Rice University

DSCI 435

Spring 2020

Belvedere Final Report

Authors:
Ye Chen
Seth Kimmel
Ankit Narasimhan
Weili Nie
Jordan Pflum
Yifan Zhang

?? May 2020

Contents

1 Introduction

Future contracts are a unique, exciting, and important financial instrument which allows a someone to buy or sell a commodity at a specific price at some specified time in the future. As an example, consider an airline which believes that fuel prices will increase over the next several months. They may wish to "lock in" a lower price by purchasing a future contract for jet fuel with a lower expiry price than they forecast the price to be in several months. If they forecast correctly, they can take delivery of "discounted" fuel at the time the contract expires. If they did not, they may be buying fuel for a much higher price than the current market value. Thus, futures contracts provide a very important service to the economy and a liquid, efficient market is critical to these participants. The most active futures market is the Chicago Mercantile Exchange with an average daily volume of over 1 million contracts per day across various types of commodities, currencies, indexes, and more.

Thus, there is a high demand for accurate predictions of the futures market. With increasingly granular future data becoming available to interested parties, data science tools can easily be applied to assist in accurate forecasting of the futures market.

One way to make predictions about prices of financial instruments is to understand how they correlate with one another. If you know that two financial instruments are closely positively tied in price, then an increase (respectively decrease) in one of the instruments will likely lead to the other increasing (respectively decreasing) as well. If there is a strong negative correlation, then an increase (resp. decrease) of one of the instruments will likely lead to a decrease (respectively increase) in price of the other. Thus, if one can accurately predict correlations at a given time - as they are constantly changing - one can better be able to predict price changes as a result.

Companies which take interest in the future price of financial instruments, such as our industry sponsor Belvedere Trading, are vital in ensuring a healthy futures market. Belvedere uses future contracts as a means of hedging other assets in their portfolio. They would like to get a better understanding of liquidity of prices in the futures market to better be able to lower their costs to hedge and pass those savings on to other market participants. Therefore, Belvedere Trading has tasked our teams to explore the correlations between a provided set of future contracts in relation to market conditions and generate useful insights. For example, knowing that an increase in a certain technical indicator such as ADX on a weekly time-scale leads to a higher correlation between two futures could better allow Belvedere to make price predictions and trading decisions. This objective will be expanded upon in Section 3.

2 Background

2.1 Belvedere Trading

Belvedere Trading is a proprietary trading firm based in Chicago Illinois which specializes in software development and market-maker trading strategies. Due to their ultramodern

proprietary technology and risk management capabilities, Belvedere is able to quickly capitalize on inefficiencies in the marketplace. Additionally, their trading models and software systems are continually re-engineered, optimized, and maintained to stay on top of the industry. [1] Our team was tasked with analyzing the future markets in order to identify market conditions (observable features or combination of features from the futures markets) where strong and unique correlations (both positive and negative, as well as cyclical trends) between futures emerged. This information would enable to Belvedere to better capitalize when key market conditions materialize.

2.2 Future Contracts

Futures contracts are financial derivatives that require the buyer to purchase an underlying asset (or the seller to sell that asset) at a predetermined future price and date. Therefore, a futures contract allows an investor to speculate on the future price of a security, commodity, or a financial instrument using leverage. Futures are commonly used to hedge the price movement of the underlying asset to help prevent losses from unfavorable price change. [2]. While there are many specific future contracts, some common ones include index futures (Emini S&P 500, Emini DJIA), currency futures (Euro to US Dollar, British Pound to US Dollar), and commodity futures (gold, silver, oil).

A simple example helps those previously unfamiliar with future contracts understand the concept more clearly. Consider the price of corn cereal. Generally, its shelf price will remain the same, week-to-week. However, the price of one of the main ingredients in corn cereal, corn, changes daily. So why does the cost of processed food stay stable even though the crops that go into them fluctuate? Partly, thanks to the futures market.

The producer of corn, a farmer, is always looking to sell their corn at a high price. Conversely, the user of corn, a cereal company, is always looking to buy their corn at a low price. The farmer has a problem because their whole crop is harvested at once. Because farmers all harvest around the same time in the year, the market is saturated with the supply of corn, which sends the price falling. Similarly, the cereal company does not want to buy all the corn they will use for the year at once, because they would have to store a huge quantity of corn, which is expensive. This is where the advantages of a future market kicks in. Buyers and sellers of can buy and sell future contracts on corn instead of buying and selling the actual corn. Therefore, the underlying asset, corn, rarely changes hands.

Additionally, a future contract provides a hedge against a change in price, commonly refereed to as risk management. This way neither side is stuck with only whatever the market price is when they want to buy or sell. Farmers will not sell all of their corn in futures, only enough to ensure that a low price at harvest wont ruin their business. The future contract provides that security. Even if the price of corn rises at harvest and the farmer is losing money on their future contract, they can offset their losses by selling the remaining corn at a higher price. Similarly, the cereal company uses future contracts to protect from a high price during harvest.

This example illustrates one of the main benefits the futures market serves and risk management. It doesn't seek to maximize profit. Instead, it focuses on balance, and in this way it keeps producers and suppliers in business and the price of consumer processed goods stable.

2.3 Previous Semester's Work

In order to determine the dependence of futures, last semester's team built contemporaneous and non-contemporaneous graphical models and estimated correlations between futures. They applied graphical models to examine contemporaneous associations between futures over different time periods (minutes, hours, days, months). They also applied Granger causal graphical models to examine lagged associations between futures. With these models, they were able to visualize and evaluate how price of futures relate to one another holding all other futures in the dataset constant.

We exported last semester's team's output into a series of correlation matrices that could be used for our purposes. We used their original R code to compute the correlations for all 26 futures on their open-close change price, open price and close price over four different time periods (weekly, biweekly, monthly, and seasonal). The specific method used is Pearson's correlation coefficient, which measures the linear dependence between two variables. When computing the Pearson's correlation in our time series data, we are observing the correlation between two variables during a specific time range. We specify a time interval using two values: the first day in the interval, and the number of days in the interval (including that first day).

