Stock Market Forecasting using Data Mining Technique

Presenter:

Marina Shah binti Muhammad Zabri Tan (WQD 180011)

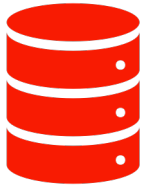
Objective

The objective of this report is to present framework of analysis based on clustering algorithm, decision tree and time series analysis to forecast stock market price.

Framework

Objective

To analyze and forecast stock market data



Data acquisition

Crawl data from websites

store data in MySQL database

Data cleaning

Clean the data and convert it for further analysis

Analysis & Modeling

Descriptive

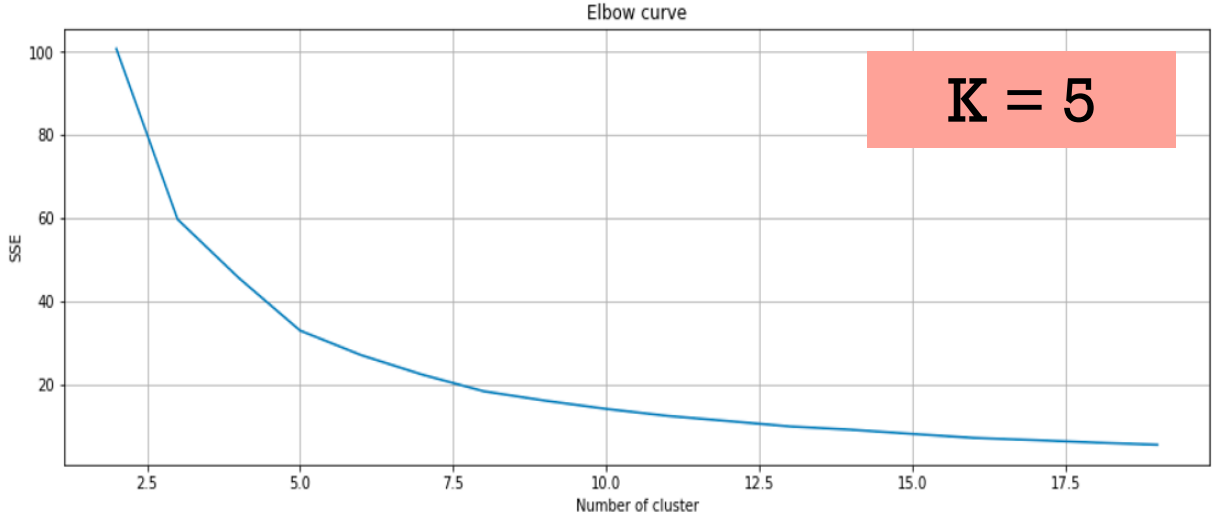
