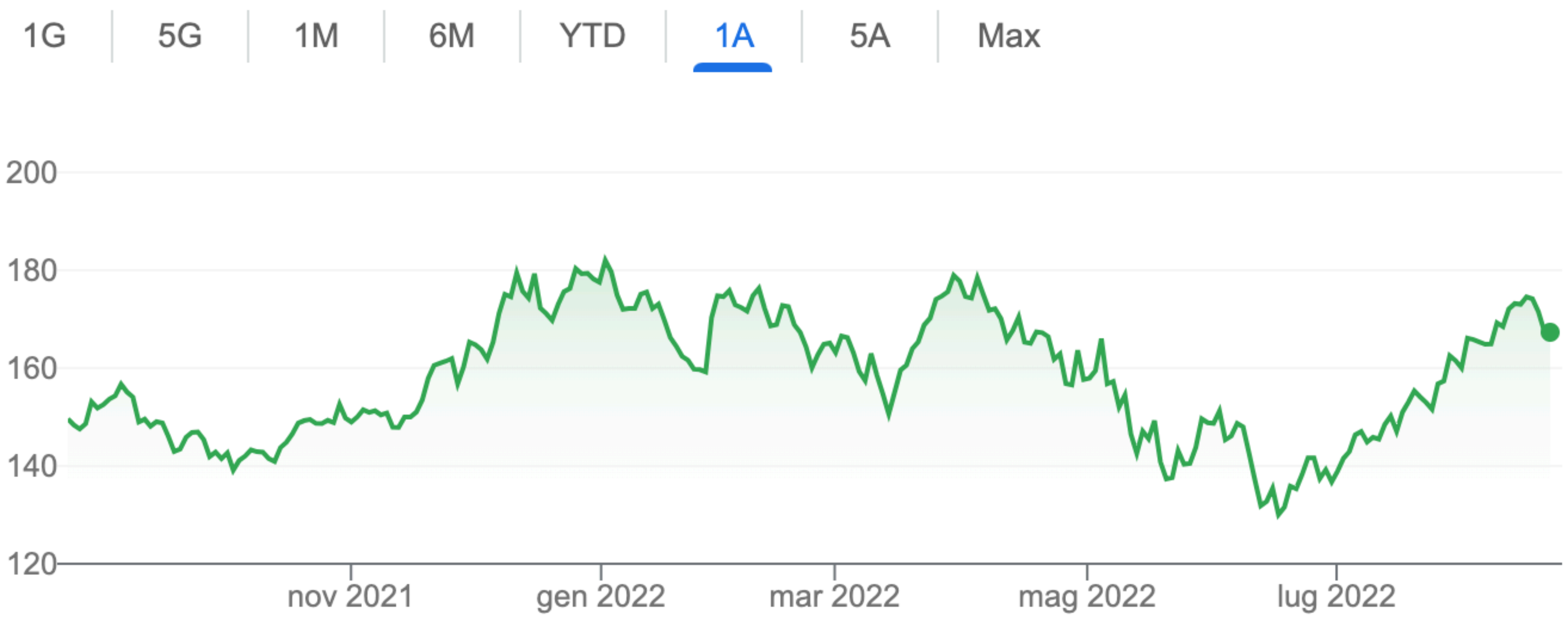
Stock market prediction

Big Data project

Benigno Ansanelli

Introduction



Building the dataset

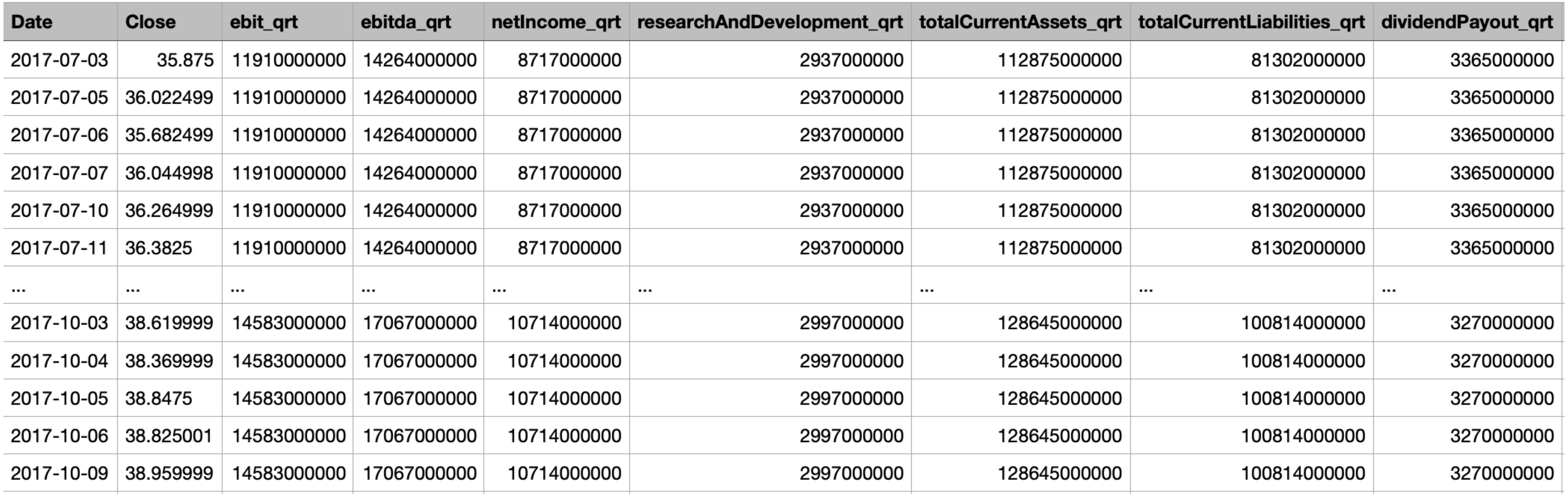
Simple dataset

- 5646 rows for 783 stocks (20 years of data)
- Sampled every day
- Open, Close, High, Low, Volume

Building the dataset

Dataset with financial indicators

- 1012 rows (or more), 1218 stocks, 5 years of data



Date	Close	ebit_qrt	ebitda_qrt	netIncome_qrt	researchAndDevelopment_qrt	totalCurrentAssets_qrt	totalCurrentLiabilities_qrt	dividendPayout_qrt
2017-07-03	35.875	11910000000	14264000000	8717000000	2937000000	112875000000	81302000000	3365000000
2017-07-05	36.022499	11910000000	14264000000	8717000000	2937000000	112875000000	81302000000	3365000000
2017-07-06	35.682499	11910000000	14264000000	8717000000	2937000000	112875000000	81302000000	3365000000
2017-07-07	36.044998	11910000000	14264000000	8717000000	2937000000	112875000000	81302000000	3365000000
2017-07-10	36.264999	11910000000	14264000000	8717000000	2937000000	112875000000	81302000000	3365000000
2017-07-11	36.3825	11910000000	14264000000	8717000000	2937000000	112875000000	81302000000	3365000000
...	...	...	...	...	...	...	...	...
2017-10-03	38.619999	14583000000	17067000000	10714000000	2997000000	128645000000	100814000000	3270000000
2017-10-04	38.369999	14583000000	17067000000	10714000000	2997000000	128645000000	100814000000	3270000000
2017-10-05	38.8475	14583000000	17067000000	10714000000	2997000000	128645000000	100814000000	3270000000
2017-10-06	38.825001	14583000000	17067000000	10714000000	2997000000	128645000000	100814000000	3270000000
2017-10-09	38.959999	14583000000	17067000000	10714000000	2997000000	128645000000	100814000000	3270000000