Their model validated some of the obvious relationships as well as discovered some interesting ones. For example, their graphical model showed that gold and silver were related to each other much more in the month of November than any other month - during which the correlation was essentially non-existent. You can see the relationship over the months in the polar plot below, in which the correlation is much stronger in the month of November.

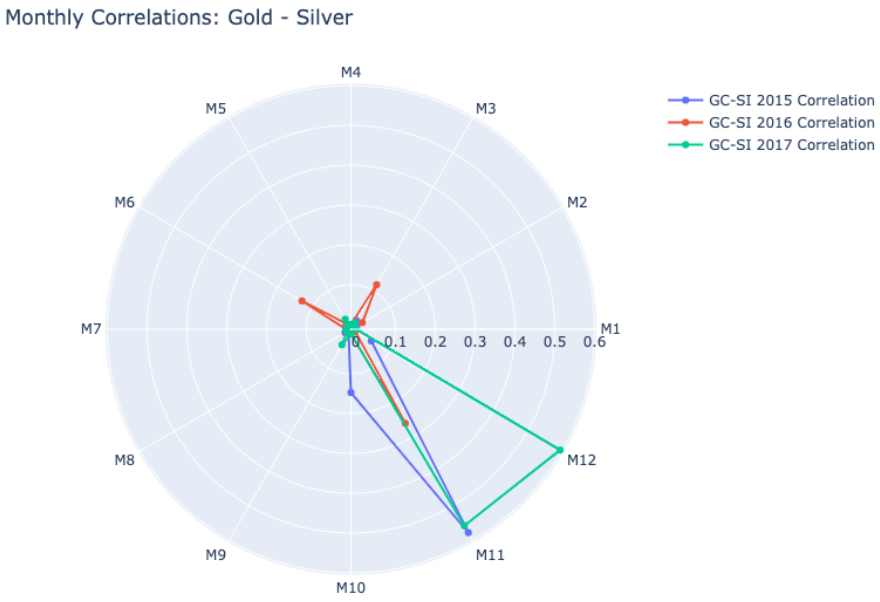

Figure 1: Monthly correlations of Gold and Silver future contracts. Data was generated via graphical model developed by last semester's team and plotted using our polar visualization scheme.

This semester, we dive deeper into determining when these correlations exist and the market conditions which influence them.

2.4 Literature Review

Naturally, we are not the first to be applying data science tools to market prediction. There is an abundance of work done in this area every day, but a small amount of it is done in an academic (non-industrial) setting. Most of the work in quantitative trading and finance is kept behind closed-doors, as the information pertaining to methods used is proprietary and often lends itself to a competitive edge. That being said, there is a good amount of scholarly research on general time-series market analysis and prediction using data science tools. As this is the baseline of our project, it is important to consider the prior work done in this area.

In [3], the authors apply a Maximum a Posteriori Hidden Markov Model to predict prices in the stock market. An even earlier attempt to use a hybrid HMM for stock market prediction comes in [4], which blends a HMM with genetic algorithms and artificial neural networks. Even more generally, in [5] the authors create a Fuzzy-HMM for time-series prediction with enhanced performance over other fuzzy models. Perhaps even more directly applicable to our project comes in [6] which evaluates the use of dynamically adaptive trading strategies, predicting different market regimes using State Switching Markov Auto-regressive models. Another paper concerned with econometric analysis and general regime models comes in [7] where HMM usage is evaluated.

In addition to the common use of Hidden Markov Models in market prediction, correlation is a widely used approach to predict prices. [8] uses a novel method which leans on correlations between companies to predict stock price movements. Another well-cited paper [9] in the financial literature deals with understanding trading volume and serial correlation in stock returns. Another very useful reference is concerned with the opportunities and limitation of using correlation strategies in trading [10].

Finally, there is a good amount of work done already in linear forecasting models. For example, [11] uses ARIMA models to forecast next-day electricity prices. [12] is a popular paper which is among the first to consider ARIMA (linear) models coupled with ANN's (non-linear) for the purposes of time-series forecasting. Finally, returning to the subject of financial markets, [13] utilizes a hybrid of ARIMA models with support vector machines for stock price forecasting. There are also papers concerned with simpler moving average models such as [14] which considers the profitability of using such trading rules. There are also simple means to improve moving average trading rules. In [15] they employ boosting and statistical learning methods to do so.

Clearly, there is a lot of prior work done in the methods we are using for this project. Many even deal with financial markets directly, although most are generally used for time-series forecasting. Regardless, all of this work serves to guide us and inspire ideas pertaining to our project.

2.5 Novelty

As we mentioned in the literature review, we are not the first group to use data science tools for market prediction. Clearly, we are also not the first to use Hidden Markov Models or ARIMA/moving average models for time-series forecasting or financial market prediction. We are certainly not the first to use correlations as a means to drive price predictions.

However, we are studying a new market (in light of published works) - the futures market - using such methods. Additionally, the accessibility and ease of use of data science tools allows us to more quickly explore, prototype, and test our ideas on real data. There is a good chance this will allow us to accomplish more than our predecessors.

Finally, it is worth noting again that there may have been other attempts to do exactly what we are working on, using the same methods, but behind closed-doors. Given the competitive landscape of the industry in which Belvedere works, this is simply how it works. However, we can hope that we are either working on something novel, have taken an alternative spin on the idea, made substantial improvements in the way that we are modeling, or that Belvedere can deploy the work more efficiently than its competitors.