Explorative

Machine learning

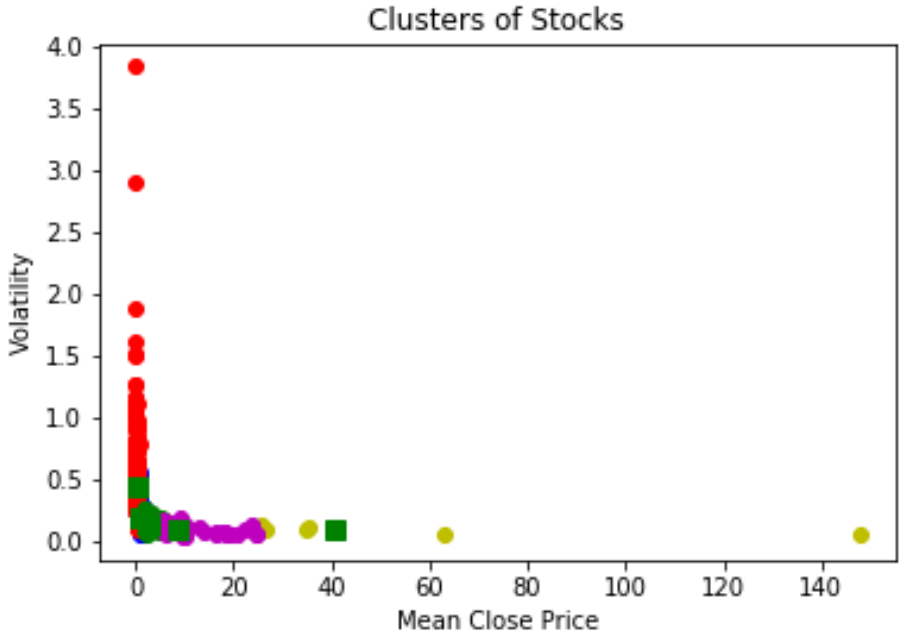
Visualization

Tables, graphs

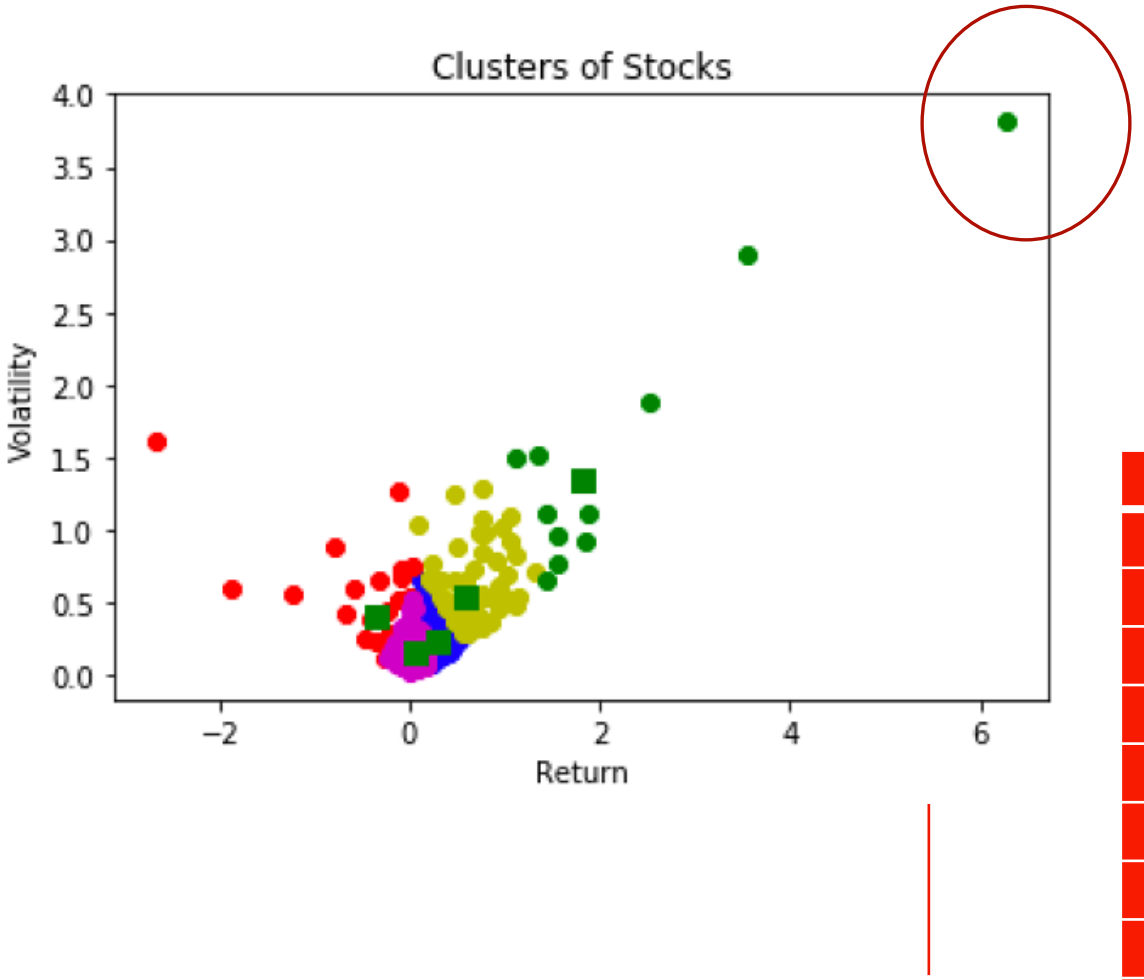
Clustering



K = 5



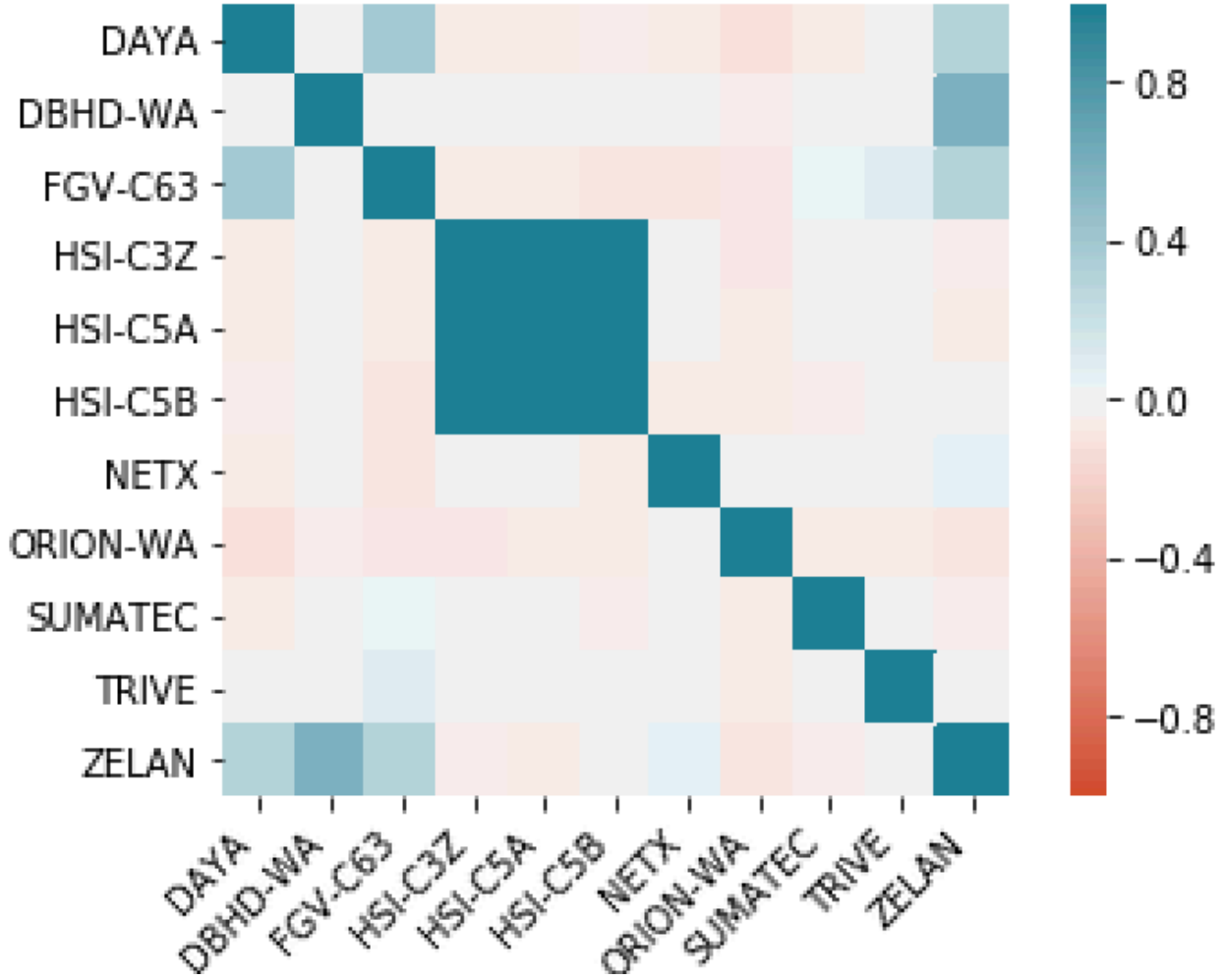
Name	Close Price	Volatility
NESTLE	147.667	0.045
SUMATEC	0.0092	3.8428



