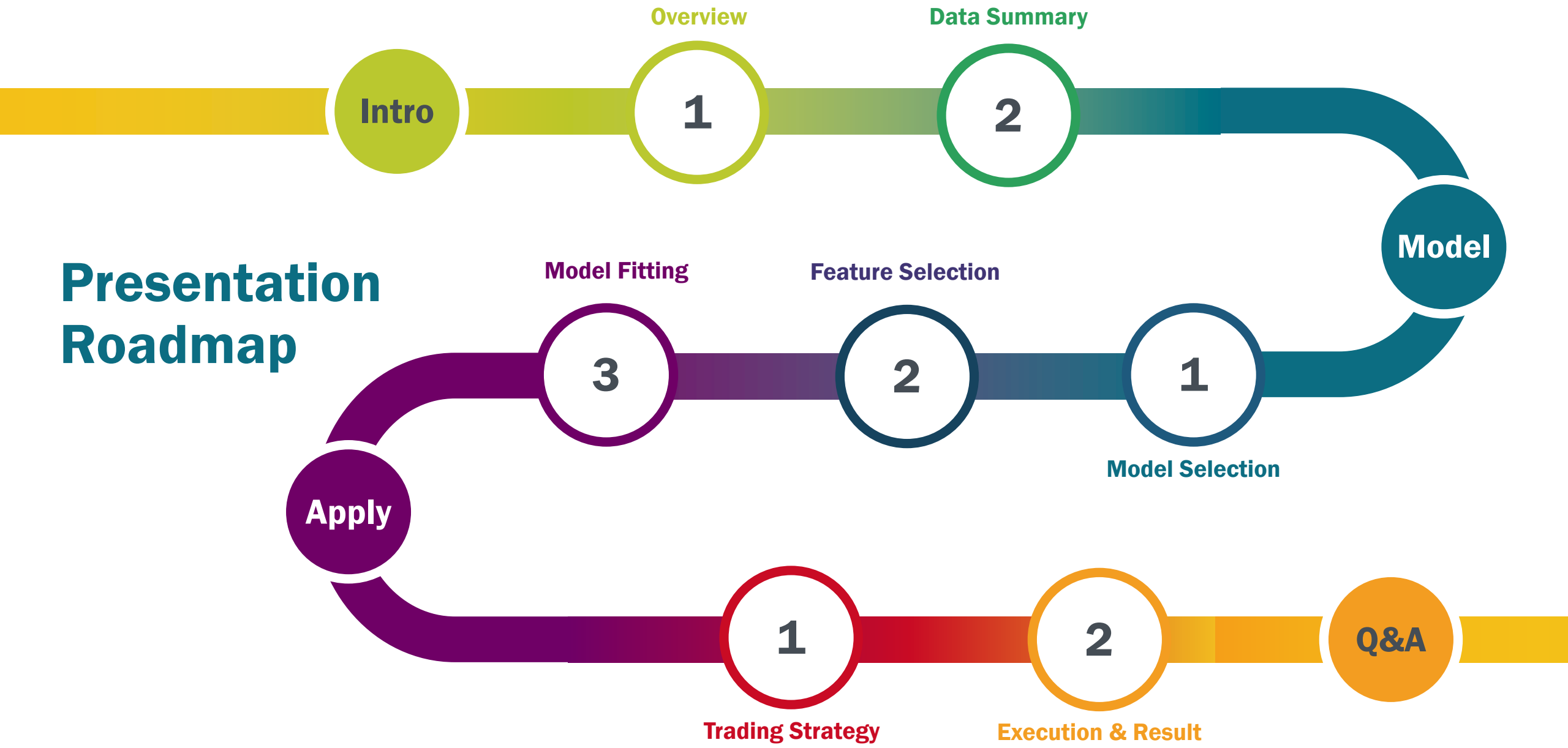
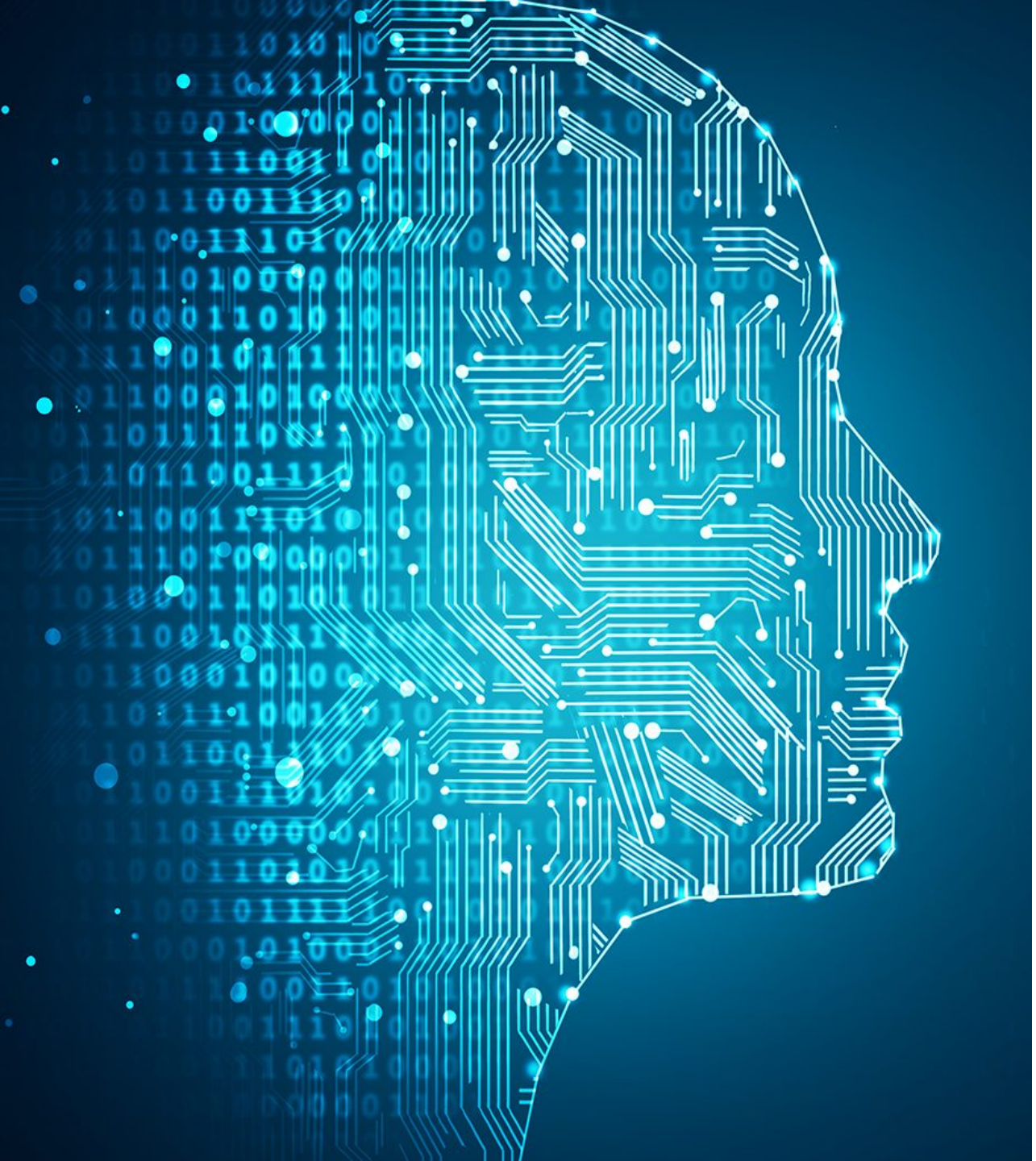
Stock Price Prediction based on Machine Learning Technique and Two Applications

Group 1
Junru Liu, Yachun Zhang, Biyao Wang, Yiduo Zhang, Chen Huang, Zelai Yu

Presentation Roadmap



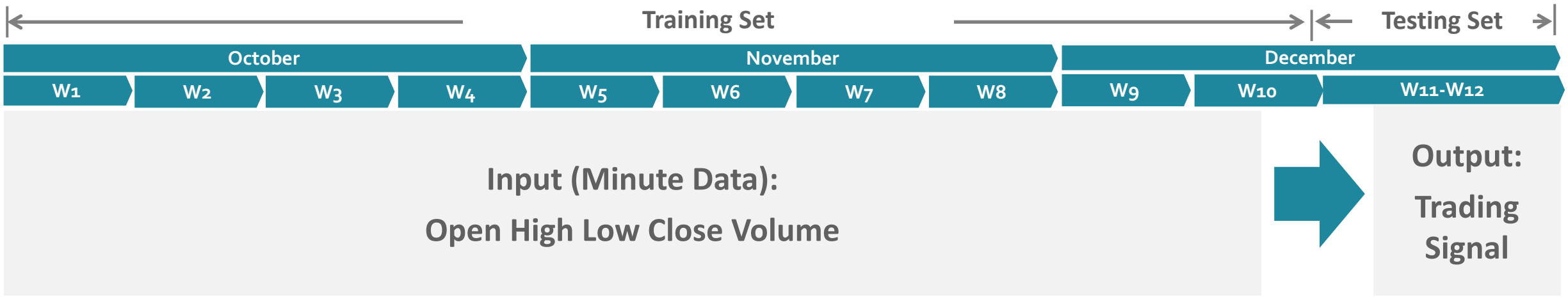


Aim to build a machine learning model to **predict** direction of price movement, and then use the results to **construct** trading strategy and **improve** execution of large trade order

Machine Learning

to develop a deeper understanding of **market microstructure**

Data Summary



Target: 10 US large-cap stocks (source: Minute data for S&P 500 composite stocks)

Selecting Criteria: High Volume, High Liquidity

Processing

Calculate the average of open, high, low and close prices

Using average price to determine price movement direction

Prediction Model

Technical Indicators

Indicators	Formula	Corresponding Categorical Criteria
MA **Moving Average**	$MA(n)_t = \frac{\sum_{i=t-n}^{t} P_i}{n}$	When MA > Close, set it to 1, otherwise -1.
VMA **Volume-Moving Average**	$VMA(n)_t = \frac{\sum_{i=t-n}^{t} P_i * Volume_i}{n}$	When VMA > 0, set it to 1, otherwise -1.
MACD	$MA(Win_1)_t - MA(Win_2)_t$	When MACD > 0, set it to 1, otherwise -1.
Williams'%R	$\frac{H_n - C_t}{H_n - L_n} \times 100\%$	NA
RSI	$RS(N)_t = \frac{\frac{1}{n}\sum_{i=0}^{n-1} UP_{t-i}}{\frac{1}{n}\sum_{i=0}^{n-1} DW_{t-i}}, RSI(N)_t = 100 - \frac{100}{1 + Rs(N)_t}$	NA
PSY **Psychological Line**	Ratio of the number of Rising Periods over the Total number of Periods	NA
ADO	$\frac{H_t - C_{t-1}}{H_t - L_t}$	NA
KDJ	$K_t(n) = \frac{C_t - LL_t(n)}{HH_t(n) - LL_t(n)}, D_t = \frac{1}{n}\sum_{i=1}^{n} K_{t+1-i}.$ $, where\ LL_t(n)\ is\ the\ lowest\ low\ price\ in\ n\ periods\ and$ $HH_t(n)\ is\ the\ highest\ high\ price\ in\ n\ periods.$	Criteria 1: when Kt > Dt set it to 1, otherwise -1. Criteria 2: when Kt > 80 set it to 1, when Kt < 20 set it to -1, otherwise 0.
CCI	$CCI = \frac{M_t - SM_t}{0.015 D_t}, where M_t = \frac{H_t + L_t + C_t}{3}, SM_t = \frac{\sum_{i=1}^{n} M_{t-i+1}}{n}, and\ D_t = \frac{\sum_{i=1}^{n} \lvert M_{t-i+1} - SM_t \rvert}{n}$	NA

Parameters in Indicators

Indicator	Parameters	Parameters Used	Notation
MA	N	3, 5, 10, 15, 20, 30	MA_N, MA_CAT_N
VMA	N	3, 5, 10, 15, 20, 30	VMA_N, VMA_CAT_N
MACD	(N1, N2)	(3, 5), (5, 10), (5, 15), (5, 30), (10, 15), (10, 20), (10, 30), (15,30)	MACD_N1_N2, MACD_CAT_N1_N2
RSI	N	10, 20, 30	RSI_N
PSY	N	10, 15, 30	PSY_N
KDJ	(N, M)	(5, 3), (15, 3)	K_N, D_N_M, K_D_CAT_N_M, K_D_CAT_B_N_80_20
CCI	N	5	CCI_N

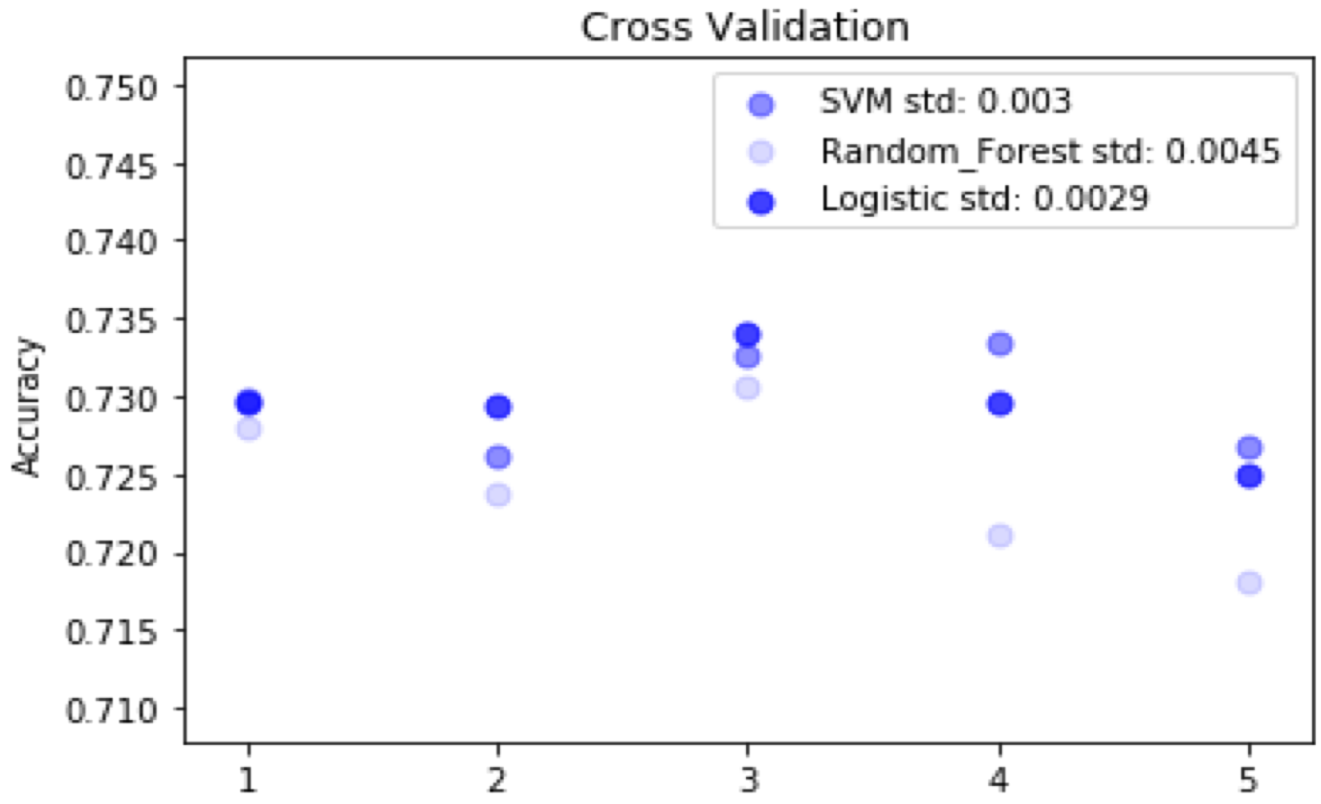
Cross Validation

5-Fold Cross Validation:

- Applied to training set

Candidate Models:

- **SVM:** linear model
- **Random Forest:** 100 trees, depth limit 5
- **Logistic Regression**



Model Selected: Logistic Regression

Consistency

Robust results from cross-validation, with higher and more consistent out-of-sample prediction accuracy

Significance

Generating probability prediction, which can be used as a proxy for signal significance

Stepwise

Stepwise process is applicable for feature selection

Cost

Lower computational costs

Feature Selection(1/3)

Full model:

- **Original Indicators** for William%R, AD Oscillator, RSI, PSY and CCI, and
- **Categorical indicators** for MA, VMA, MACD and KDJ, with selected parameters

Selection process:

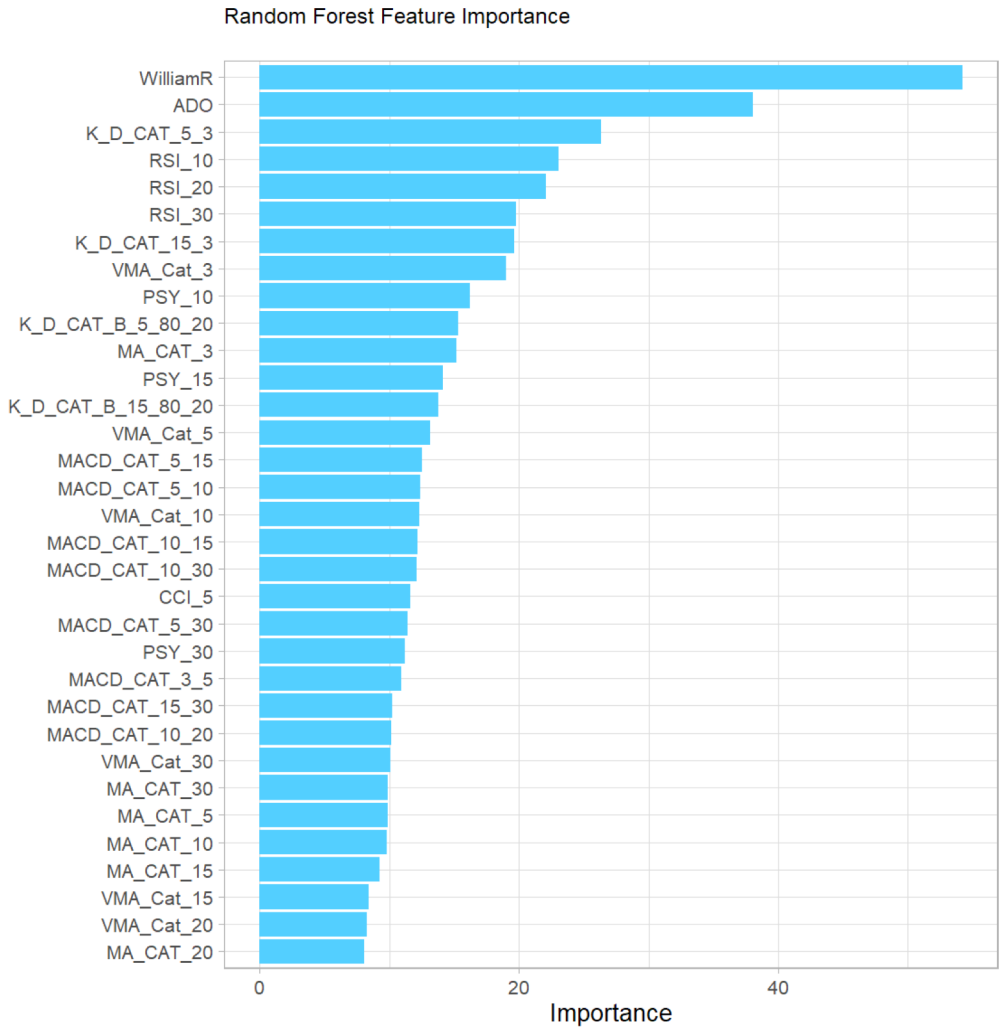
- Applying stepwise process with training set data, using AIC as selection criteria

Example: AAPL

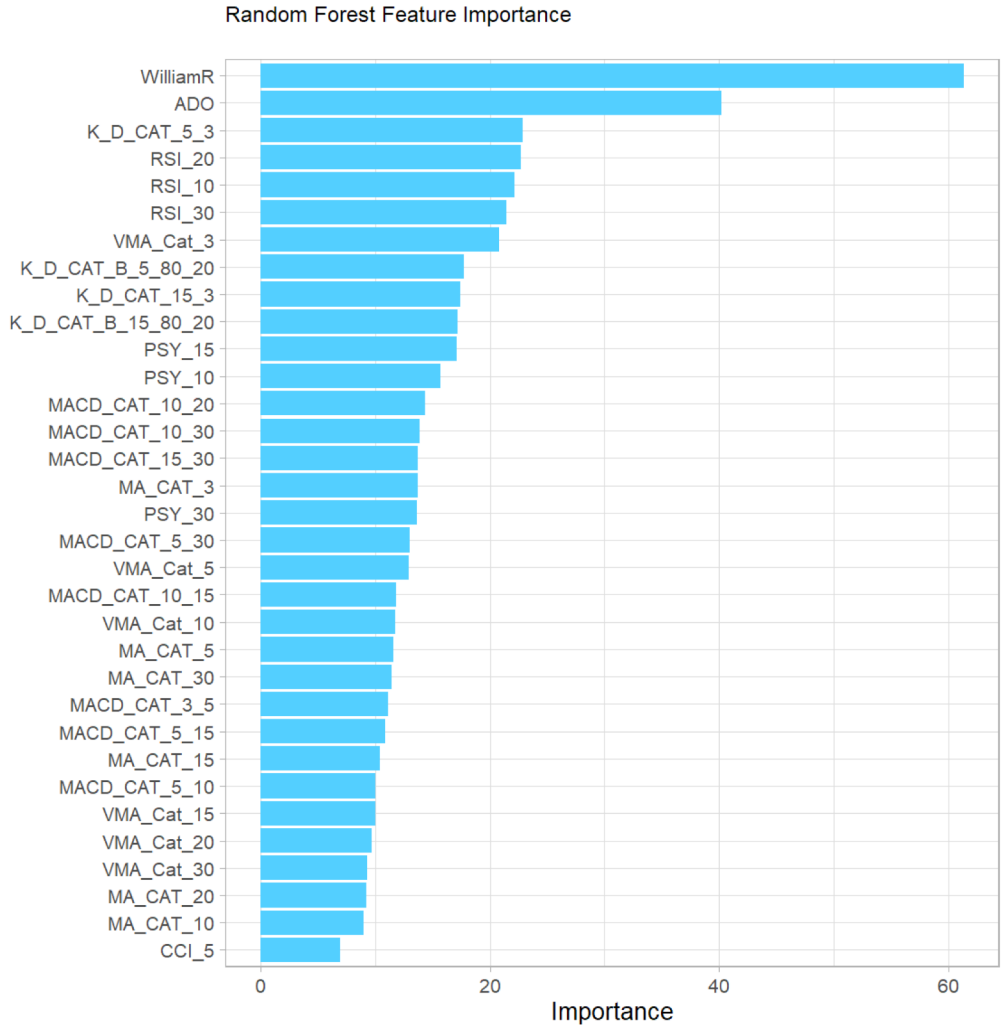
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.102723   0.155715   7.082 1.42e-12 ***
MA_CAT_20         -0.076194   0.052242  -1.458  0.14471
VMA_Cat_3          0.058869   0.028845   2.041  0.04127 *
VMA_Cat_10         0.060165   0.027253   2.208  0.02727 *
VMA_Cat_20         0.157657   0.049878   3.161  0.00157 **
VMA_Cat_30        -0.065844   0.032637  -2.017  0.04365 *
MACD_CAT_3_5      -0.066424   0.022094  -3.006  0.00264 **
MACD_CAT_5_15     -0.077561   0.029522  -2.627  0.00861 **
MACD_CAT_10_15    -0.051229   0.025338  -2.022  0.04319 *
MACD_CAT_15_30    -0.039179   0.025480  -1.538  0.12414
WilliamR          -3.351291   0.081335 -41.204  < 2e-16 ***
RSI_20             0.003463   0.002421   1.430  0.15262
PSY_15            -0.646963   0.219768  -2.944  0.00324 **
PSY_30             0.518289   0.283767   1.826  0.06778 .
ADO                0.807206   0.061338  13.160  < 2e-16 ***
K_D_CAT_15_3      -0.043988   0.029281  -1.502  0.13303
K_D_CAT_5_3        0.064230   0.025895   2.480  0.01312 *
K_D_CAT_B_15_80_20 0.182365   0.044491   4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28584  on 20621  degrees of freedom
Residual deviance: 22380  on 20604  degrees of freedom
AIC: 22416