3 Project Objectives and Applications

3.1 Project Objective

In our project, we aim to identify specific market conditions where unique and strong relationships between futures emerge. Our work will extend the work done by last semester's team which implemented a graphical model to identify interactions (correlations) between futures throughout a set time period. This semester, we seek to also find the specific market conditions (market features) for when certain correlation "regimes" (for example: low, neutral, and high) between future contracts may exist in the future. Even more important than finding a useful model for predicting these correlation regimes is understanding when and why they arise. Therefore, we seek to understand which market features generated by past future contract price and volume data most significantly assist predicted future correlations. Not only will will feature selection improve our model's accuracy, it will also provide valuable information to Belvedere Trading when deciding which features to generate and include in their own future contract models. For these reasons, at the end of our project we aim to provide Belvedere with a carefully vetted group of features to use when unique market conditions emerge.

3.2 Project Application

While there are many potential applications of our project, we believe our output will help Belvedere trading in two primary functions. First, our selected features should validate, if not improve, existing models currently employed by Belvedere. Second, most trading firms aim to have a balanced portfolio, meaning they aim to balance their investment earnings against the risk of losing money to a volatile and often unpredictable market. A useful tool to limiting ones risk to the market is by hedging ones investments, reducing

the risk of adverse price movements of a given asset. For example, historically the price of oil is inversely related to the price of the U.S. dollar. FINISH EXAMPLE

4 Data Science Pipeline

When designing and implementing a data science project, a pipeline outlining the goals and corresponding techniques is essential in producing a successful deliverable at the finish line. Figure 2 details the pipeline created for our team's project.

Belvedere Data Science Pipeline

<table>
<tr><td colspan="2">Main Goal: Identify market conditions where unique and strong correlations between futures emerge.</td></tr>
<tr><td rowspan="3">Data Wrangling</td><td>Goals</td></tr>
<tr><td>• Create a combined dataset for all futures
• Pull in and format output data from previous team project (R) into Python</td></tr>
<tr><td>Techniques</td></tr>
<tr><td></td><td>• Time series data
  ○ The futures' features will be the columns while the rows will be ordered according to time
  ○ Account for formatting of dates and times
• Previous groups output data
  ○ Run R files and generate output correlation data
  ○ export csv files to read in</td></tr>
<tr><td rowspan="4">Data Exploration</td><td>Goals</td></tr>
<tr><td>• Understand how each of the futures' features behave and determine important new features
• Find possible groupings between different futures
• Identify trends in previous team's correlation data</td></tr>
<tr><td>Techniques</td></tr>
<tr><td>• Create data visualizations to better understand the original futures data and correlations (i.e. polar visualizations)
• Implement various clustering methods such as k-means and hierarchical clustering, and graphical models
• Create feature visualizations, potentially using principal component analysis and pattern detection</td></tr>
<tr><td></td><td>Challenges</td></tr>
</table>

	• Wave varying density of data availability, with some segments having consecutive minutes and others having no data for long periods.
Feature Engineering	Goals
	• Generate features from original dataset to aid modeling • Visualize features to identify trends • Identify linear and non-linear relationships with correlation data
	Techniques
	• Use ta-lib to automatically generate 150+ relevant features for each future contract • Plot features against correlation data to visually identify trends • Use autoregressive and other simple models to identify meaningful relationships
	Challenges
	• Automatic feature generation is not always suitable for intended modeling purposes • Subtle relationships between features and correlations may be hard to detect without more complicated means
Modeling	Goals
	• Identify unique market conditions • Predict futures price/correlations movements via movements in other futures or features (MA, Volatility, etc.) • Benchmark using simple linear models to predict futures price/correlation movement (pattern recognition)
	Techniques
	• Use Hidden Markov Models to predict changes in correlation states • Use simple linear models such as moving averages and ARIMA models as benchmarks for more complex models • Use LASSO to enhance variable selection of the model

	Challenges
	• Resolution of the data • Changes based on backfilling
Validation	Goals
	• Determine how well the predictive model performs • Generalizing of data-driven discovery
	Techniques
	• Use validation techniques such as train/test split and cross validation
Communication	Goals
	• Display findings and results
	Techniques
	• Create a report and poster with data visualizations and detailed explanations that are both technical and business related.

Figure 2: Data Science Pipeline

5 Data Description

Our original dataset comes in the form of 28 individual files, each respectively containing the open, high, low and closing prices as well as and traded volume of the future contract they represent, in one-minute intervals. For each contract, there is data for every day of trading (9:00 am to 3:00 pm) for all days in 2015 through 2017. In the merged dataframe, this is approximately 250,000 rows of data and 4 columns for each future, plus a timestamp column, a total of roughly 28 million data points. Examples of such future contracts include Lean Hogs, E-mini S&P 500 Futures, and Copper.

We removed two futures - PA (palladium metal) and RTY (Russell 2000 minis) due to a lack of trading data over our period of study, and the fact that the Fall semester's team did so as well and did not compute correlations for them.

Additionally, we use data produced by the Fall semester's team which investigated correlations between these futures contracts. We use the values they have derived for the correlations as our target variables which we hope to explain using our engineered features. See Section 2 for a more detailed explanation of how these correlations are generated. We use correlations occurring over discrete time windows in weekly, bi-weekly, monthly, and quarterly time periods. This choice was made due to the fact that Belvedere specified weekly intervals as the shortest period of time over which a correlation would be relevant to trading. There are 650 pairs of futures (when you exclude pairings of the same future), and four correlation periods, this means the target dataframe contains 2,600 columns. There is a row for each minute, meaning there is a total of roughly 650 million data points in the target dataframe.

Below is a sample of the original dataset. This is for E-mini S&P 500 (ES) Futures Contracts.