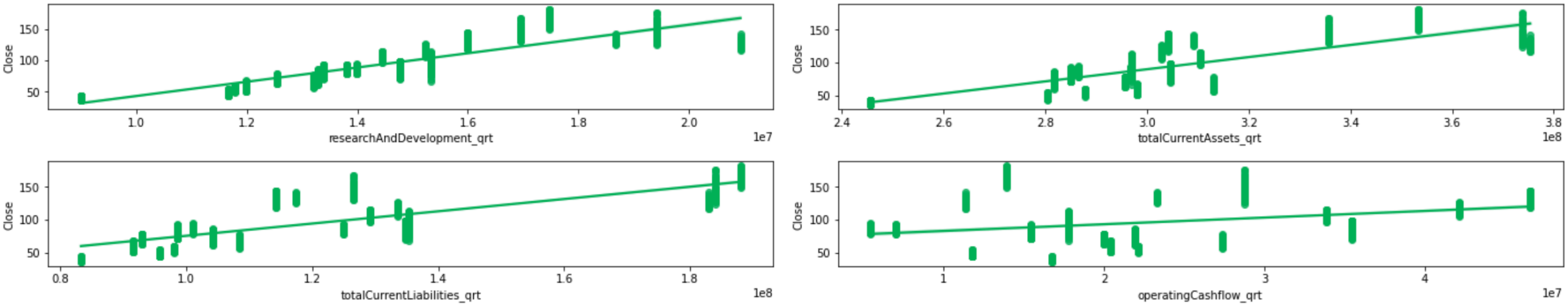
Building the dataset

Intraday Dataset

- 360 rows (or more) for 1359 stocks
- Only 2 months for each stock
- Used for prediction of the next 30 minutes

