Intraday Pairs Trading Strategy Using Random Forest

PRIYANSHU BANSAL, AMAN JAIN, RAMANUJAM NARAYANAN, GOKUL RAMANATHAN, ALANKRIT VARMA

Agenda

- Introduction
- Data
- Methodology
- Results
- Next Steps

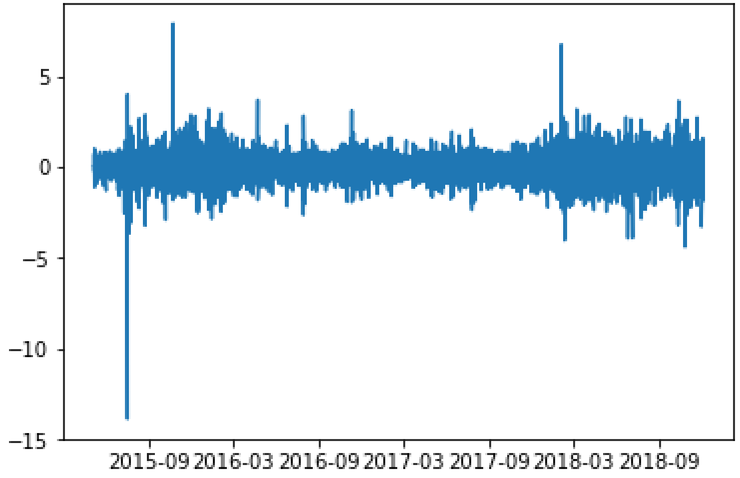
Introduction

- Pairs trading is a popular trading strategy in the last three decades after it was first used by Morgan Stanley in 1980s
- Pairs trading means to utilize a pair or a bag of related financial instruments to make profits by exploiting their relations
- In this project, we will use the spread model, the O-U mean reverting model, and Random Forest to build a trading strategy and apply the strategy to GOOG/GOOGL
- The two shares classes trade at different prices because GOOG shares have no voting rights, while GOOGL shares do

Data

- The data for GOOG and GOOGL shares is downloaded from Bloomberg
- The data set contains the OHLC for both securities and ranges from 5/1/2015 to 12/3/2018 at a 15-minute frequency
- 37 (plus their lags) technical indicators are created from this data as trading signals / features for our analysis
- Some of the major indicators are Accumulation/Distribution Index (Volume), Bollinger Bands (Volatility), Moving Average Convergence Divergence (Trend), Relative Strength Index (Momentum), and Daily Log Return etc.

The Spread



Methodology

- The focus of our model is to predict whether the spread between the two classes of shares is positive or not in the next 15-minute interval
- A time series of the difference in prices of the securities is created. We first-difference this time series to obtain the change in spread over time
- The dependent variable (target) is created as follows:

$$y = \begin{cases} 1, \text{ if the first-difference of spread is} > 0 \\ 0, \text{ otherwise} \end{cases}$$

- This essentially converts the prediction problem into a classification one
- Now the model is built to predict the direction (sign) of the change in spread using Random Forest in python

Methodology

O-U Process:

The canonical pairs trading spread model is as follows:

$$\frac{dA_t}{A_t} = \alpha dt + \beta \frac{dB_t}{B_t} + dX_t \tag{1}$$

$$dX_t = \theta(\mu - X_t)dt + \sigma dW_t \tag{2}$$

By integrating (2), we have

$$X_{t+1} = a + bX_t + \epsilon_{t+1} \tag{3}$$

O-U process equation:

$$\theta = -log(b) \times \frac{1}{\Delta t}$$

$$\mu = \frac{a}{1-b}$$

$$\sigma = \sqrt{\frac{Var(\epsilon)2\theta}{1-b^2}}$$

$$\sigma_{eq} = \sqrt{Var(X_t)} = \frac{\sigma}{\sqrt{2\theta}} = \sqrt{\frac{Var(\epsilon)}{1-b^2}}$$