Number of Fisher Scoring iterations: 4
```

Feature Selection(2/3) - Feature Importance

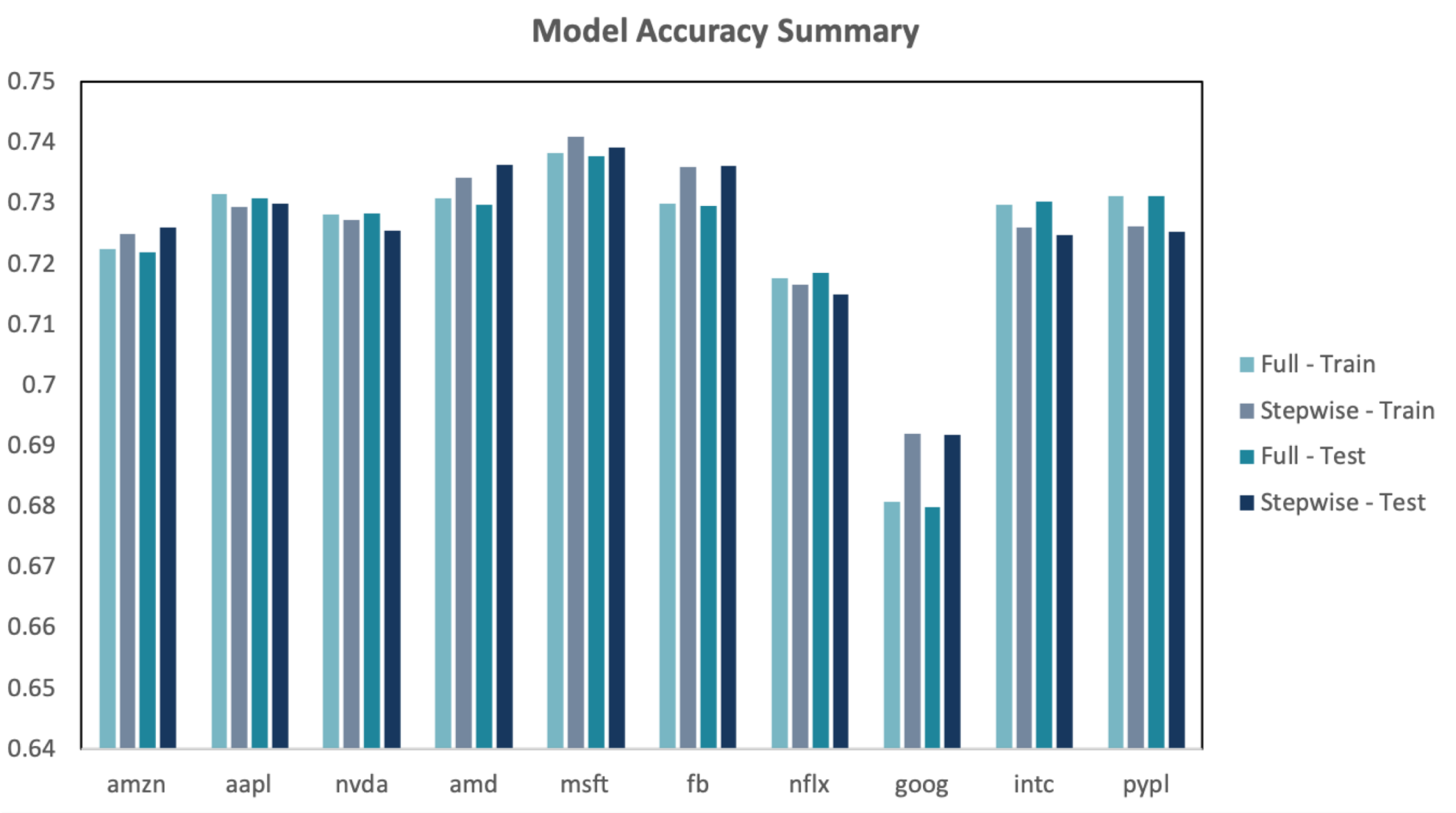
Example: AMZN



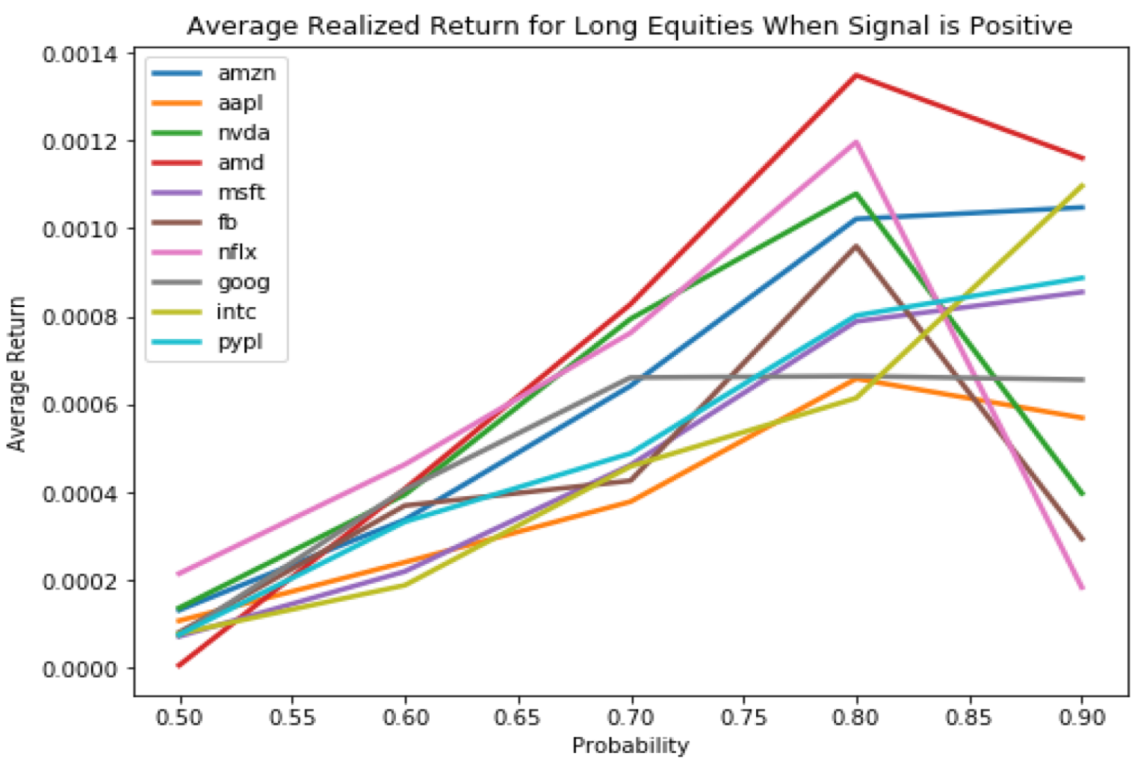
Example: AAPL

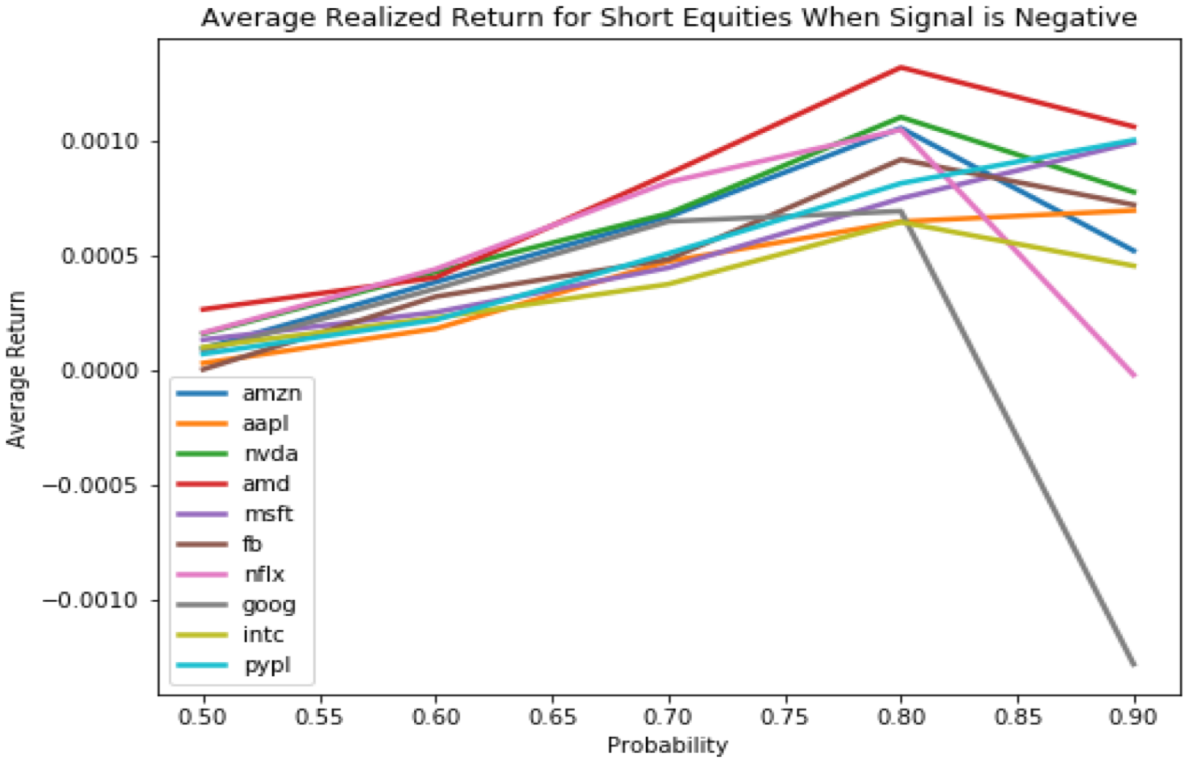


Feature Selection (3/3) - Result



Probability Prediction from Logistic Regression





Trading Strategy

Rationale

Basic Rule:

Open a long position when predicted direction change is 1, open a short position when it is -1

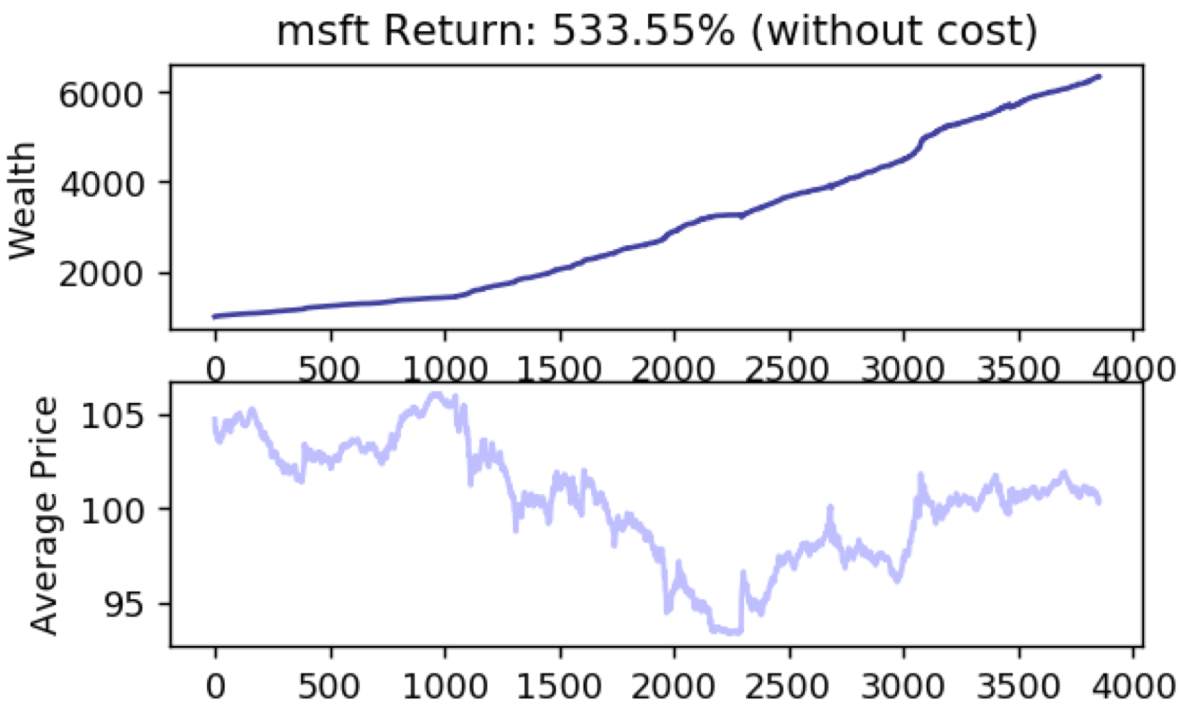
Assumptions:

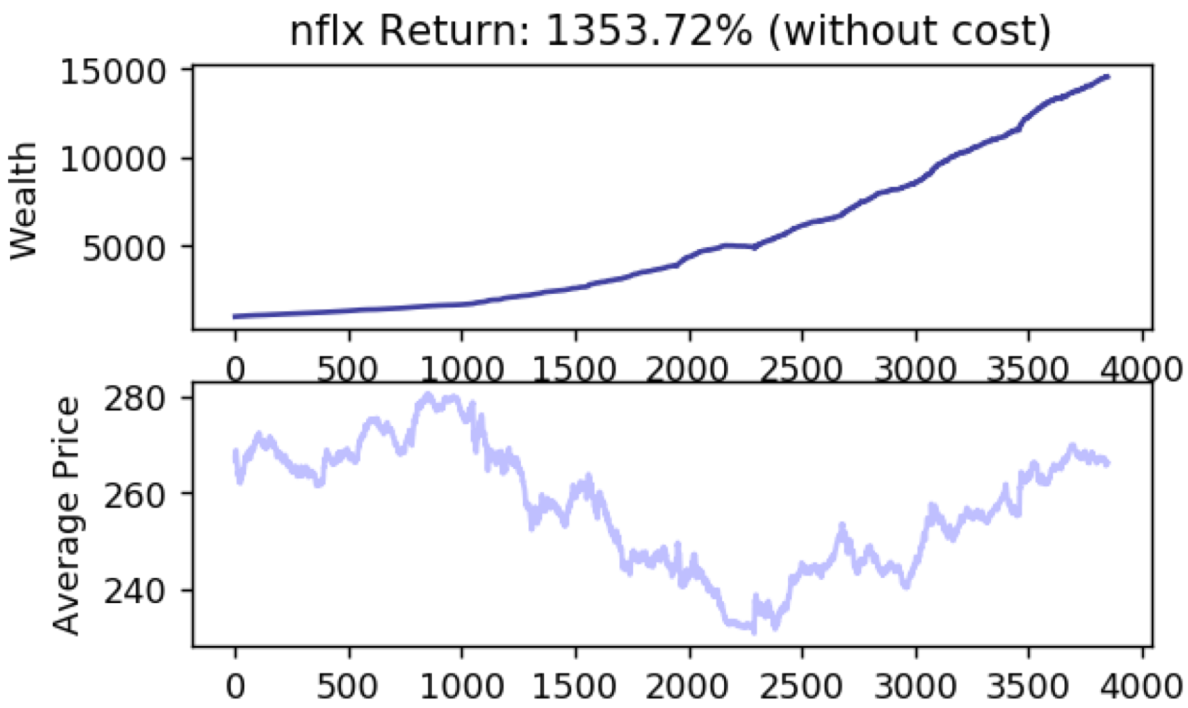
1. Invest all the capital when open a long position
2. 100% margin requirement for short position
3. Transaction cost is simplified as a percentage of nominal trading amount

Backtest Results (*Assuming 0.03% transaction cost)

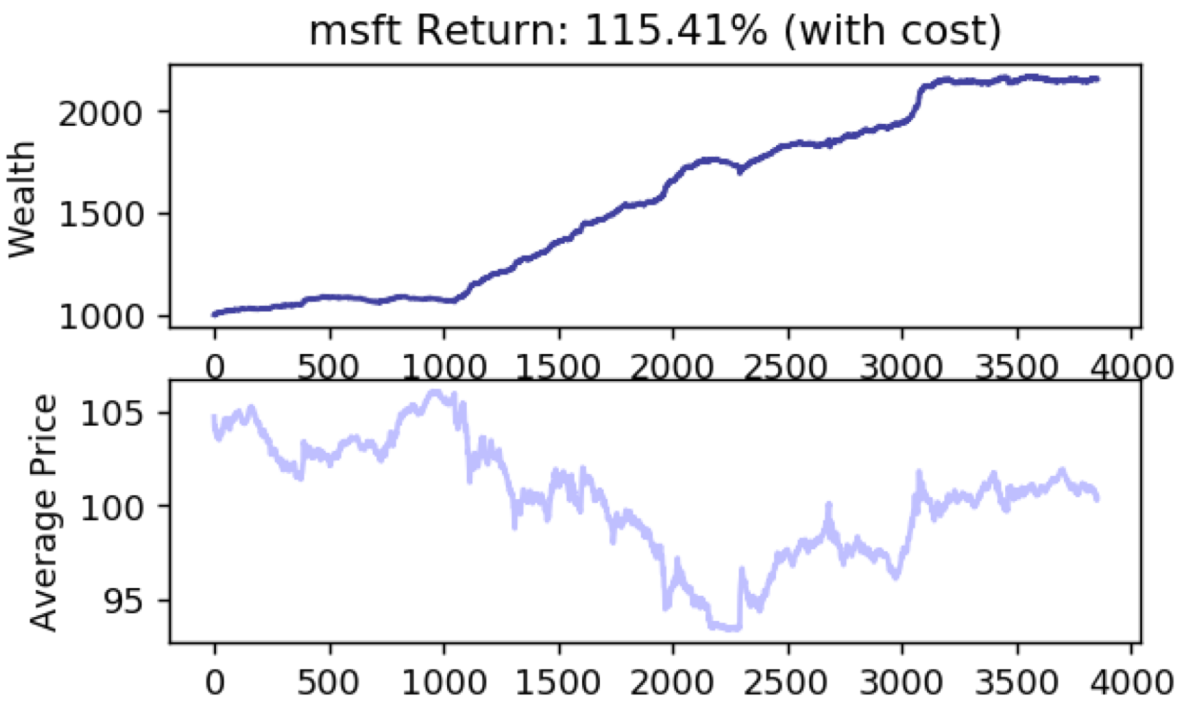
Stocks	Return (10 days)	Sharpe Ratio	Drawdown	Drawdown Period	Calmar Ratio
AAPL	64.07%	7.56	7.63%	155 min	8.39
AMZN	246.30%	12.00	21.70%	152 min	11.35
NVDA	338.51%	14.27	18.55%	113 min	18.25
AMD	636.44%	15.81	17.70%	117 min	35.96
MSFT	115.41%	10.16	7.03%	101 min	16.42
FB	165.18%	11.70	9.97%	119 min	16.57
NFLX	393.95%	15.11	19.07%	126 min	20.65