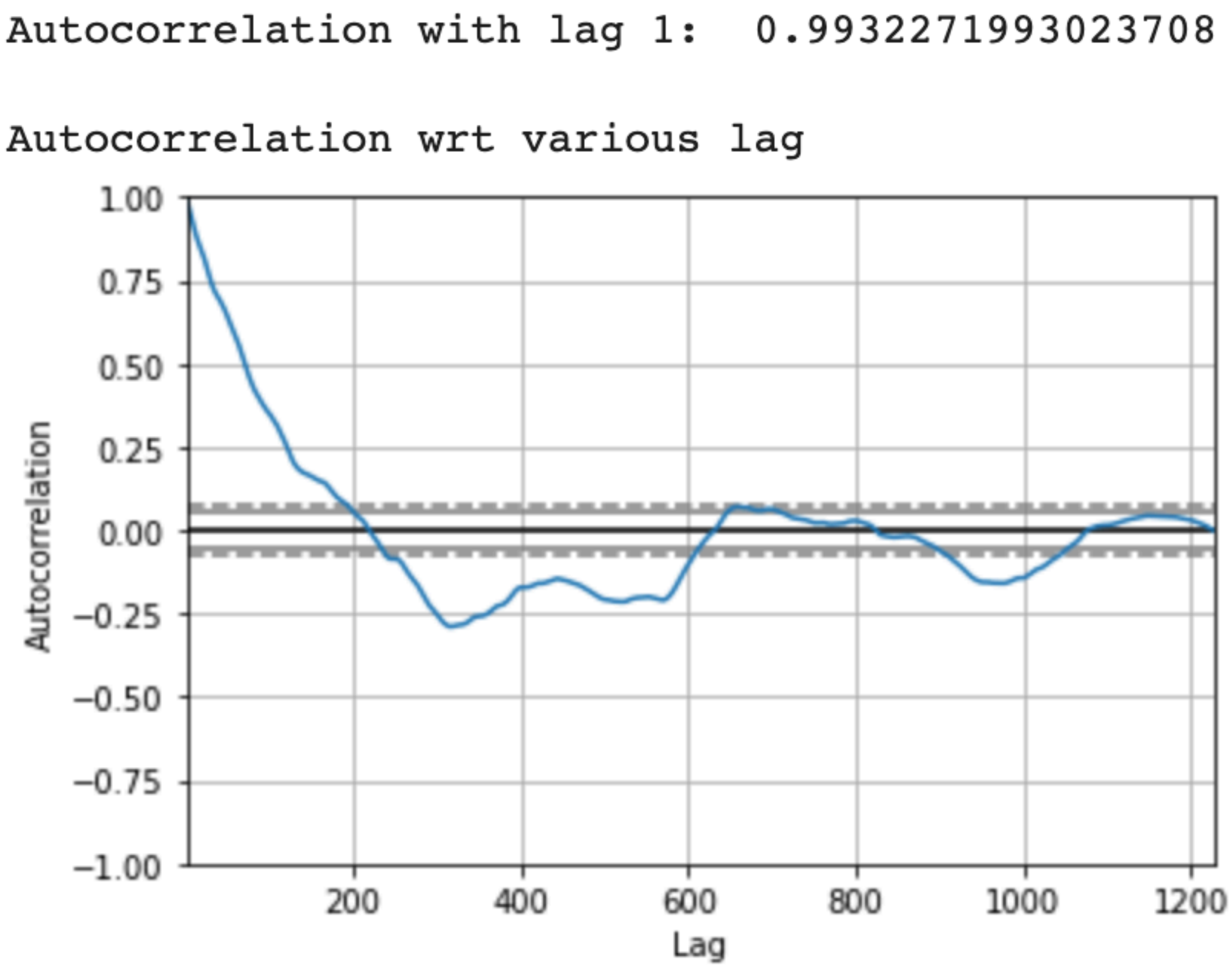
Features engineering

Correlation between financial indicators and target



Features engineering

Autocorrelation



```
Autocorrelation with lag 1:  0.9932271993023708

Autocorrelation wrt various lag
```

Features engineering

Building the features

- Lagged features
- Averaged features
- Differenced features

Models

- Linear Regressor

- Random Forest Regressor

- Gradient Boosted Tree Regressor

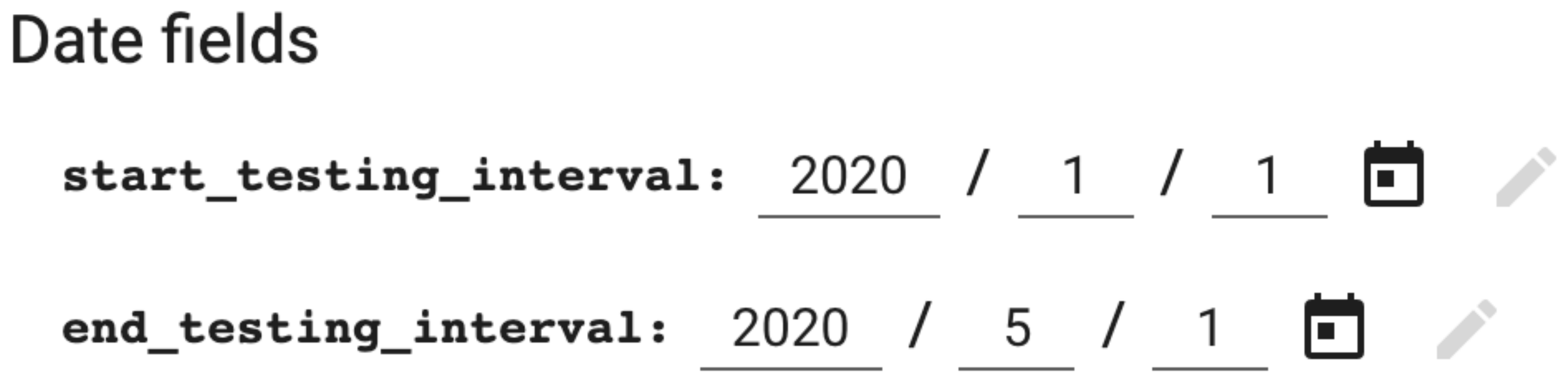
- Neural Network

Testing

- Period of testing: COVID crisis (1/01/2020 - 1/05/2020)

