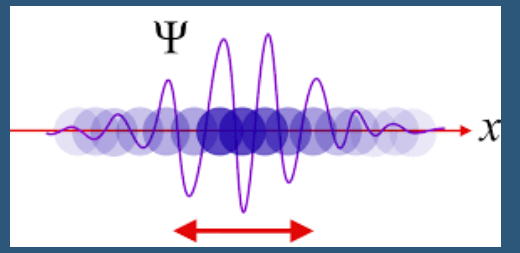
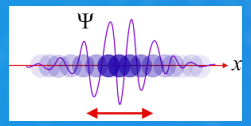


QUANTUM CAPITAL LLC

ENERGY FUND PROSPECTUS

"DRIVING CAPITAL MARKETS THROUGH ALGORITHMIC DESIGN"

CONTENTS



Executive Summary

EXECUTIVE SUMMARY

The Grab

• Quantum Capital merges cutting edge data science research with fundamental commodities trading strategies. Take advantage of the unique skills of Quantum Capital to say goodbye to passive investing and achieve above-market returns. Our team deploys a model-based approach to automated trading so you can generate positive returns in the oil sector no matter if the market is up or down!

Problem/ Opportunity

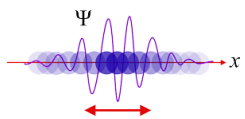
• In 2017, for the first time in 17 years, the hedge fund industry experienced more fund closures than openings – a trend that continued in 2018. The commodities hedge fund space for new entrants is not nearly as crowded as it has been. The shift in closures outpacing openings has been accompanied by a shift in hedge fund trading strategies to favor quantitative approaches. In the eight years leading up to 2017, the market value of quant funds nearly tripled, and they also grew in relative terms to their non-quant peers. Today, while quant shops comprise only 27% of all hedge funds, five of the six largest hedge funds rely on algorithmic trading. Quantum Capital is primed to take advantage of this shift in strategy and opening in the competitive landscape.

Solution/ Product

• Our aim is to develop profitable trading strategies in the commodities space that also have an attractive risk profile. The models developed using machine learning techniques will be the foundation for the strategies. The end will be a multi-strategy investment vehicle with AUM $20 MM. The first strategy will consist of trades in crude oil, and crude oil byproducts futures contracts. The second strategy will target gasoline-specific instruments using proprietary stitching of futures contracts and feature extraction. Quantum Capital provides an on-the-go solution for monitoring trades and performance by way of an online dashboard, and mobile phone application.

Potential Upside

• Our strategy employed in the oil space yielded annualized returns of **26.11%** and **16.8%**, compared to their underlying commodities at *23.12%* and *13.4%* respectively**.** The gasoline strategy yielded an annualized return of **43.9%** relative to its underlying commodity at *23.8%* for the same period. Both strategies out-performed the broader markets considerably, with the S&P 500 Equal Weighted Energy ETF returning 3.4% for the period.



EXECUTIVE SUMMARY

Competition

- Aside from direct competitors in the hedge fund space, the biggest threat to active management is the flow of investment into passive strategies, such as index funds. The top three Oil and Gas ETFs have combined assets of $14.67 BB, but their weighted average YTD return is quite abysmal at 0.47%. The problem with these strategies is that they rely too heavily on the performance of the underlying in oil and gas related companies and are one-directional (long). The energy markets are simply too volatile to yield above-market returns when using a one-directional trading strategy. Our strategy takes advantage of the volatility and focuses more on futures pricing instead of individual company performance.

Execution

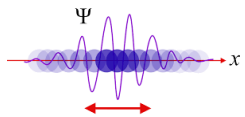
- To execute our strategy, we will download futures contract pricing data, along with data from the Department of Energy. We will then analyze futures contract curves, accounting for contango and backwardation. During our EDA we will also explore various technical indicators, along with correlations. During data preparation we will remove outliers and fill in missing values. The trading strategy will be executed using automated trading algorithms. We will display and track results in a dashboard and mobile application.

Financials

- The team is seeking a commitment for the initial capital raise at $10 MM, per strategy.

The Team

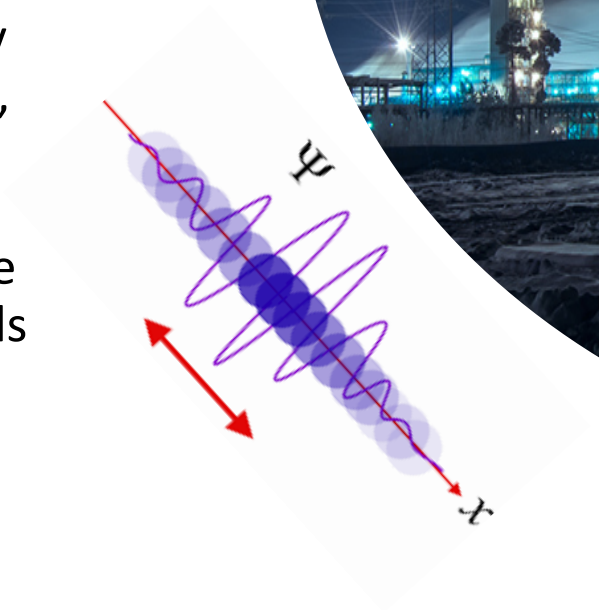
- The team brings a wide range of experience in finance, and analytics. Andrius, Brandon, Josh, and Tate all met at the prestigious MS in Data Science program at Northwestern University, and will all graduate at the end of this year.



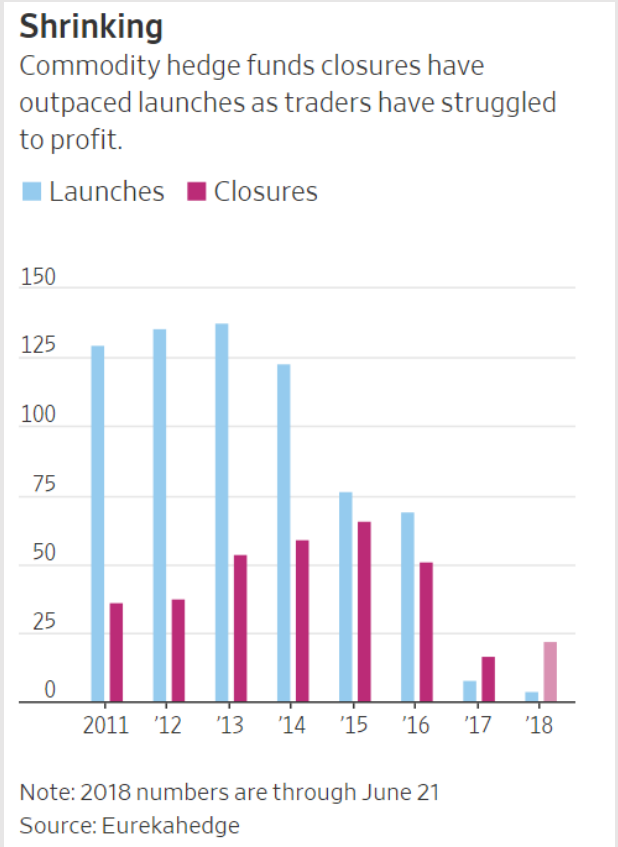

Introduction & Project Goals

COMPANY BACKGROUND

Quantum Capital Management LLC is a new private fund management company in fundamental commodity strategies with a specialization in the oil and energy complex. Quantum Capital is led by a team of data scientists. The investment strategy targets absolute returns with an asymmetric upside, via detailed supply and demand forecasting, fundamental, macro economic and physical market information combined with various technical market indicators to generate fair values, forecasts and trading signals for energy and commodities.



OPPORTUNITY – PROPRIETARY COMMODITY TRADING SPACE IS SHRINKING…



In 2017, closures of commodities hedge funds outnumbered launches for the first time in data going back to 2000, according to data provider Eurekahedge—a trend that has continued into this year.

The reason? Fund managers and traders say, investors who were burned by the severe two-year market rout that started in 2014 aren't rushing back despite prices of commodities, including oil, copper, lumber and cotton, all rebounding to multiyear highs.

Commodity futures offer tremendous upside if one can manage volatility.

COMMODITIES TRADERS ARE INCREASINGLY ADOPTING ALGORITHMS

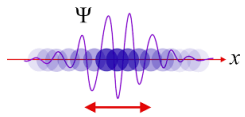
THE REASON

One of the reasons commodities are attracting new players and new strategies is that the markets are not as mature as equities. ***There is greater opportunity in commodities for electronic strategies*** to take advantage of market inefficiencies. Also, with more and more institutional money flowing into commodities, both through exchange-traded funds (ETFs) and listed futures and options, managers are looking for new ways to generate alpha in commodities.

According to estimates, about 15 percent to 20 percent of all futures trading occurs in commodities futures contracts. Within commodities, there are various sub-types, including metals; agricultural products, such as sugar, soybean futures and grains; and energy. Automated trading is being applied to the most highly volatile commodities, ***and active subsets of that tend to be energy futures contracts and specifically crude oil products.***

THE OPPORTUNITY

According to a recent survey, hedge funds implementing algorithms that support trading decision and risk management all in one integrated process that is supported by algorithmic decision making are highly sought by investors. While the overall hedge fund industry has performed poorly since the 2008 – 2009 financial crisis, the bright spot in the industry are quant-based funds.

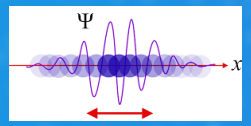


INITIAL GOAL: SEED CAPITAL

In order to establish a solid track record and solicit additional investors via a marketing campaign.

Potential investors in our new fund, will be funds of funds, endowments, pensions, family offices and high-net-worth individuals.

Our commercial ask of you is to green light the project; we need an initial injection of capital to open the fund and build out an analytic and machine learning infrastructure.



THREE CHARACTERISTICS OF OUR FUND

- **Story:** Identify key signals in scenarios that correlate strongly with the probability that the market has continuously mispriced futures prices in these situations, creating opportunities to earn market returns but with significantly less risk.
- **Process:**
  - Extract, track and archive market activity in the energy sector.
  - Apply various data transformation and feature extraction techniques.
  - Fit various models using algorithmic techniques to predict market behavior.
  - Back-test strategy methodology and forecast results.
- **Performance:** Working in a team of four using multiple strategies with $10 million in assets under management, we will meet or exceed an annualized return of 12%.

INVESTMENT STRATEGIES

- Data containing all continuous energy futures contracts, global index data, the open, high, low, and closing prices of select crude oil, natural gas, heating oil, and gasoline suppliers and the United States Energy Information Administration.
- All available energy futures contracts will be used to produce a statistical arbitrage strategy.

INVESTMENT STRATEGIES CONTINUED…

- Additional data from global indexes and Department of Energy will be weaved into the data set.
- Long/Short Trading Strategy will be developed based on price forecast conclusions based on proprietary models.
- Back testing framework will feature artificial intelligence.

OUR METHODOLOGY

Data Collection

- Data warehouse server to archive global index data.
- Repeatable process to download future contracts from the Department of Energy data.
- This will be a bespoke proprietary data set and will be owned by Quantum Capital Management.

Data Analysis & Understanding

- Analyze future contract curves.
- Analyze for contango and backwardation - potential feature development.
- Fit additional feature sets such as Bollinger Bands and Exponential Moving Averages.
- Analyze correlations (Pearson's, Spearman's and Tau's).
- Initial data set was over 4,000 attributes, for the purpose of this project we will be significantly pruning.

Data Preparation

- Outlier Detection.
- Standardization and Normalization for Linear and Neural Network analysis
- Fill missing values
- Output to csv files for further processing

Modeling

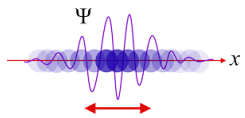
- Neural Network(specifically LSTM) – price activity
- Random Forest for classification of signal direction
- Logistic Regression for signal direction
- Linear Regression for price activity

Evaluation

- Standard model fit parameters
- Build backtesting framework for real world signal and pattern evaluation

Visualization

- Dashboard Development
- Mobile Application Development

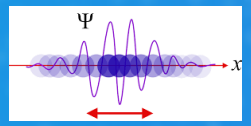


Trade Modeling Process Flow

We will need $22 million in cap ex funds for this project to deliver the initial phase of this project.

- Data Collection
- Data Analysis & Understanding
  - Statistical Summary
  - Time Series
  - Pattern Matching
  - Date Type Handling
- Data Preparation
  - Outer Detection
  - Normalization
  - Missing Value Imputation
- Modeling
  - Neural Network
  - Random Forests
  - Logistic Regression
  - Linear Regression
- Evaluation
  - ROC Tables
  - Lift Table
  - Confusion Matrix
  - R-Squared
  - MSE
- Visualization
  - Dashboard Development
  - Mobile Application Development

This will be an iterative process.

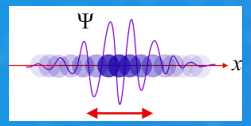


Data Preparation & Tools

TOOL / TECHNOLOGY OVERVIEW

- Data Set
  - Department of Energy Weekly Report
  - Open, High, Low, Close(OHLC) Energy Future
  - Global Index Data
  - Proprietary Stitching of Future Contracts
  - Feature Creation
- Jupyter Notebook
- Python
  - NumPy, SciPy, Pandas
  - Matplotlib
  - XGBoost
- RStudio
- Vertica Database Platform

"Data Science is a street fight…."



BESPOKE DATA SET CREATION

We will be creating our own data sets for modeling. Up until recently it was rather difficult and expensive to obtain consistent futures data across exchanges in frequently updated manner. However, certain new platforms make acquiring this data possible.

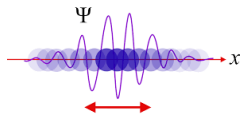
Data stitching and munging still needs to occur to make the data relevant for modeling purposes. Our data is downloaded and manipulated with the final version stored in a data warehouse.

They are the following data sets;

1. Historical futures contracts for a wide variety of energy related continuous contracts
    - *The main difficulty with trying to generate a continuous contract from the underlying contracts with varying deliveries is that the contracts do not often trade at the same prices. This will be explained in greater detail in subsequent slides.*

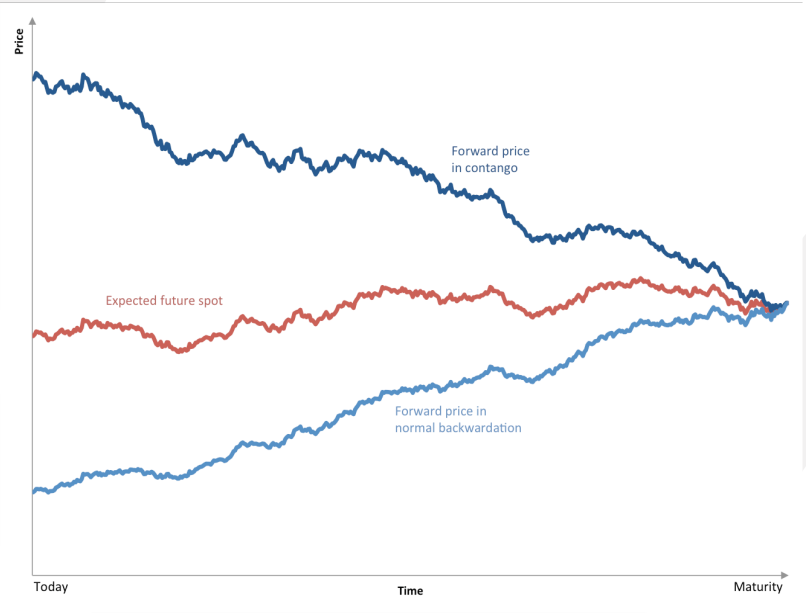
2. Weekly Department of Energy(EIA) reports summarizing the storage of products and refinery outputs.

3. Global Indexes

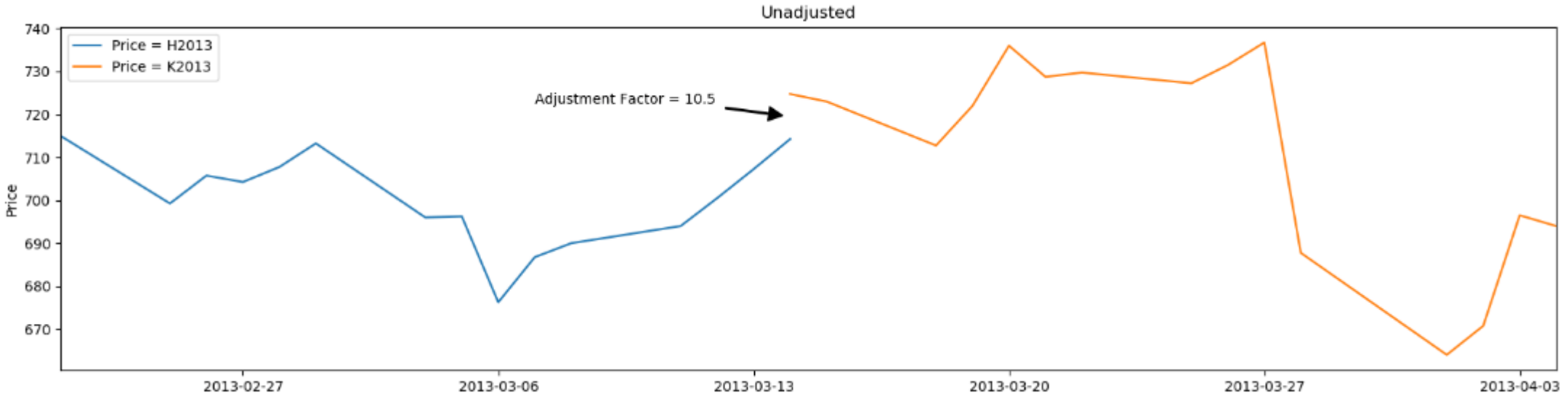


FORMING A CONTINUOUS FUTURES CONTRACT

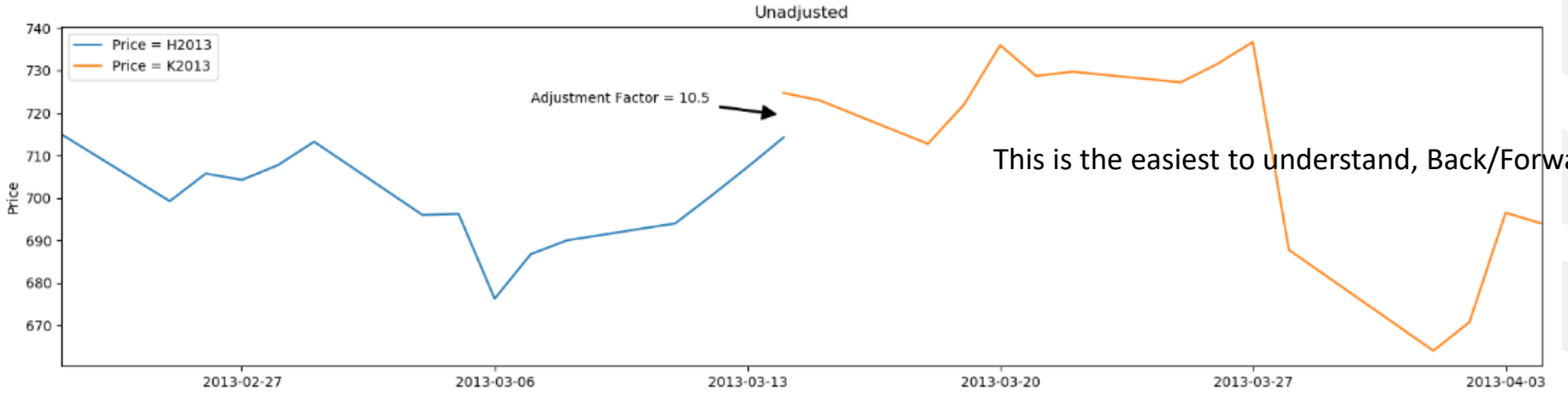
The main difficulty with trying to generate a continuous contract from the underlying contracts with varying deliveries is that the contracts do not often trade at the same prices. Thus, situations arise where they do not provide a smooth splice from one to the next. This is due to contango and backwardation effects.



