PRINCETON UNIVERSITY

Applications of Machine Learning to Statistical Arbitrage Pairs Trading

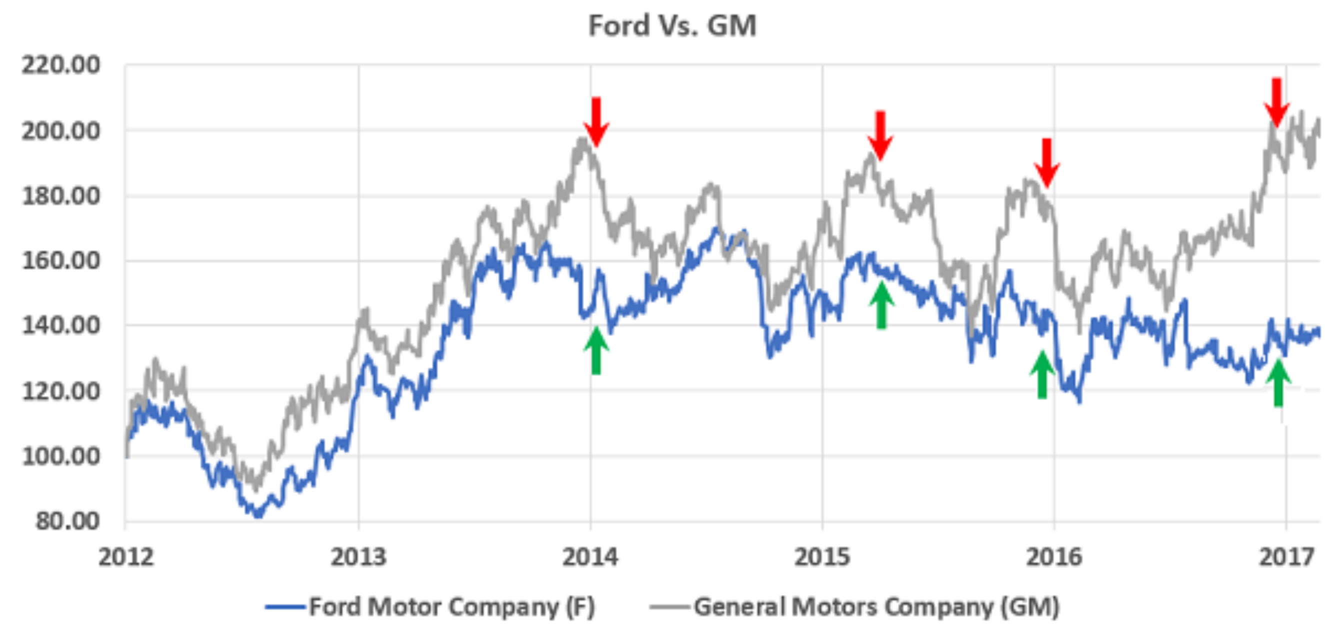
Nicholas Kim

ABSTRACT

- Trading in financial markets is an ever-present part of modern society, driving the fundamental aspects of price discovery and liquidity. Countless strategies for efficient trades exist today, with one of the most popular being a form of statistical arbitrage known as "pairs trading".
- A pairs trading strategy aims to find stocks that are highly cointegrated, and uses the mean reverting nature of such pairs to predict price movements
- While deceptively simple, pairs trading in practice is quite difficult to profitably implement.
  - Many approaches rely simply on observing the difference between the pair's price on a given day, the historical difference, and an implementation of a manual threshold
  - Such approaches have high propensity for false trading signals, and fail to make use of a plethora of other highly rich data sources
- We seek to explore applications of machine learning to generate more reliable trading signals. We present an analysis of two strategy approaches - one based on difference halving, the other on velocity reversals

