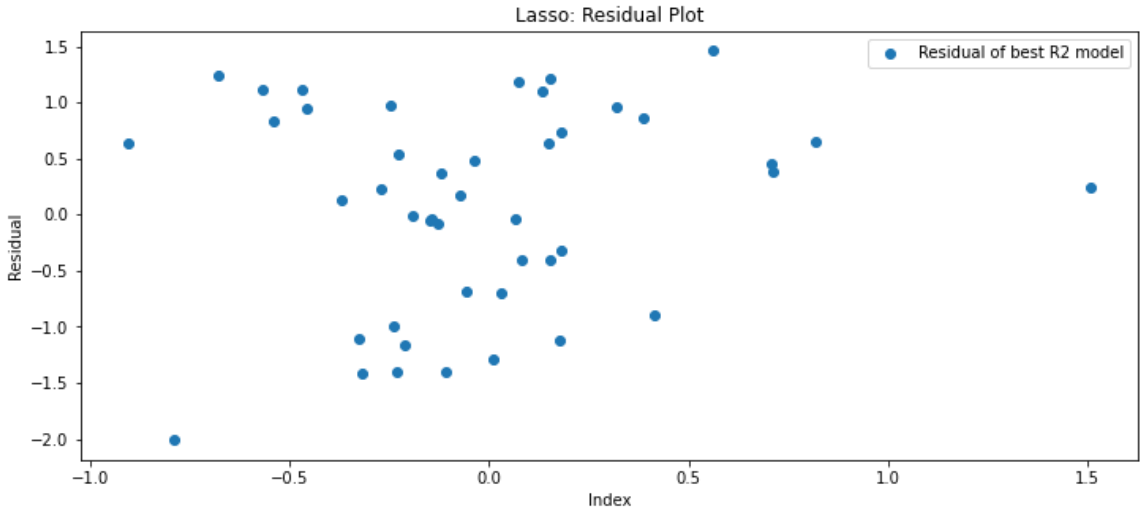
# Lasso Regression with Hyperparameter tunning

```
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
```

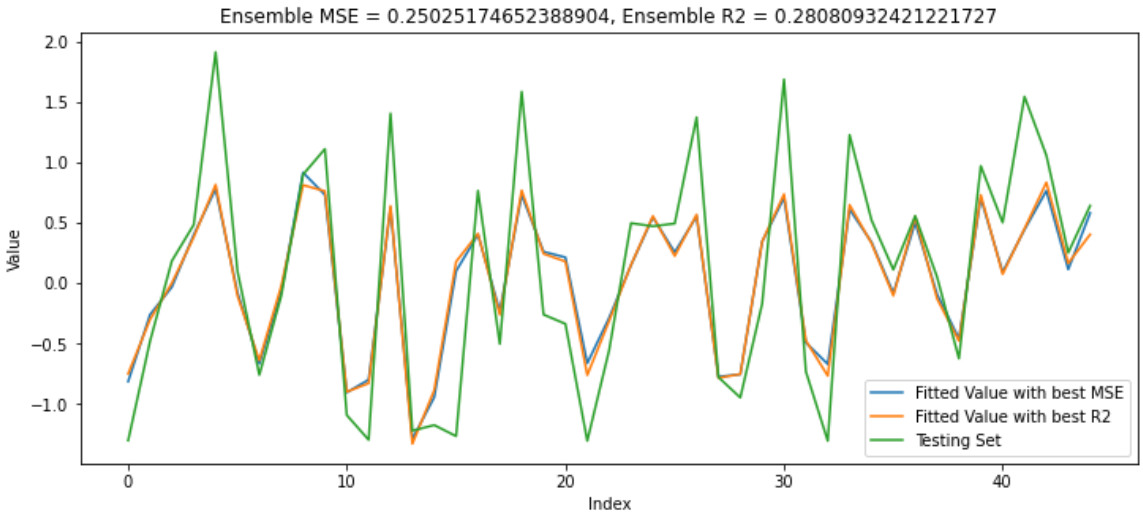





# 5) Ensembling




# 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