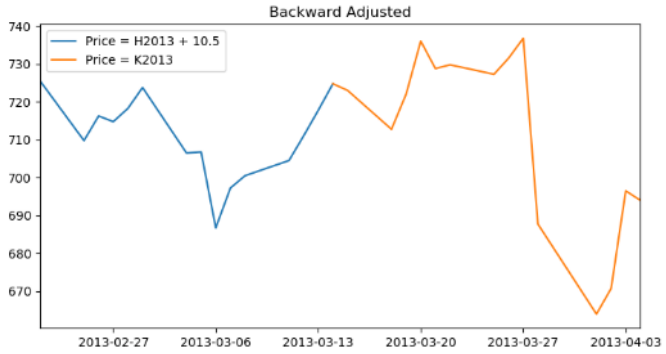
Here is another view of a March delivery contract rolling over to April – Price versus Time

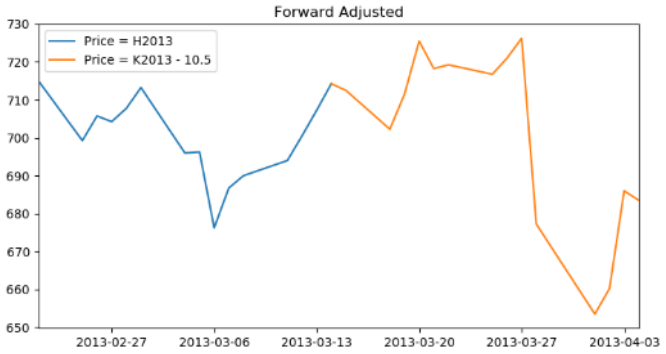


FORMING A CONTINUOUS FUTURES CONTRACT(CONTINUED)



This is the easiest to understand, Back/Forward Adjustment

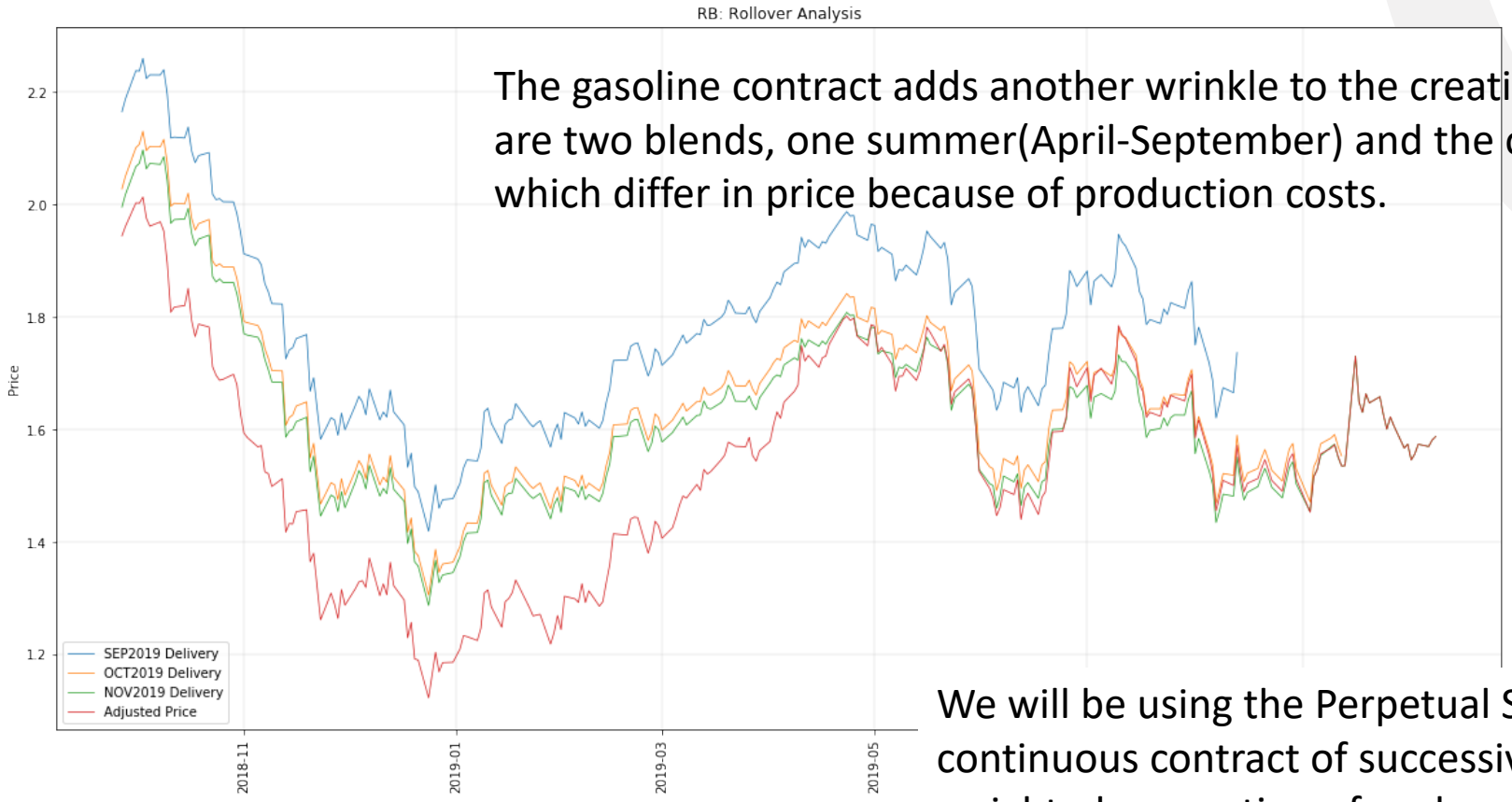




Backward/Forward Adjusting Example

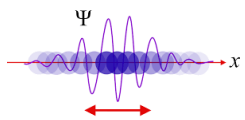
The key problem with these methods includes the introduction of a trend bias, which will introduce a large drift to the prices. This can lead to negative data for sufficiently historical contracts. In addition there is a loss of the *relative* price differences due to an absolute shift in values. This means that returns are complicated to calculate (or just plain incorrect).

FORMING A CONTINUOUS FUTURES CONTRACT(CONTINUED)

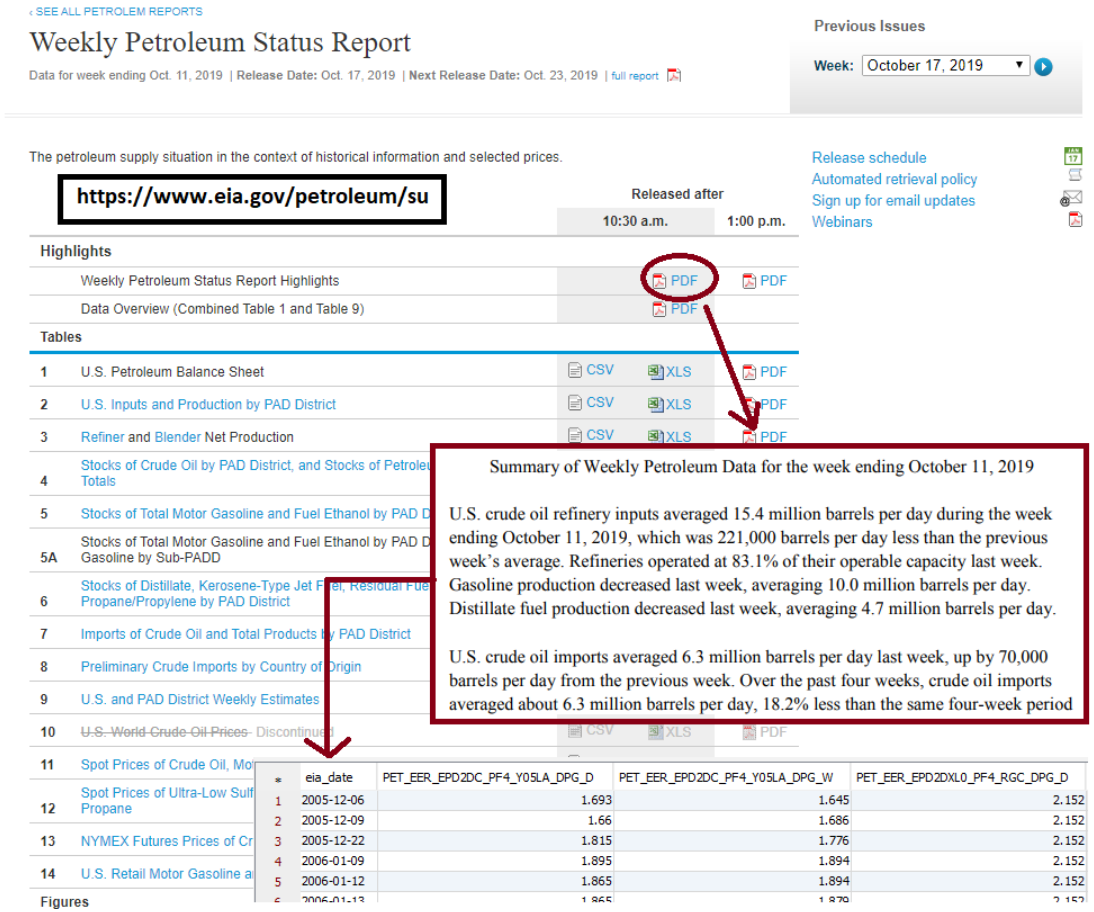


The gasoline contract adds another wrinkle to the creation of a continuous contract. There are two blends, one summer(April-September) and the other winter(October – March) which differ in price because of production costs.

We will be using the Perpetual Series approach to create a continuous contract of successive contracts by taking a linearly weighted proportion of each contract over a change in open trading interest to ensure a smoother transition between each. We will concentrate on implementing the perpetual series method as this is most appropriate for backtesting.



DEPARTMENT OF ENERGY WEEKLY REPORT DATA



Downloaded and stitched together weekly petroleum data from EIA website into usable format for modeling.

GLOBAL INDEX-CONTINUOUS ENERGY FUTURE CONTRACTS-FEATURE DEVELOPMENT

SPX	S&P 500 Index					10/22/2019
LAST:		CHANGE:	OPEN:	HIGH:	ASK:	VOLUME:
		↓ 10.73	3,011	3,015	0	0
2,996		CHANGE(%):	PREV:	LOW:	BID:	OPEN INT:
		0.36	3,007	2,996	0	0

CHART

Period: End of Day Display

RECENT END OF DAY PRICES

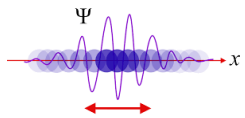
Date	Open	High	Low	Close	Volume	Open Interest
10/22/19	3,011	3,015	2,996	2,996	0	0
10/21/19	2,996	3,007	2,995	3,007	0	0
10/18/19	2,997	3,000	2,976	2,986	0	0

OHLC prices for data point

A snapshot of a data points that will be fed into our modeling process. In subsequent slides a data dictionary and feature definition list will be provided.

Feature development, such as twenty data moving average, Bollinger bands, etc. .

Some of the data points stored in the database.

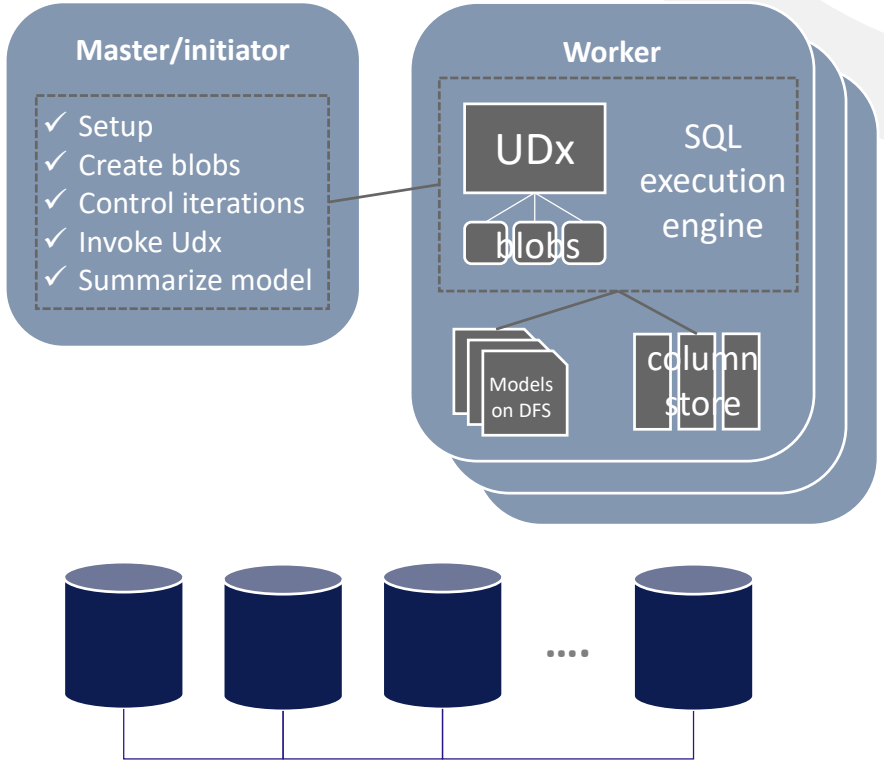



Model Analysis
• Explore Data Details
• Feature Development
• Initial Model Exploration
MICRO FOCUS

EXPLORE DATA DETAILS

Initial data exploration will be performed using a couple different platforms. The data is stored in a distributed data platform called Vertica Analytics Database. From Jupyter Notebooks client we will interface with the machine learning interface that Vertica provides via a library.

- **In Memory processing**
- **Optimized parallelism**
- **Distributed model storage**
- **State-of-the-art Distributed Machine Learning Algorithms**



MPP distributed cluster

Linear Regression
Logistic Regression
K-Means
Naïve Bayes
SVM
Random Forest

Custom Model (C++, Java, Python, R)

Train, predict and evaluate

Model management

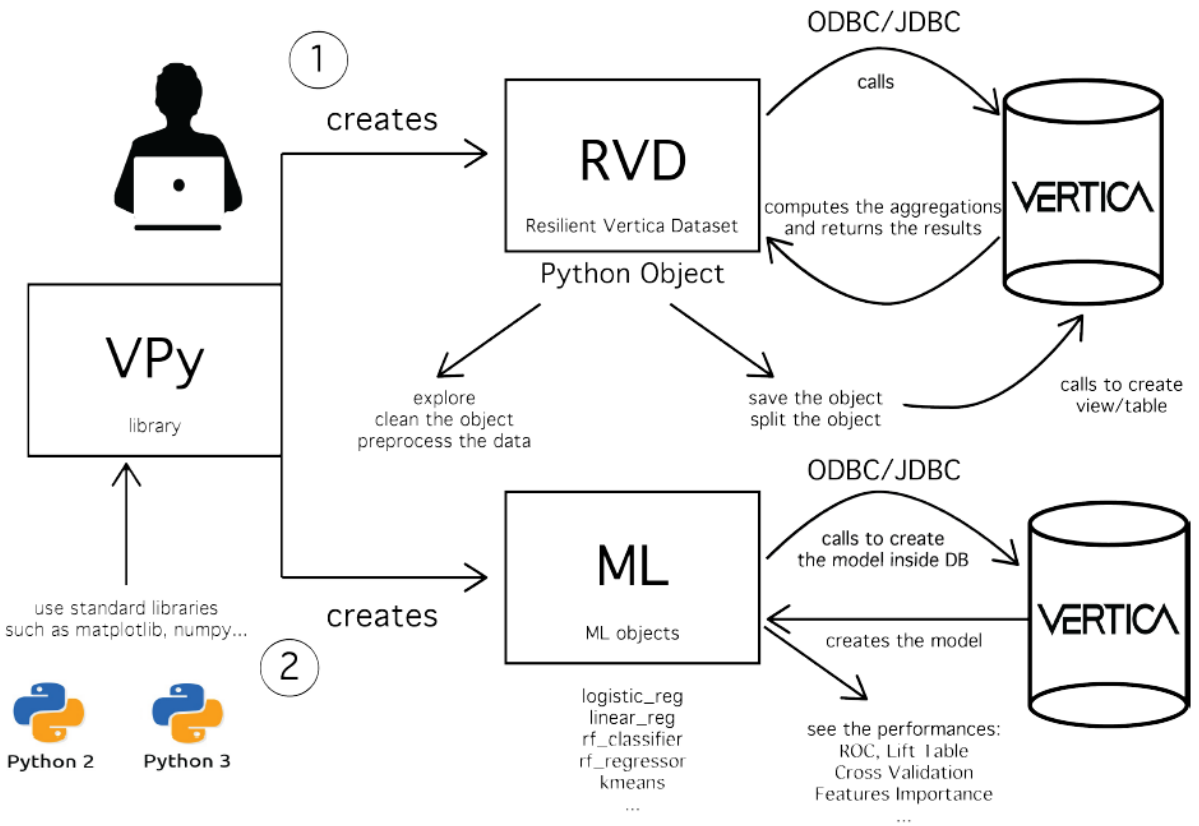
Data preparation

EXPLORE DATA DETAILS

The library abstracts and streams data science functionality to manipulate large data sets stored in Vertica by taking advantage of what Vertica is known for – speed and built-in analytics and machine learning capabilities.

From data preparation to model scoring and deployment, Vertica supports the entire machine learning process. Users can prepare data with functions for normalization, outlier detection, sampling, imbalanced data processing, missing value imputation and more. Machine learning models can be created, trained and tested on massive data sets, and then evaluated with model-level statistics such as ROC tables and confusion matrices.

Using the Vertica Analytics Platform to support the machine learning and exploratory data process via a python interface. The computation and fitting is done leveraging a massively parallelized environment with the results sent back to the Jupyter Notebook.



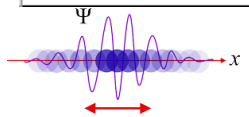
DATA SET
COMPLETE LIST OF RAW ATTRIBUTES

SYMBOL	DESCRIPTION
CL	Crude Oil WTI
SC	Crude Oil Brent
HO	ULSD NY Harbor (Heating Oil - Ultra Low Sulfer)
RB	Gasoline RBOB
HP	Natural Gas(F)
MT	Gulf Sour Crude Oil
NG	Natural Gas
QA	Crude Oil Brent(F)
QG	Natural Gas Mini
VE	Crude Oil VIX

GLOBAL INDEX	
symbol	description
DSEN	DJ US OIL&GAS
DSOG	DJ US OILGAS
DSOI	DJ US OILEQPSRV
DSOL	DJ US INTGOILGAS
DSOQ	DJ US OILEQPSRV
OILBR	Crude Oil Brent NYMEX
OILSW	Crude Oil Light Sweet NYMEX
OIV	CBOE/NYMEX WTI Volatility Index
OSX	Phlx Euro Style Oil Svc Index
OVX	CBOE CRUDE OIL VOLATILITY INDEX
SG3I	S&P GSCI Crude Oil Index
SG4A	S&P GSCI Crude Oil Index Total Return
SG5I	S&P GSCI Heating Oil Index
SPX	S&P 500 INDEX
SS1J	S&P 500 EQUAL WEIGHTED Energy [Sector]
UOI	US Oil Iopv

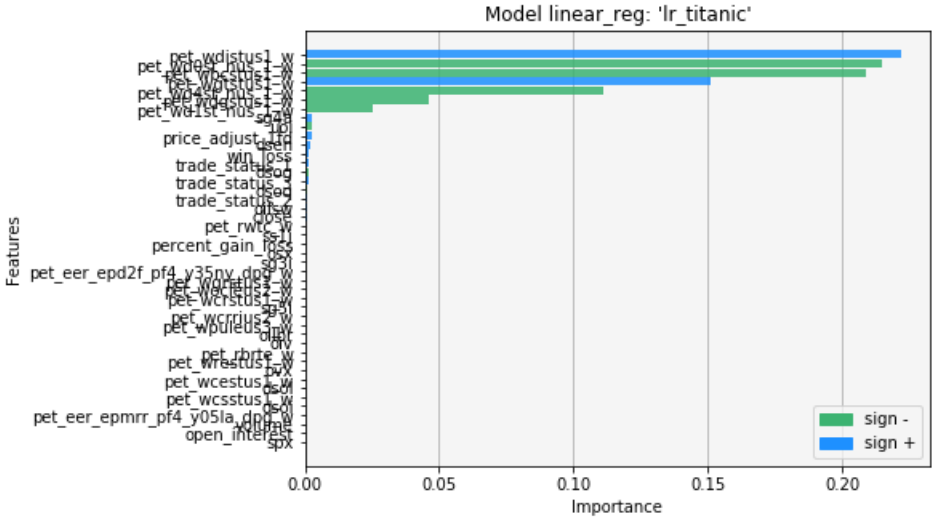
Department of Energy	
code_id	code_description
PET_WRESTUS1_W	U.S. Ending Stocks of Residual Fuel Oil, Weekly
PET_WPULEUS3_W	U.S. Percent Utilization of Refinery Operable Capacity, Weekly
PET_WOCLEUS2_W	U. S. Operable Crude Oil Distillation Capacity, Weekly
PET_WGTSTUS1_W	U.S. Ending Stocks of Total Gasoline, Weekly
PET_WGRSTUS1_W	U.S. Ending Stocks of Reformulated Motor Gasoline, Weekly
PET_WG4ST_NUS_1_W	U.S. Ending Stocks of Conventional Motor Gasoline, Weekly
PET_WDISTUS1_W	U.S. Ending Stocks of Distillate Fuel Oil, Weekly
PET_WDGSTUS1_W	U.S. Ending Stocks of Distillate Fuel Oil, Greater Than 500 ppm Sulfur, Weekly
PET_WD1ST_NUS_1_W	U.S. Ending Stocks of Distillate Fuel Oil, Greater than 15 to 500 ppm Sulfur, Weekly
PET_WD0ST_NUS_1_W	U.S. Ending Stocks of Distillate Fuel Oil, 0 to 15 ppm Sulfur, Weekly
PET_WCSSTUS1_W	U.S. Ending Stocks of Crude Oil in SPR, Weekly
PET_WCRSTUS1_W	U.S. Ending Stocks of Crude Oil, Weekly
PET_WCRRIUS2_W	U.S. Refiner Net Input of Crude Oil, Weekly
PET_WCESTUS1_W	U.S. Ending Stocks excluding SPR of Crude Oil, Weekly
PET_WBCSTUS1_W	U.S. Ending Stocks of Gasoline Blending Components, Weekly
PET_RWTC_W	Cushing, OK WTI Spot Price FOB, Weekly
PET_RBRTE_W	Europe Brent Spot Price FOB, Weekly
PET_EER_EPMRR_PF4_Y05LA_DPG_W	Los Angeles Reformulated RBOB Regular Gasoline Spot Price, Weekly
PET_EER_EPD2F_PF4_Y35NY_DPG_W	New York Harbor No. 2 Heating Oil Spot Price FOB, Weekly

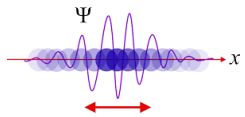
In total, for the statistical arbitrage portfolio, there are 86 predictors being used for modelling.