- Metrics: RMSE, R2, score
- Score = RSME / mean of the stock price

- Intraday vs interday models



Date fields

start_testing_interval: 2020 / 1 / 1

end_testing_interval: 2020 / 5 / 1

Results

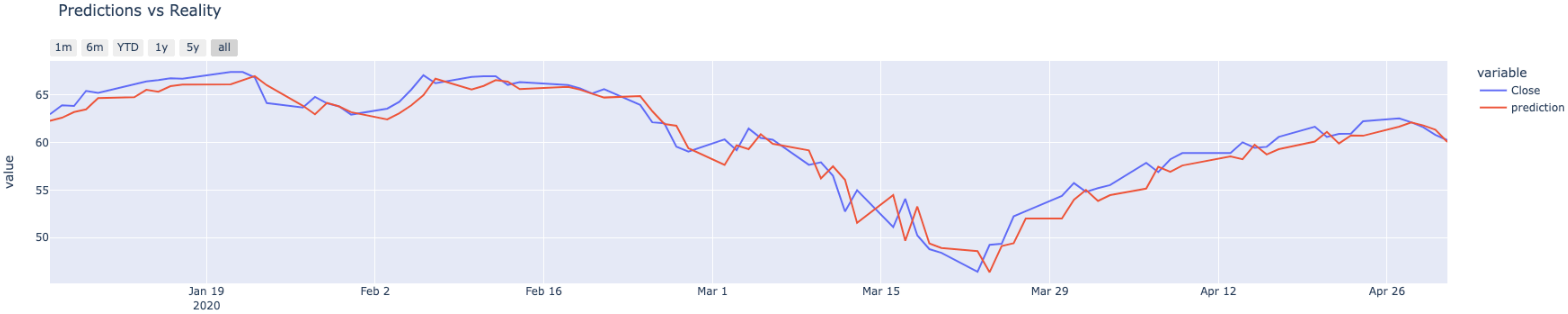
Random forest regression

Predictions vs Reality
1m
6m
YTD
1y
5y
all
150
140
130
120
value
Jan 19
2020
Feb 2
Feb 16
Mar 1
Mar 15
Mar 29
Apr 12
Apr 26
variable
Close
prediction

Stock: Kimberly-Clark
Score: 0.0179
Validation - RMSE:1.96 R2:0.83
Testing - RMSE:2.47 R2 0.89

Results

Linear regression



Stock: Bristol-Myers Squibb
Score: 0.024
Validation - RMSE: 0.94 R2: 0.96
Testing - RMSE: 1.49 R2: 0.92

Results

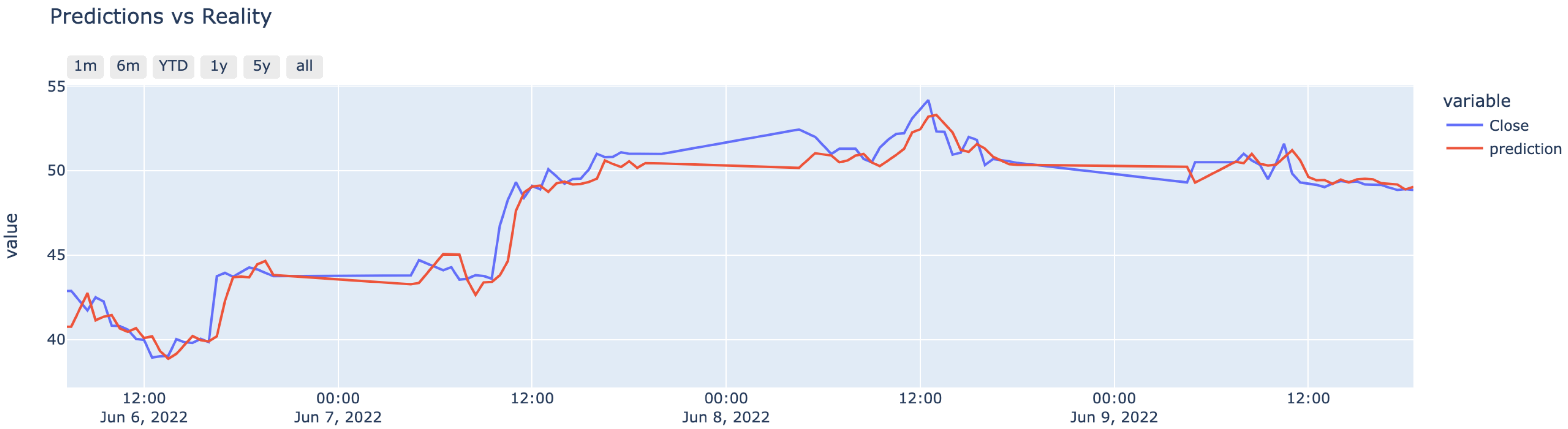
Neural network

Predictions vs Reality
1m
6m
YTD
1y
5y
all
value
34
32
30
28
26
Jan 19
2020
Feb 2
Feb 16
Mar 1
Mar 15
Mar 29
Apr 12
Apr 26
variable
Close
prediction

Stock: Silgan Holdings
Score: 0.026
Validation - RMSE: 0.33 R2: 0.97
Testing - RMSE: 0.80 R2: 0.86