BACKGROUND AND APPROACH

A Simple Example of an Optimal Pairs Trading Strategy for Ford and GM

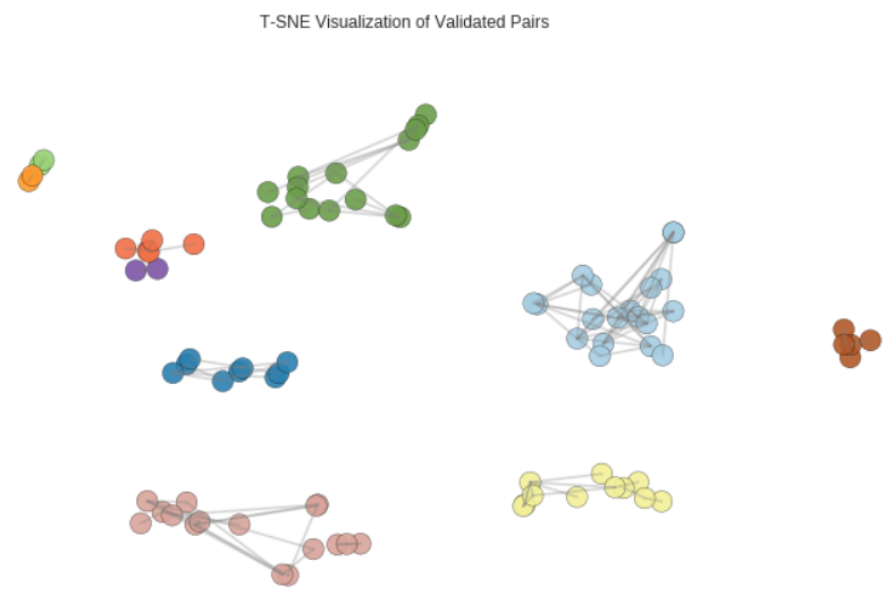


➢**Optimal trades occur at the “maximum split” points, right before the prices begin to correct. The question becomes how to best predict the start of a reversal.**

How to predict profitable trading times with ML technques?

- **Dataset: Quantopian Morningstar Pipeline** – Includes thousands of highly traded US equities
- **First Task – Deriving a dataset of highly cointegrated stocks**
  - Many potential ways to go about. For instance, raw cointegration tests could be computed between the equities in the dataset, and some percentage threshold of cointegration could be used
  - However, approach runs into many issues. 1) Computationally expensive 2) Arbitrary Separations 3) Lack of global consistency. Too many idiosyncrasies in comparing only specific pairs one at a time
- **Solution – DBScan Clustering (from open source Quantopian code)**

- Convert prices into returns to capture percentage changes
- Apply principal component analysis to reduce dimensionality of data (since there are prices for hundreds of days, this creates extremely high dimensional data)
- Apply DBSCAN algorithm, which automatically finds an optimal number of clusters and does not include data points that do not fit well into any cluster.



COMPARING AND EVALUATING

- <u>**Sample Strategies for Comparison**</u>
- **1) Buy and Hold –** How does a given strategy compare to a simple hold of the two equities in a pair?
- **2) Rudimentary Algorithmic Pairs Trading** – Find a historical mean difference, and then establish a valid trading signal whenever the mean difference is more than a certain % (in this case 30) away from the historical

- **METRICS**
  - **Returns –** The percentage change in value of one's portfolio
  - **Profitability –** Percentage of trades that were profitable
  - **Volatility –** Standard deviation of returns
  - **Sharpe Ratio** – Risk adjusted returns, defined as (returns – risk-free returns)/Volatility
  - **Maximum Drawdown –** The worst trade executed by a given strategy

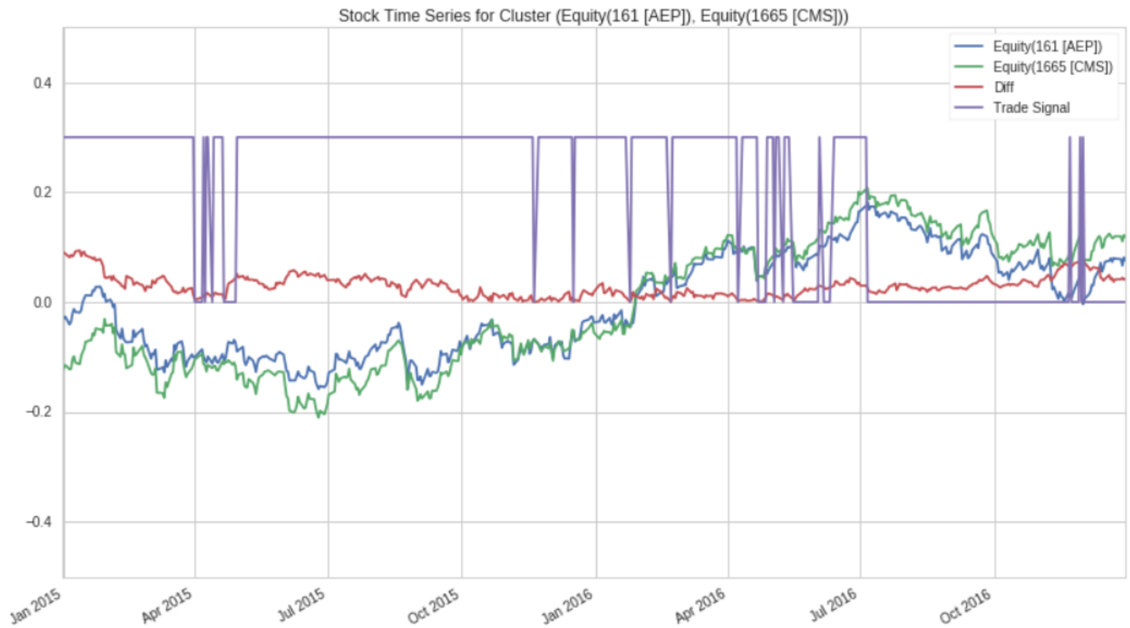
STRATEGY 1) – SVM Applied to Return Difference Velocity