INITIAL VISUAL OF THE DATA SET

We examine the correlation between all the attributes as the first pass of analysis. We can clearly see clusters of attributes that are highly correlated. We quickly call the linear regression function and produce an initial model to examine the statistics. We form a visual graph of the most important features seen below;

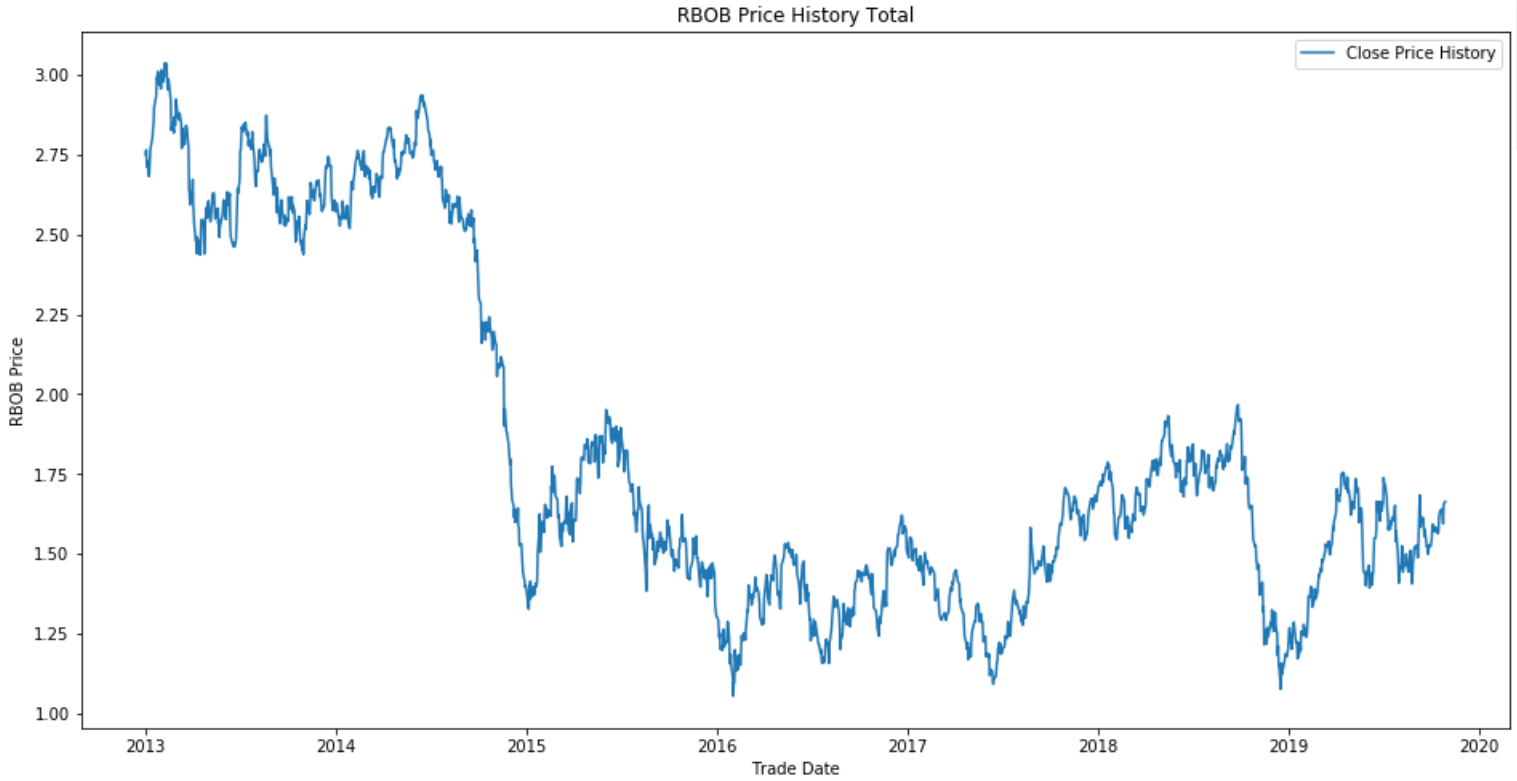




INITIAL LINEAR REGRESSION AND FEATURE DEVELOPMENT

The initial investigation will comprise of discovering a model to price the RBOB(gasoline) front month contract. However, please note, the objective is not the accuracy or best fitting model, but the best strategy to monetize this information.

Taking a look at the price action of the RBOB front month contract, we can see quite a few different patterns and perhaps linear regression might not be the best model, it is a good one to start with in determining next steps in the learning process.

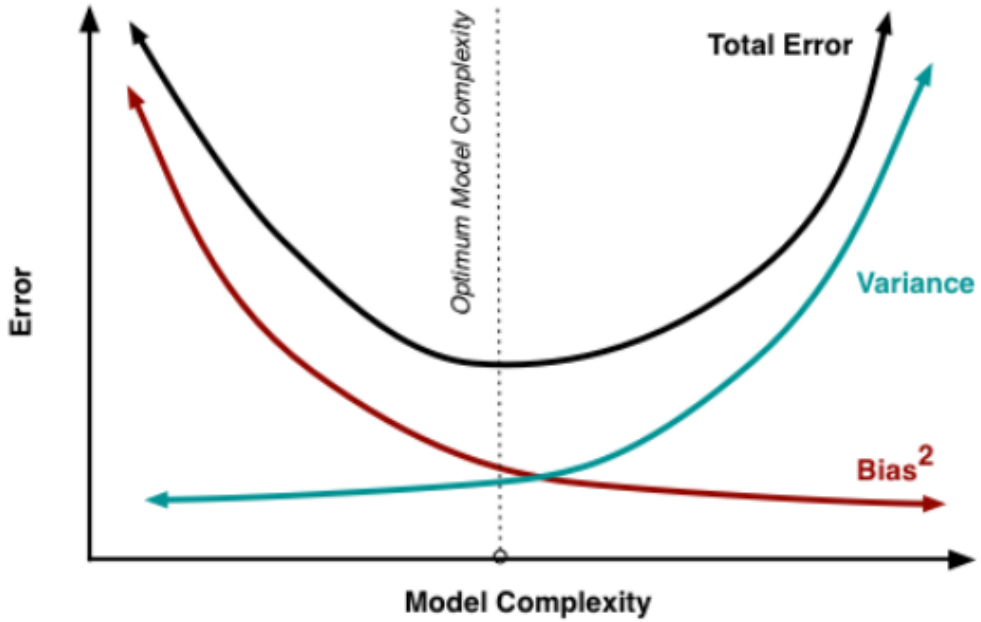


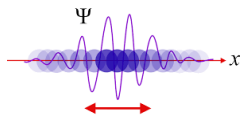
LINEAR REGRESSION – THE ONE MOST UNDERSTAND…

The most basic machine learning algorithm that can be implemented on this data is linear regression. The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable. The equation for linear regression can be written as:
Here, x1, x2,….xn represent the independent variables while the coefficients θ1, θ2, …. θn represent the weights.

$$Y = \theta_1 X_1 + \theta_2 X_2 + \ldots \theta_n X_n$$

Here, we have Y as our dependent variable (future price of front month of gasoline), X's are the independent variables and all thetas are the coefficients. Coefficients are basically the weights assigned to the features, based on their importance. When we have a high dimensional data set, it would be highly inefficient to use all the variables since some of them might be imparting redundant information. We would need to select the right set of variables which give us an accurate model as well as are able to explain the dependent variable well. There are multiple ways to select the right set of variables for the model. First among them would be the business understanding and domain knowledge. For instance while predicting sales we know that marketing efforts should impact positively towards sales and is an important feature in your model. We should also take care that the variables we're selecting should not be correlated among themselves.

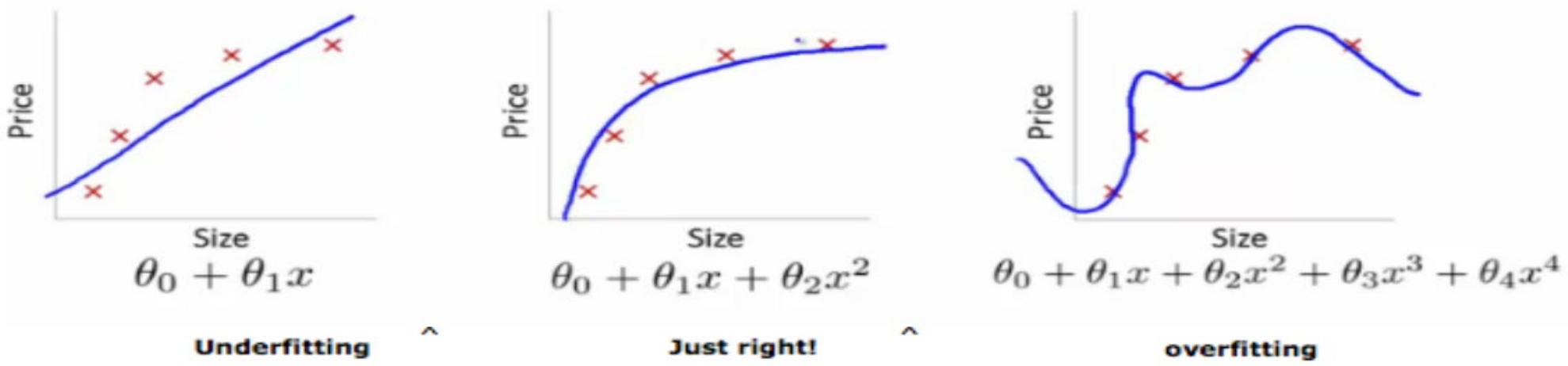




SELECTING THE RIGHT FEATURES FOR YOUR MODEL

As we add more and more parameters to our model, its complexity increases, which results in increasing variance and decreasing bias, i.e., overfitting. So we need to find out one optimum point in our model where the decrease in bias is equal to increase in variance. In practice, there is no analytical way to find this point. So how to deal with high variance or high bias? To overcome under fitting or high bias, we can basically add new parameters to our model so that the model complexity increases, and thus reducing high bias. Now, how can we overcome Overfitting for a regression model? Basically there are two methods to overcome overfitting;

- Reduce the model complexity
- Regularization

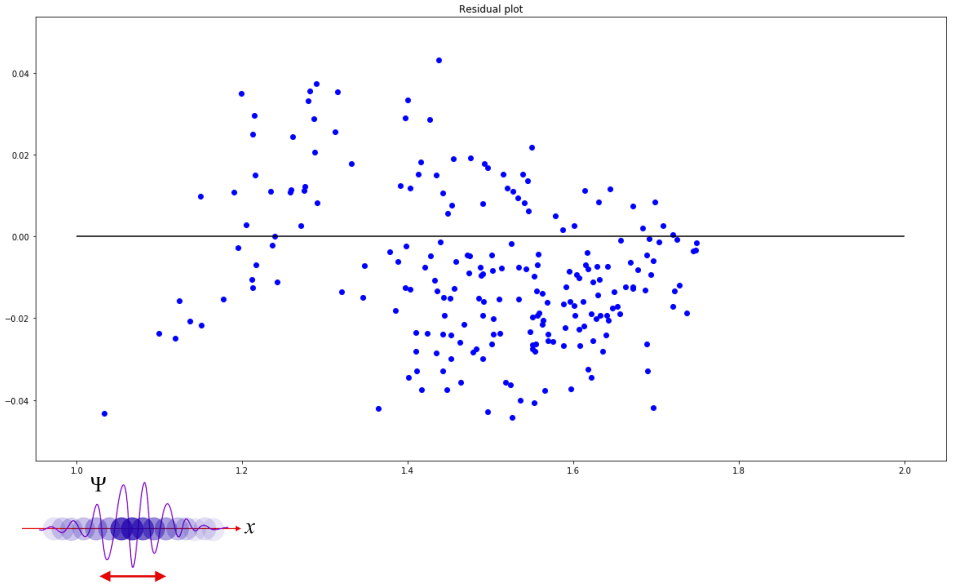


LINEAR REGRESSION MODEL SUMMARY

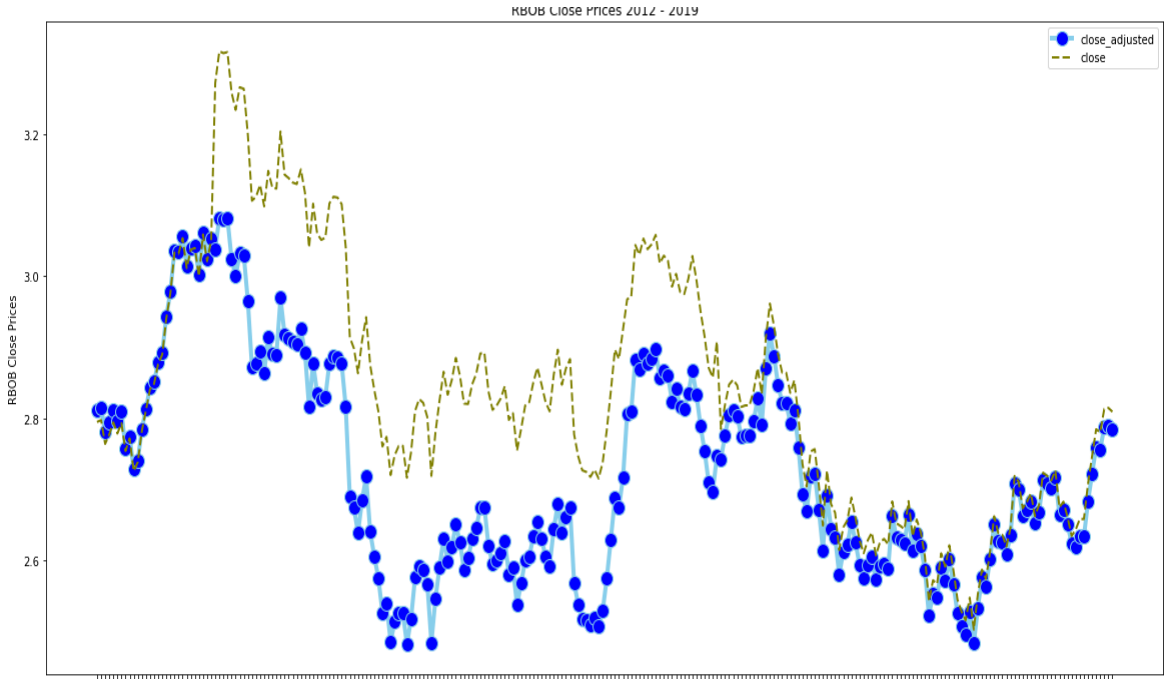
We analyzed several linear models; Linear Regression, Polynomial, Ridge and Lasso.
All these models had a RSquared >98% and MSE ~0.00.

We have discussed Linear regression and Polynomial introduces a more controlled fit depending on the degree put into the model. Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso go with l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values of the coefficients. It's a shrinkage towards zero using an absolute value (l1 penalty) rather than a sum of squares(l2 penalty). Ridge regression can't zero coefficients. You either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the coefficients of collinear variables.

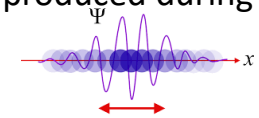
Error residual looked good…



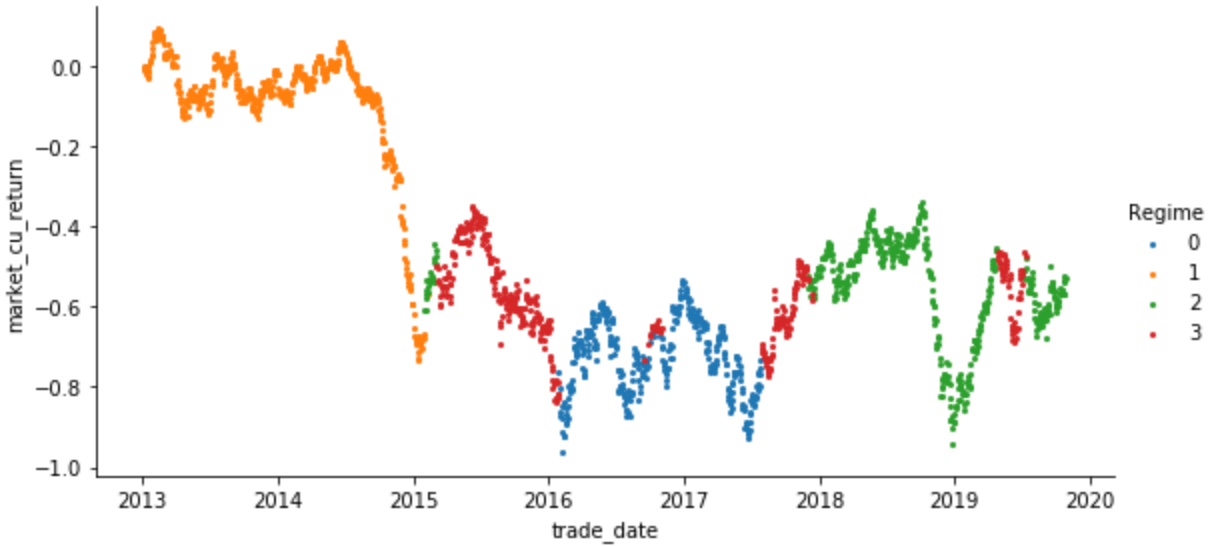
INITIAL FEATURE CREATION – GAUSSIAN MIXTURE MODELS



Examining the data set, we actually see an additional pattern start to emerge. The front future contract price of gasoline(in fact most energy sector instruments) is highly volatile. While k-means clustering might not be the right way of viewing this volatility clustering, we will take a look at Gaussian mixture models (GMMs), which can be viewed as an extension of the ideas behind *k*-means, but can also be a powerful tool for estimation beyond simple clustering. Note the graph above is smoothing out the close price, due to two different blends of gasoline produced during the year.



A Gaussian mixture model (GMM) attempts to find a mixture of multi-dimensional Gaussian probability distributions that best model any input dataset. In the simplest case, GMMs can be used for finding clusters in the same manner as *k*-means. Though GMM is often categorized as a clustering algorithm, fundamentally it is an algorithm for *density estimation*. That is to say, the result of a GMM fit to some data is technically not a clustering model, but a generative probabilistic model describing the distribution of the data. Below, we break down the volatility of the pricing action and create "regimes" of volatility in the pricing action.

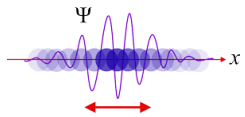


Therefore we will be including these volatility regimes in our next modeling effort; SVM

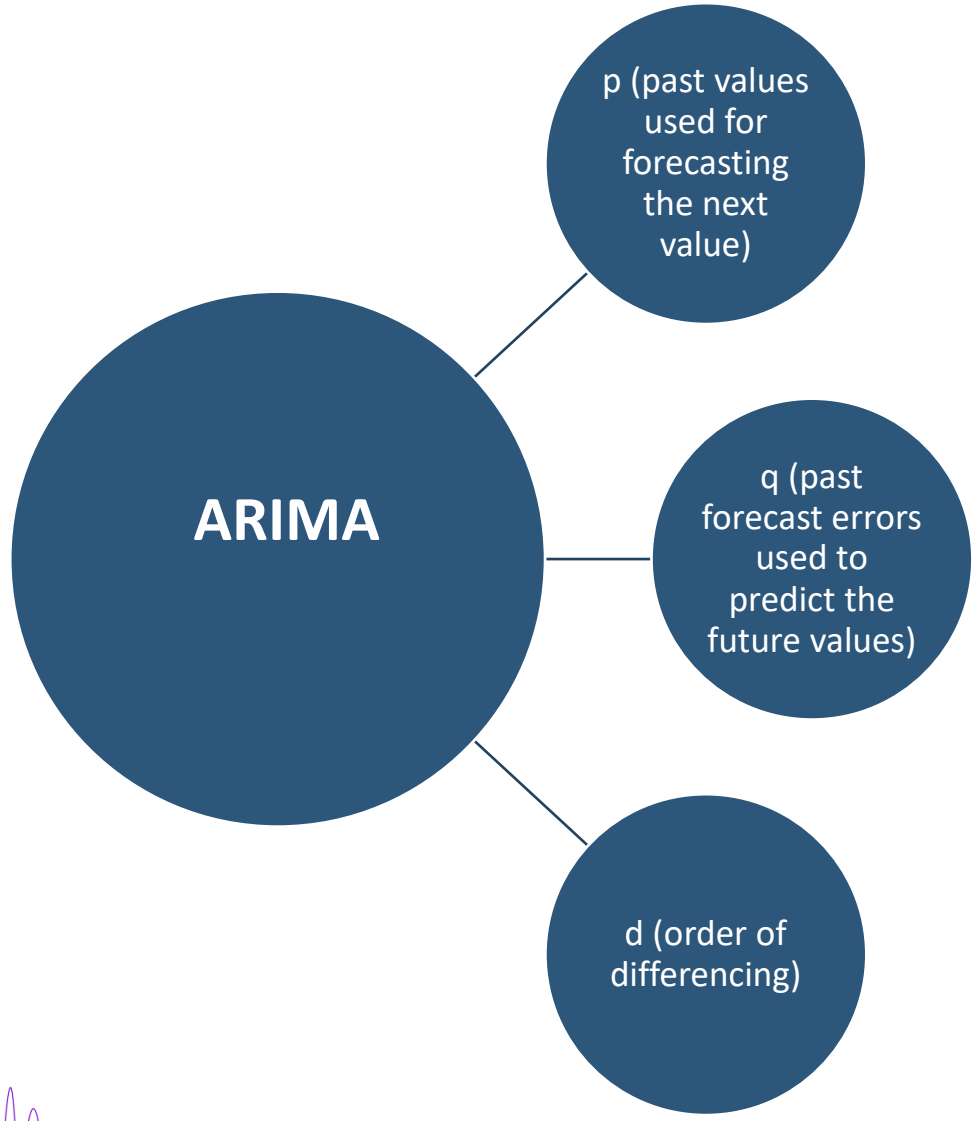
ARIMA

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a form of regression analysis whose goal is to predict future values by examining the *differences between values* in a series instead of through actual values. It can be better understood by looking at its three parts:

- Autoregression (AR) - By regressing a variable on past values of itself, AR causes autocorrelations to gradually decay. Assuming recent events have greater influence than former events on the next outcome, the AR portion applies a weight to each of the past observations based on how recent they are and applies those values forecasts future outcomes.
- Integrated (I) - To reduce seasonality and make a time series stationary, data values are replaced by the difference between the current values and the previous values. This *differencing* eliminates the seasonality.
- Moving Average (MA) - To remove the random movements from a time series the Moving Average portion of ARIMA uses error terms from previous values to predict future observations. MA uses a fixed window and weights that are relative to the time making them more volatile and responsive to current events.