Timestamp	OPEN	HIGH	LOW	CLOSE	TRADED_VOLUME
2015-01-02 08:27:00-06:00	2062.5	2063	2062.5	2062.5	683
2015-01-02 08:28:00-06:00	2062.5	2062.75	2062.5	2062.75	271
2015-01-02 08:29:00-06:00	2062.5	2062.5	2061.5	2061.75	744
2015-01-02 08:30:00-06:00	2061.75	2062	2061	2061	1357
2015-01-02 08:31:00-06:00	2061.25	2062	2060.25	2061.25	7911
2015-01-02 08:32:00-06:00	2061.25	2062	2060	2061.75	4809
2015-01-02 08:33:00-06:00	2061.75	2064	2061.5	2063.75	6824
2015-01-02 08:34:00-06:00	2063.75	2064.75	2063.25	2064.5	8626
2015-01-02 08:35:00-06:00	2064.5	2064.75	2063.25	2063.25	4566

Figure 3: A sample from the E-mini S&P 500 futures data.

5.1 Data Wrangling

While the full dataframe is large and encompassing, this is not the final data we are working with. First, we engineer features from the original data (specific features to be discussed) and append new columns to the dataframe representing these features. There can be several hundred features generated per original data column, leading to a far greater-sized dataframe.

However, after features are generated at a minute-level granularity, the dataframe is condensed to a daily-level granularity. This reduces the number of rows to roughly 800. This allows us to more easily work with the correlations which occur over longer time intervals. The method by which this aggregation is done is meant to fully capture the minute-level data.

It is also worth mentioning that the original data provided from Belvedere only contained rows of data for those minutes in which trading occurred (and hence volume was greater than 0). In those minutes which trading did not occur the prices for open, low, high, and close remained the same as the previous minute in which trading occurred, and we had to back-fill the data to account for all minutes in our desired time period.

6 Data Exploration

Due to the cyclical nature of our time-series data in weekly, biweekly, monthly, and seasonal periods, we decided to employ polar visualizations to help us better understand our data. As an example, we have chosen three grouped futures - Soybeans (ZS), Soybean Meal (ZM), and Soybean Oil (ZL). When mutually paired, we can produce polar visualizations showing their correlations using the Fall semester's model to reveal interesting structural patterns. The below plot shows the pairing of Soybean Meal and Soybeans, and their weekly correlations over three years.

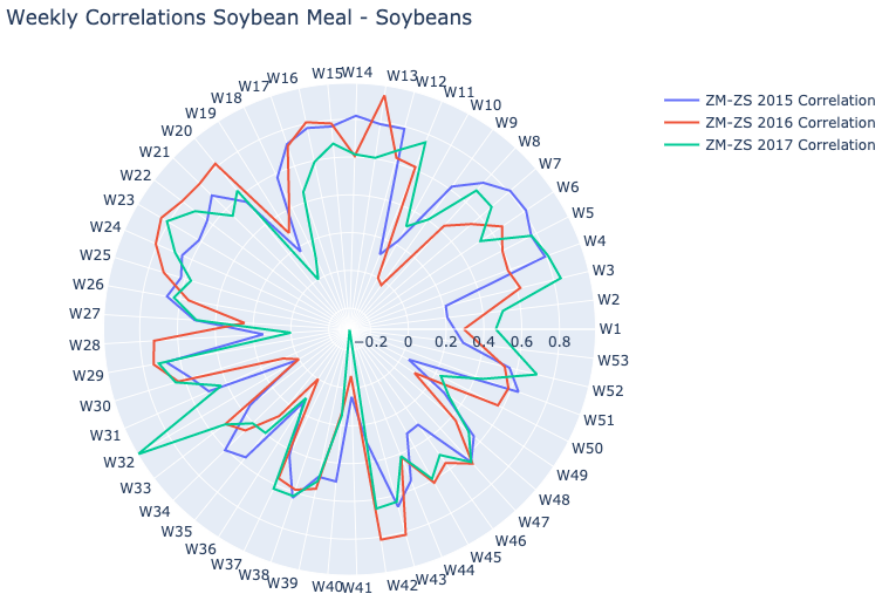

Figure 4: Weekly correlations of Soybean Meal and Soybean future contracts.

It is quite interesting to see how closely the three years behave, with high correlations lasting roughly five weeks at a time, taking a sharp dive for a period of roughly three weeks, and then returning to a high correlation again. This structure appears more apparent in the first six months of the year, but exists for all three years. This is the kind of information

that Belvedere can use to anticipate changes in their ability to trade certain futures. It also gives us insight and direction as data scientists to study such phenomena and use models to determine its causes.

We ultimately want to capture large "regime shifts" when correlations change significantly. We can instead plot the correlations differences from the previous week to see these correlation differences - effectively weekly derivatives. Below is a plot of this for Soybean Oil and Soybean Meal.

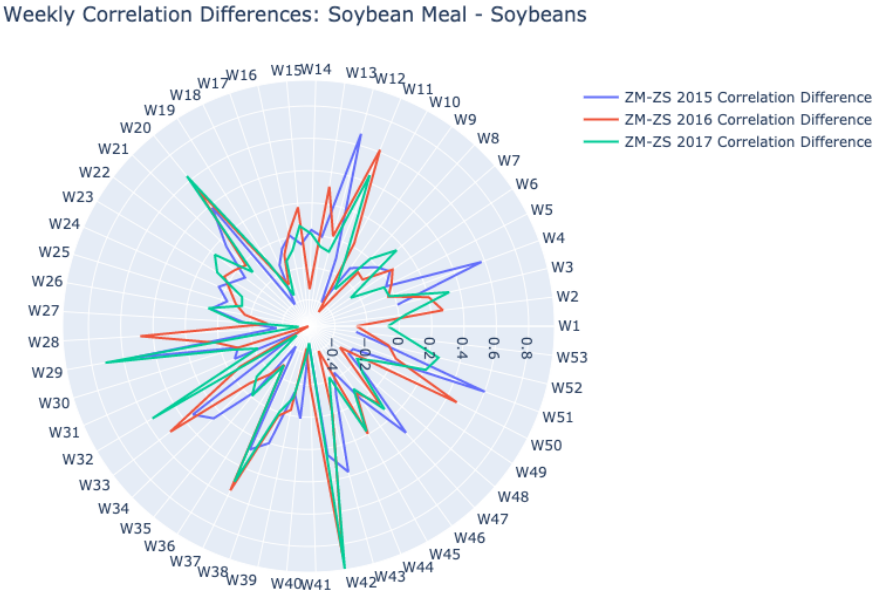

Figure 5: Weekly correlation differences of Soybean Meal and Soybean future contracts.