GOOG	70.40%	6.66	5.84%	329 min	12.05
INTC	46.30%	6.48	4.65%	269 min	9.96
PYPL	124.26%	10.50	4.46%	262 min	27.88

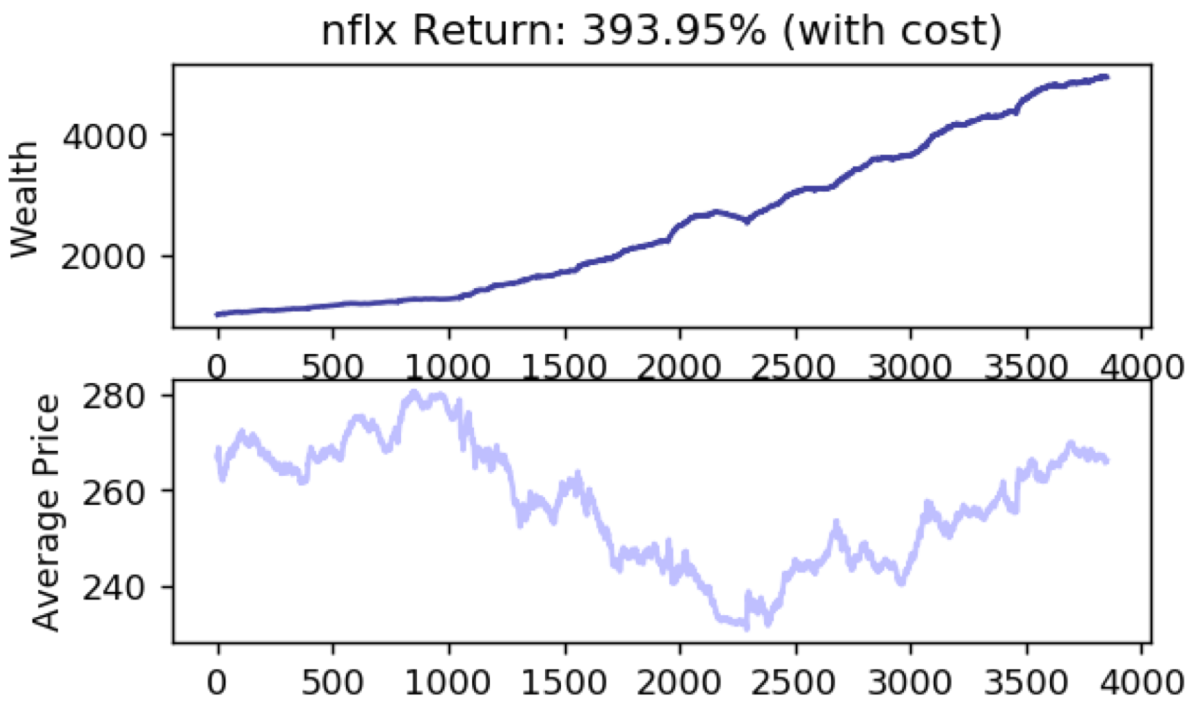
Example: No Transaction Cost



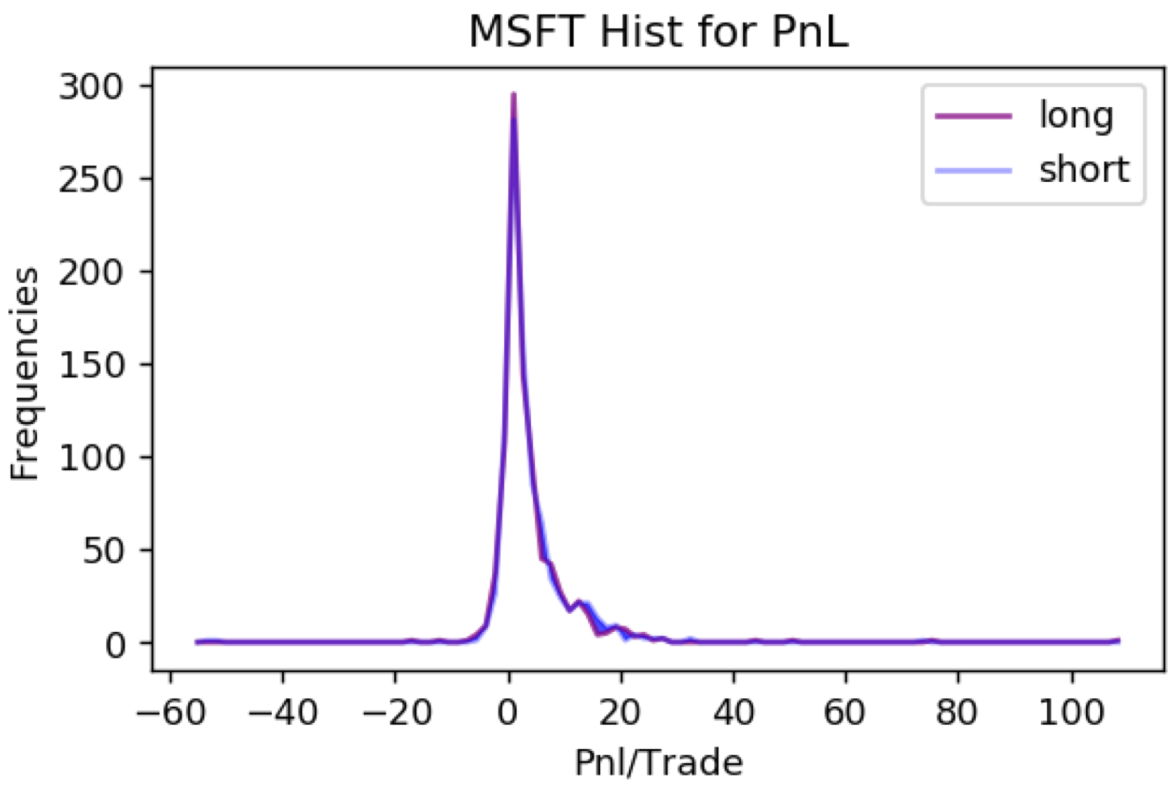


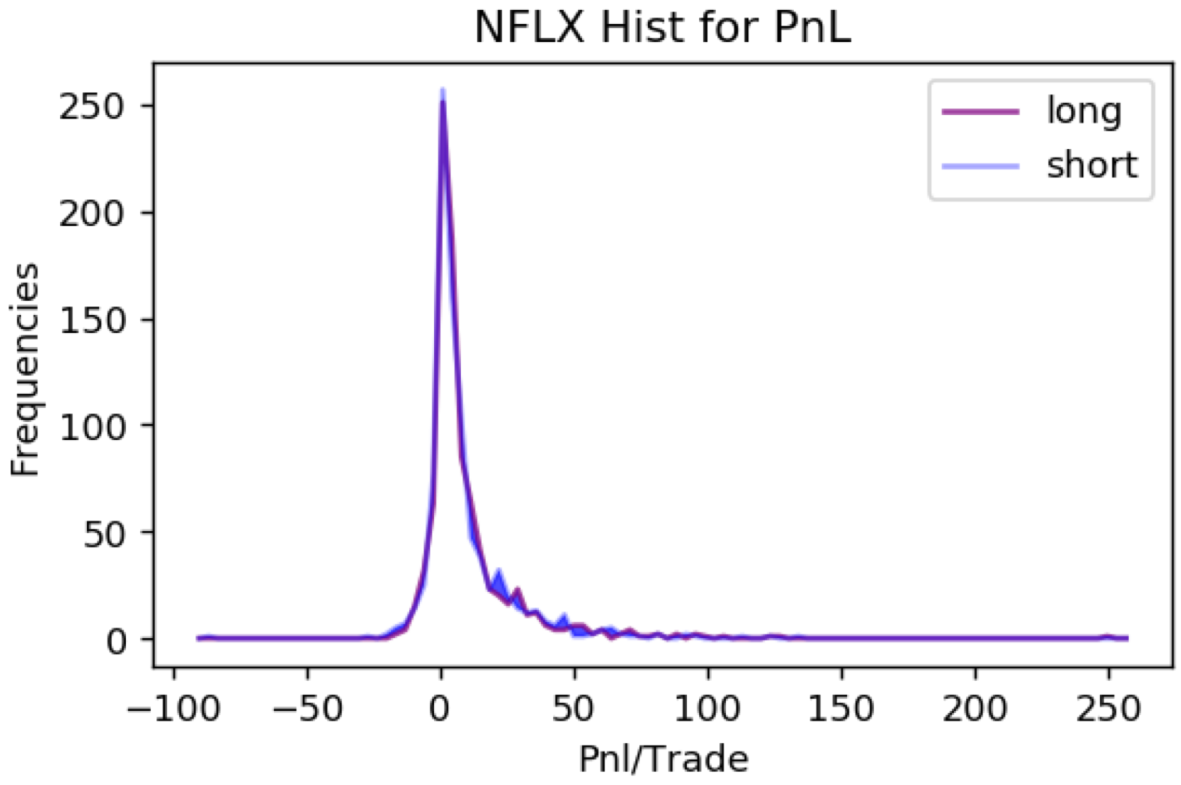
Example: 0.03% Transaction Cost



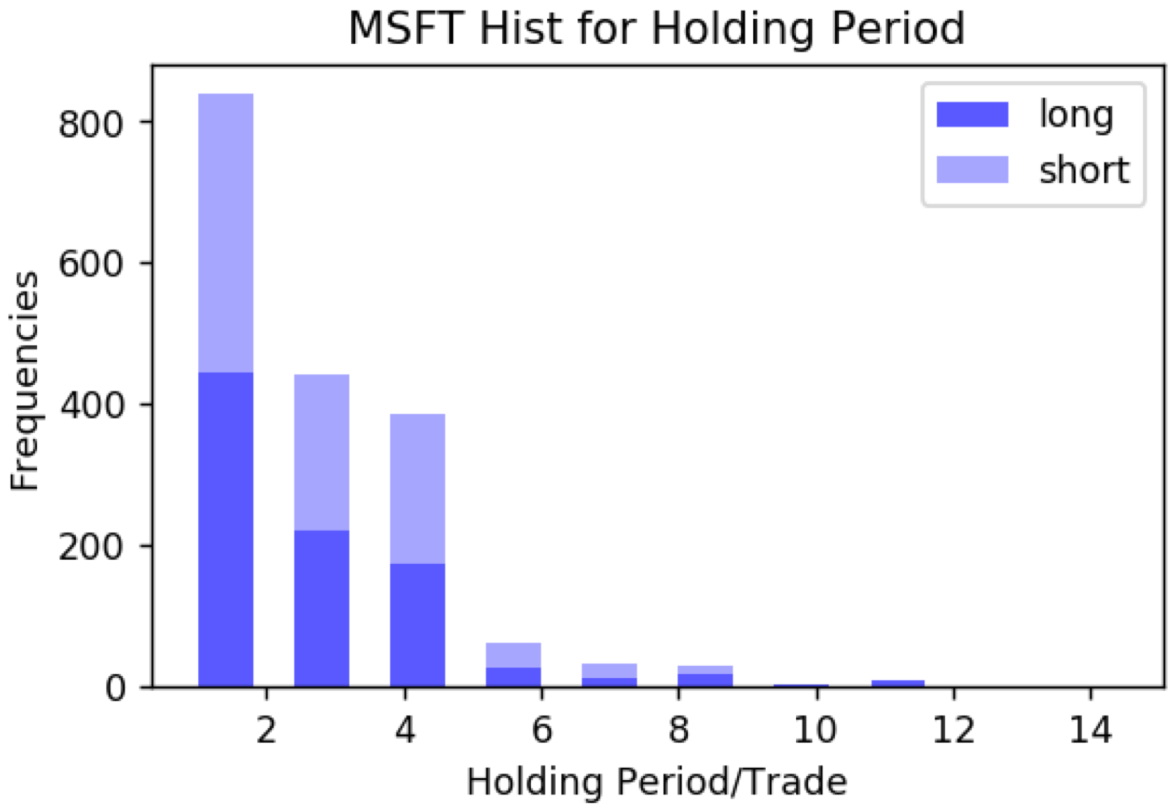


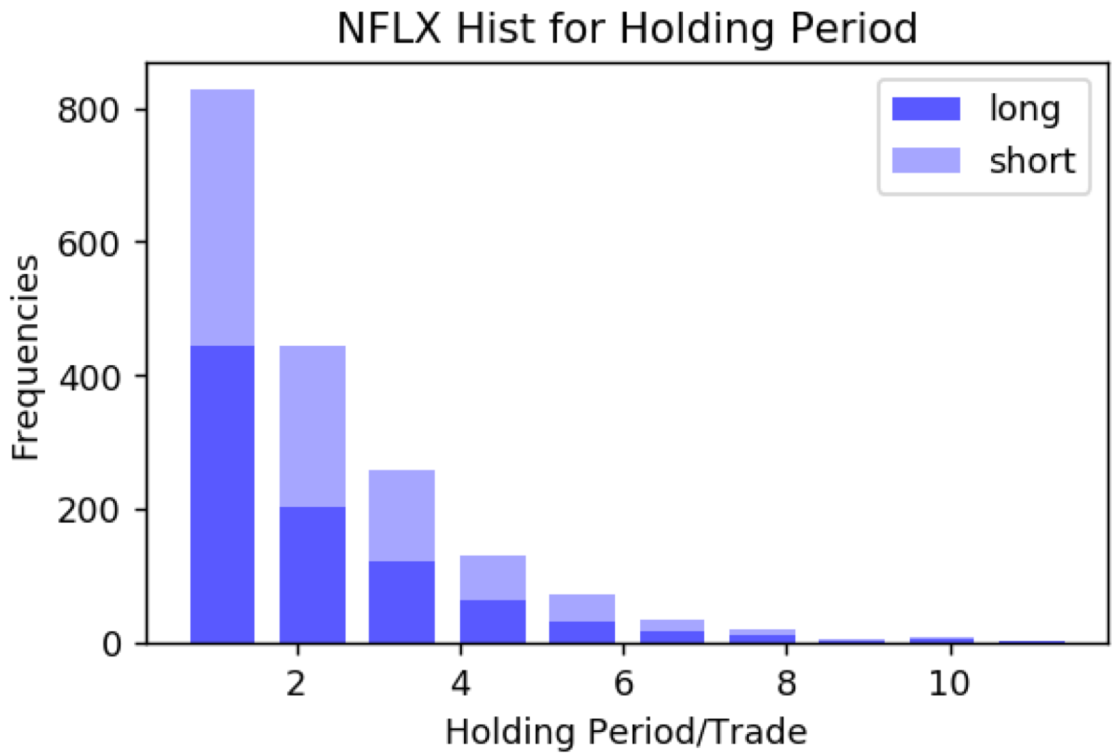
P&L per Trade


MSFT Hist for PnL
long
short
Frequencies
Pnl/Trade
300
250
200
150
100
50
0
−60
−40
−20
0
20
40
60
80
100

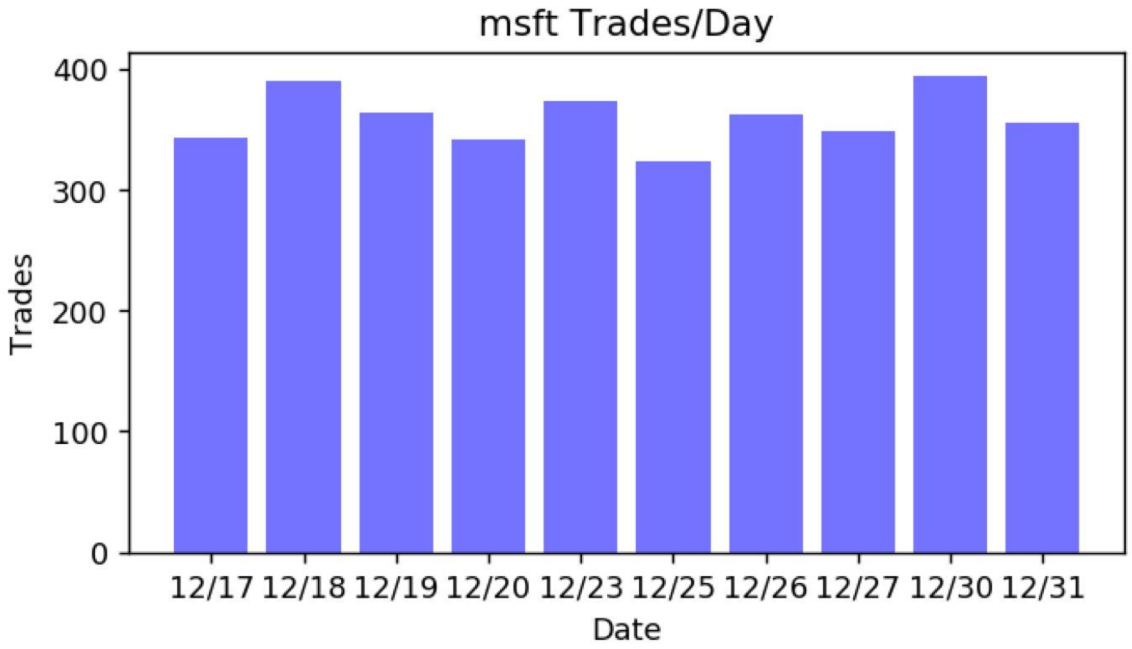

NFLX Hist for PnL
long
short
Frequencies
Pnl/Trade
250
200
150
100
50
0
−100
−50
0
50
100
150
200
250

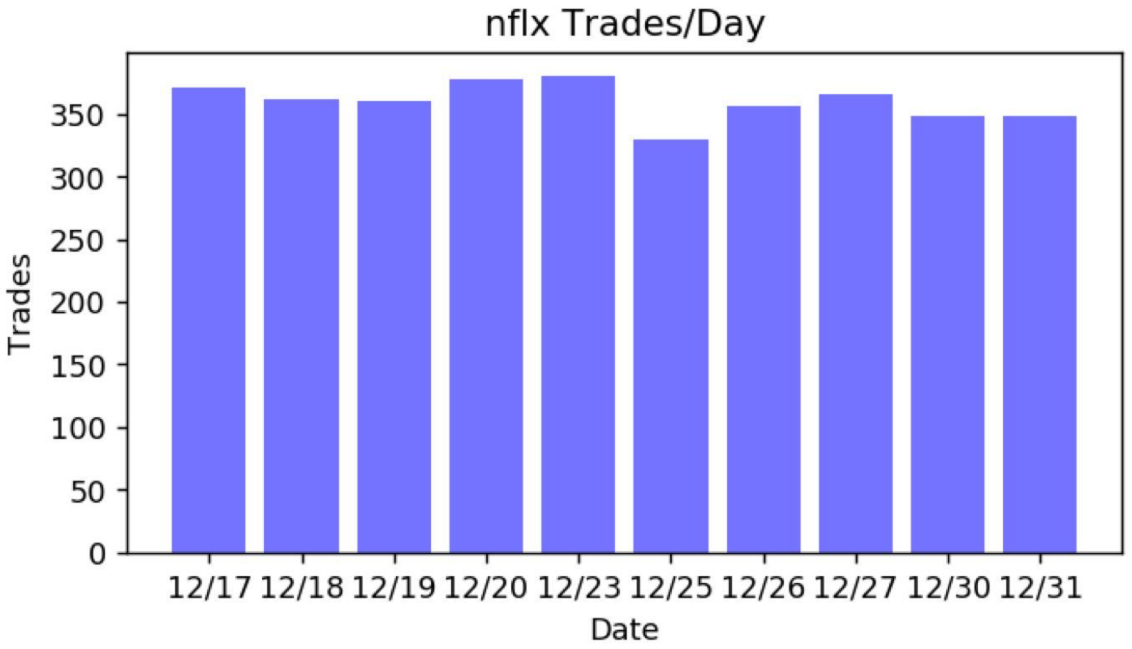
Holding Period



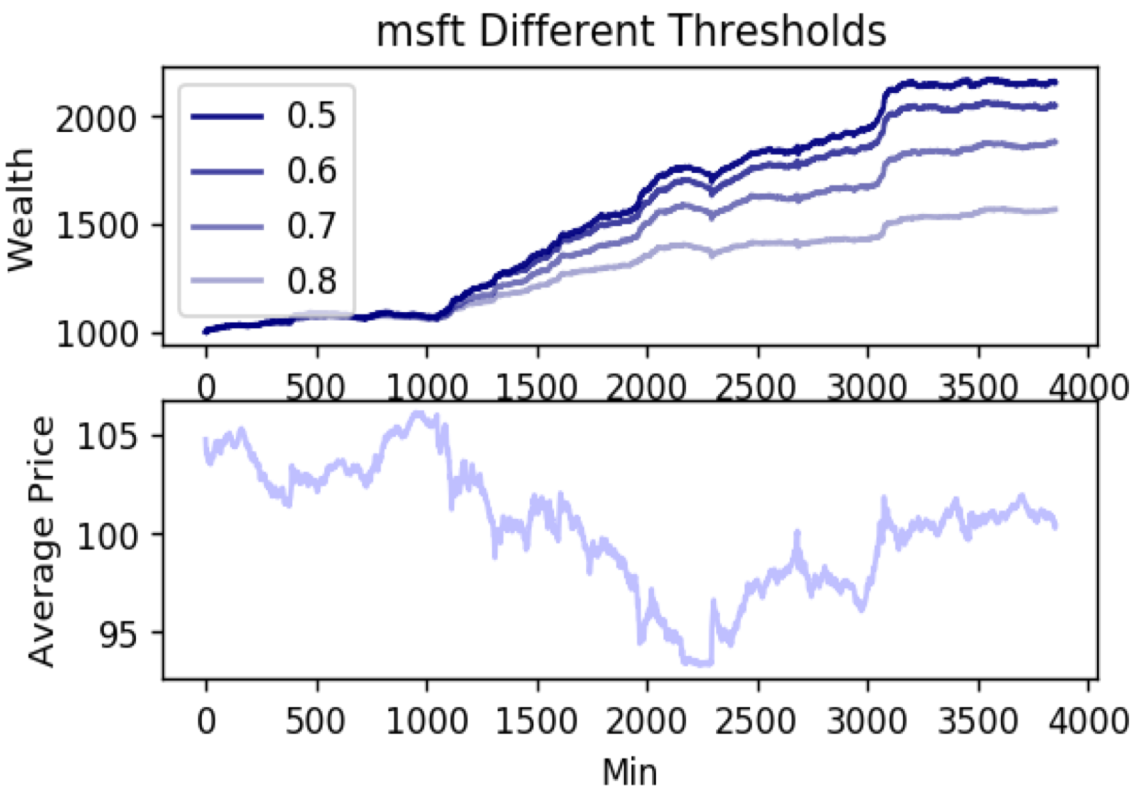


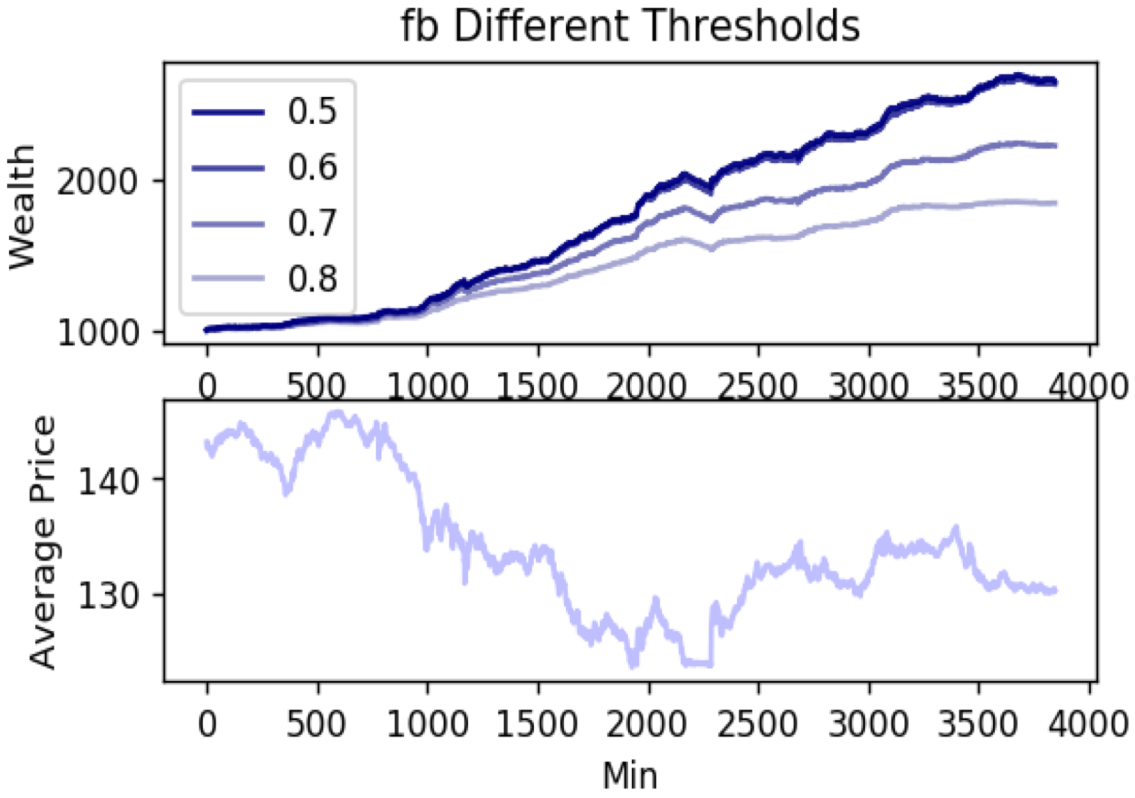
Number of Trades



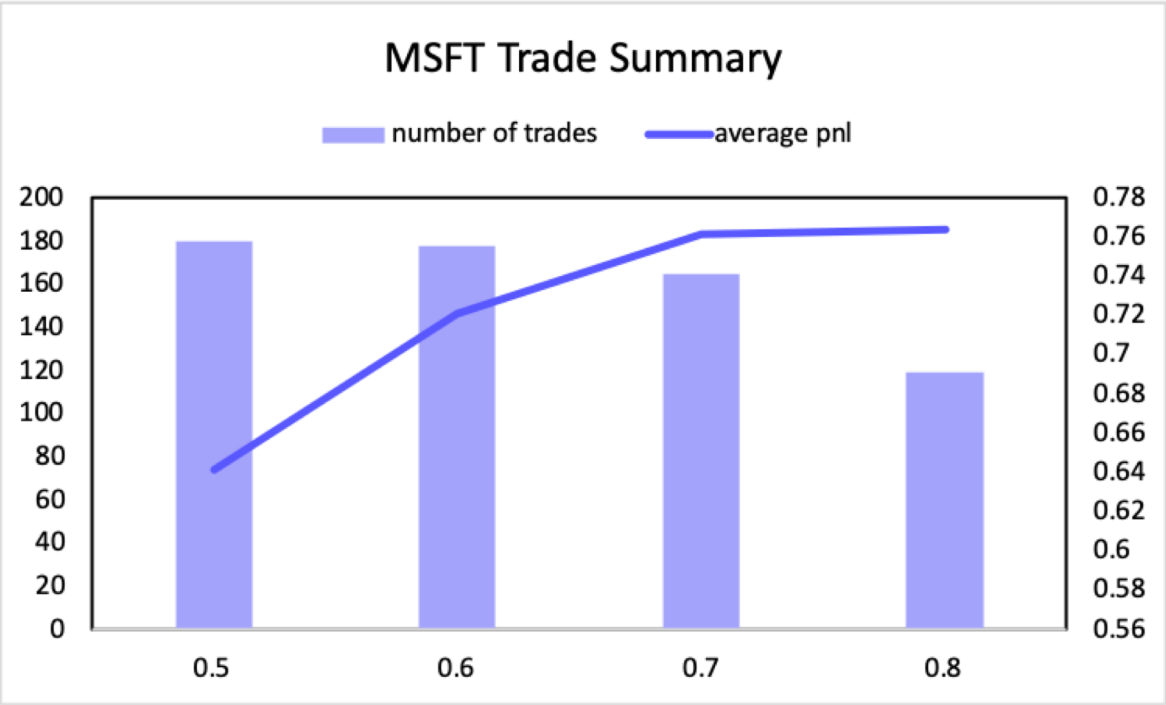


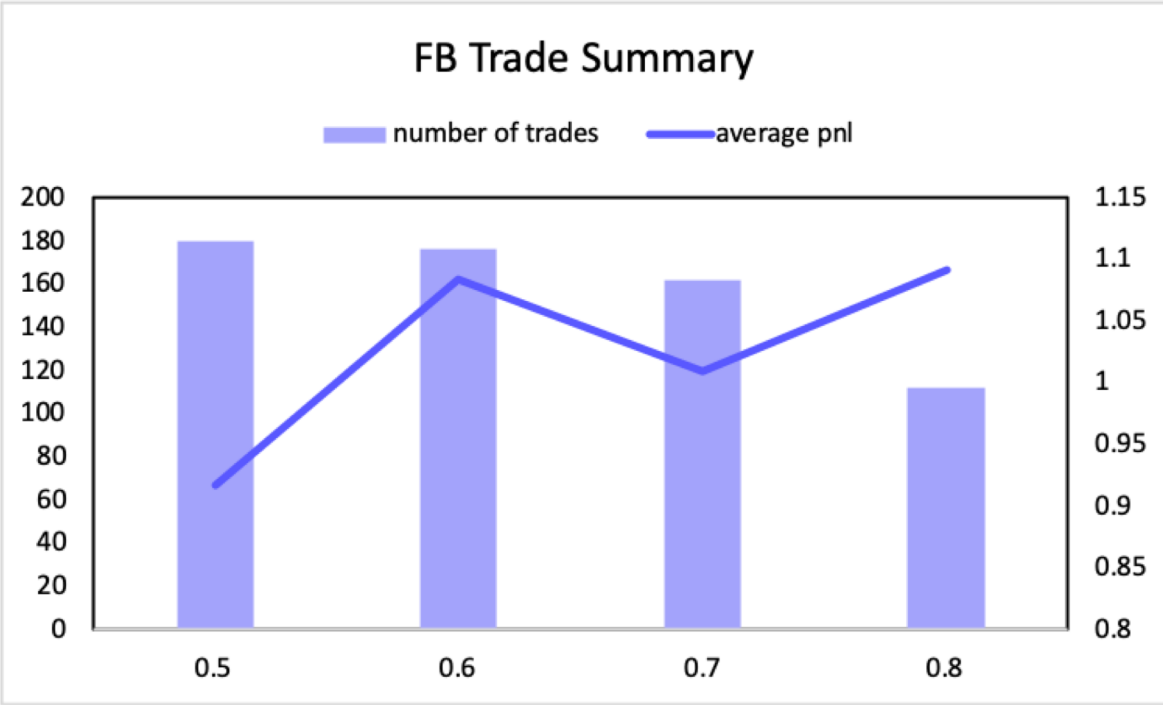
Probability Threshold(1/2)



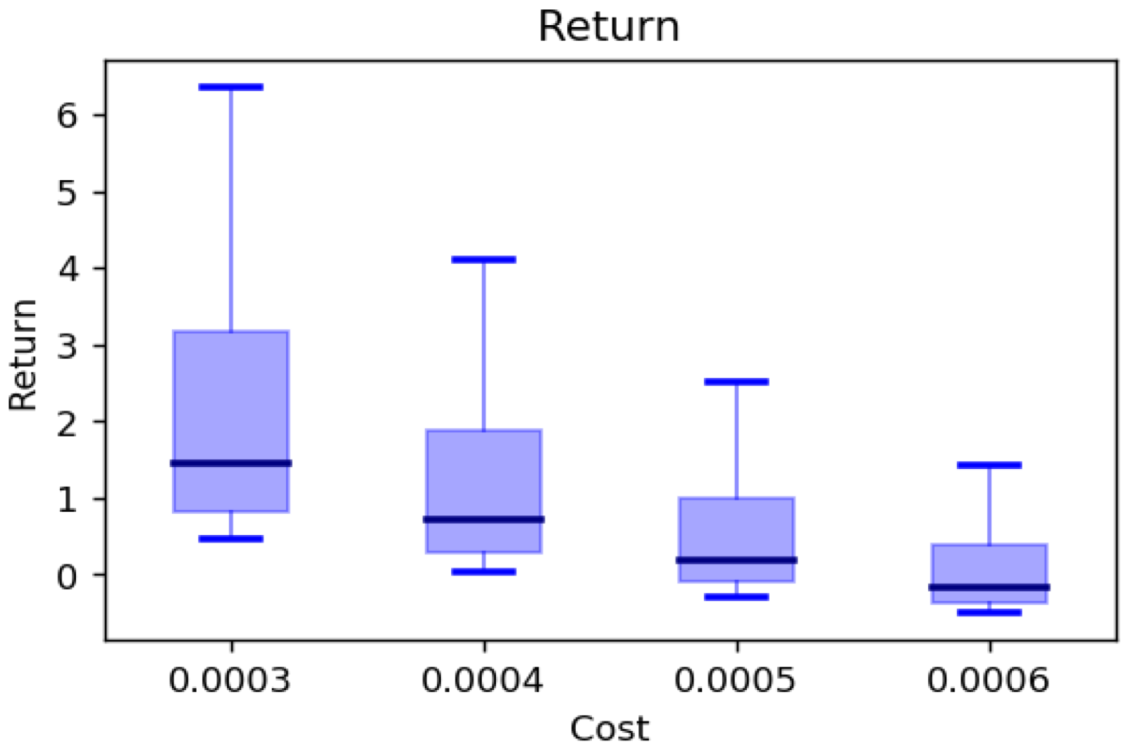


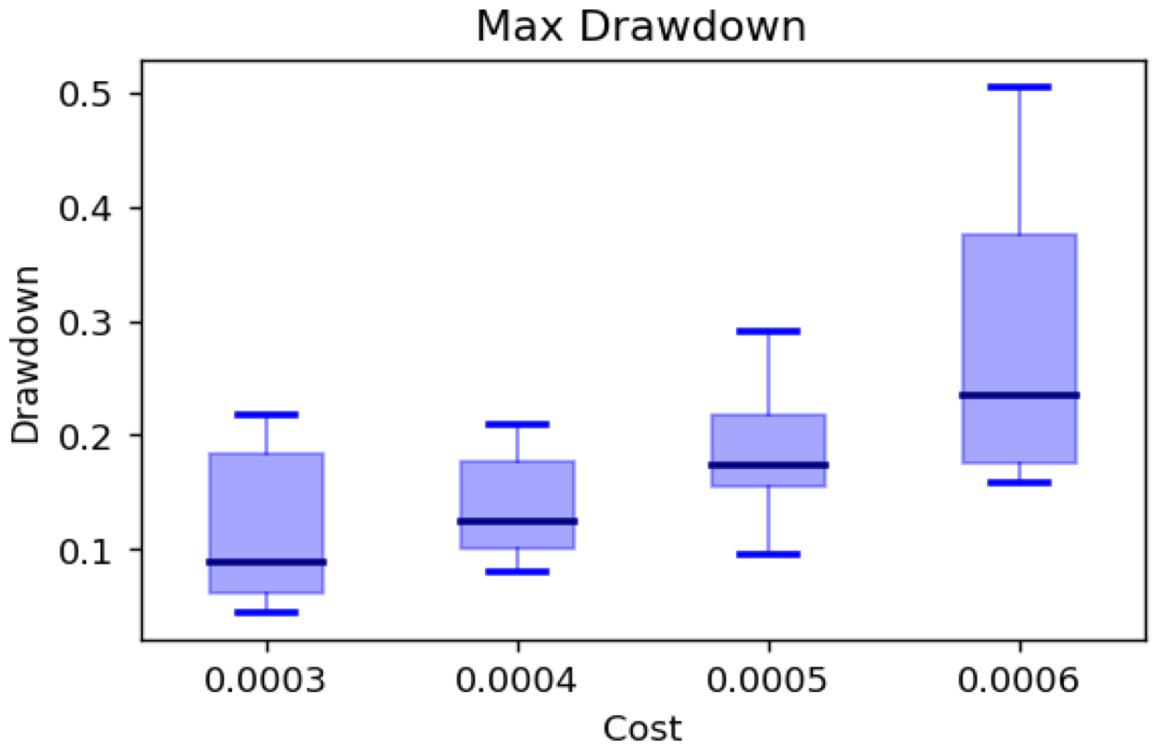
Probability Threshold(2/2)





Transaction Costs





Summary

- Before considering transaction, our strategy generates outstanding risk and return profile
- Sensitive to transaction cost due to high trading frequency