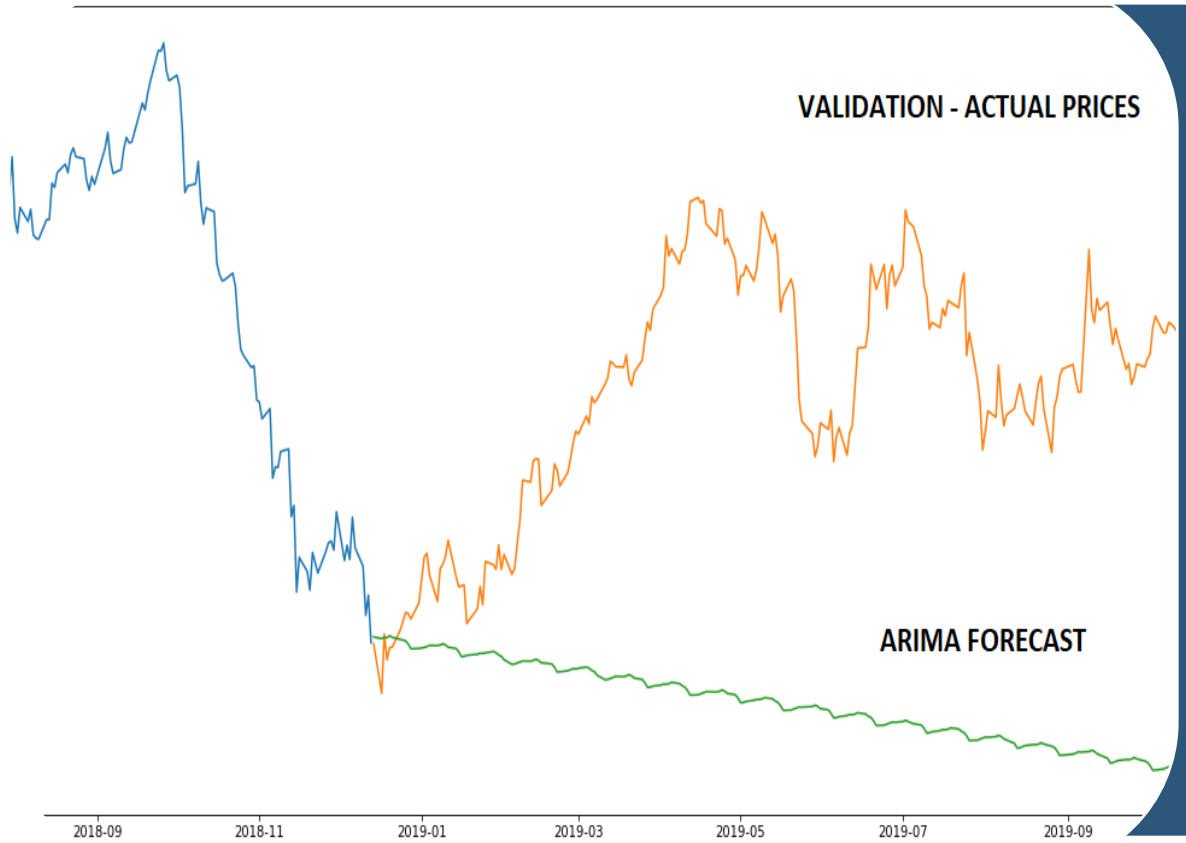
ARIMA PARAMETERS



ARIMA Model Assumptions Include:

- Previous timepoints influence behavior of present timepoints
- Data is free from anomalies
- Model parameters and error term are constant
- Series is stationary

ARIMA



Parameter tuning for ARIMA consumes a lot of time so auto ARIMA was used, which automatically selects the best combination of (p,q,d) that provides the least error.

Note: while the graph indicates this is a poor strategy, domain expertise and more advanced types of this modeling technique will make up a good part of our trading strategy. More on this in the next section.

GENERALIZED AUTOREGRESSIVE CONDITIONAL HETEROSCEDASTICITY (GARCH) MODEL

GARCH, with roots in econometrics, is an approach for estimating volatility in financial markets.

GARCH models are particularly useful with financial instruments, such as commodities futures, that have ever changing volatility. Oil and gasoline are particularly volatile as they can have periods of relative stability followed by periods of extreme volatility due to, for example, geopolitical turmoil. Once the turmoil subsides, volatility may return to normal levels, or after a period result in even greater volatility. Heteroscedasticity describes the aforementioned pattern of variation in volatility of an error term, or variable, in a predictive model. Where there is heteroscedasticity, clustering tends to persist versus linear patterns. GARCH helps to capture this variation in volatility, offering an improvement over simple linear regression.

GARCH CONTINUED…

3 Steps of GARCH Models

1 – estimate best-fitting autoregressive model

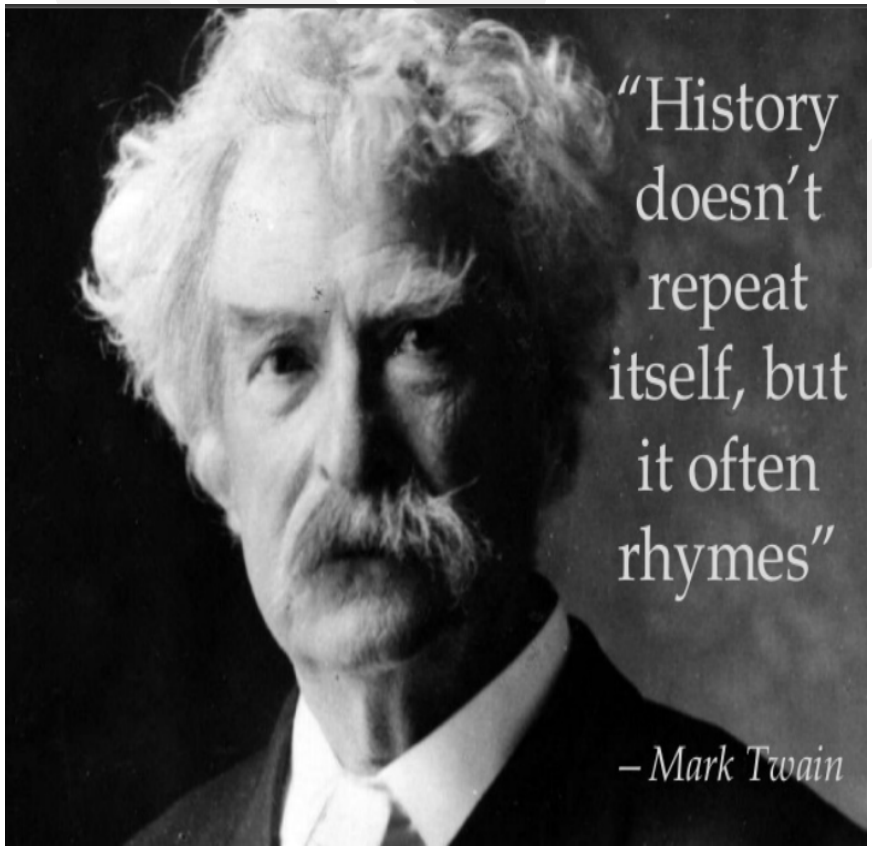
GARCH models are autoregressive: this means they predict future values based on past observations

2 – compute autocorrelations of the error term

Autocorrelation measures the degree of similarity between a given time series and a lagged version of the same series over sequential time periods

3 – test for significance

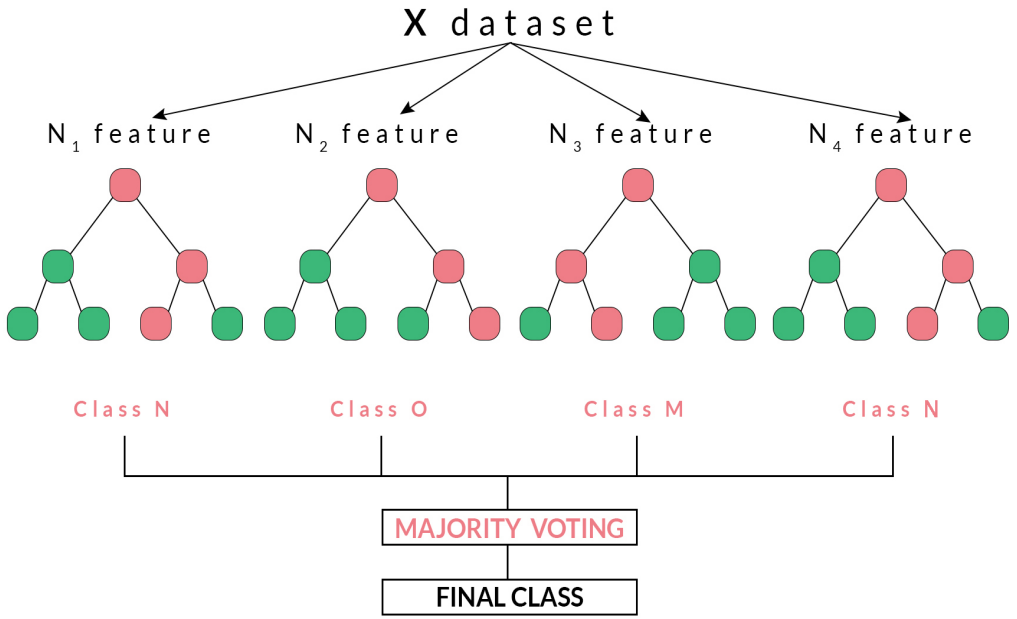
Testing to determine the extent to which the predicted values are explained by the lagged values

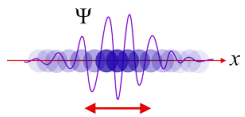


RANDOM FOREST ALGORITHM IN TRADING

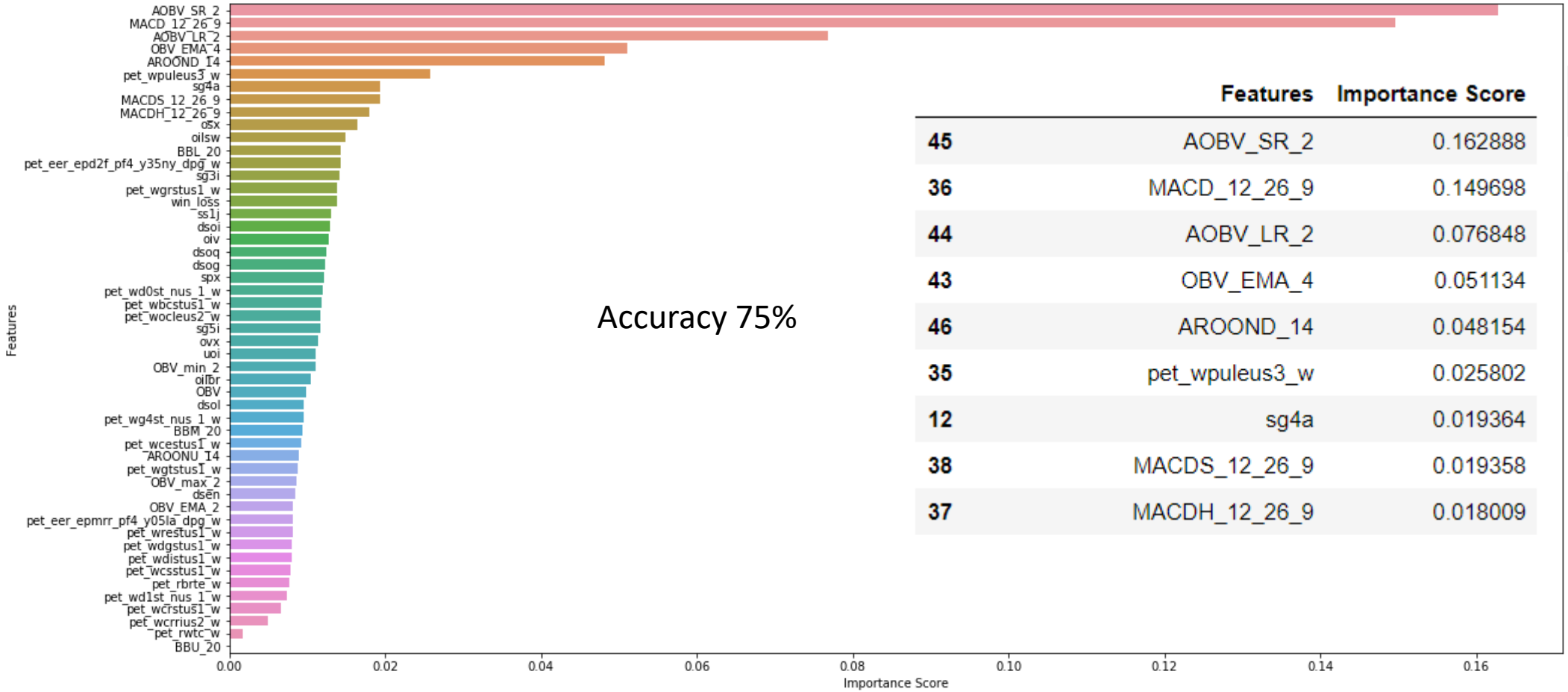
Random forest is a supervised classification machine learning algorithm which uses ensemble method. Simply put, a random forest is made up of numerous decision trees and helps to tackle the problem of overfitting in decision trees. These decision trees are randomly constructed by selecting random features from the given dataset. Random forest arrives at a decision or prediction based on the maximum number of votes received from the decision trees. The outcome which is arrived at, for a maximum number of times through the numerous decision trees is considered as the final outcome by the random forest. Random forests are based on ensemble learning techniques. Ensemble, simply means a group or a collection, which in this case, is a collection of decision trees, together called as random forest. The accuracy of ensemble models is better than the accuracy of individual models due to the fact that it compiles the results from the individual models and provides a final outcome.

Features are selected randomly using a method known as bootstrap aggregating or bagging. From the set of features available in the dataset, a number of training subsets are created by choosing random features with replacement. What this means is that one feature may be repeated in different training subsets at the same time. For example, our dataset contains a number of features and subsets of 5 features are to be selected to construct different decision trees then these 5 features will be selected randomly and any feature can be a part of more than one subset. This ensures randomness, making the correlation between the trees less, thus overcoming the problem of overfitting. Once the features are selected, the trees are constructed based on the best split. Each tree gives an output which is considered as a 'vote' from that tree to the given output. The output which receives the maximum 'votes' is chosen by the random forest as the final output/result or in case of continuous variables, the average of all the outputs is considered as the final output.





RANDOM FOREST ALGORITHM RESULTS



Accuracy 75%

	Features	Importance Score
45	AOBV_SR_2	0.162888
36	MACD_12_26_9	0.149698
44	AOBV_LR_2	0.076848
43	OBV_EMA_4	0.051134
46	AROOND_14	0.048154
35	pet_wpuleus3_w	0.025802
12	sg4a	0.019364
38	MACDS_12_26_9	0.019358
37	MACDH_12_26_9	0.018009

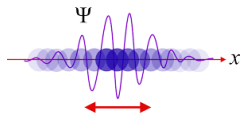
FUTURE CONTRACT PREDICTION WITH XGBOOST

We are going to change the approach future contract price prediction as a classification problem where we will try to predict whether future contract, on the n+5 day, will go up or down(5% winner), using historical future contract data. We will enrich the data to use technical indicators as features for our problem. We have a very good data set for this type of modeling since

A couple considerations before we start modeling the dataset that we need to think about before starting;

Data Set and Feature Creation
Initially we will only be looking at global index and government energy report summary numbers. We may need to create features.

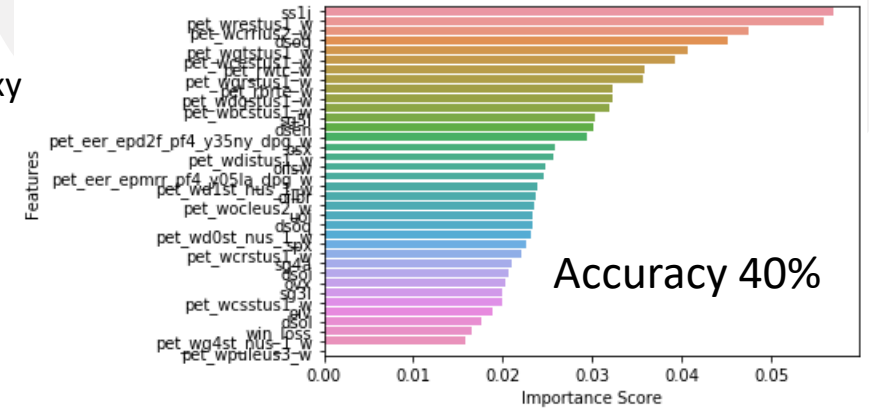
Class Imbalance
Class imbalance can be a nuisance while training and also give a false feeling of high accuracy. Fortunately, after checking the data set, it is a fairly even distribution(50%/50%) between winners and losers. If this wasn't the case, we would need subsample the data corpus.



XGBOOST RESULTS

We are not shuffling data before splitting as we really want to predict prices in future by training our model on past data. Caution should be taken while training and evaluating time series data as there can be a high chance of overfitting (and don't use cross-validation for evaluation).

Our initial accuracy on just the dataset of index and energy department data was around 40%. Technical indicators are good predictors for future contract prices and hence serves as a good proxy for features in a parametric model (as XGBoost and other similar algorithms don't capture time dependency of features), though there are better practices that can be tried. However, this approach serves as a good beginning for stock prediction nonetheless.

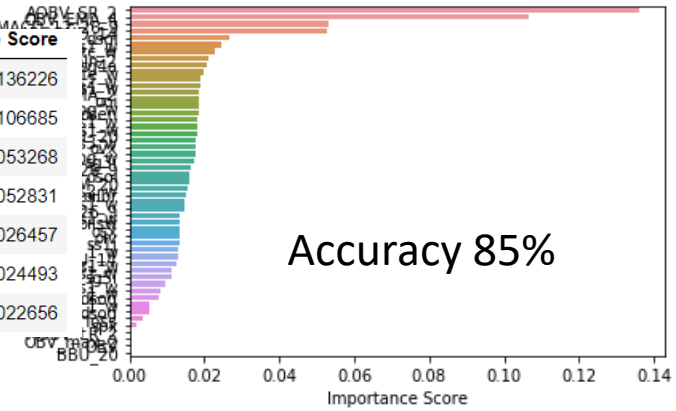


Therefore we created a number of ocillators and technical indicators based on price. Below is the list of features added to the dataset;

- Moving Average Convergence Divergence (MACD),
- Hull Exponential Moving Average (HMA),
- Bollinger Bands (BBANDS),
- On-Balance Volume (OBV),
- Aroon Oscillator (AROON)

So, we are able to get some performance with best accuracy of 85%. Since, forecasting future contract prices is quite difficult, framing it as a 2-class classification problem is a good way to start and we can then go more granular by turning it into a multi-class problem by defining a label for different ranges of price movement and if we succeed at that then we can finally move to price prediction (a regression problem).