We are interested in predicting the large positive and negative spikes in this plot, when correlations jump from high to low, and vice versa. In our actual modeling, using this data will likely be more relevant.

However, this previous example of Soybean Meal and Soybeans serves as an "easier" modeling task due to its structural repetition. If we can train a model to capture this data well, we can move to predicting more irregular data. This may be even more useful to Belvedere, as trends shown in the plot above may easily be captured by less sophisticated modeling techniques at competing firms. An example of a more difficult modeling task is Soybean Oil and Soybean Meal. A polar plot of these weekly correlations is shown below.

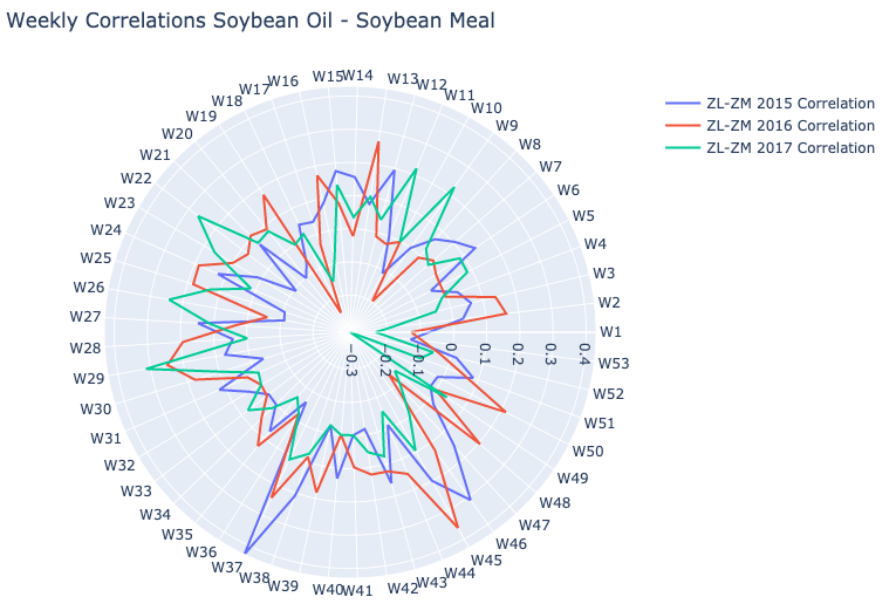

Figure 6: Weekly correlations of Soybean Oil and Soybean Meal future contracts.

There are very few structural similarities over the years in this example which will lend itself to a more difficult modeling task. If we can train a model to learn this type of situation well, we have effectively created a generalized model to learn on any set of similar data. This will allow Belvedere to apply our work to any futures they wish to trade.

So far, we have identified five groups of future contracts we would like to study. These are: Soybeans (Soybeans, Soybean Meal, Soybean Oil), Livestock (Lean Hogs, Feeder Cattle, Live Cattle), Energy (Crude Oil, Heating Oil, Natural Gas), Agriculture (Corn, Soybeans, Wheat), and Metals (Gold, Silver, Platinum). We have generated polar visualizations for all of the groupings to see and manually identify pairs of interest.

7 Feature Engineering

Feature engineering is the process of using domain knowledge to extract features from raw data in order to improve the performance of machine learning algorithms. [16] When analyzing price data, feature engineering is commonly employed to better represent the current market conditions, beyond merely the price of assets. For example, features can be engineered that capture the momentum, relative strength, and future volume of assets. Additionally, when analyzing the futures market, approximate interest rates can be derived by analyzing the price of particular futures, such as the Euro-Dollar or the 2 and 10 year treasury bonds.

7.1 TA-Lib

For this project, we are using TA-Lib [17], an open-source technical analysis library which is capable of generating over 150 features and indicators from the original data that Belvedere has provided. These include volume, strength, interest, and moving average indicators, among many others. After we have automatically generated our features, we can test each for their relationship to our target variables (correlation and correlation differences). This will ultimately allow us to try multiple modeling approaches incorporating different combinations of these features. After the modeling, we will use LASSO regularization (to be discussed in Section 8.3) to evaluate whether the feature has added considerable benefit to the model. Strong and specific features will be discussed in detail in the final report for the project.

8 Modeling

8.1 Hidden Markov Models

Hidden Markov models (HMM) are excellent tools to capture time dependencies and to make predictions for time series data. HMM's are statistical Markov models which are used to model state transitions when the output depending on the state is visible, but the states themselves are not directly visible to the observer. They are popular in applications in which the future states depend only on the current state but not on the events which occurred before it (that is, when the Markov property can be assumed).

HMM's are widely used as a prediction model for financial time series data, especially in stock market prediction. [18] Since financial prices depend on multiple market conditions which keep fluctuating, the time series is usually non-stationary. Due to the ability of HMM's to model dynamical systems, it is very useful in handling non-stationary financial time series data.

Therefore, our goal is to implement an Hidden Markov Model to estimate the movement of correlations in the future.