- **A Classification Problem Approach - Determining whether or not it would be profitable to enter a trade at a given time**
- **Issues – 1) Target Labels 2) Features**

- **Target Label – Return difference halves in the next 80 trading days.**
  - **Observe the difference between the log returns of the two equities in the pair**
  - **If this difference halves in the next 100 trading days, then we consider it safe to enter into a trading position**

- **Main Feature – 20-day Trailing Velocity of Return Difference**
  - **First, the difference in log returns is derived between the two pairs**
  - **Then, the percentage change in this difference is calculated for each day**
  - **Then, for any given day where we are trying to predict the trading signal, we use the past 20 days of trailing returns**

Illustration of Derived Features for Pair AEP, CMS



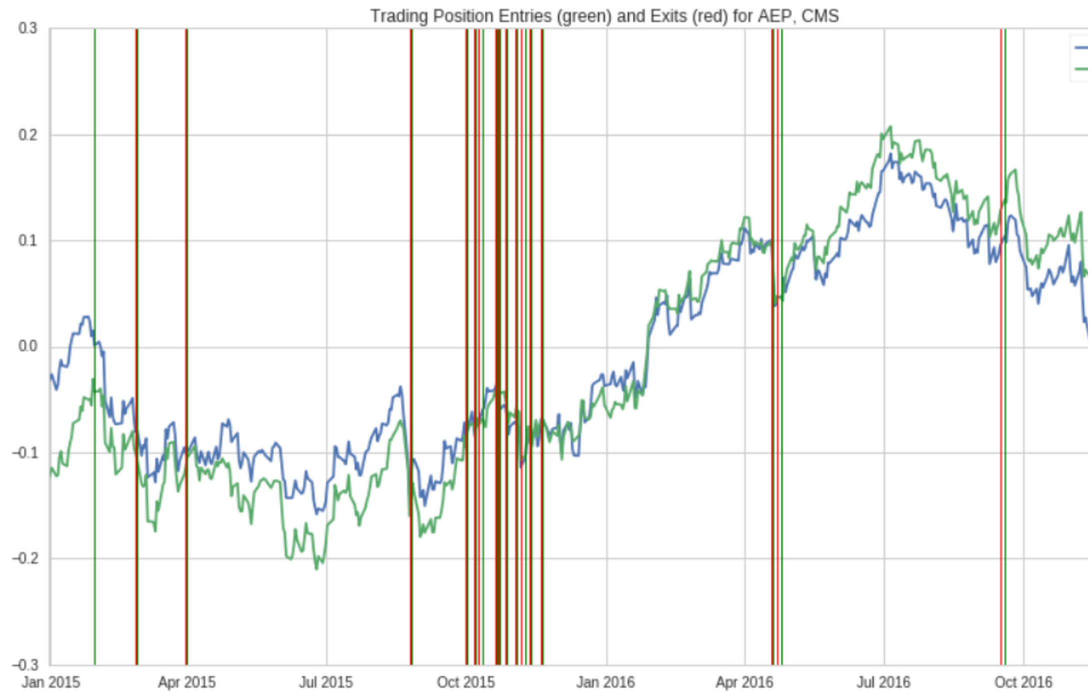
- **SVM** was applied to the classification problem with an 80:20 train:test split applied to the valid pairs
- In the test set, **SVM** managed to achieve **77.67%** Accuracy in determining "difference halving" scenarios

Converting Signals Into A Testable Trading Strategy

- Now for a given time series, we are able to sequentially predict whether or not it is safe to open a trade. However, this is far from a testable strategy, since one does not have infinite money to trade

- An algorithm to test was decided with a minimalistic framework. This was applied and averaged over all equity pairs in the test set to obtain evaluation metrics.
  - 1) Check to see if it is valid to open a trade
  - 2) If it is valid, then the strategy assumes that all of one's initial capital is dumped into the trade
  - 3) If the return difference halves within the next 100 days, close the trade and collect the returns
  - 4) If the return difference fails to half, then the trade is automatically closed after 100 days to cut potential losses