	Features	Importance Score
45	AOBV_SR_2	0.136226
43	OBV_EMA_4	0.106685
36	MACD_12_26_9	0.053268
46	AROOND_14	0.052831
4	dsol	0.026457
18	pet_wcestus1_w	0.024493
31	pet_rwtc_w	0.022656



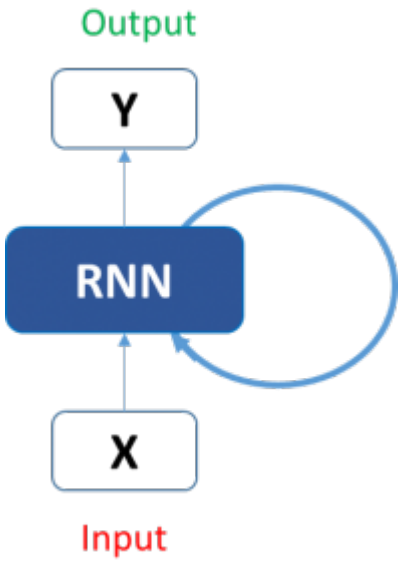
ESSENTIALS OF DEEP LEARNING : INTRODUCTION TO LONG SHORT TERM MEMORY(LSTM)

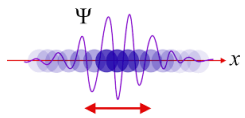
Sequence prediction problems have been around for a long time. They are considered as one of the hardest problems to solve in the data science industry. With the recent breakthroughs that have been happening in data science, it is found that for almost all of these sequence prediction problems, Long short Term Memory networks, LSTMs have been observed as the most effective solution.

LSTMs have an edge over conventional feed-forward neural networks and RNN in many ways. This is because of their property of selectively remembering patterns for long durations of time. The purpose of this article is to explain LSTM and enable you to use it in real life problems.

Take our example of sequential data, which is the future contract price data. A simple machine learning model or an Artificial Neural Network may learn to predict the future prices based on a number of features: the volume of the future contract, the opening value etc. While the price of the contract depends on these features, it is also largely dependent on the contract values in the previous days. In fact for a trader, these values in the previous days (or the trend) is one major deciding factor for predictions. In the conventional feed-forward neural networks, all test cases are considered to be independent. That is when fitting the model for a particular day, there is no consideration for the stock prices on the previous days. This dependency on time is achieved via Recurrent Neural Networks.

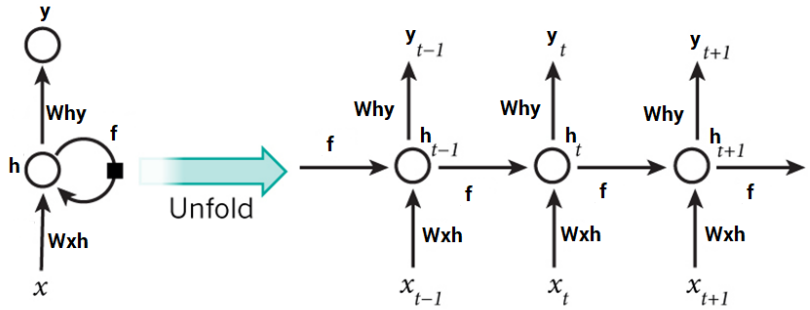
A typical RNN looks like:





A LOOK INTO RECURRENT NEURAL NETWORKS (RNN)

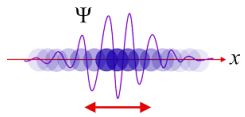
Now it is easier for us to visualize how these networks are considering the trend of contract prices, before predicting the contract prices for today. Here every prediction at time t (h_t) is dependent on all previous predictions and the information learned from them.
RNNs can solve our purpose of sequence handling to a great extent but not entirely. RNNs are great when it comes to short contexts, but in order to be able to build a story and remember it, we need our models to be able to understand and remember the context behind the sequences, just like a human brain. This is not possible with a simple RNN.



LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:

The input gate: The input gate adds information to the cell state
The forget gate: It removes the information that is no longer required by the model
The output gate: Output Gate at LSTM selects the information to be shown as output

For now, let us implement LSTM as a black box and check it's performance on our particular data.

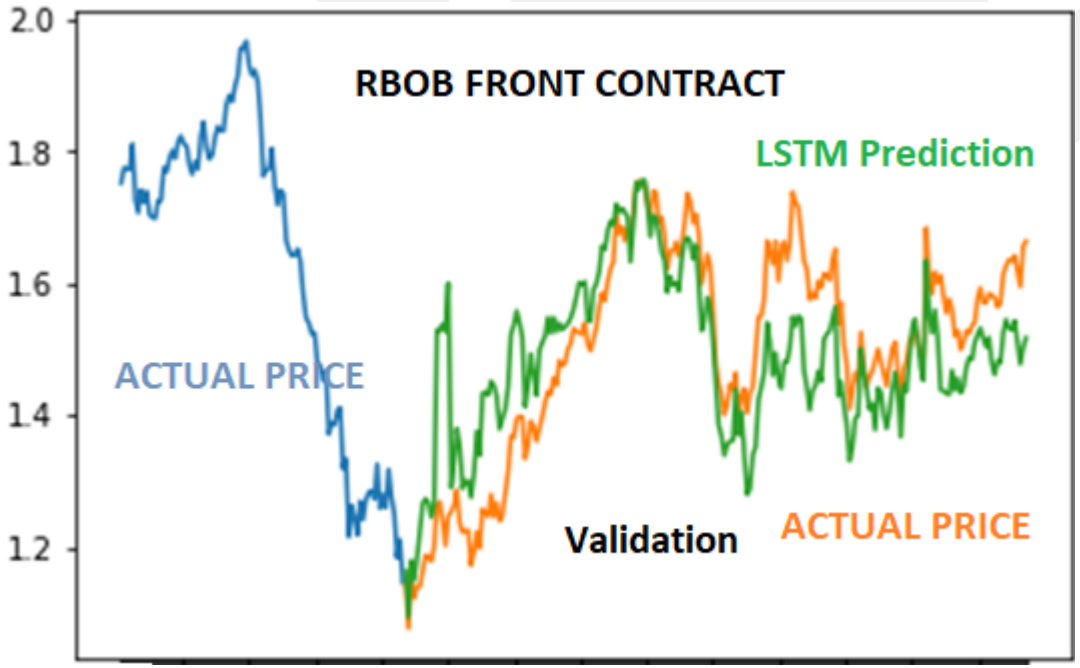


DEEP LEARNING RESULTS

This modeling was done only on the close price being the variable. Much further work needs to be done to develop a more reasonable model. However, the prediction just on one variable is remarkably good, given all the other modeling techniques up to now have been poor and needed additional attributes or feature development.

One important thing to note. The model's hyperparameters are extremely sensitive to the results you obtain. Below I listed some of the most critical hyperparameters;

1. The learning rate of the optimizer
2. Number of layers and the number of hidden units in each layer
3. The optimizer. We found Adam to perform the best

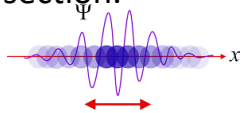


INITIAL MODELING CONCLUSTIONS

Just a quick summary of algorithms fitted on the RBOB front month contract;

1. Linear Regression, Polynomial, Ridge and Lasso(Price Prediction/Direction).
    - All these models had a RSquared >98% and MSE ~0.00.
    - NOTE: performed SVM – Regression and results were similar, not discussed in above material.
2. Gaussian Mixed Mixture (Price Prediction/Direction).
    - Did a great job of clustering and feature development for volatility segments in the price movements
3. ARIMA (Price Prediction/Direction).
    - As the graph shows, very poor in predicting actual prices and direction. HOWEVER, in the next section, using domain expertise and data transformation, moving averages plays an important part in strategy development.
4. Random Forest(Classification – Winner or Loser).
    - 75% Accuracy
    - Needed additional feature development to get better accuracy, initial results we in the 40% range.
5. XGBoost (Classification – Winner or Loser).
    - 85% Accuracy – again needed feature development.
6. Deep Learning (Price Prediction/Direction).
    - Very initial stages of development, however, very promising in terms of predicting direction of prices.

Here are several takeaways of this analysis;
Stock price/movement prediction is an extremely difficult task. Personally I don't think any of the future contract price prediction models out there shouldn't be taken for granted and blindly rely on them. However models might be able to predict contract price movement correctly most of the time, but not always. Several models shows predictions curves that perfectly overlaps the true future contract prices(linear regression, for example). This can be replicated with a simple averaging technique and in practice it's useless. A more sensible thing to do is predicting the future contract price movements. This is the basis of the analysis in the next section.



Market Stylized Facts & Modeling Details

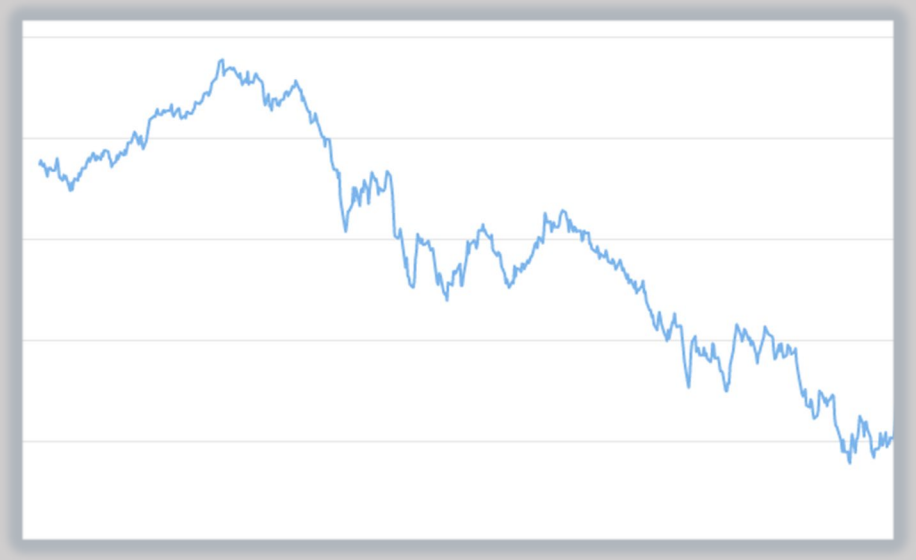
INVESTMENT STRATEGY CONCEPTS

- The commodities daily closing prices are converted to a return series in order to **normalize** the movements of the assets and represent them on an equal scale.
- We then model each series **independently**, such that we capture as much of the uniqueness of that series as possible.
- A strategy is then devised by modeling the movements of that series, looking for **statistically significant** signals that exhibit a clear deviation from its normal behavior.
- These signals define our entry and exit points for each position. The holding period for each strategy is also unique to the series, based upon the implied **volatility** of the asset.

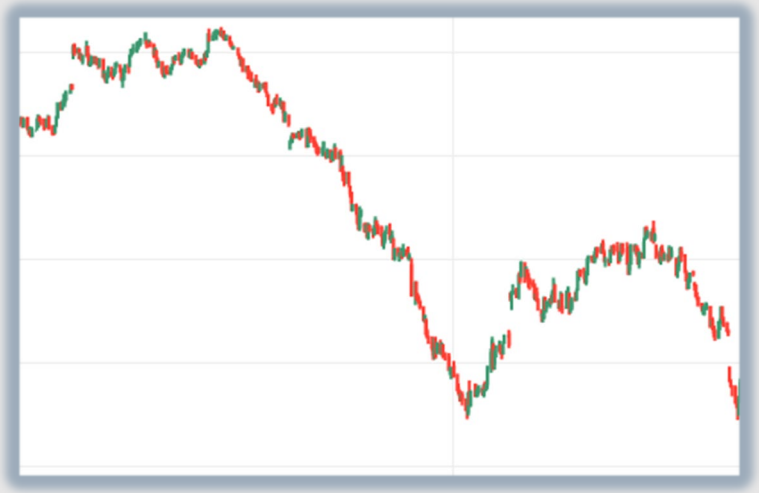
RISK PROFILE Underlying Assets

EXTREME VOLATILITY IS OBSERVED IN THESE SERIES.

BENCHMARKS HAVE A SIMILAR RISK PROFILE.

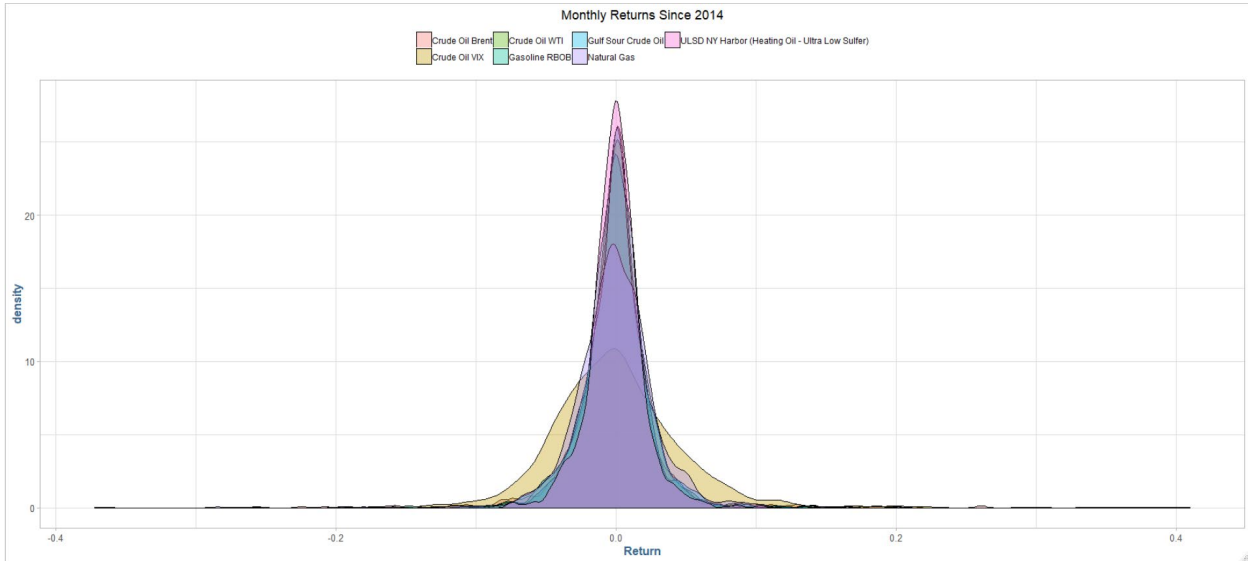


OUR OPPORTUNITY IS TO EXPLOIT THIS BEHAVIOR.



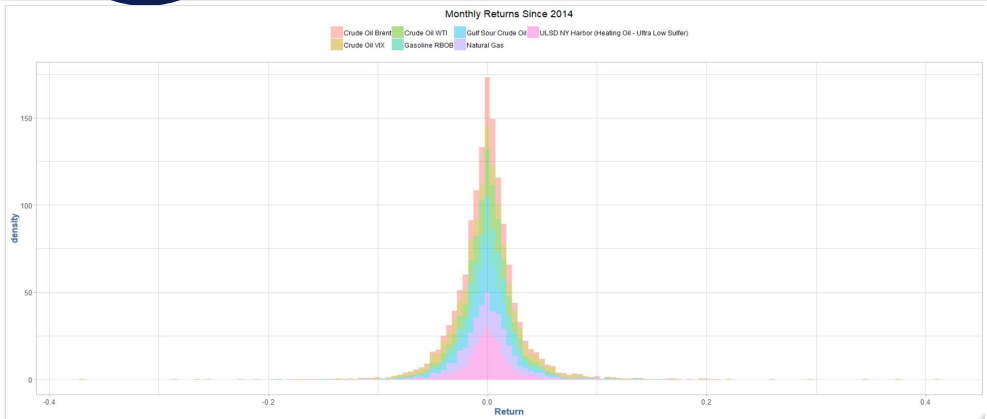
RISK PROFILES: UNDERLYING ASSETS CONTINUED

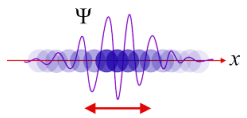
EXTREMELY VOLATILITY



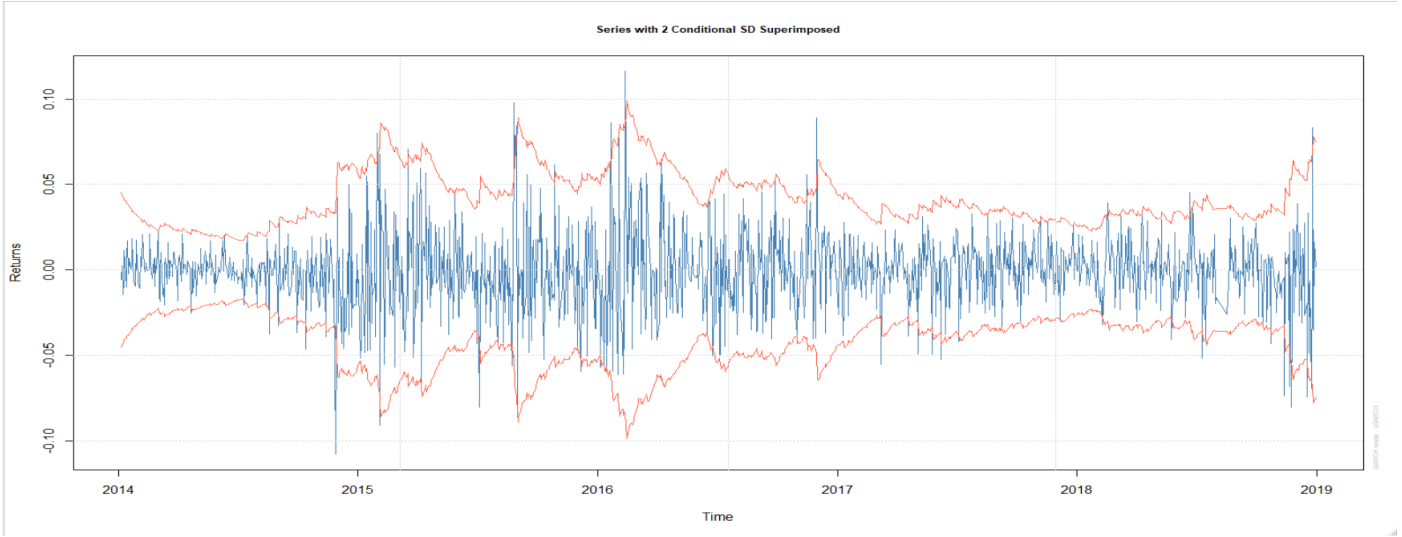
Each commodity has a unique volatility profile, one that we will look to model and exploit for our trading strategies.

UNIQUE CLUSTERING PROFILE

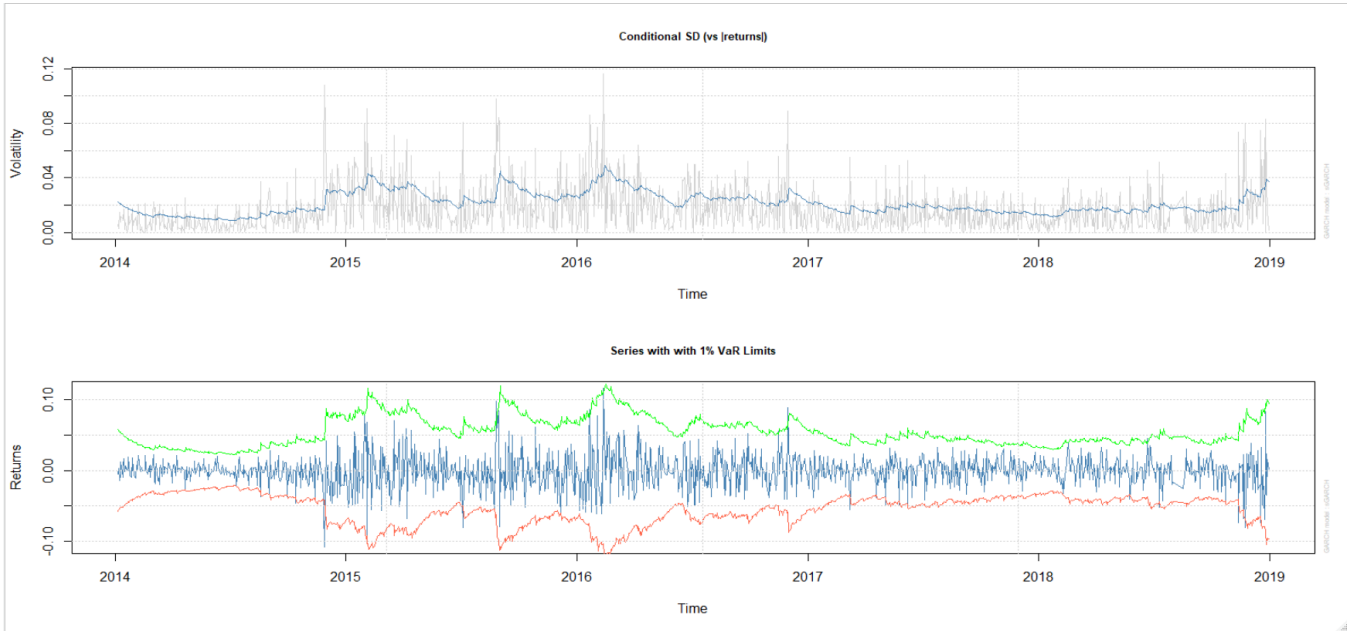




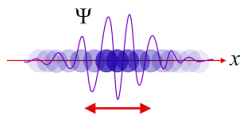
MODELING VOLATILITY



The returns, superimposed with a 2 standard deviation fit, reveal how extreme volatility is in this series.