- Largest affordable cost is about 6 bp
- Optimal choice for probability threshold is 0.5

Execution

ML-Enhanced VWAP Strategy

Main Goal

Utilizing **short term price** prediction to gain a better execution price.

Idea

Considering we are executing the bid order.
If the market is expected to:

- Go up → Buy more now (Aggressive)
- Go down → Buy less now (Passive)

Tracking Error Limitation

For VWAP Strategy, we use our historical average volume distribution as our tracking goal. We have to ensure our tracking error do not exceed a limit.

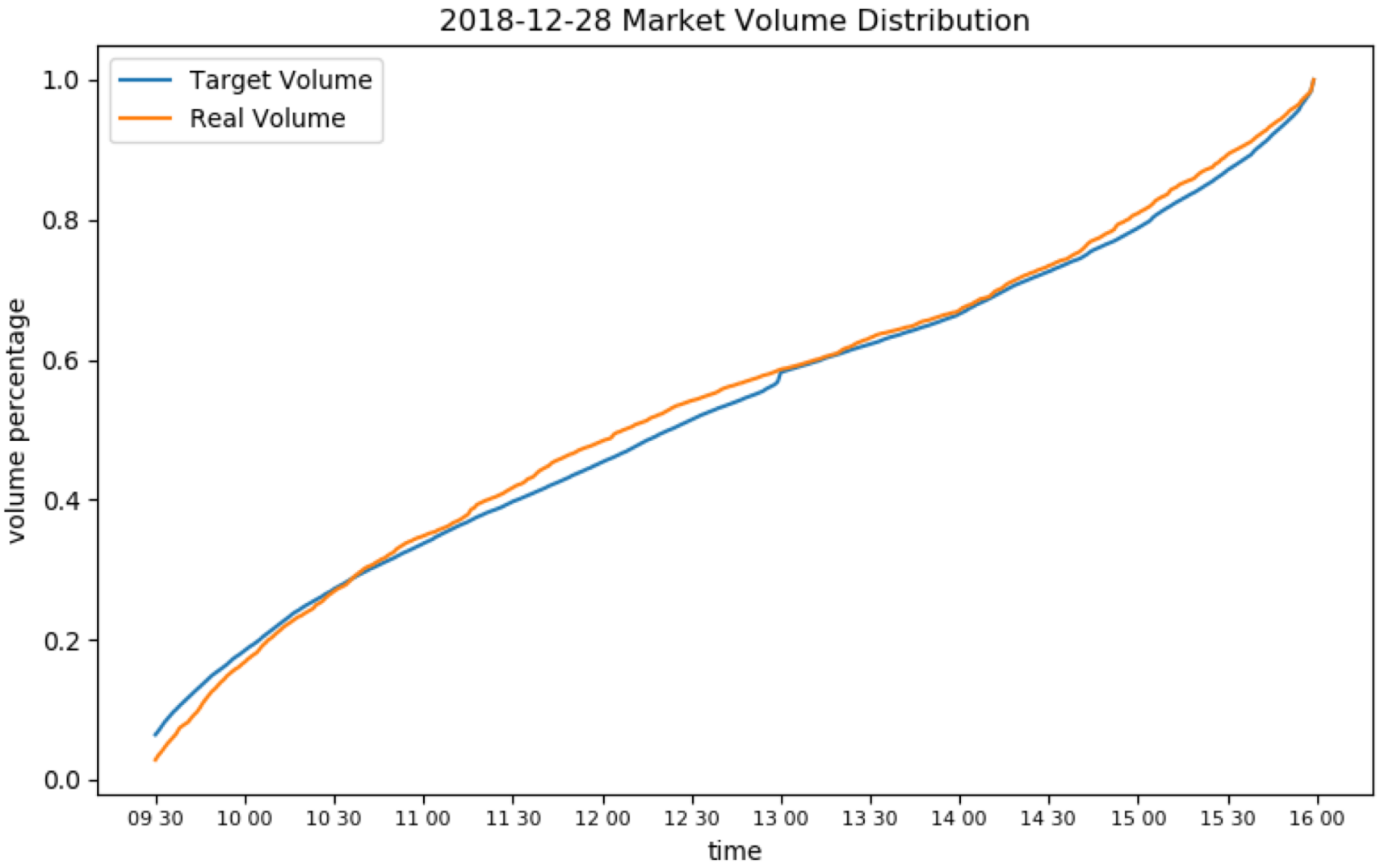
Reason:

If we execute our trade too early(Late), we can not capture the market change later, thus increasing the uncertainty(Risk) of return.

In our strategy, we use relative tracking error:
(V: Volume traded, G: goal)

$$0.97 \leq \frac{V}{G} \leq 1.03$$

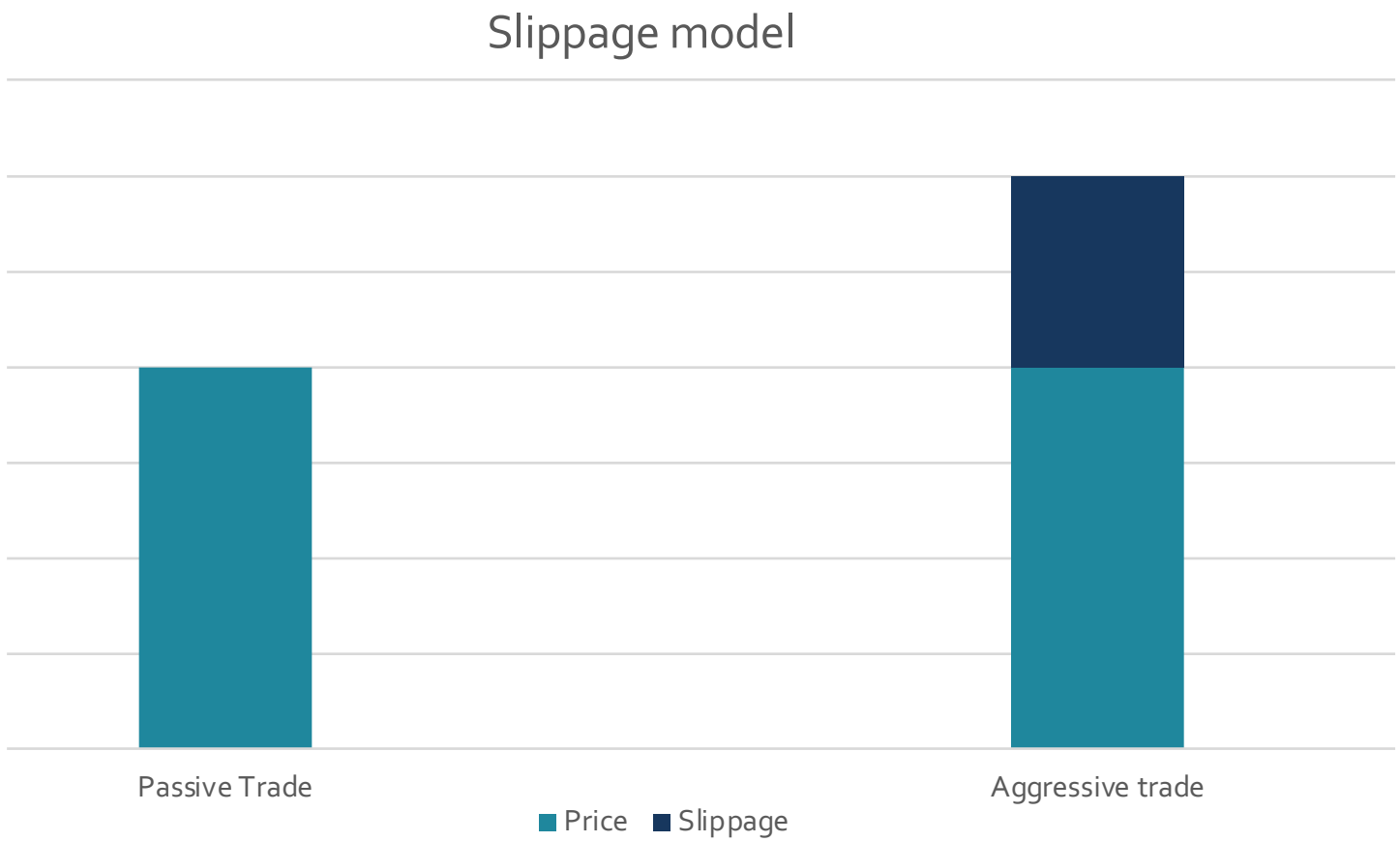
Historical Volume versus Realistic Volume



Our Model and Strategy

Slippage(Market Impact Model):

Consider the more we trade, the more market impact we will have.



Market Impact Model(1/3)

Assume we can trade at mean(Open, Close, High, Low) if there is no market impact considered.