SUMATEC	R: 6.269	V: 3.812

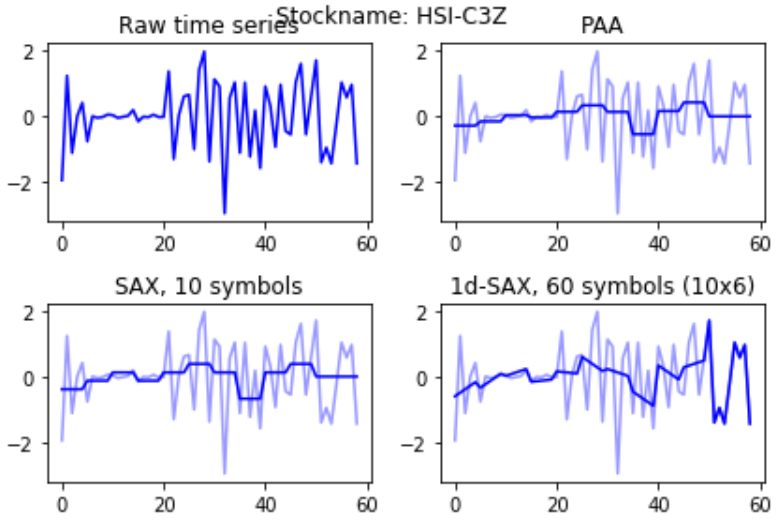
No	name	Returns	Volatility
1	SUMATEC	6.269	3.812
2	DAYA	3.558	2.890
3	DBHD-WA	2.524	1.875
4	ZELAN	1.878	1.120
5	FGV-C63	1.862	0.929
6	HSI-C5A	1.558	0.962
7	HSI-C5B	1.549	0.775
8	ORION-WA	1.447	0.650
9	HSI-C3Z	1.429	1.110
10	NETX	1.356	1.518
11	TRIVE	1.101	1.508

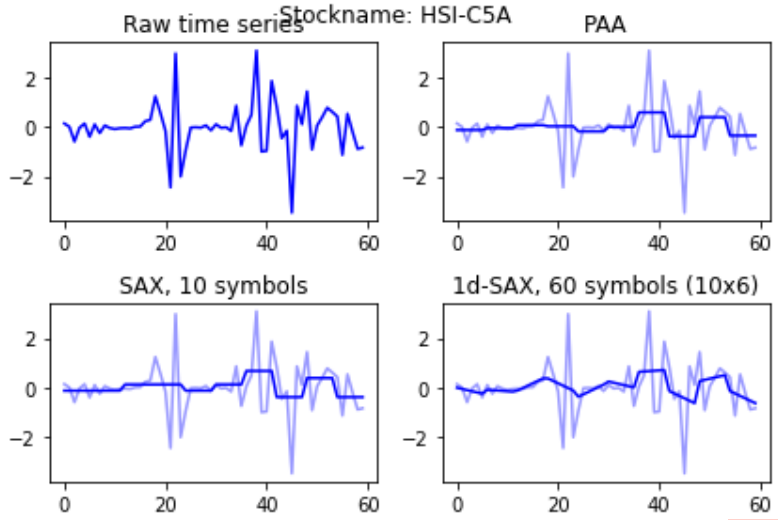
Clusters of stock: Volatility against Return



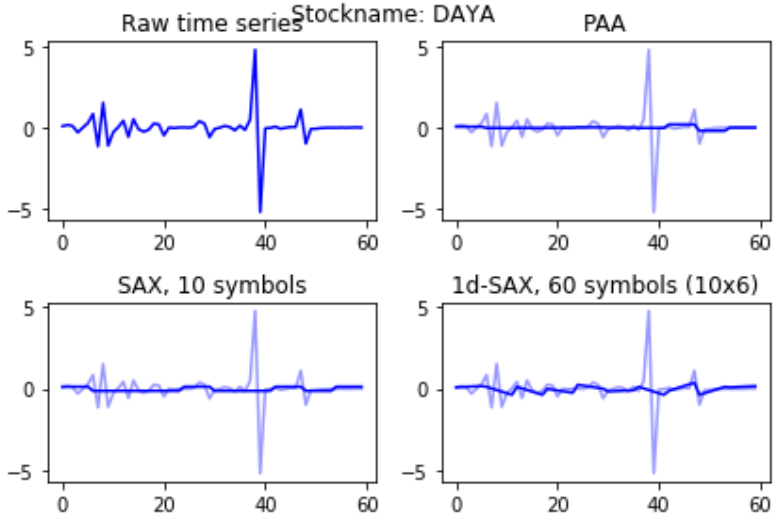
Most Positive correlation:	HSI-C3Z	HSI-C5A	0.998
Most Negative correlation:	DAYA	ORION-WA	-0.106