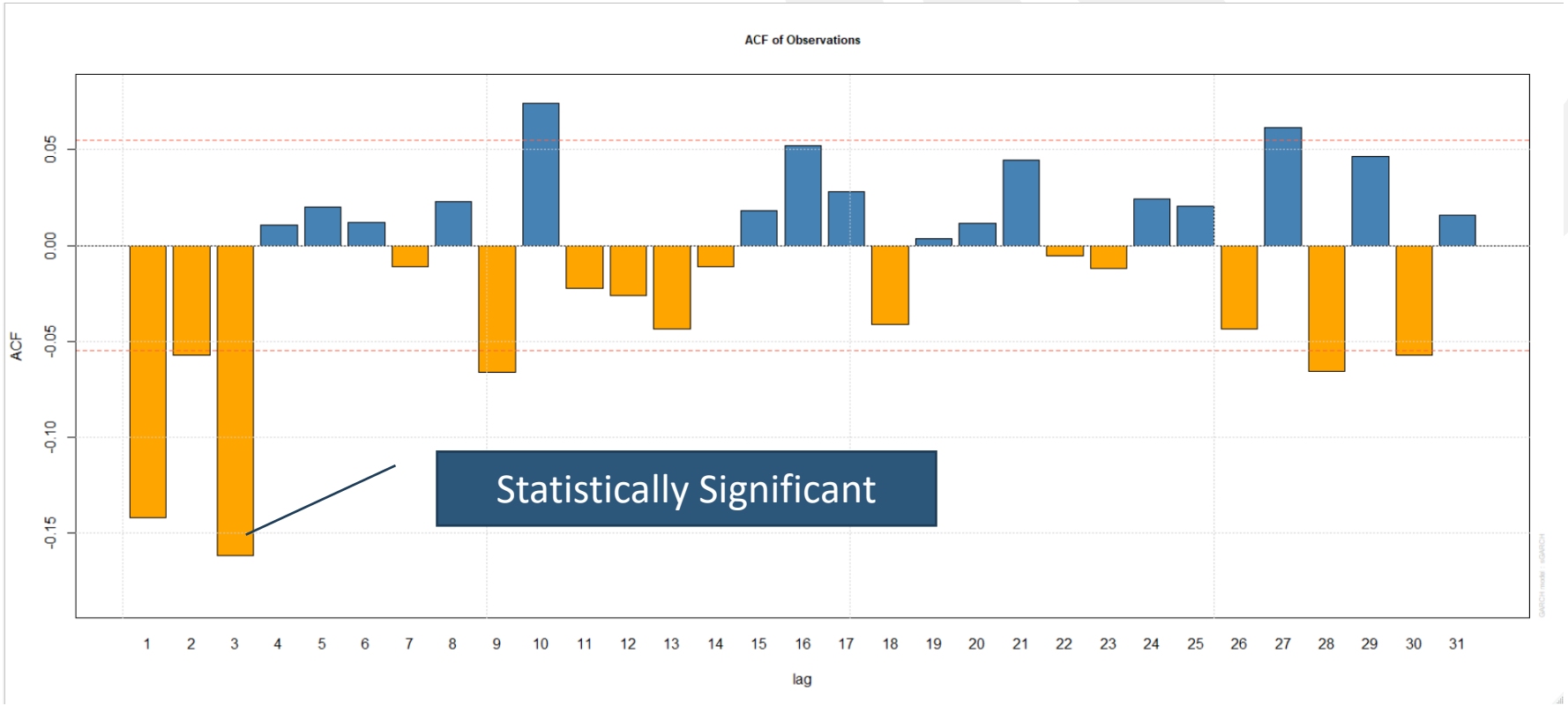


An advantage can be gained from the volatility if a model is built that considers the explosive momentum seen in large periods of the instrument's prices.

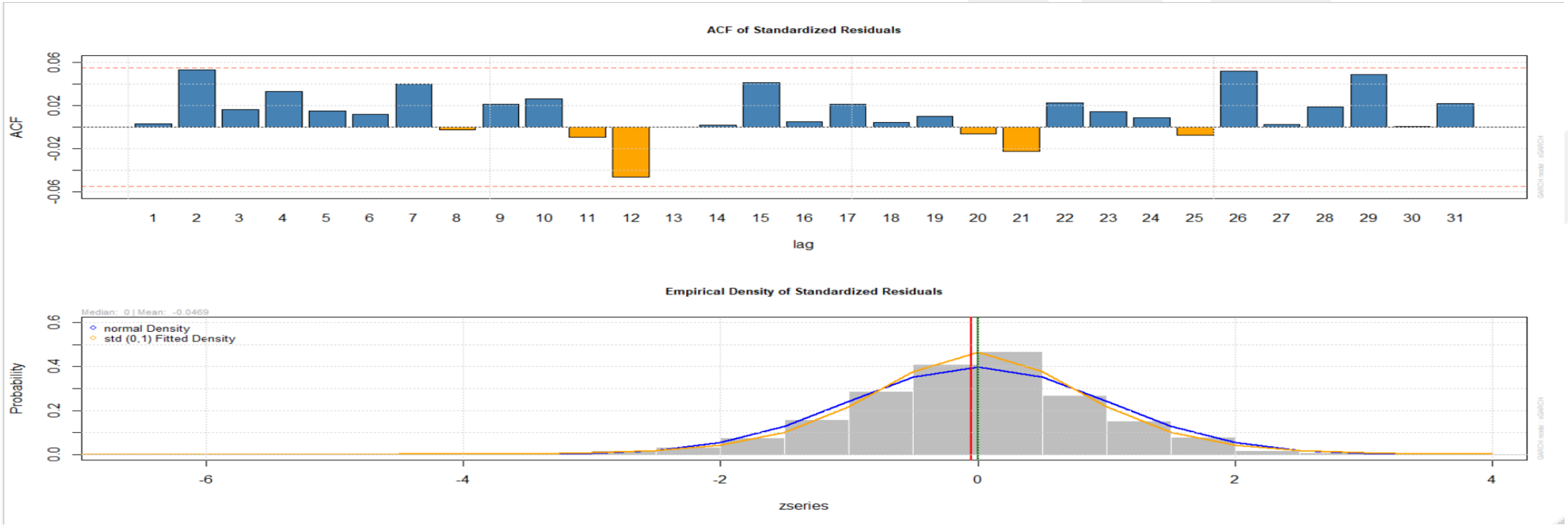


TIME SERIES ANALYSIS & STATISTICAL CORRELATION

- When viewing the autocorrelation of this series, the characteristics of a stationary process are observed at lag = 3.
- A Ljung-Box test at lag = 3 yields a p-value beyond the level of strongly significant at the .01 level.
- The null hypothesis of a random-walk process is thereby rejected, and it is concluded that this series is self-correlated at the lag interval 3.
- This will determine our holding period for this strategy.



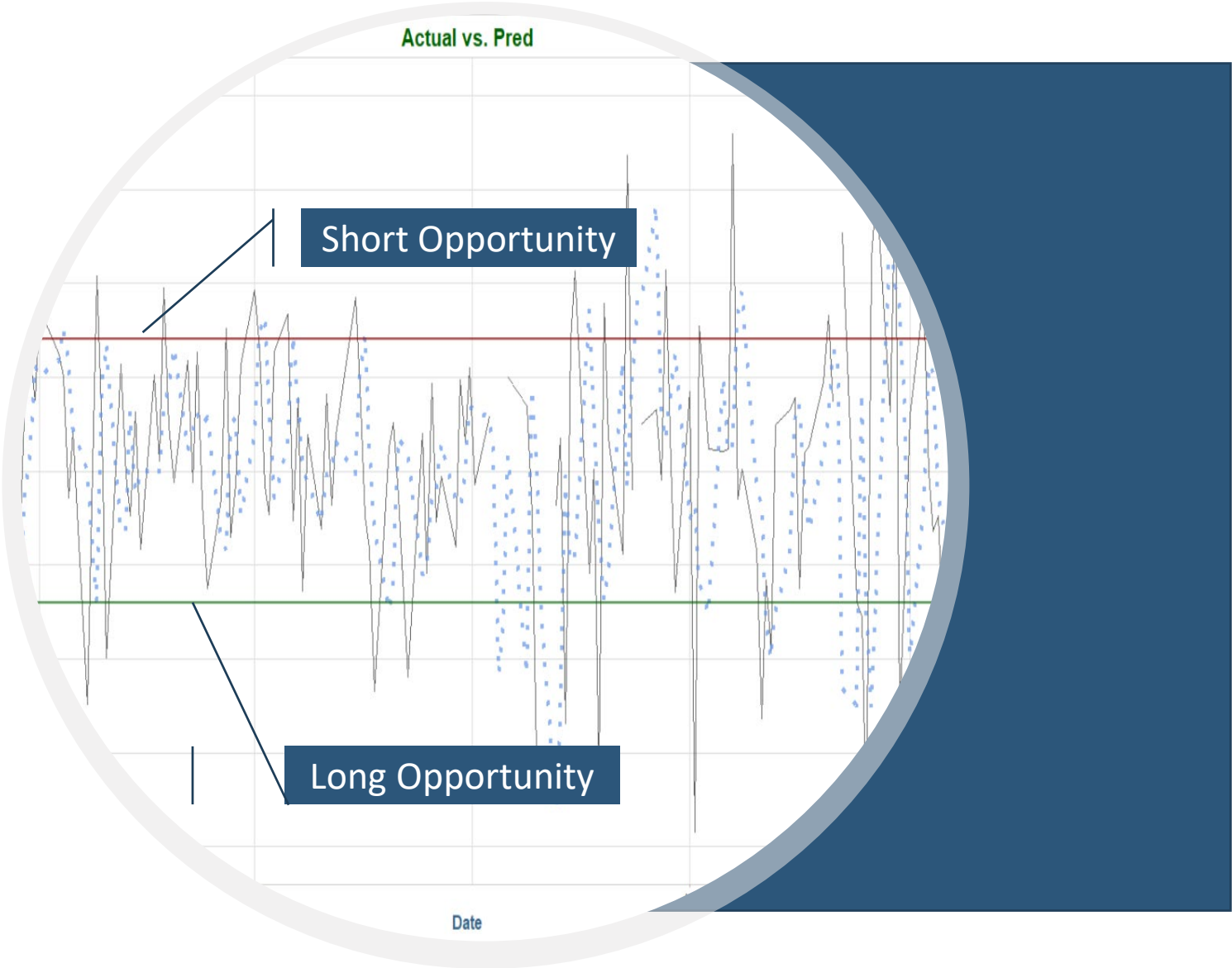
ASSESSING MODE FIT



Once a statistical model has been fitted to our series, we assess the model fit by observing the standardized residuals.

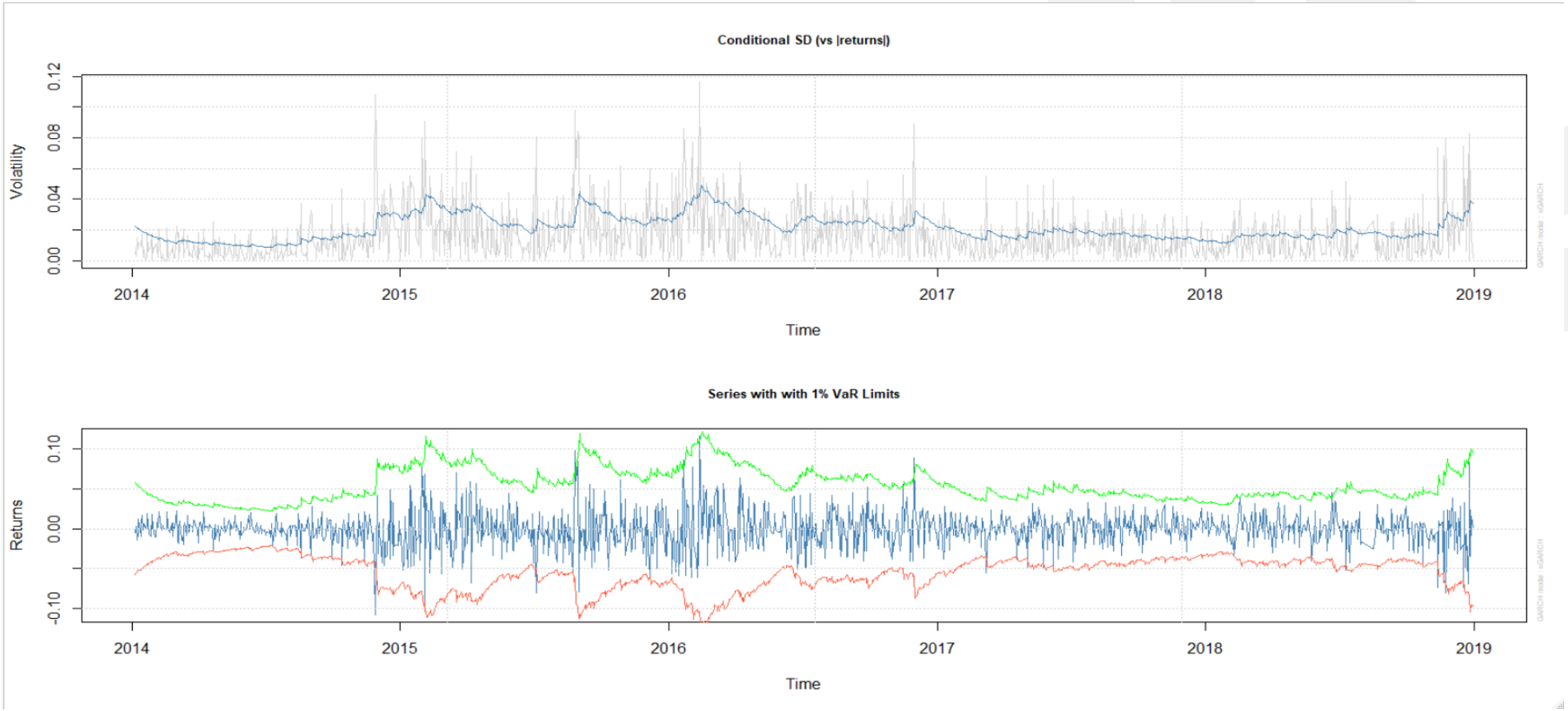
PREDICTIVE MODELING

- Here we see the return series of a commodity (black).
- Superimposed is our model predictions for the period (blue).
- Based upon our strategy threshold, defined by the auto-correlating behavior of the series, we see opportunity for profit from the large deviations from the expectation.



MODELING TAIL RISK

A GARCH process is used to capture the extreme movements in these assets.



The historical volatility of WTI, with 2 SD conditionals & 1% value-at-risk lines superimposed on the training data, along with the empirical density of the model residuals

Strategy Results

TRADING STRATEGY OVERVIEW

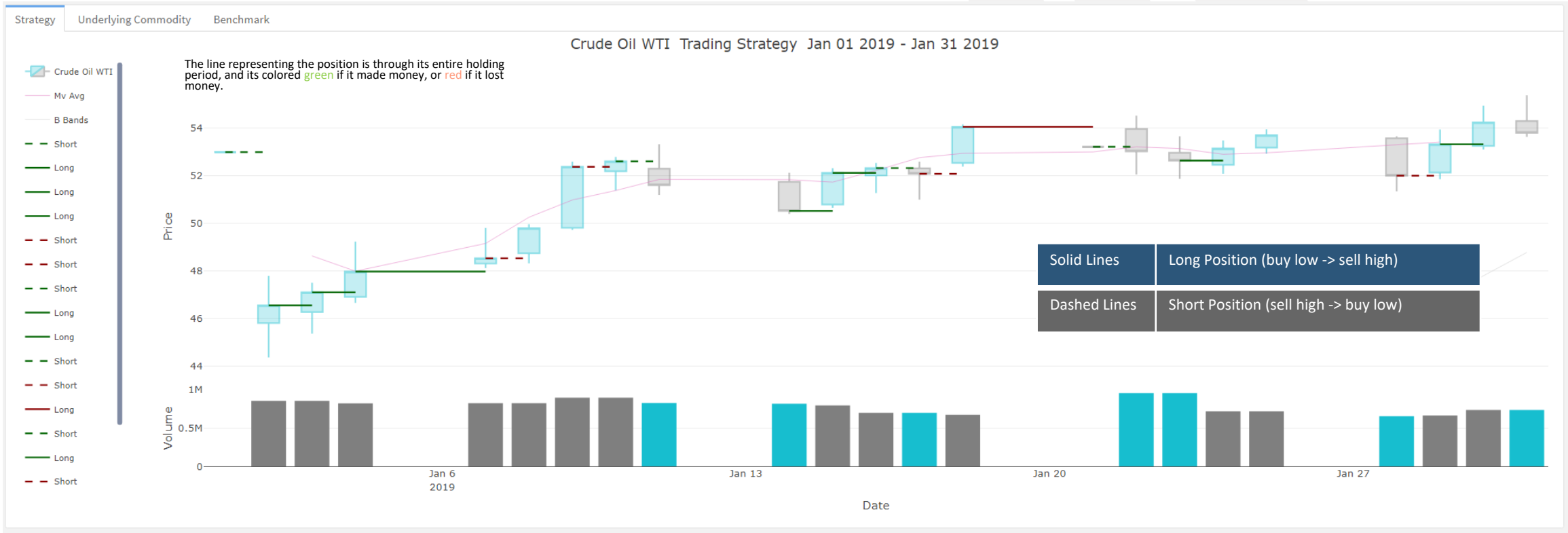
Using our **model predicted** data, our trading strategy is simple: we look for peaks and troughs in the returns, areas that are relatively large outliers.

The idea is that because we have created a stationary time-series using multiple asset returns, when one returns take a sharp up/down-turn, we enter the position (the direction is the **opposite** from the sign of the predicted return: positive, we **sell**, negative we **buy**).

Since there is a statistically significant **autocorrelation at a given lag interval**, we will take the inverse side and close out our position in *x* trading days from entry (*determined by strategy*).

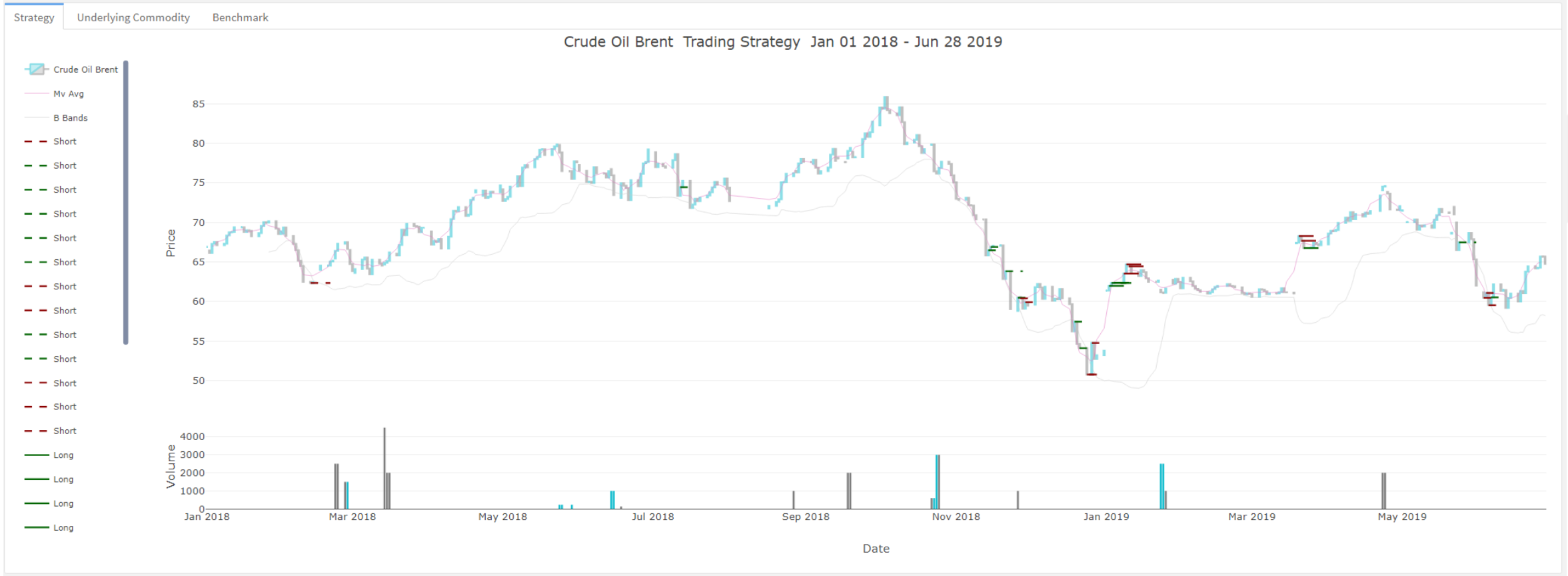
STRATEGY EXECUTION: DEEP DIVE

As an example of the strategy in action, multiple positions were entered over the period in January 2019.



We enter in the position on 1/1 with a sell-short on WTI Crude at $52.98, and then issued a cover short to close out our position at $46.56, yielding a net profit of $6.44 per barrel. Assuming 44,000 barrels per contract, this transaction netted our portfolio a gross profit of $241,120, or a 12.1% return on invested capital.