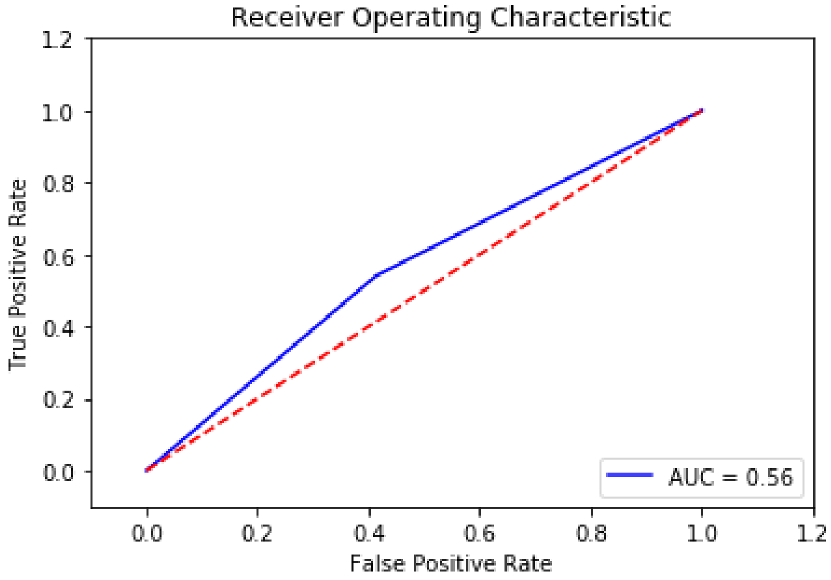
Methodology

XGBoost and Feature Engineering:

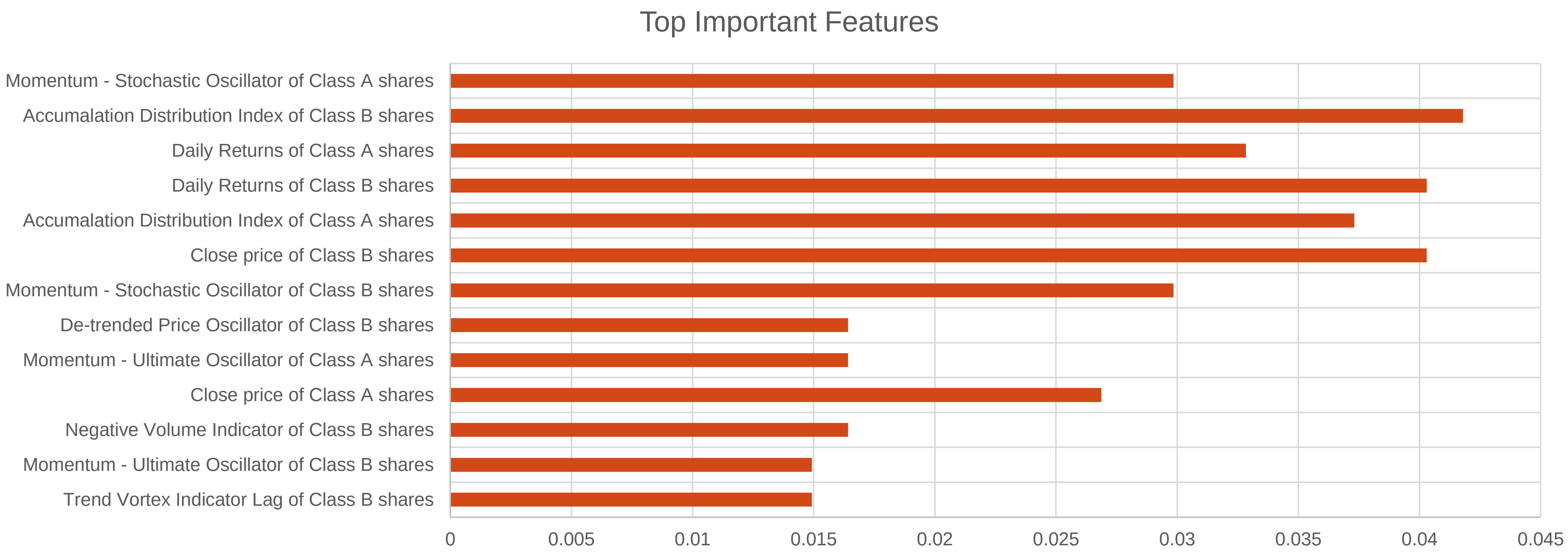
- Features are common technical indicators and their lagged values used
- XGBoost - Gradient boosting framework uses an ensemble of weak decision trees to produce a prediction model with strong prediction properties
- Ensemble trained in a stage-wise fashion to progressively improve performance
- L1 and L2 regularization used to penalize model complexity
- Advantage - Better performance (accuracy) than bagging, in general
- Disadvantage - Can overfit due to the stage-wise construction of ensemble