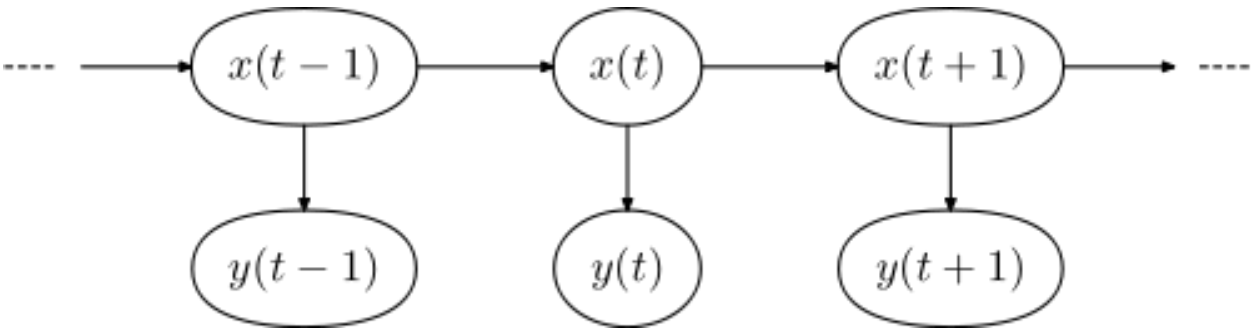

Figure 7: An example of a time-series Hidden Markov Model showing three states/observation variables.

In our first attempt to use Hidden Markov models, we attempt to predict weekly correlation changes of Soybean Meal and Soybeans. This was one of our "easy" modeling examples, and we would like our model to accurately predict these easy examples before moving to more complex ones. Below is a plot of the actual and predicted correlation changes for the weeks set aside as test data. It is also important to note that this model only uses prior correlation change data as features, rather than our original data or engineering features. Hence this model is more geared towards getting our "feet wet" with Hidden Markov Models and seeing a baseline of performance.

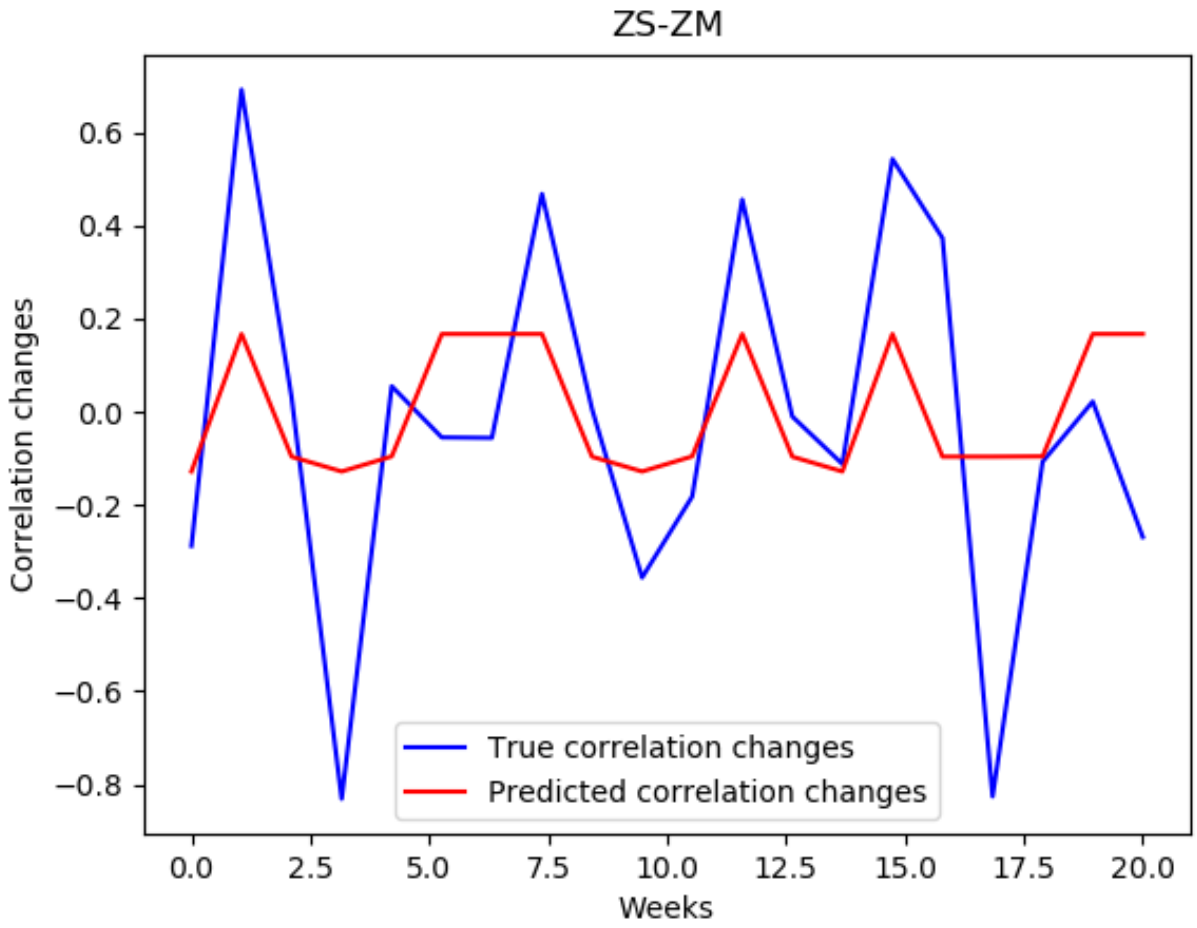

Figure 8: Actual and predicted correlation changes values for Soybean Meal and Soybean future contracts in our first attempt to use Hidden Markov Models.

The above model has a directional accuracy of 51.0%, an average squared error of 3.43, and a proximity accuracy of 35.0%. Clearly this model is not accurate enough for our purposes, but performs quite well given the lack of information it has available for training.

When the model is extended to predict correlation change values for all pairs rather than just Soybean Meal and Soybeans, we obtain even worse results. This is expected, as it includes the more difficult pairs as well as the easy modeling pairs. Below is a plot of the actual and predicted correlation change values for this model.