TRADING STRATEGY AND PERFORMANCE Crude Oil Brent



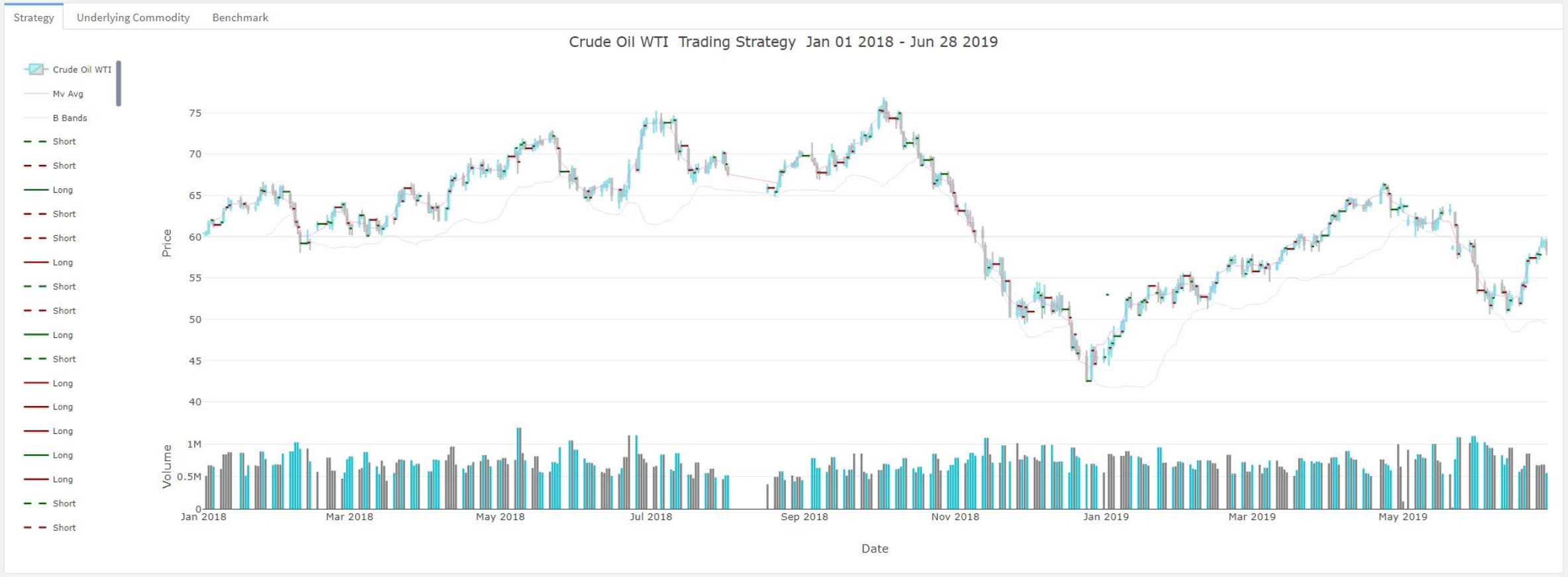
With a low trading threshold from the models predicted values, and a short-period, explosive volatility time series, the resulting strategy is one with a 1 day holding period per transaction. The annualized return on the strategy is **29.27**%, resulting from 97 holdings over the period.

STRATEGY SUMMARY PERFORMANCE

Crude Oil Brent – 1/1/19 – 10/30/19

	Commodity	Strategy	Benchmark
Periodic	13.45%	16.8%	3.74%
Annualized	17.94%	22.4%	4.99%

STRATEGY PERFORMANCE Crude Oil WTI



Year to date in 2019, our testing data range, we issue exactly 16 positions (32 transactions), leaving us with a net zero position at the end of the holding period, for an annualized return of **35.81%**.

STRATEGY SUMMARY PERFORMANCE

Data pre-processing, normalization and augmentation.

Risk profiles for each of the various assets & corresponding parametric model estimates.

Preliminary time series analysis & statistical correlations for **Crude products**

A predictive model for WTI, and corresponding trading strategy (**mean-revision**).

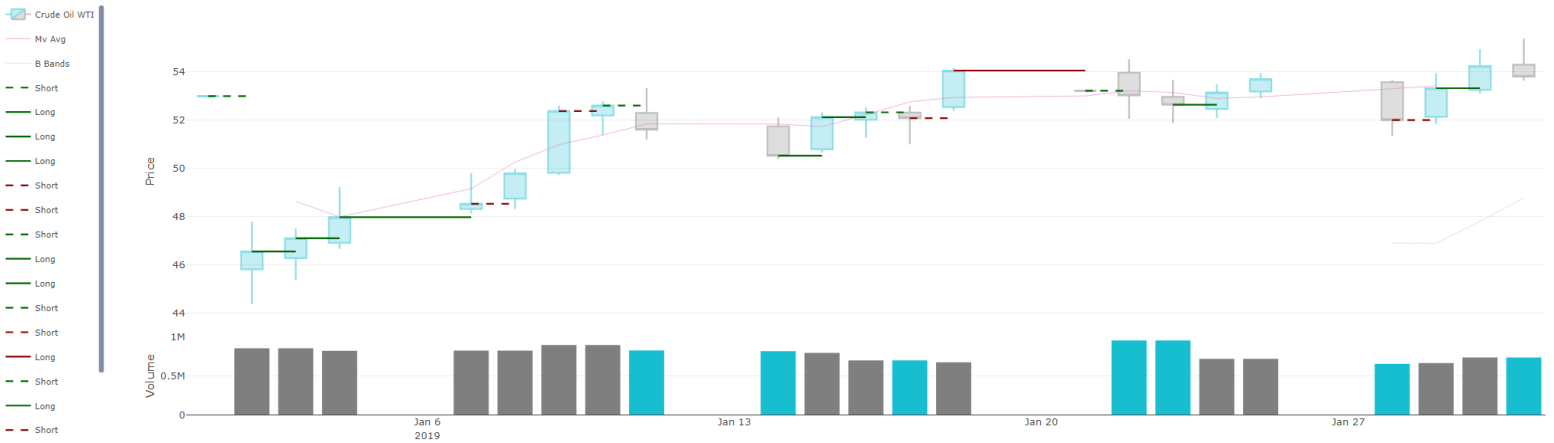
Initial backtest results **yield net positive PnL and returns** for the testing period (2019).

WTI mean revision strategy performance is **26.11%** year to date.

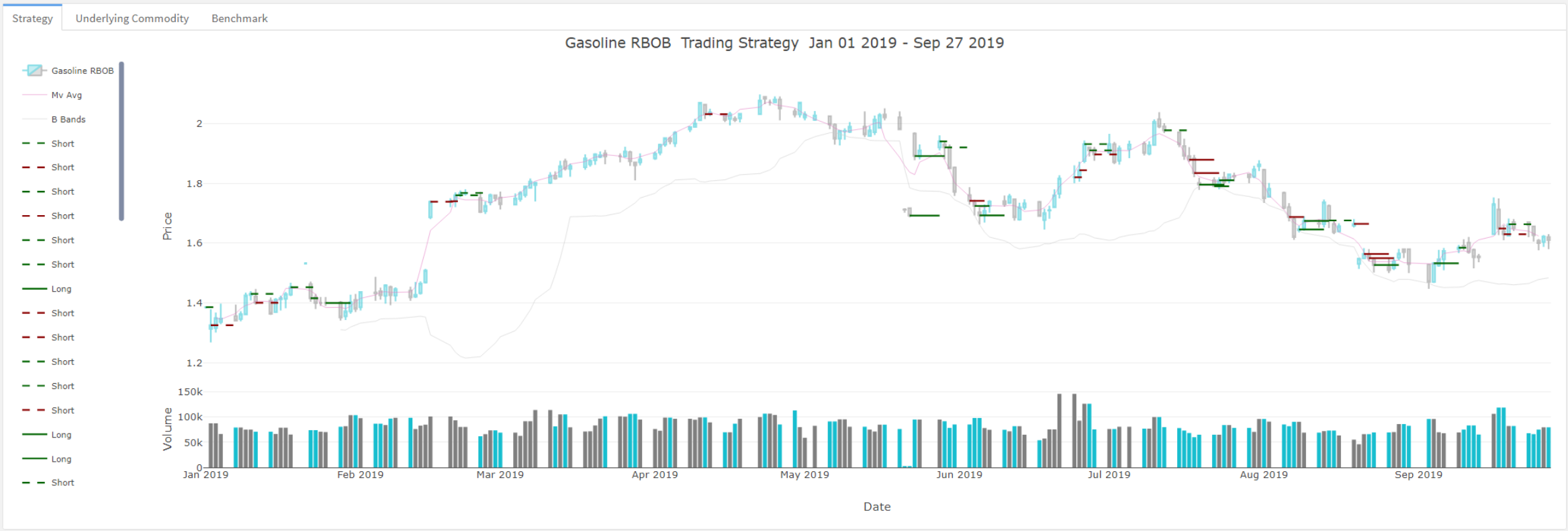
The trade execution / PnL **dashboard** is functional locally.

Crude Oil WTI – 1/1/19 – 10/30/19

	Commodity	Strategy	Benchmark
Periodic	23.1%	26.11%	3.74%
Annualized	30.8%	34.81%	4.99%

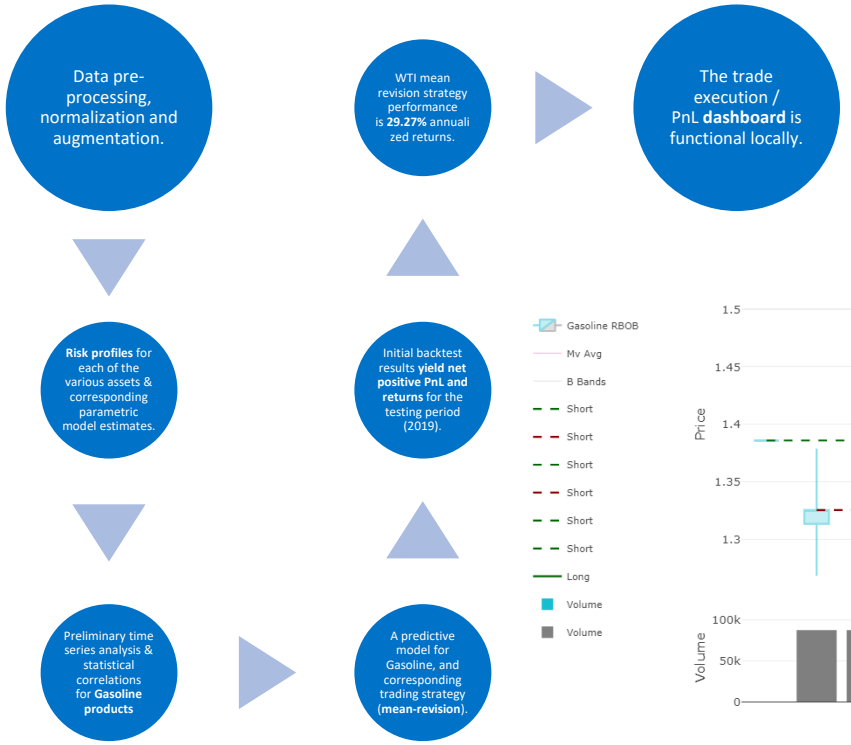


TRADING STRATEGY AND PERFORMANCE Gasoline



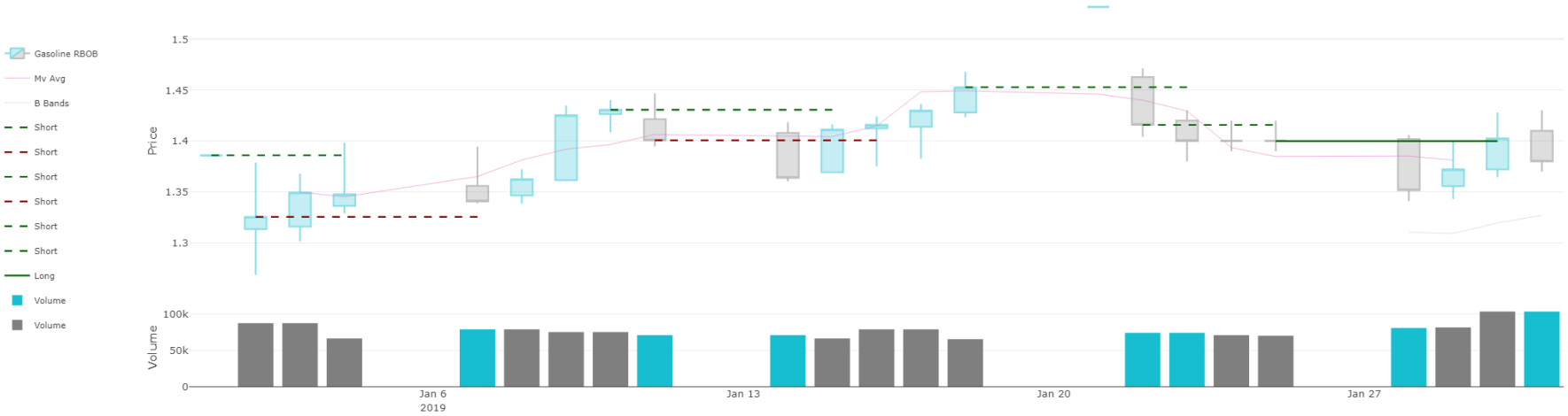
With a low trading threshold from the models predicted values, and a short-period, explosive volatility time series, the resulting strategy is one with a 3-day holding period per transaction. The annualized return on the strategy is **58.45**%, resulting from 76 holdings over the period.

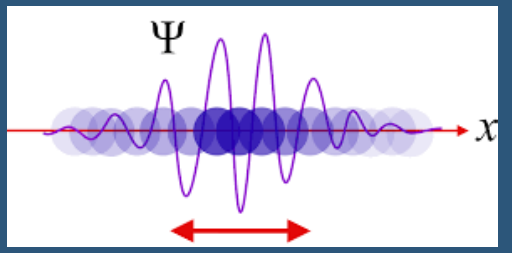
STRATEGY SUMMARY PERFORMANCE

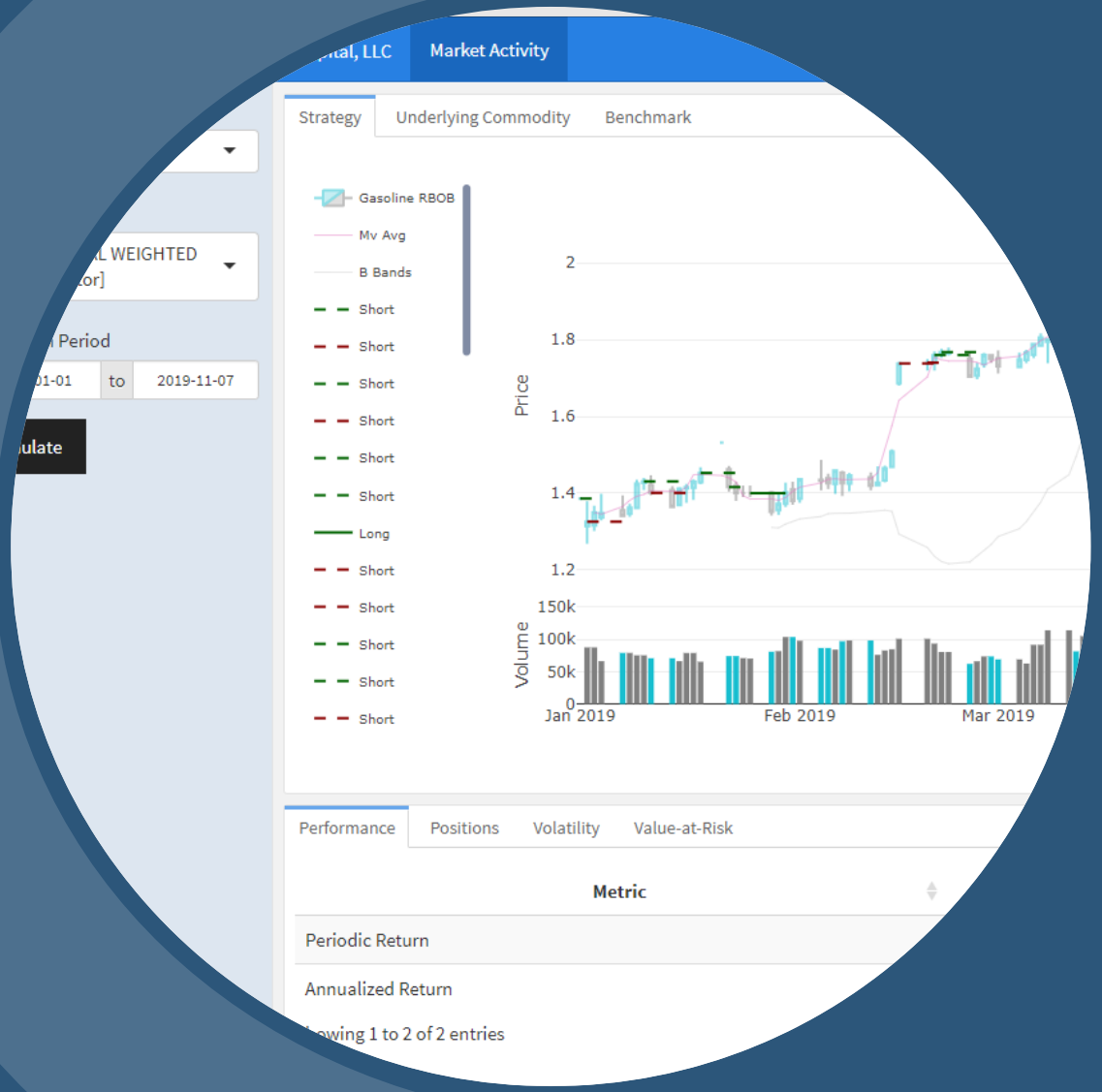


Gasoline – 1/1/19 – 10/30/19

	Commodity	Strategy	Benchmark
Periodic	23.8%	43.85%	3.74%
Annualized	31.8%	58.45%	4.99%







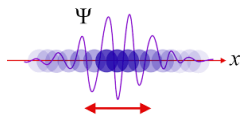
INTERACTIVE DASHBOARD

SEE OUR STRATEGIES IN ACTION, VISIT OUR LIVE DASHBOARD AND SEE HOW WE PERFORM.

Project Schedule and Next Steps

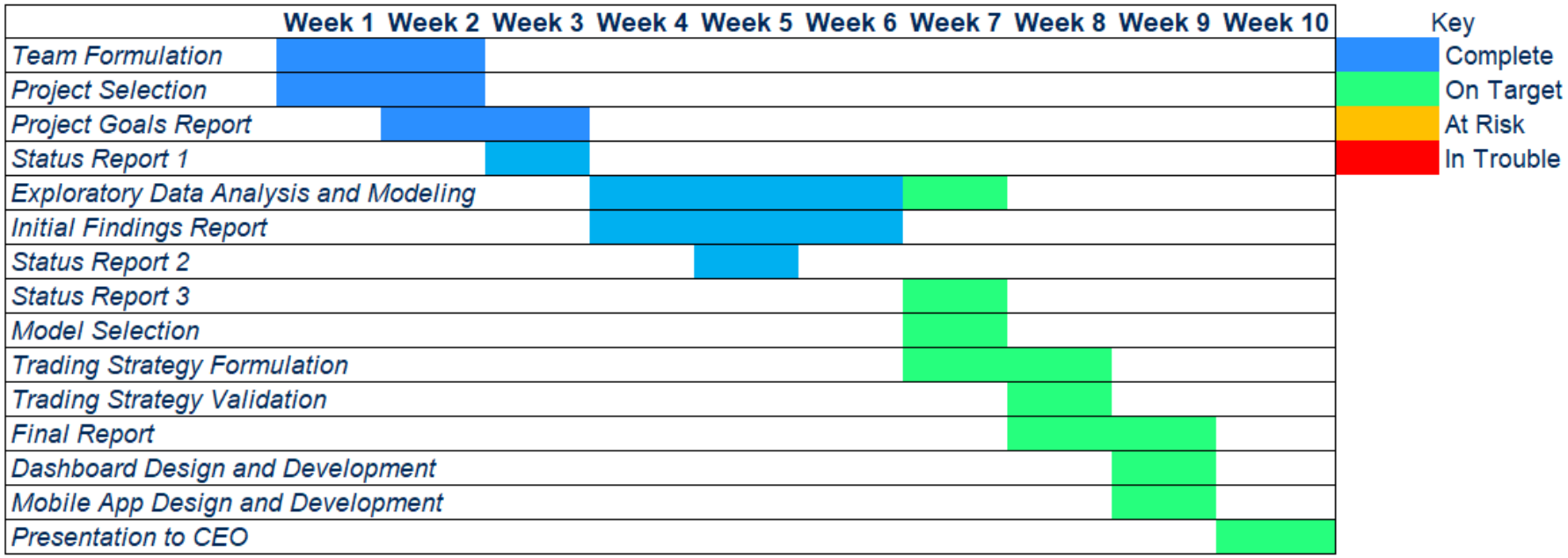
PROJECT DELIVERABLES

- Week 1-2 | Project Definition and Scope
- Week 3 | Project Goals
- Week 4-6 | Initial Findings
- Week 7-9 | Final Recommendations and Executive Summary
- Week 10 | Presentation to Project Stakeholders



PROJECT TIMELINE

The project will be broken down into 10 weeks. All deliverables are due by the end of each week, except for the Final Report, which will be delivered earlier in Week 9 on Tuesday, November 19, 2019.
The first key deliverable is the Goals Report, due at the end of Week 3. Next, there are three status reports that will be delivered at the end of Weeks 3, 5, and 7. The Initial Findings Report will be turned in at the end of Week 6, and the Final Report, Dashboard, and Mobile Application will be completed on or before the end of Week 9. At the end of Week 10 the team will deliver a presentation to the CEO. Any changes to the schedule will require a change request, signed by the project sponsors (CEO).