Results

- The in-sample accuracy of prediction is 66.85%. While, the model exhibits an out-of-sample accuracy of 56% (shown by the ROC chart below)
- In comparison to the traditional OU or AR1 model, which achieved an accuracy of 52%, the ML algorithm gives us a 4% edge (out-of-sample)

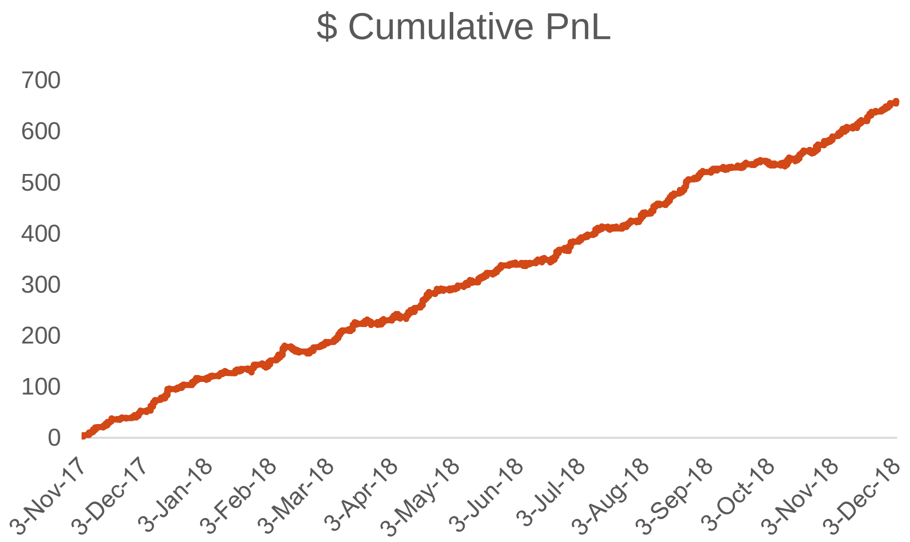


Top Important Features



Trading strategy based on signals

Assuming no transaction cost



- The strategy goes long the spread when the predicted signal is 1 and short the spread when the signal is -1
- The adjacent graph assumes no transaction costs
- Sharpe Ratio (without slippage/ funding costs): 12
- Sharpe Ratio (with slippage of 1bp/ funding cost of 4%): -2

Funding costs/ Slippage	0.005%	0.010%	0.020%
2%	8.40	4.81	-2.57
4%	8.19	4.61	-2.78
6%	7.99	4.40	-2.99
8%	7.78	4.19	-3.20
10%	7.58	3.99	-3.41

Next Steps

- This approach can be tested for other firms in the market with distinct share classes, like VIACOM
- We can test the accuracy of the forecasts using different machine learning methods like Support Vector Machines (SVM) or Neural Networks, or Gradient Boosting
- Going forward, we can explore more features through better feature engineering

Thank You

Appendix

- The **Negative Volume Index** is a technical indication line that integrates volume and price to graphically show how price movements are affected from down volume days
- The **Ultimate Oscillator** is a range-bound indicator with a value that fluctuates between 0 and 100. Similar to the Relative Strength Index (RSI), levels below 30 are deemed to be oversold, and levels above 70 are deemed to be overbought.
- The **detrended price oscillator** (DPO) is an indicator in technical analysis that attempts to eliminate the long-term trends in prices by using a displaced moving average so it does not react to the most current price action.
- **Momentum Stochastic Oscillator**: In technical analysis of securities trading, the stochastic oscillator is a momentum indicator that uses support and resistance levels
- **Accumulation Distribution Indicator** or ADL (Accumulation Distribution Line) is a volume based indicator which was essentially designed to measure underlying supply and demand