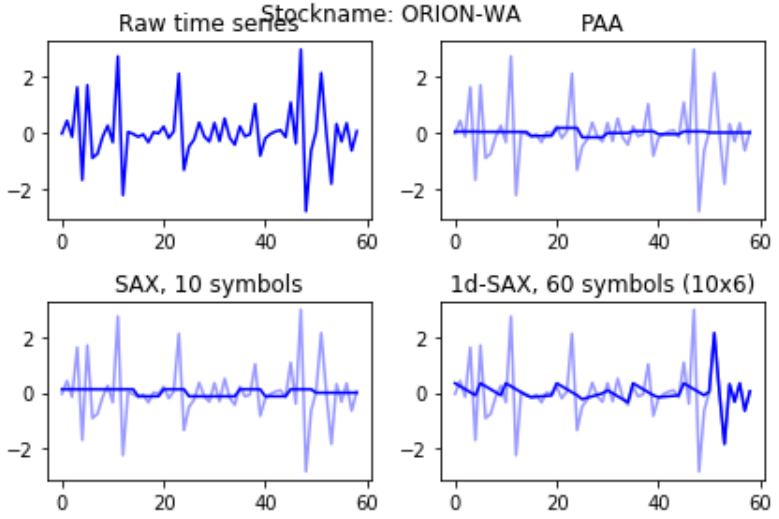
HSI-C3Z vs HIS-C5A



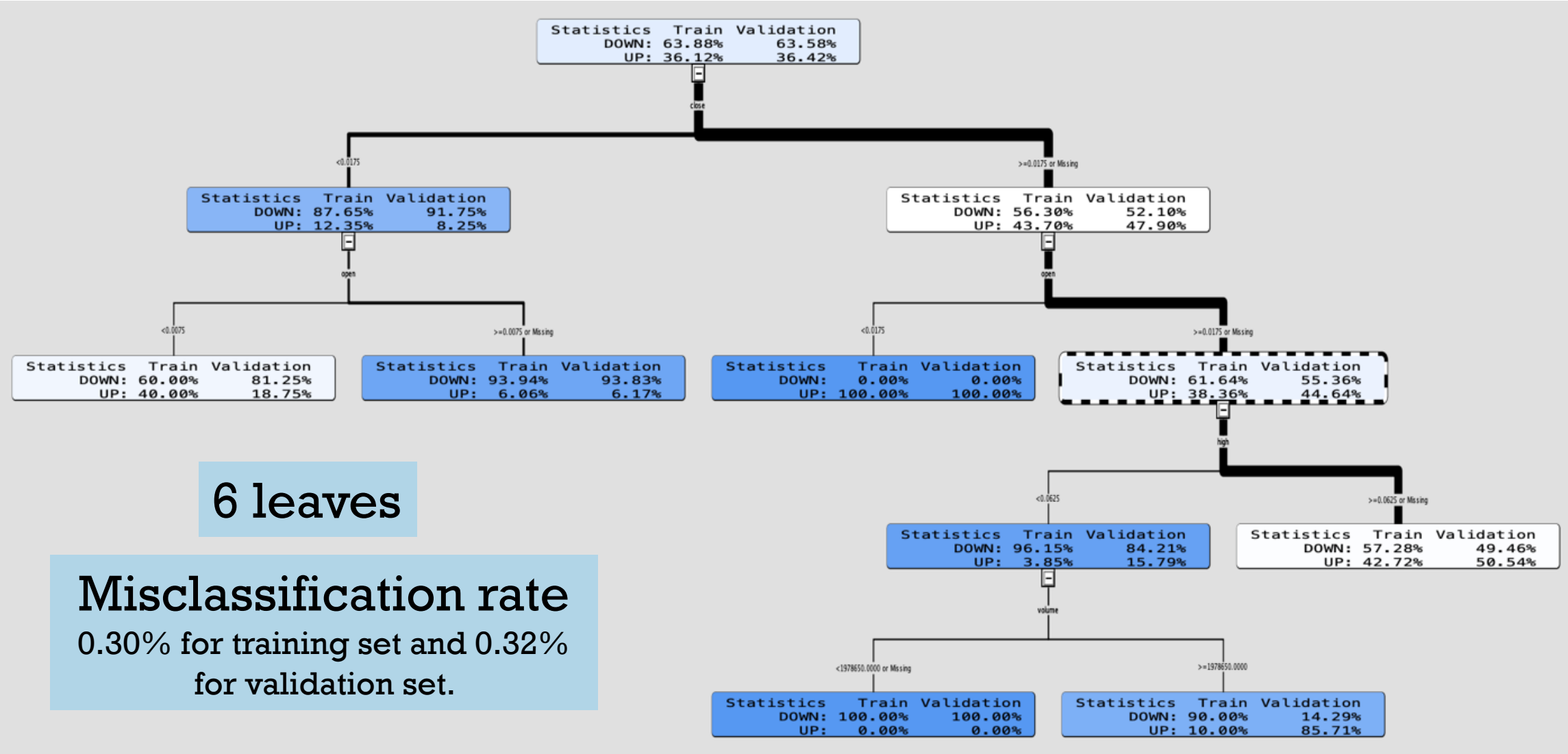


Daya vs Orion WA