- **Trading Strategy Avg Returns** 1.0671
- **Buy and Hold Avg Returns** 1.2196
- **Risk-Free Returns** 1.0066
- **Percentage of Profitable Trades** 0.9071
- **Volatility of Returns** 0.04733
- **Worst 10 Trades Avg** -0.05806
- **Single Worst Trade** -0.1067
- **Sharpe Ratio** 1.4149
- **Simple Algorithmic Pairs Returns –** 1.0821



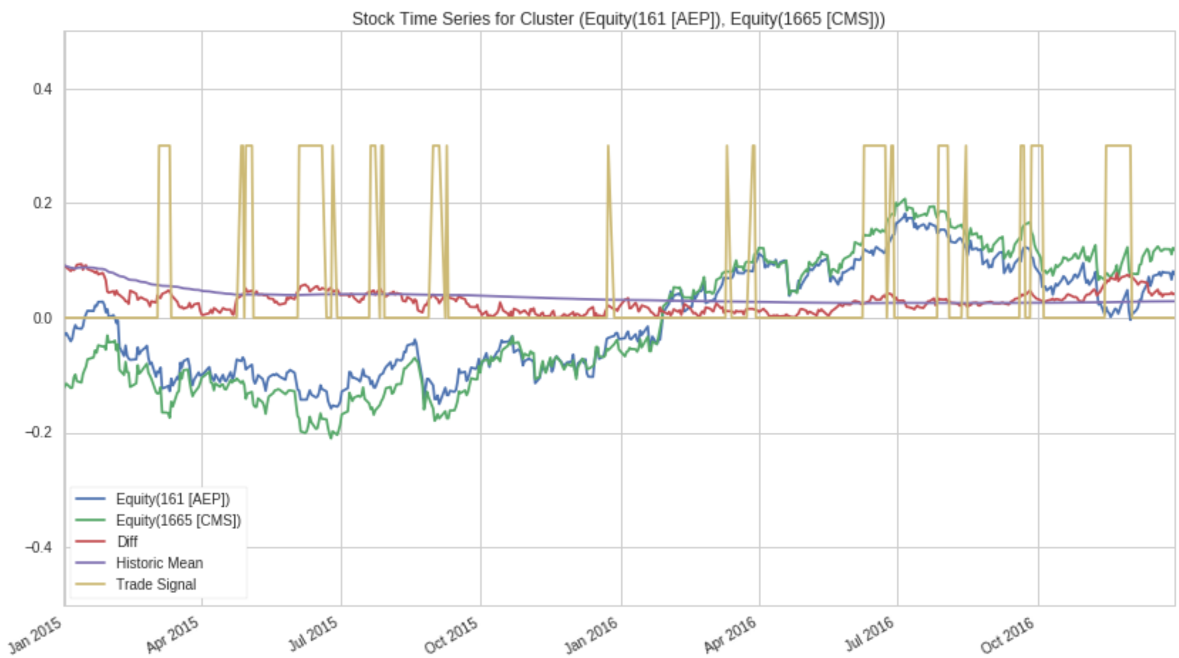
STRATEGY 2) – Decision Tree Classifier Applied to Difference Moving Average Reversal

- **<u>Modified Problem Approach</u>** - **Determining whether or not it would be profitable to enter a trade at a given time, purely based on velocity factor**
- **Issues – 1) Target Labels 2) Features**

- **Target Label – The average 20-day trailing difference velocity is positive, and the future 20-day velocity is negative.**
  - **Observe the difference between the past 20 days of "difference returns", or percentage changes in differences**
  - **If at any point there is a positive past and negative future, then we consider it safe to enter into a trading position**

- **Main Features – Historic Mean Diff, Difference, 20-day trailing velocity, Difference Acceleration**
  - **Calculate mean of all past instances of return differences**
  - **Acceleration comes from taking the second derivative of difference movements**