Results

Linear regression with intraday data



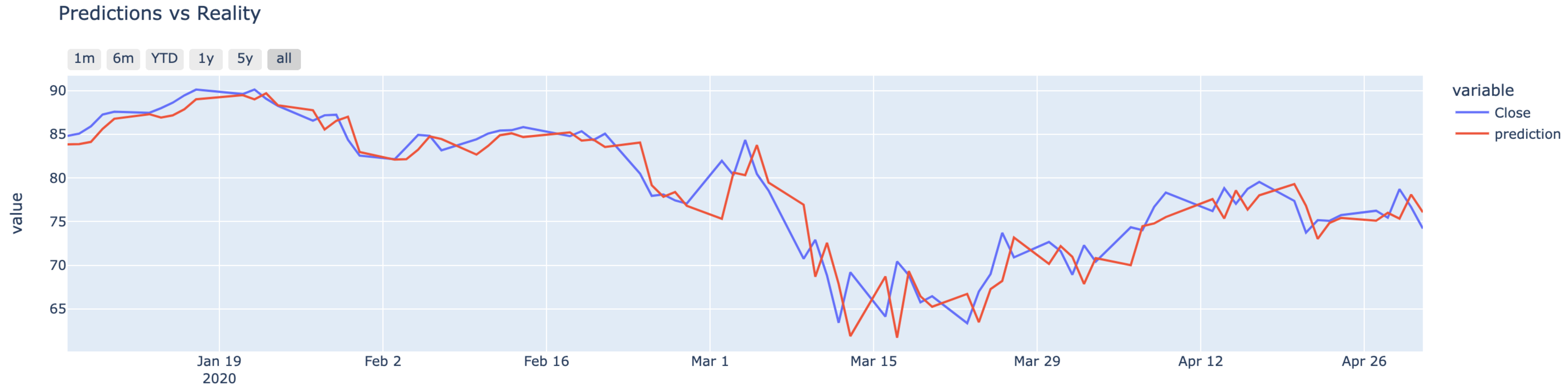
Score: 0.019

Validation - RMSE: 0.96 R2: 0.95

Testing - RMSE: 0.89 R2: 0.96

Conclusion

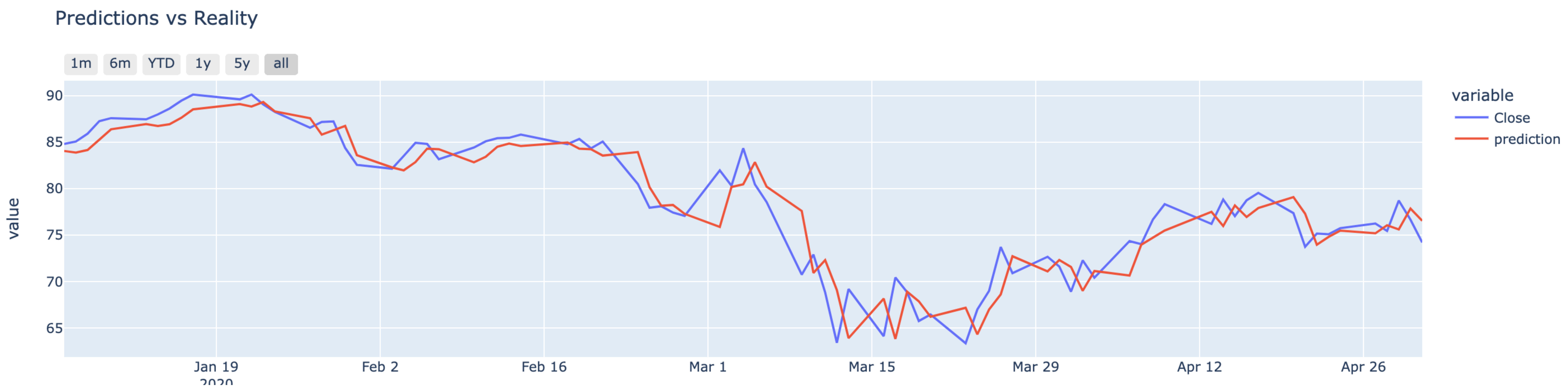
Use of financial indicators



Validation - RMSE: 0.94 R2: 0.99
Testing - RMSE: 2.61 R2: 0.86

Conclusion

Use of financial indicators



Validation - RMSE: 1.21 R2: 0.93
Testing - RMSE: 2.40 R2: 0.88

Conclusion

- Train the model really close to the prediction
- Difficult to generalize
- Autocorrelation and next day prediction
- Hyperparameters tuning

Future improvements?

- Using LSMT and RNN
- Intraday predictions
- Live prediction
- One model for multiple stocks