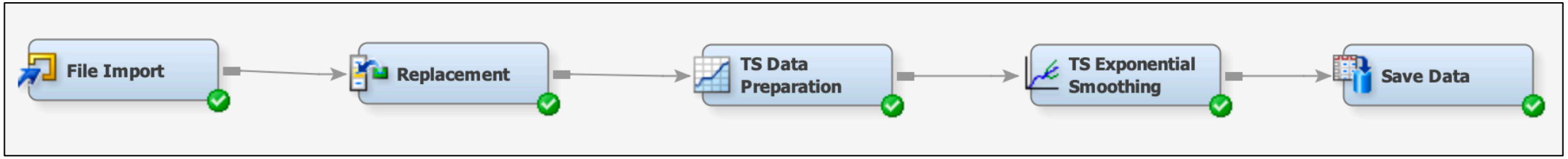
SAX & PAA

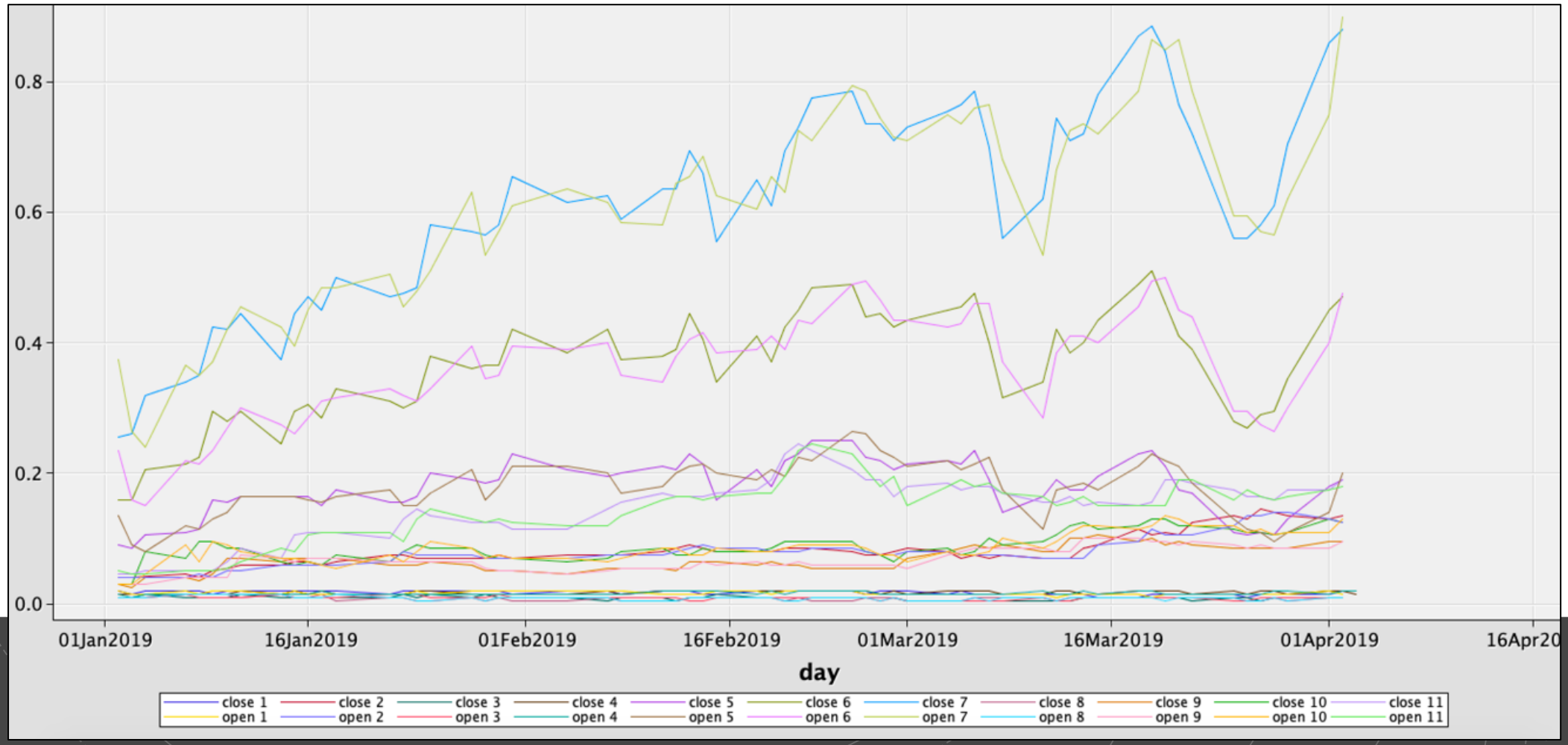


6 leaves