Illustration of Derived Features for Pair AEP, CMS



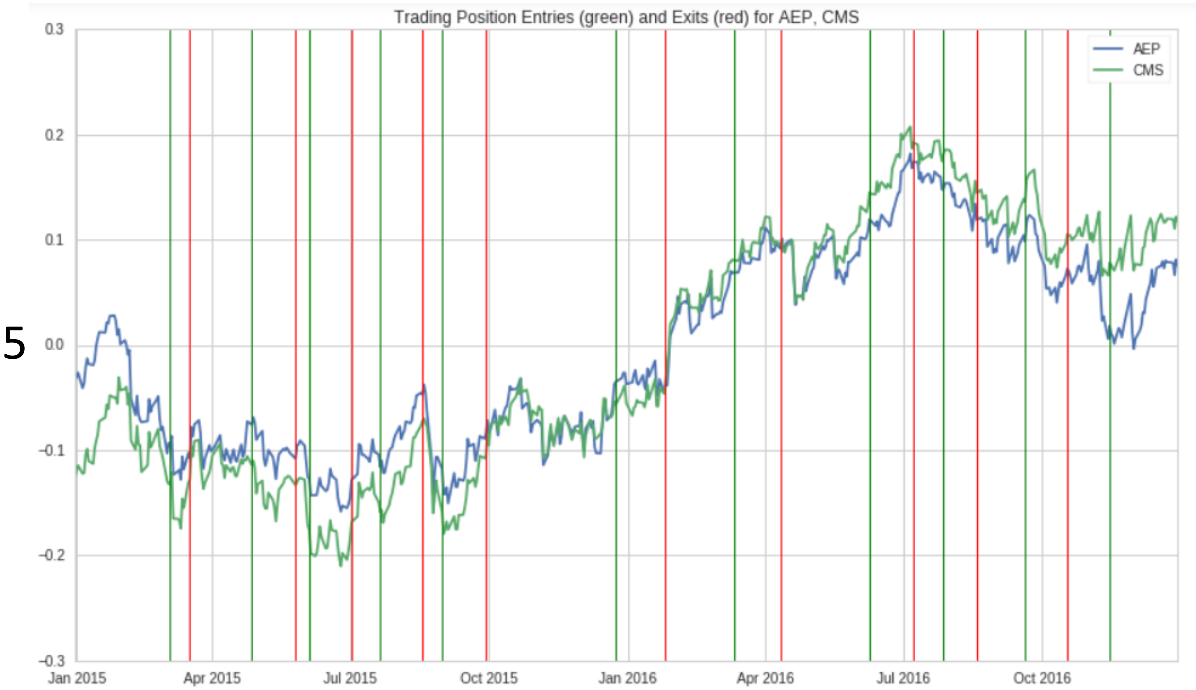
- **DT** was applied to the classification problem with an 80:20 train:test split applied to the valid pairs
- In the test set, **DT** managed to achieve **91.90%** Accuracy in determining "Velocity Reversing" scenarios

Converting Signals Into A Testable Trading Strategy

- Now for a given time series, we are able to sequentially predict whether or not it is safe to open a trade. However, this is far from a testable strategy, since one does not have infinite money to trade

- An algorithm to test was decided with a minimalistic framework. This was applied and averaged over all equity pairs in the test set to obtain evaluation metrics.
  - 1) Check to see if it is valid to open a trade
  - 2) If it is valid, then the strategy assumes that all of one's initial capital is dumped into the trade
  - 3) If the past 20 days have a negative rate of movement, then it is safe to trade.
  - 4) If the return difference fails to half, then the trade is automatically closed after 100 days to cut potential losses

- **Trading Strategy Avg Returns** 1.114
- **Buy and Hold Avg Returns** 1.2196
- **Risk-Free Returns** 1.0066
- **Percentage of Profitable Trades** 0.915
- **Volatility of Returns** 0.04023
- **Worst 10 Trades Avg** -0.00515
- **Single Worst Trade** -0.00882
- **Sharpe Ratio** 2.54
- **Simple Algorithmic Pairs Returns –** 1.0821



CONCLUSIONS

- Machine Learning shows some promise in being able to derive profitable trading strategies, even with relatively simple indicators
- The results, while safe, seem to not generate high amounts of excess returns, and often seem to fall short of buy-and-hold strategies
- Trading based off of a velocity reversal indicator seems superior to pair halving, with higher returns, safer maximum drawdowns, better risk-adjusted returns, and a higher percentage profitability. We note the incredibly low levels for maximum drawdown as well.
- Work remains to be done in automating more of the feature and label engineering process. Numerous other trading strategies also ought to be explored.
- Additional studies could include modifications for the assumptions made. Key flawed assumptions to explore include – full entry into each trade, lack of transaction costs, dividend payments.

REFERENCES

- 1) Chen, Ren, Lu. Machine learning in Pairs Trading Strategies, Stanford University. 2012.
- 2) Van Der Have. Pairs Trading Using Machine Learning: An Empirical Study, Erasmus University. 2017.
- 3) Larkin, Jonathan. Pairs Trading with Machine Learning. Quantopian, 2017