**Machine Learning in Finance Lab – Week 05**

**Liao, Yu-Ching**

**Part 1. EDA:**

* **Scatter Plot Metrics:**

A picture containing text

Description automatically generated

* **Shape of Data:**

The number of Columns is 31 .

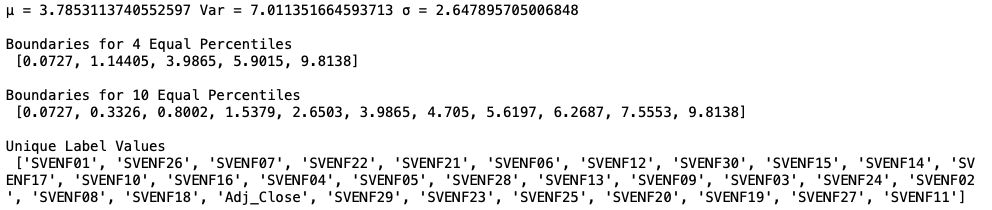
The number of Rows is 8071 .

* **Nature (truncated):**

**Table

Description automatically generated**

* **Summary of Statistics:**



* **QQ Plot:**

**Chart, line chart

Description automatically generated**

* **Summary of data:**

**Table

Description automatically generated**

* **Plot out data:**

**Chart, line chart

Description automatically generated**

* **Cross Plotting Pairs:**

**Chart, line chart, scatter chart

Description automatically generated**

* **Correlations (truncated):**

**Table

Description automatically generated**

* **Correlations Visualization:**

**A picture containing chart

Description automatically generated**

**Part 2. PCA:**

Table

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence Chart

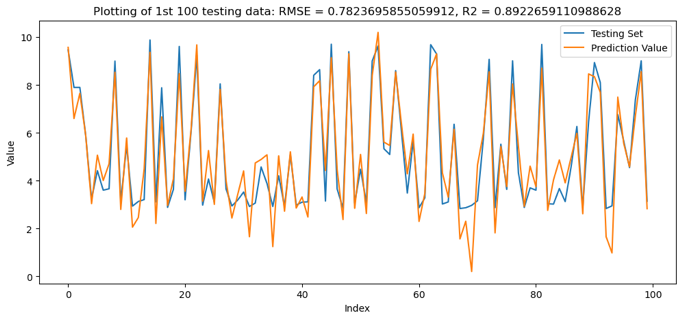
Description automatically generated

**Part 3. Linear Regression VS SVM:**

* **Linear Regression with all features:**

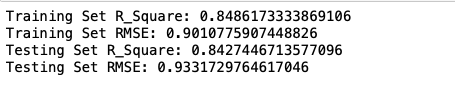
**Text

Description automatically generated**

Chart, line chart

Description automatically generated

* **Linear Regression with PCA:**



Line chart

Description automatically generated with medium confidence Chart, line chart

Description automatically generated

* **SVM** **with all features:**

Text

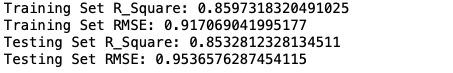
Description automatically generated

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

* **SVM with PCA:**

****

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

**Part 4. Conclusion:**

From the R\_square and RMSE, we can notice that, with PCA, the performance of the model **will not necessary be better**. In this case, I think it is because, reducing the features from 30 to 3, we may have as well reduced the pivotal signal of data.

And to compare Linear Regression and SVM, we can notice that **none of them is significantly better**. However, there is **large gap in their learning times (as shown below)**.

As a result, I would still use linear regression if there are no necessity to use SVM, and I would test several different number of components so that we can still retain pivotal signal in our model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Reg. (all) | Linear Reg. (PCA) | SVM (all) | SVM (PCA) |
| R\_Square (Train) | 0.89 | 0.84 | 0.88 | 0.85 |
| R\_Square (Test) | 0.89 | 0.84 | 0.89 | 0.85 |
| RMSE (Train) | 0.77 | 0.90 | 0.81 | 0.91 |
| RMSE (Test) | 0.78 | 0.92 | 0.82 | 0.95 |
| Times | 323ms | 282ms | 12.9s | 10.9s |

**Part 5. Appendix:**

Github:<https://github.com/yu7yu7/IE517_Machine-Learning-in-Finance-Lab/blob/main/IE517_SP23_HW5/ML_Week05_HW.ipynb>

**Part 6. Sign:**

My name is Yu-Ching Liao.

My NetID is ycliao3.

I hereby certify that I have read the University policy on Academic Integrity and that I am not in violation.