Misclassification rate

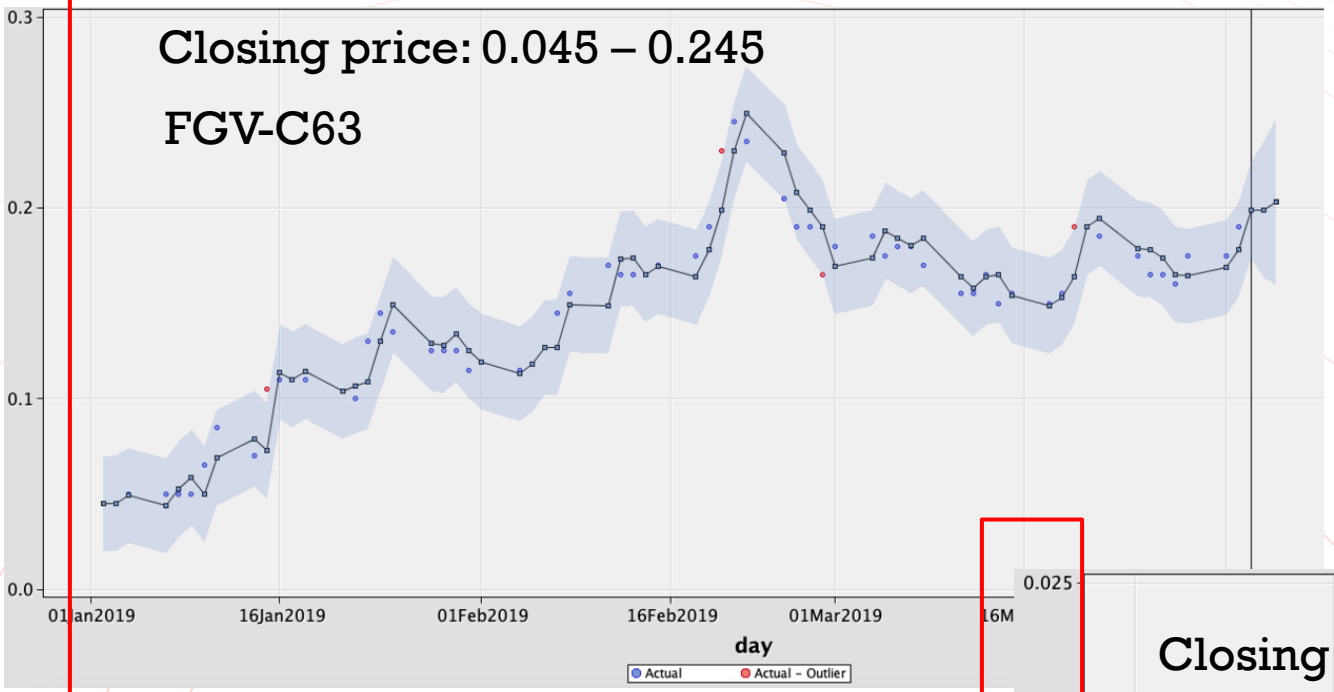
0.30% for training set and 0.32% for validation set.