Assume market impact is a linear function of volume:

Slippage = F(volume)

For example , At time t, the market volumet is VOL and we trade 0.2*VOL at time t.

The highest price we trade will be price+ 0.2*A*(High - price).
(A is sensitive constant.)

Market Impact Model(2/3)

Market Impact Model(3/3)

We optimize our trade volume at each minute by

MAX (Expected return)

S.T market impact$_t$ ≤ Expected Profit$_t$

highest trading price$_t$ < Expected price$_{t+1}$

Strategy Performance

Split the data into training set(Before Dec 17) and test set(After Dec 17)

Alpha Per Share: Saving per share
Definition: $vwap\_price_t$ - $Average\_trade\_price_t$

Timing Alpha: Excess profit by proper timing
Definition: $volume_t$*($vwap\_price_t$ - $Average\_trade\_price_t$)

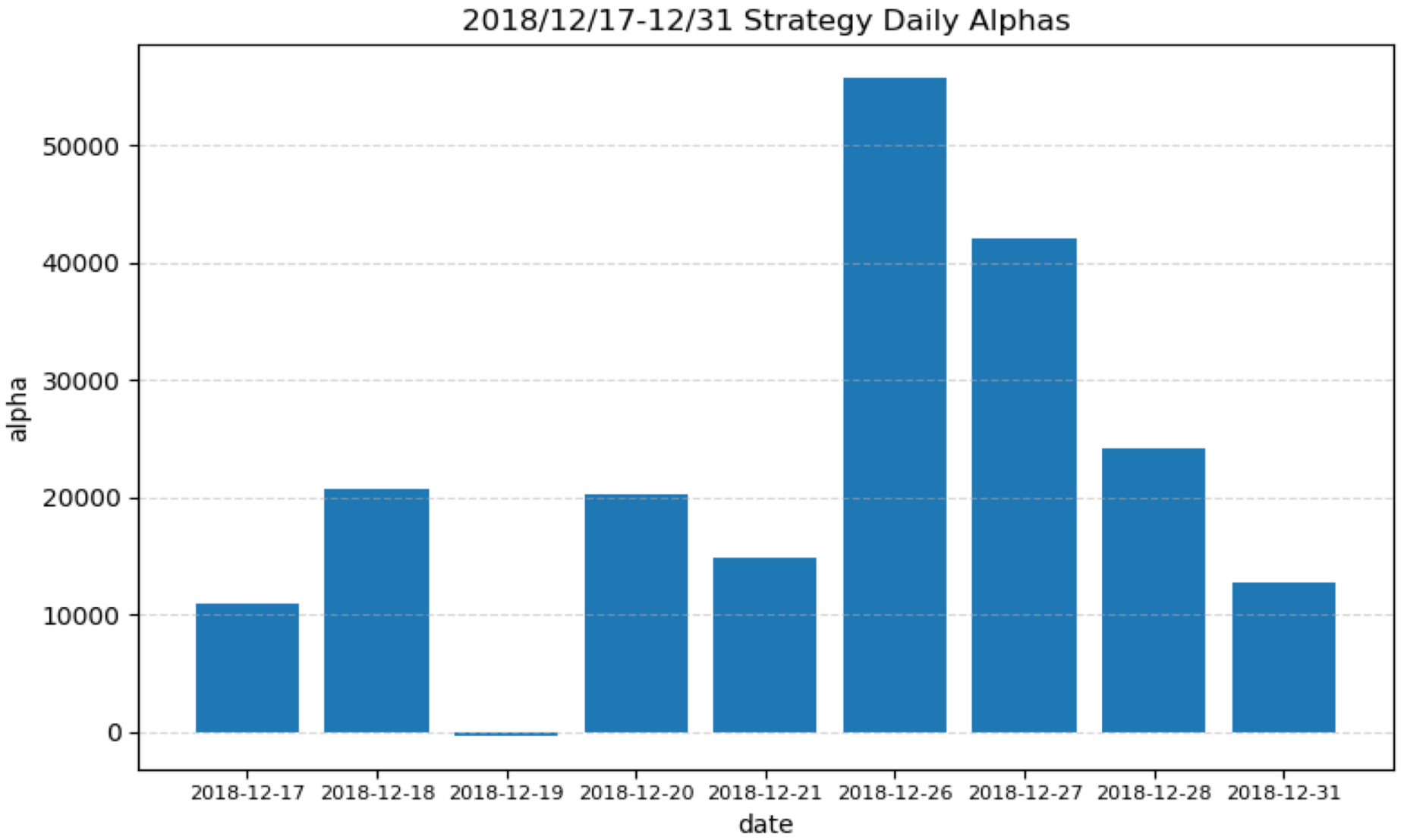
Market Impact Loss: Loss caused by slippage
Definition: $\sum volume_t$ * $Slippage_t$

Strategy

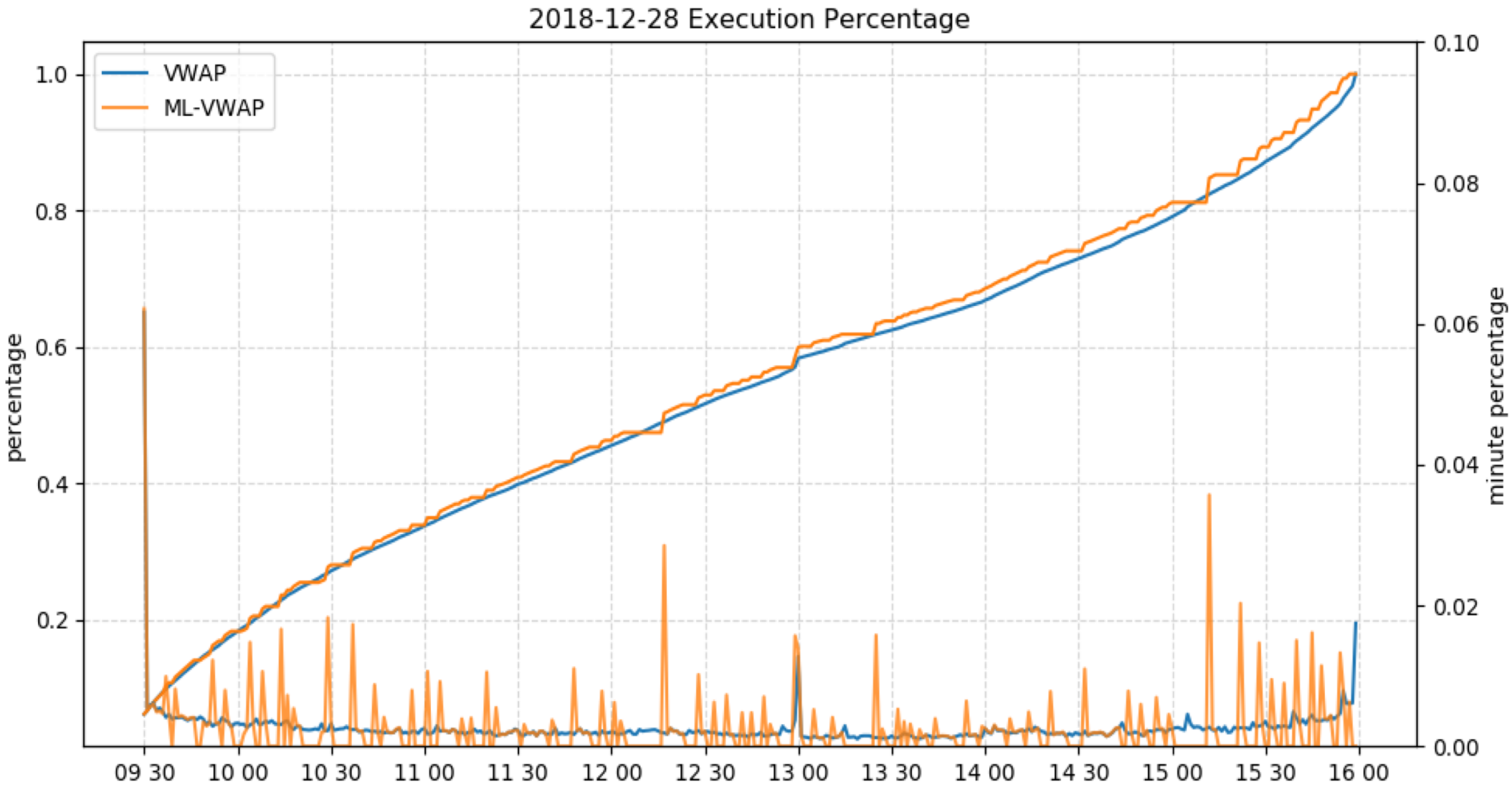
Assume we want to buy 200,000 shares of AAPL in a certain day.

Result:
We can buy at 0.11(7 bps) lower price per share compared with traditional VWAP strategy.

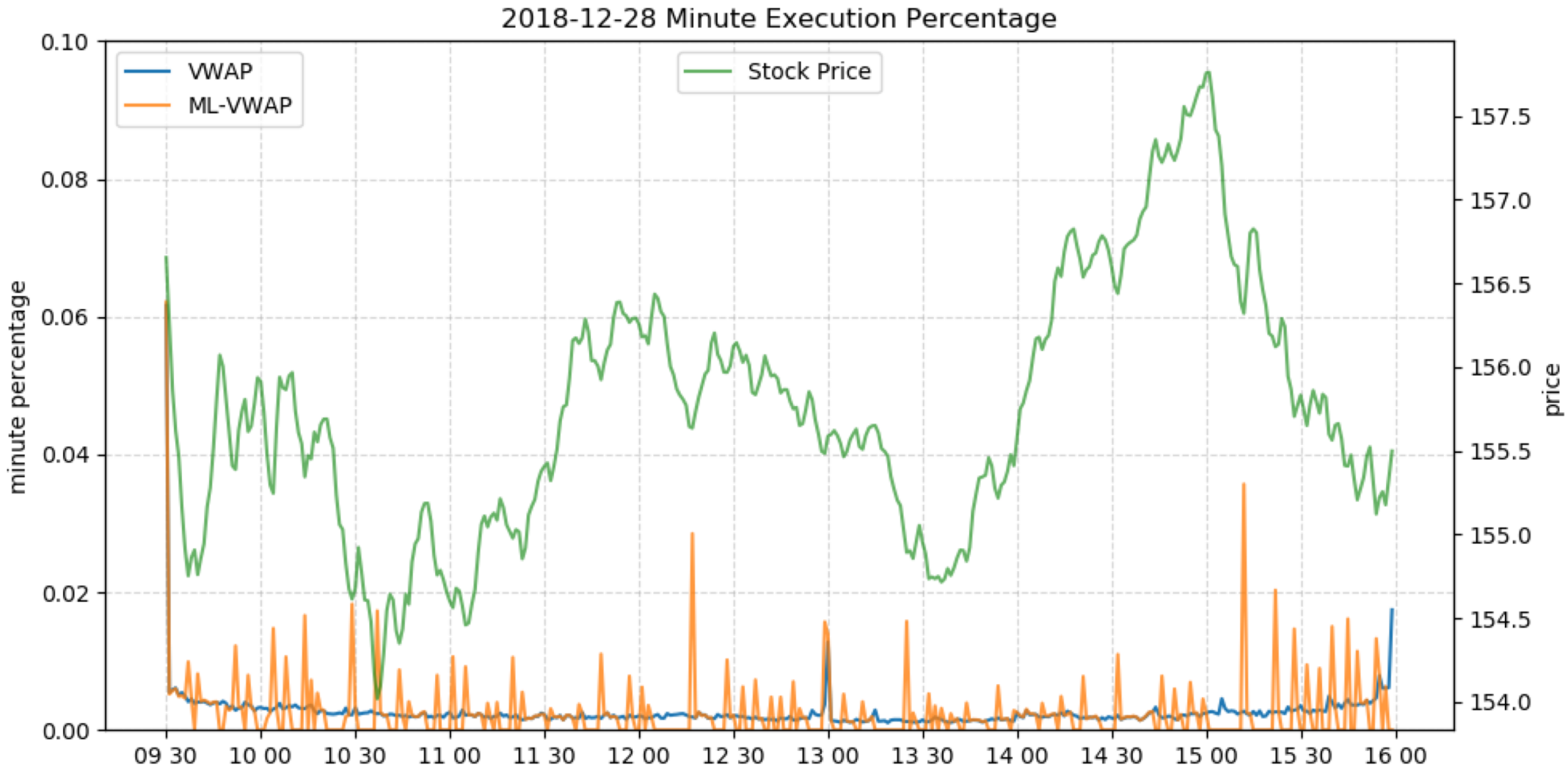
Strategy Performance in Testing Set