NEXT STEPS

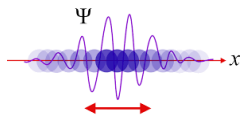
1 – continue to review strategies and find the best two based on back tested returns

2 – Polish up initial findings and create a final report for the CEO

3 – Complete mobile design and development

4 – Complete dashboard design and development

5 – Present findings to CEO



TEAM

ANDRIUS MARKVALDAS

Has over 20 years of information technology and data management experience on a variety of different platforms. He defined enterprise data strategies, as well as, built and managed world-class data capabilities for proprietary trading firms.

JOSHUA WOOD

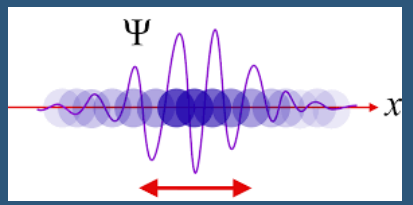
Flew for Southwest Airlines as captain of a Boeing 737 before turning to finance. His attention to detail, situational awareness and self-confidence rounds out the team dynamics.

BRANDON MORETZ

Has close to two decades of experience in the financial services industry. Most recently heading the development efforts of a proprietary in-house order management, research and risk analytics systems for a hedge with $20B in assets under management. Before that he spent time in the anti-money laundering space for some of the world's largest financial institutions.

TATE BOLLICK

For over 10 years, Tate has held a variety of roles within finance and technology. He has spent the last six years in program and project management for medium and large scale financial institutions. During that time he has worked on number of different projects for front, middle, and back office functions. He received his MBA in 2014, and has lived in Charlotte, NC since then.



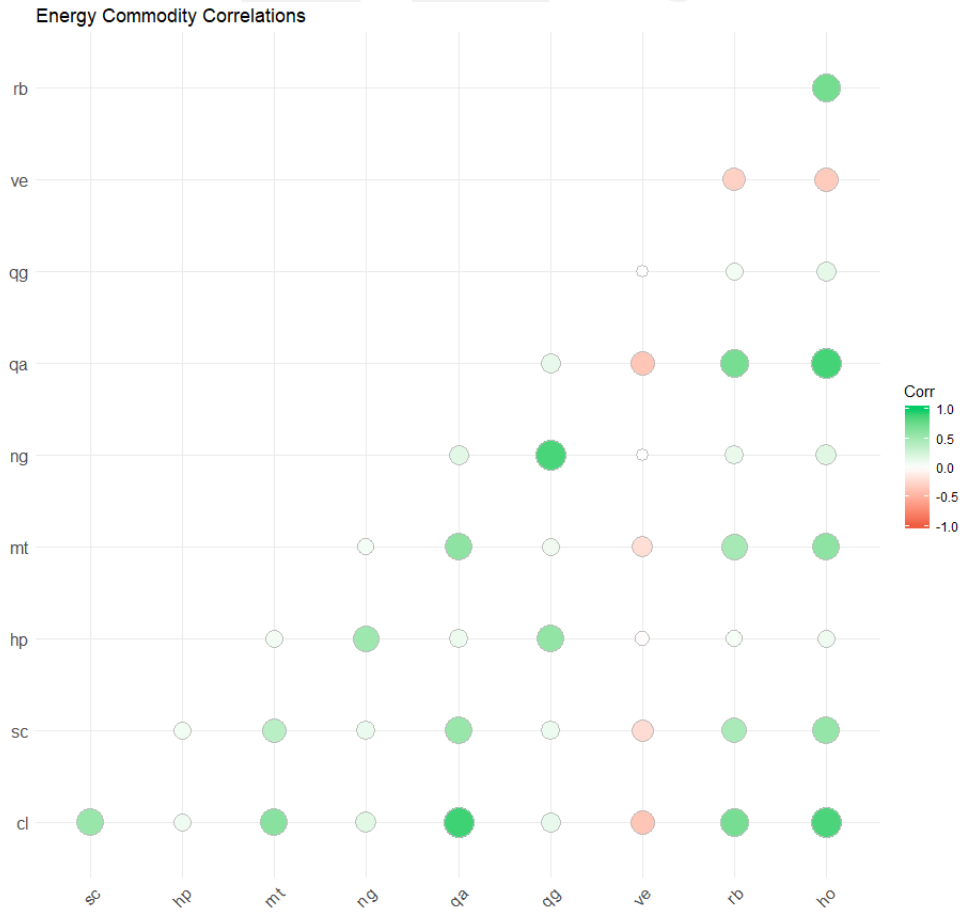
THANK YOU!

Feel free to contact any one of us…..

APPENDIX

Further Analysis

In the accompanying correlation matrix, the simple “linear” (Pearson’s) correlation only produces negative correlations between VE, which is the volatility index for crude. Incidentally, the VE is a lagging measure of the rolling volatility in crude oil products, and the negative correlation to the VE is only prominent in the four crude symbols, CL, SC, MT and QA.



CORRELATIONS

Person's Correlation:

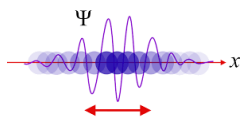
rn	sc	cl	mt	ve
sc	1.0000	0.5475	0.5311	-0.2163
cl	0.5475	1.0000	0.8589	-0.3829
mt	0.5311	0.8589	1.0000	-0.3348
ve	-0.2163	-0.3829	-0.3348	1.0000

Spearman's Rho:

rn	sc	cl	mt	ve
sc	1.0000	0.7473	0.7166	-0.3736
cl	0.7473	1.0000	0.8863	-0.4643
mt	0.7166	0.8863	1.0000	-0.4305
ve	-0.3736	-0.4643	-0.4305	1.0000

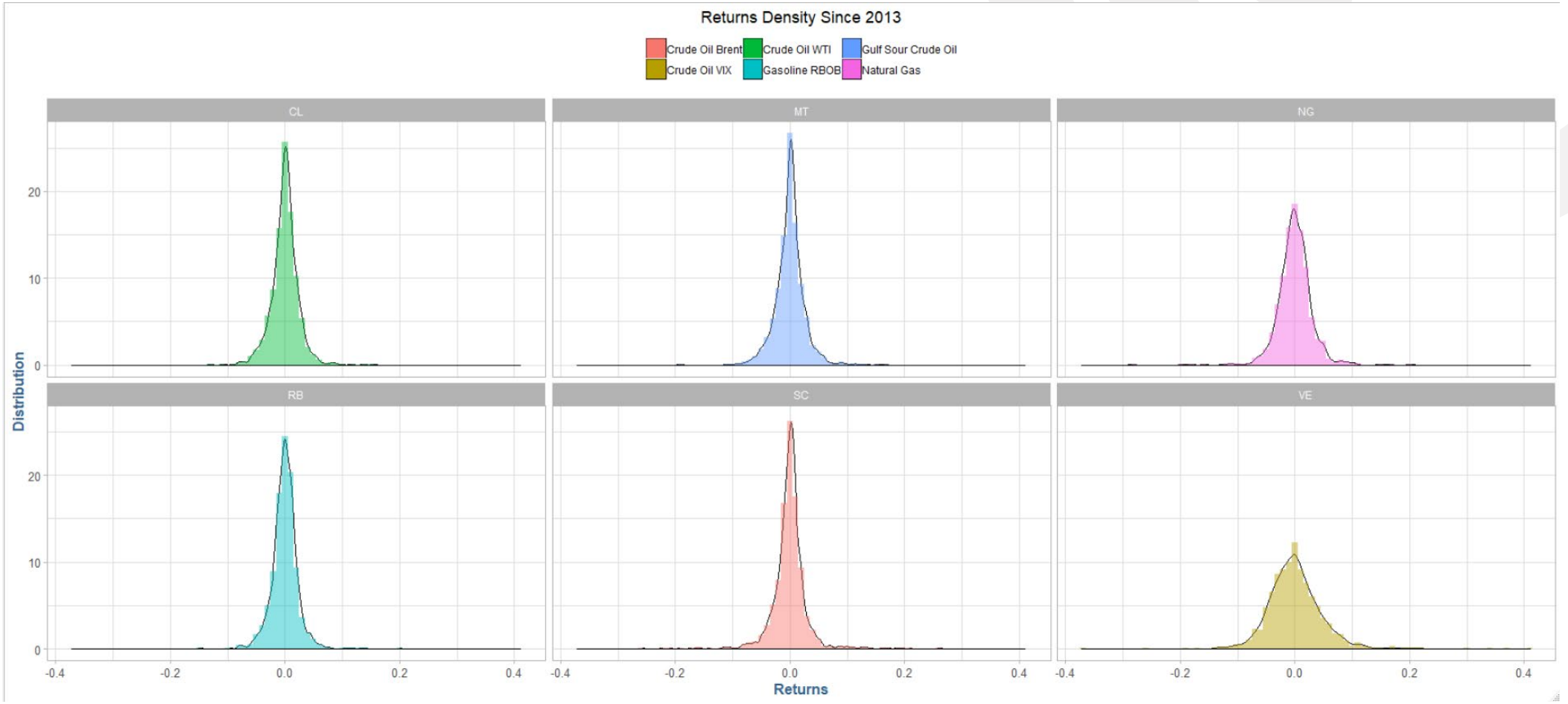
Kendall's Tau:

rn	sc	cl	mt	ve
sc	1.0000	0.6071	0.5767	-0.2630
cl	0.6071	1.0000	0.7513	-0.3313
mt	0.5767	0.7513	1.0000	-0.3055
ve	-0.2630	-0.3313	-0.3055	1.0000



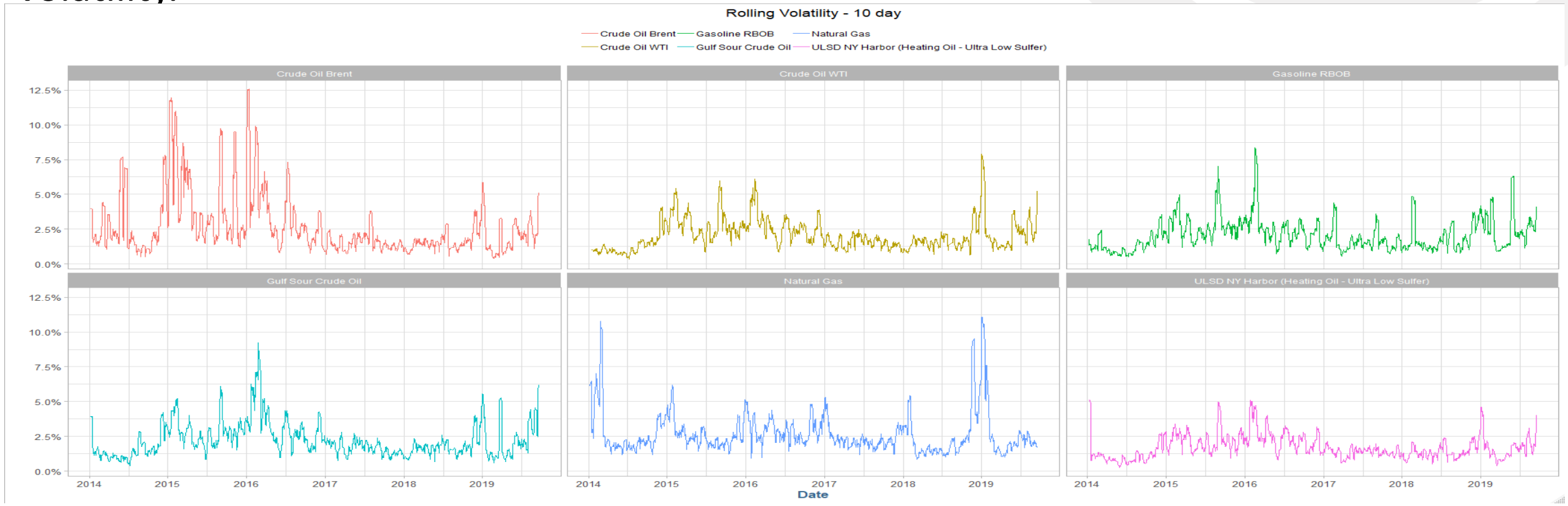
RISK PROFILES: Underlying Assets Continued

Looking at the clustering behavior of the returns shows the Brent, WTI, Gulf Sour and Gasoline appear to be the most stable assets. Natural gas and the Crude VIX exhibit a great deal of spread, which make modeling challenging.



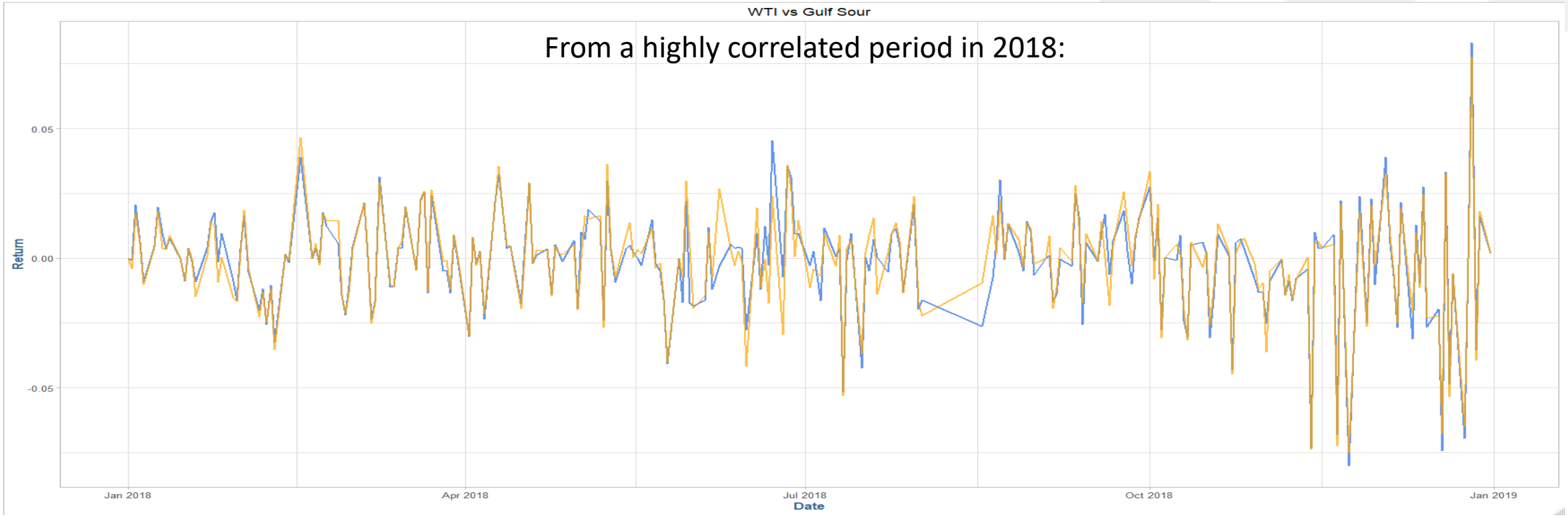
FURTHER ANALYSIS

In addition to the correlations which measure the strength, the commodities move in the same direction overall; a rolling standard deviation can be computed for a specified period. This process helps smooth out the variance over time, which cuts down on the noise and thereby increases the transparency of the relationships in volatility.



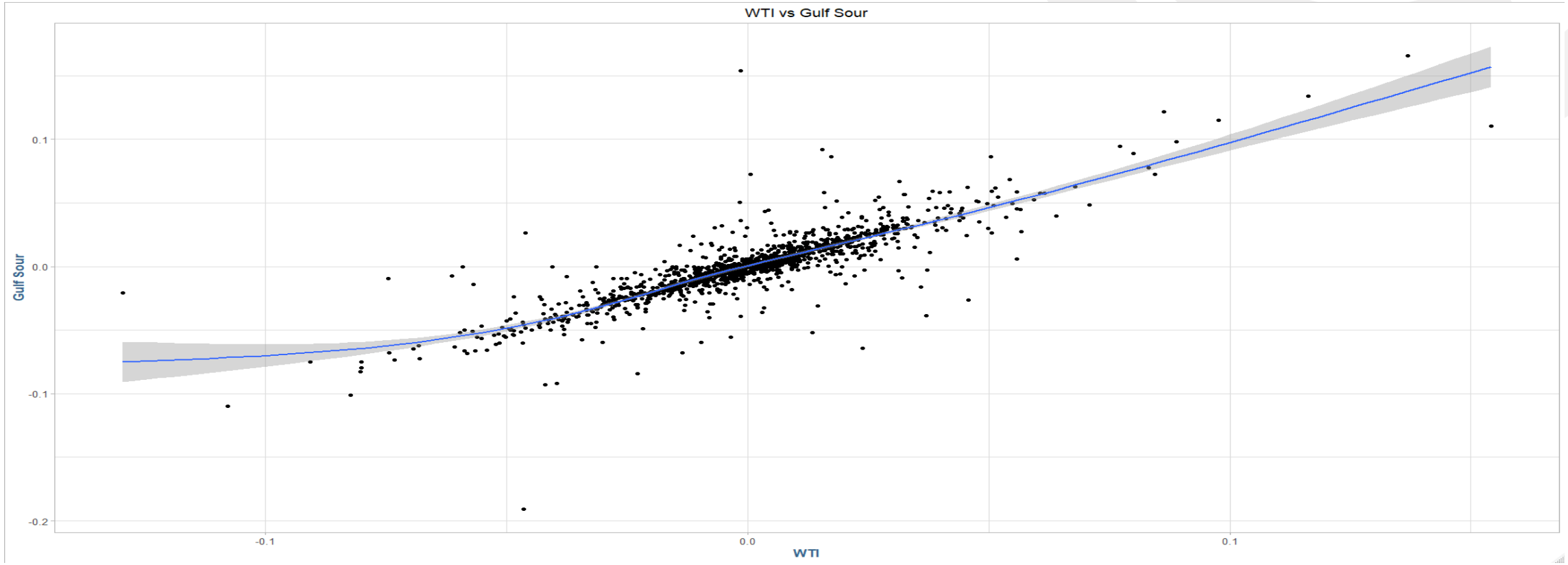
FURTHER ANALYSIS

The correlations show the Spearman's Rho for WTI and Gulf Sour seem to be directionally correlated about 89% of the time. Additionally, there is a strong linear relationship indicated by Pearson's correlation coefficient:



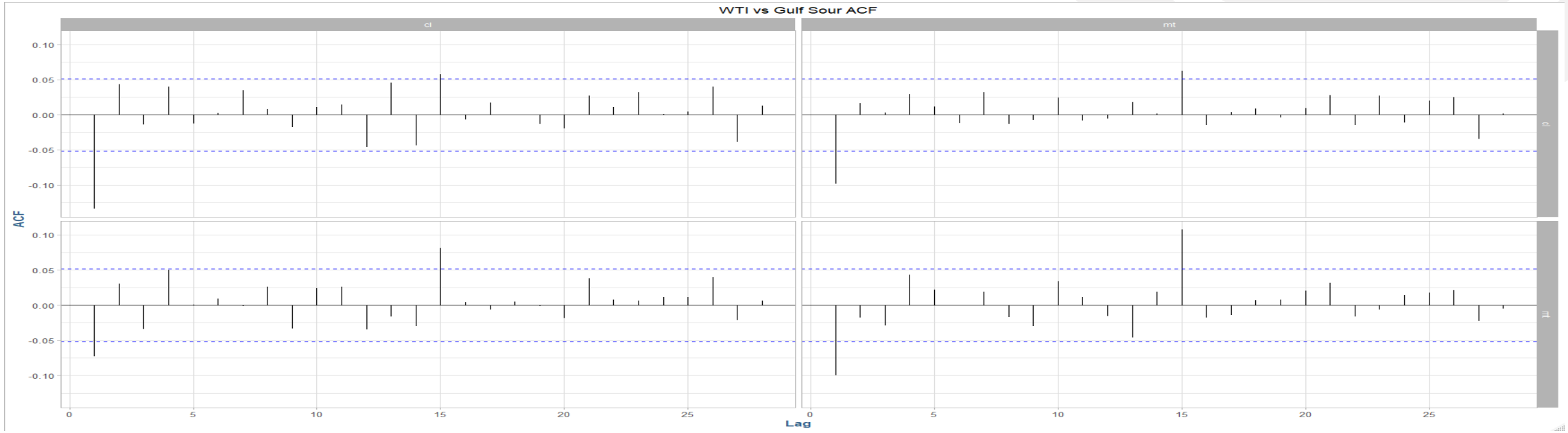
FURTHER ANALYSIS

A scatterplot of the returns reveals additional clustering, and confirms the near linear relationship:



FURTHER ANALYSIS

On a time-series basis, we see strong negative autocorrelations at 1, 4 and 15 days out, indication there could be a strong leading indicator here, therefore a transactional arbitrage opportunity:



There are additional strong relationships further out, however, the chance of those relationships being generated purely by chance are quite high.