File Import
Replacement
TS Data Preparation
TS Exponential Smoothing
Save Data

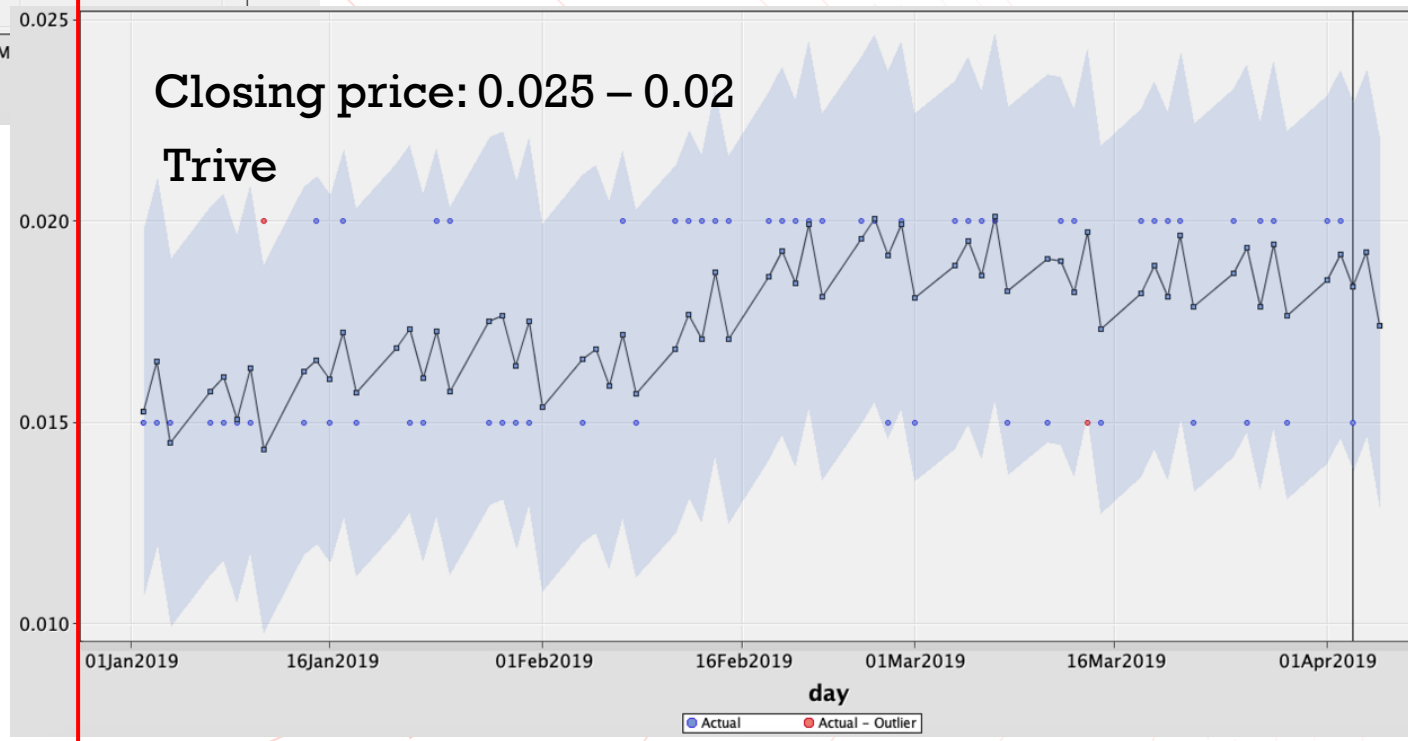

0.8
0.6
0.4
0.2
0.0
01Jan2019
16Jan2019
01Feb2019
16Feb2019
01Mar2019
16Mar2019
01Apr2019
16Apr20
day
close 1
open 1
close 2
open 2
close 3
open 3
close 4
open 4
close 5
open 5
close 6
open 6
close 7
open 7
close 8
open 8
close 9
open 9
close 10
open 10
close 11
open 11

Company Name	RMSE	R-Squared
NETX	0.0024	0.3385
ORION-WA	0.0063	0.9366
DAYA	0.0024	0.2178
TRIVE	0.0023	0.1650
HSI-C3Z	0.0236	0.6805
HSI-C5A	0.0423	0.7422
HSI-C5B	0.0590	0.8549
SUMATEC	0.0026	0.2563
ZELAN	0.0068	0.8714
DBHD-WA	0.0109	0.7599
FGV-C63	0.0124	0.9330

Model Assessment



Date	Company	Predicted	Actual	% Difference
03-Apr	FGV-C63	0.199	0.190	5%
04-Apr	FGV-C63	0.199	0.180	10%
05-Apr	FGV-C63	0.203	0.180	11%
03-Apr	Trive	0.015	0.015	0%
04-Apr	Trive	0.019	0.015	22%
05-Apr	Trive	0.017	0.020	-15%



Conclusions

- It can be observed that the model performs differently depending on the value of the dependent variable.
- The performance of the machine learning models can be further improved with more data.