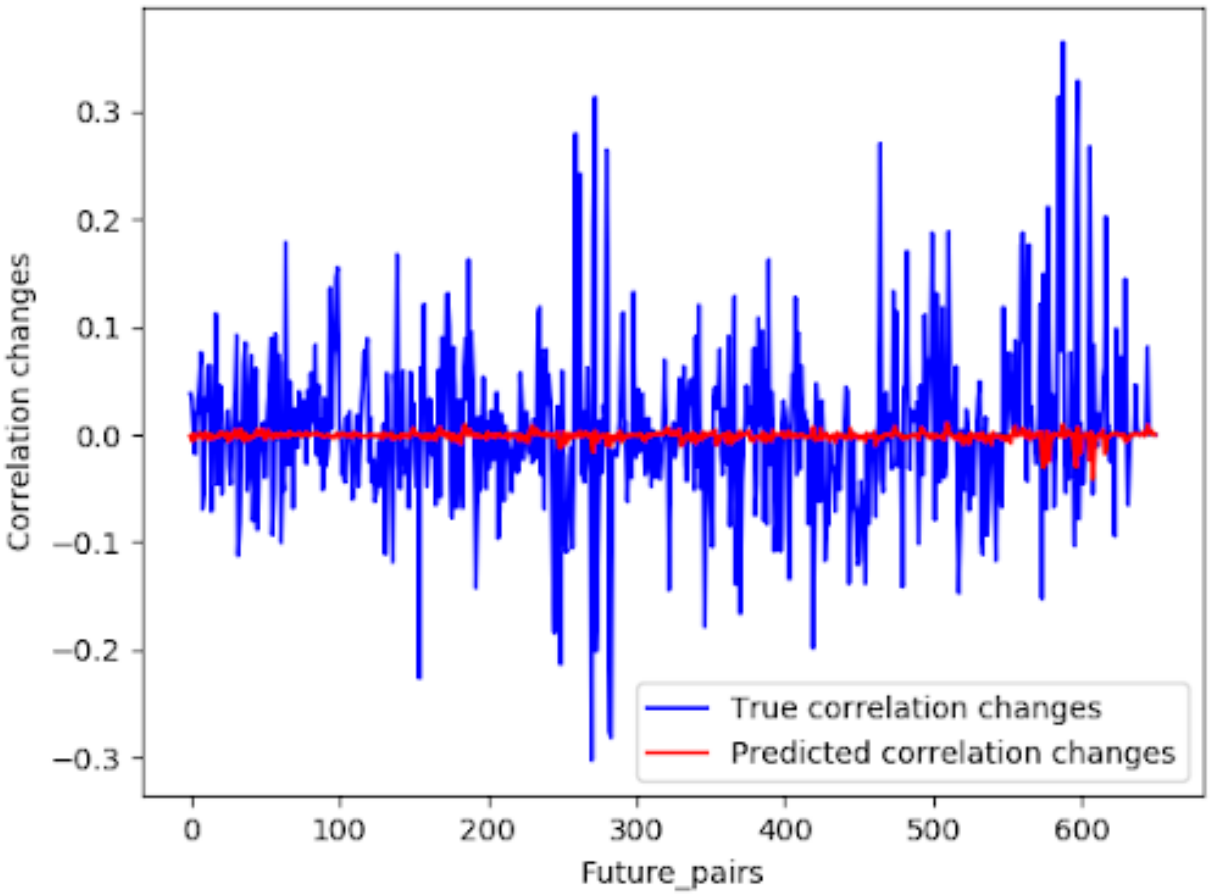

Figure 9: Actual and predicted correlation changes values for all pairs of future contracts.

Clearly, this model is not even close to being ready. We will need to incorporate far more features into the model to increase the accuracy of its predictions. However, this is still just a first attempt and gives us a start at a modeling framework using HMMs. It is also worth noting that for the objectives of the project, we do not seek to predict all correlation, correlation change, or price values simultaneously - rather only a few. Such a model using all pairs at once gives us a broad understanding of how well the model is performing across the board in one quick snapshot run.

8.2 Classification Models

An alternative approach to predicting correlations and understanding why they occur is to treat it as a classification problem. Using alternative approaches allows us to compare against other methods and models that we use. Specifically it allows us to compare against the HMM to see how well it can learn and predict correlations. Additionally, using recursive feature elimination, we can select top features from these models and compare against our other methods for doing this.

To set up the classification problem, we bucket potential correlation ranges into n-components and assign labels to these components. There are three methods for component binning that we tested including uniform splits on the minimum and maximum possible correlation range (-1.0, 1.0), uniform splits on the minimum and maximum

correlation values for the test set, and using an algorithm which attempts to match the value distribution and assign an equal number of labels to each bin.

Additional hyperparameters include the time interval to use (weekly, biweekly, monthly, seasonal), target (correlation or correlation differences from the previous interval), the option to train the model to predict this interval's value or the next interval's value, which features to include (and whether or not to use original data), and finally the model type. We tried seven different models - Naive Bayes, Support Vector Machine (with Radial Basis, Polynomial, and Sigmoid kernels), Random Forest, Gradient Boost, and Decision Tree. These models are all built into the popular data science package Scikit-Learn, which we used for this task. Additionally, Scikit-Learn supports recursive feature elimination for the last three models, allowing us to glean top features when using these model types.

We train these various models and hyperparameter values using data and label pairs, and see how well they perform using cross validation on the test set.

For our project objectives, it was important to understand how using features generated by just the target pair vs. using features generated by the target pair as well as additional futures compare to each other. For example, we chose the modeling task of predicting ZM-ZS correlations for the current interval, split into 3-components and using matched distribution to assign labels. We use original data from the ZM and ZS futures, as well as features generated from both of them. In a separate modeling test, we use the same exact parameters, but include features generated from ZC (Corn), ZL (Soybean Oil), and ZW (Wheat) futures, which are all agriculture products and could show Granger causality with respect to our target pair. Below is a table showing the 10-fold cross validation accuracy for the tasks using both weekly and monthly time intervals.