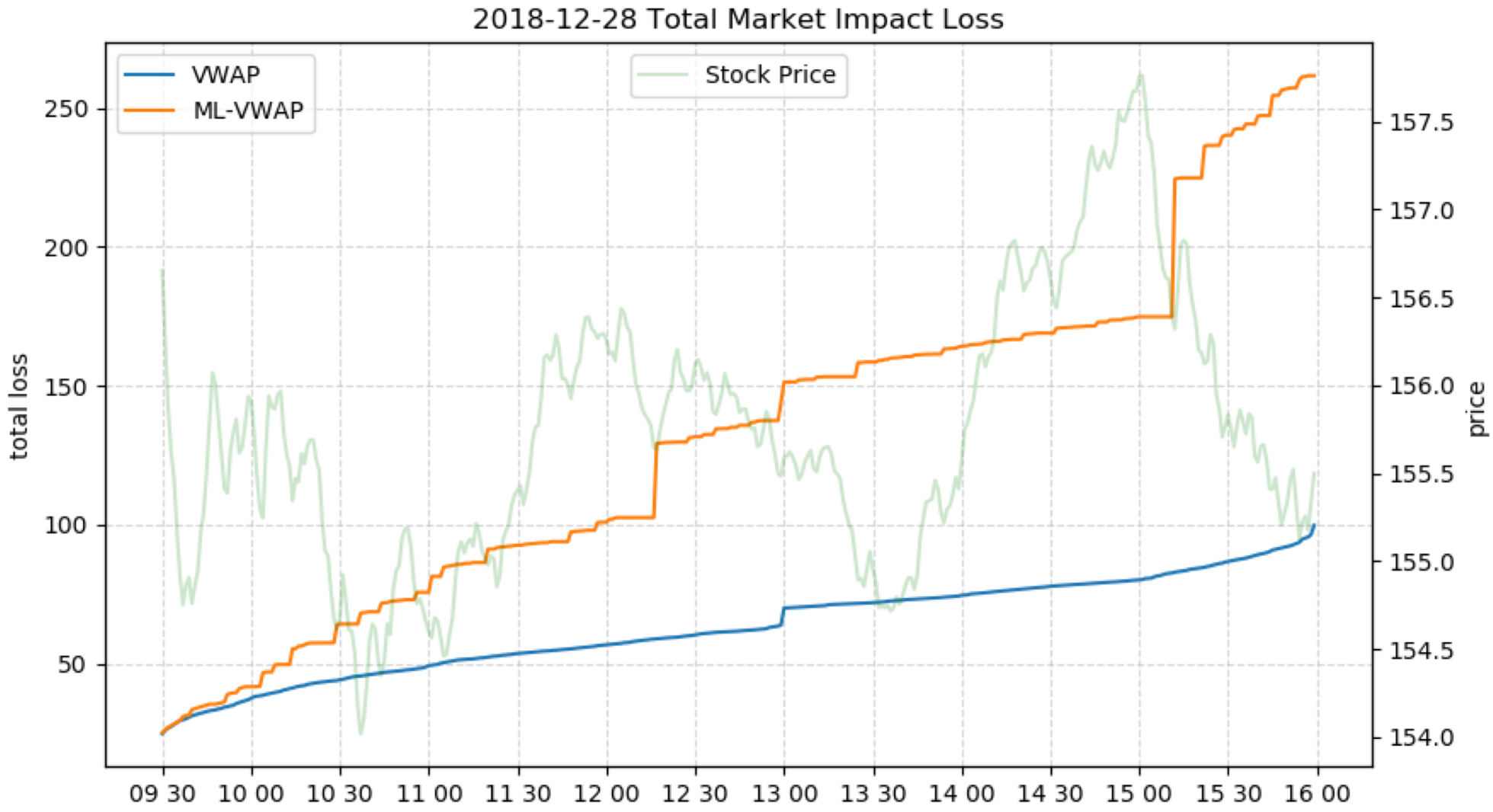
Strategy Comparation



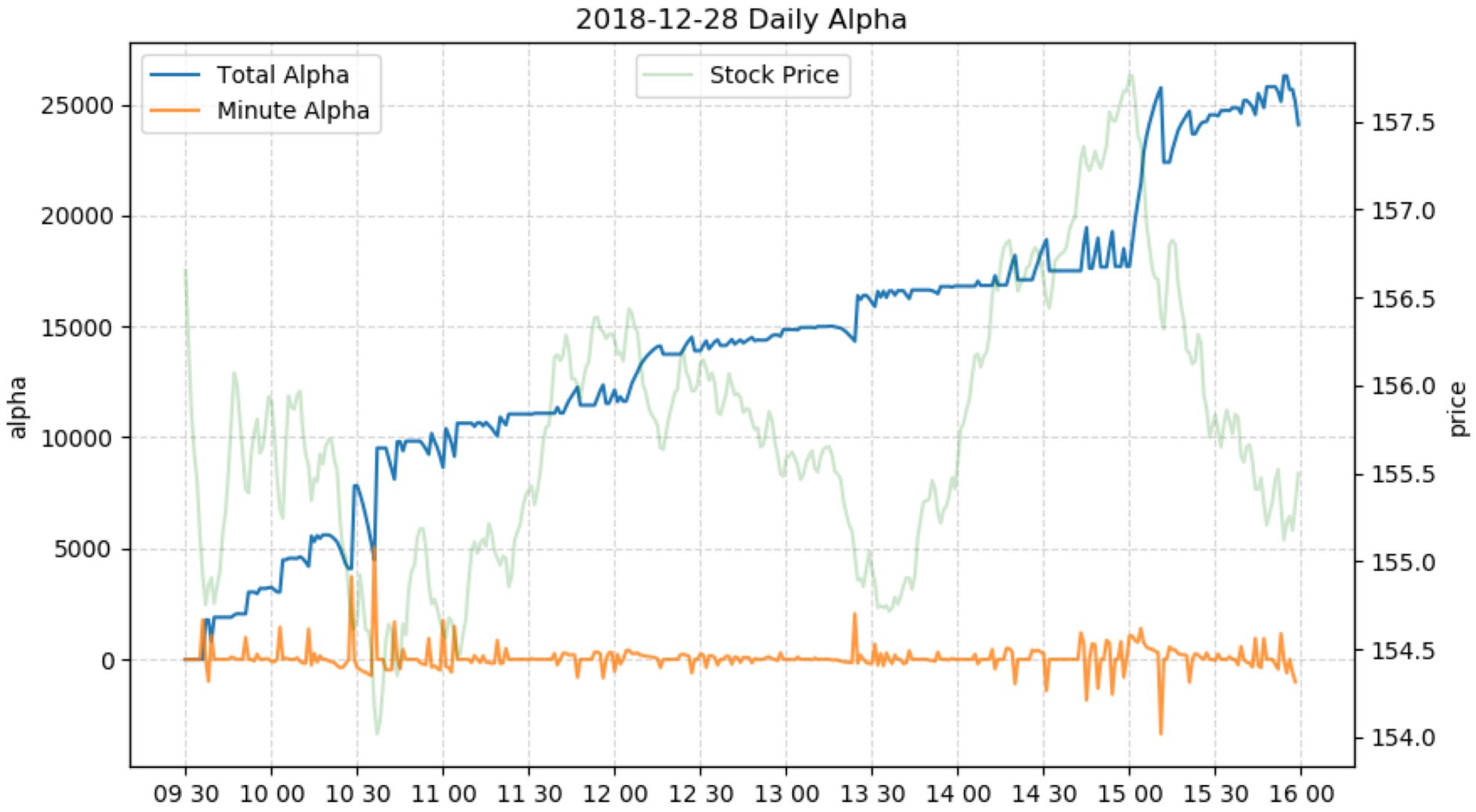
Price Change vs Optimal Execution



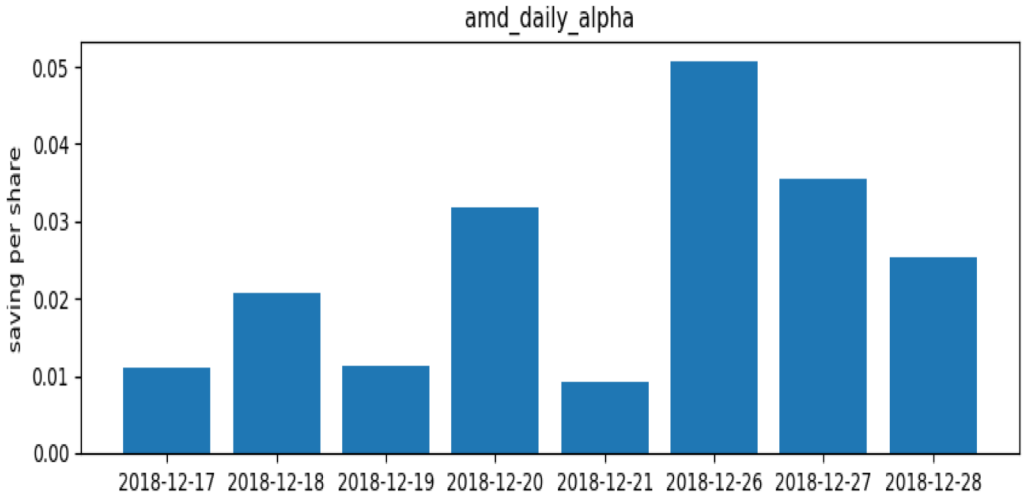
Aggressive Strategy Increases Market Impact

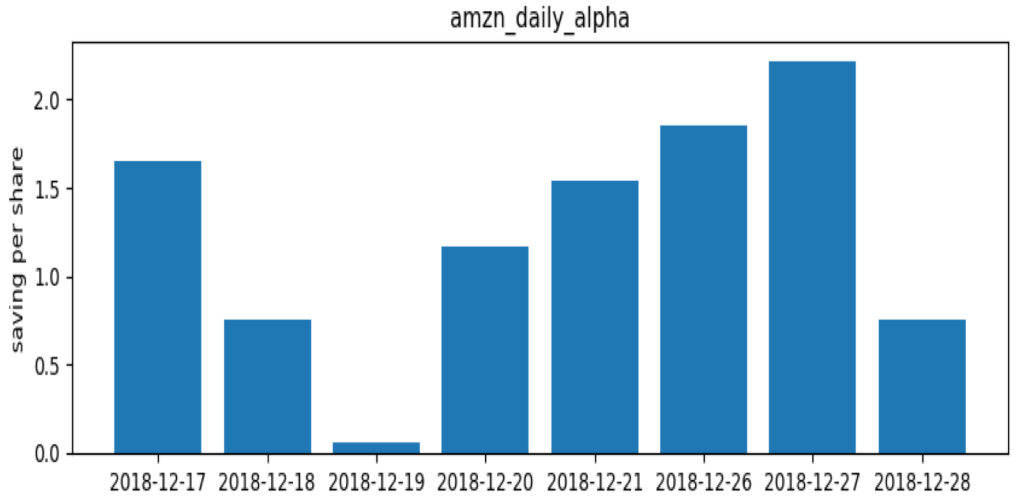


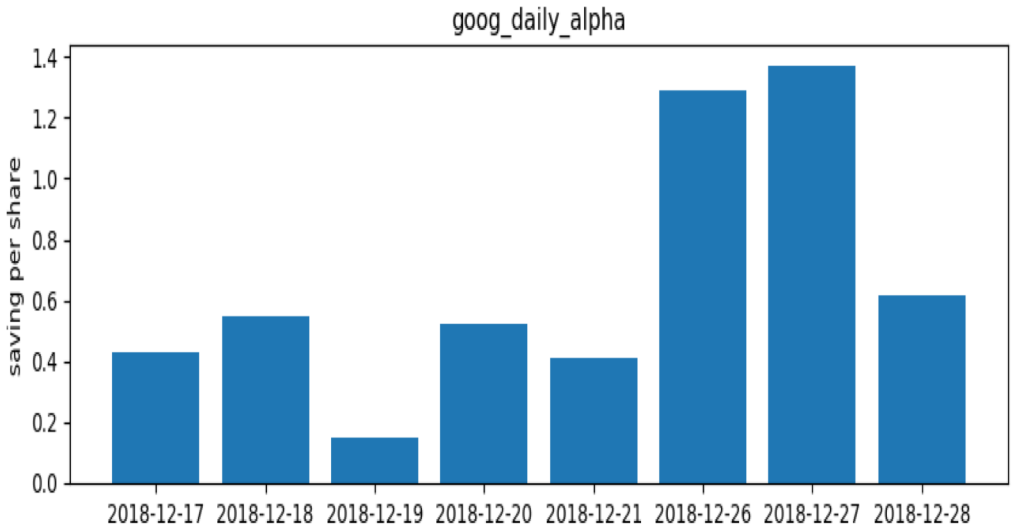
Cumulative Timing Alpha

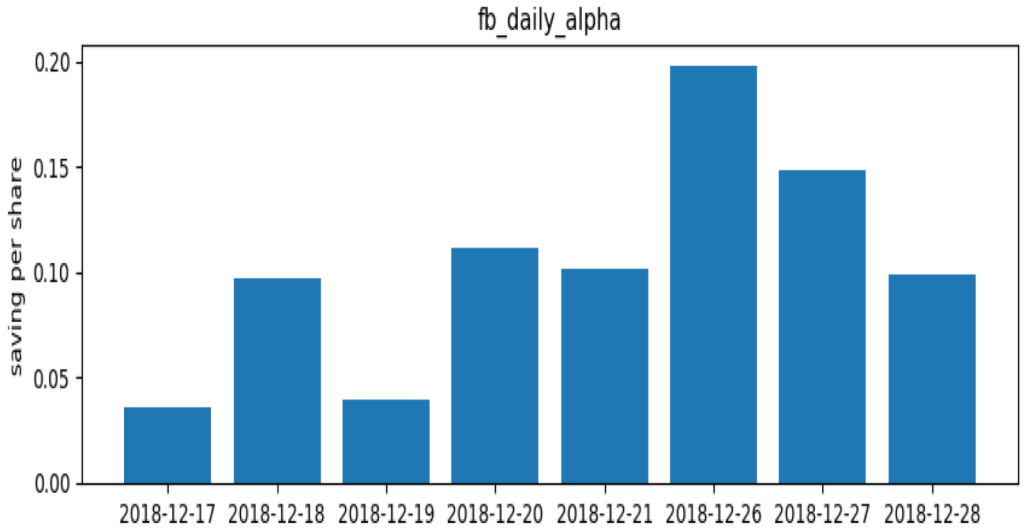


Results in Other Stocks









Where the Timing Alpha Comes from?

We find the alpha per share is proportional to the volatility of stock price.
Compare our excess return with average price change per minute:

	AAPL	AMZN	AMD	FB	GOOG
Average Saving Per Share	0.11	1.25	0.024	0.104	0.67
Average Price Change	0.096	1.57	0.022	0.114	0.761
Savings	7bps	9bps	12bps	7bps	7bps

Conclusion

- The prediction accuracy of the machine learning models is satisfactory (around 70%)
- Trading strategy based on the results has robust performance but sensitive to transaction cost
- Potential improvement ①: Increase volume prediction accuracy by using dynamic model
- Potential improvement ②: Utilizing order book data to optimize slippage model

References

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20), 7046-7056.

Chang, P. C., Fan, C. Y., & Lin, J. L. (2011). Trend discovery in financial time series data using a case based fuzzy decision tree. Expert Systems with Applications, 38(5), 6070-6080.

Chiu, D. Y., & Chen, P. J. (2009). Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. Expert Systems with Applications, 36(2), 1240-1248.

Choudhry, R., & Garg, K. (2008). A hybrid machine learning system for stock market forecasting. World Academy of Science, Engineering and Technology, 39(3), 315-318.

Hu, Y., Feng, B., Zhang, X., Ngai, E. W. T., & Liu, M. (2015). Stock trading rule discovery with an evolutionary trend following model. Expert Systems with Applications, 42(1), 212-222.

Kearns, M., & Nevmyvaka, Y. (2013). Machine learning for market microstructure and high frequency trading. High Frequency Trading: New Realities for Traders, Markets, and Regulators.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Systems with Applications, 42(1), 259-268.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.

Tsai, C. F., Lin, Y. C., Yen, D. C., & Chen, Y. M. (2011). Predicting stock returns by classifier ensembles. Applied Soft Computing, 11(2), 2452-2459.

Q & A

THANK YOU!