Classification Model	Task Accuracy: ZM-ZS Only		Task Accuracy: ZM-ZS + ZL, ZC, ZW Features	
-	Weekly	Monthly	Weekly	Monthly
Naive Bayes	0.41 (+/- 0.30)	0.53 (+/- 0.31)	0.41 (+/- 0.30)	0.53 (+/- 0.31)
Radial Basis SVM	0.34 (+/- 0.01)	0.35 (+/- 0.01)	0.34 (+/- 0.01)	0.35 (+/- 0.01)
Polynomial SVM	0.39 (+/- 0.25)	0.57 (+/- 0.39)	0.39 (+/- 0.25)	0.57 (+/- 0.39)
Sigmoid SVM	0.34 (+/- 0.01)	0.35 (+/- 0.01)	0.34 (+/- 0.01)	0.35 (+/- 0.01)
Random Forest	0.41 (+/- 0.24)	0.57 (+/- 0.41)	0.44 (+/- 0.31)	0.62 (+/- 0.36)
Gradient Boost	0.39 (+/- 0.27)	0.61 (+/- 0.37)	0.39 (+/- 0.28)	0.61 (+/- 0.37)
Decision Tree	0.40 (+/- 0.16)	0.63 (+/- 0.44)	0.38 (+/- 0.21)	0.62 (+/- 0.44)

Table 1: 10-Fold cross validation accuracy for various modeling tasks and model types.

As you can see from these results, there is no significant increase or decrease in task accuracy when features from more supposedly related futures are introduced. In general, monthly accuracy is also much better than weekly accuracy. Gradient Boost and Decision Tree appear to be the most accuracy models across the trials, but only by small margins.

This information is potentially quite valuable to Belvedere trading for several reasons. Showing little to no Granger causality between non-target pair futures (for example - ZC, ZL, and ZW) and predicting correlations for the target pair (ZM-ZS). If this is confirmed via our other models, Belvedere could save vast computational resources by excluding these additional futures in a given model. Additionally, given that weekly predictions are quite weak, whereas monthly predictions are normally more correct than incorrect, Belvedere would likely decide to build their trading strategy based on predicting longer-term correlations rather than shorter ones. Finally, while determining specific classification

models to use is not a main objective, it does not provide some starting insight into what may or may not work well for them as a firm.

8.2.1 Ensemble Feature Selection

In addition to using classification models to predict correlations, they can also be used to determine feature importance. Since this is a primary objective of our project, we wanted to take a deeper dive into this aspect of the models. To do so, we used Scikit-Learn's built-in recursive feature elimination method which is supported in the Random Forest, Gradient Boost, and Decision Tree models. We took combinations of these model types, prediction targets (this or next time interval), number of components, and type of label assignment to create many different modeling tasks with various hyperparameter variations. We then used recursive feature elimination to determine the top 10 features for each run, and added them to a running count of features, counting the frequency that they appeared as top features across the runs. This type of modeling is often referred to as ensemble modeling, where several different models are informing one final decision/understanding. Below is a table of results, analogous in task variation to the table showing task accuracy. The top 10 features are those with the highest frequency of selection by recursive feature elimination for the specific task/time interval pairing.

	Task Accuracy: ZM-ZS Only		Task Accuracy: ZM-ZS + ZL, ZC, ZW Features	
	Weekly	Monthly	Weekly	Monthly
Top 10 Features	'ZM_TRIX', 'ZM_ADXR', 'ZS_TRIX', 'ZS_ADXR', 'ZS_ADX', 'ZM_ADX', 'ZS_PPO', 'ZM_PPO', 'ZM_APO', 'ZS_APO'	'ZM_TRIX', 'ZS_TRIX', 'ZS_ADXR', 'ZM_ADXR', 'ZS_OBV', 'ZS_ADX', 'ZM_ADX', 'ZS_CLOSE', 'ZM_APO', 'ZM_PPO'	'ZS_TRIX', 'ZC_TRIX', 'ZW_TRIX', 'ZM_TRIX', 'ZM_ADXR', 'ZW_ADX', 'ZS_ADXR', 'ZL_TRIX', 'ZW_PLUS_DM', 'ZS_ADX'	'ZM_TRIX' 'ZW_TRIX', 'ZC_TRIX', 'ZL_TRIX', 'ZS_TRIX' 'ZM_ADXR', 'ZS_ADXR', 'ZS_ADX', 'ZL_OBV', 'ZW_ADXR'

Table 2: Top 10 features selected by ensemble feature selection for various modeling tasks and time intervals.

Across all four task/time interval combinations, three features were selected: ZM TRIX (Triple Exponential Average), ZS TRIX , and ZS ADX (Average Directional Movement Index). In the target-pair only task, five additional features consistently occurred: ZM ADXR (Average Directional Movement Index Rating), ZS ADXR, ZM ADX, ZM PPO (Percentage Price Oscillator), and ZM APO (Absolute Price Oscillator). In the target-pair plus additional features runs, five features also commonly occurred: ZC TRIX, ZL TRIX, ZW TRIX, ZM ADXR, and ZS ADXR.

These results show a consistency in which indicators/features are most important in predicting correlations. Belvedere can make great use of this information, building specific models tailored around these indicators. It may also be worth taking a deeper dive into these indicators to figure out why exactly this relationship may occur. For example, most of these indicators are known as oscillator indicators, and their importance in learning

this specific model may come from the structural repetition that was identified in the exploratory data exploration for the ZM-ZS pair. Taking a deep dive into these technical indicators would be somewhat outside of the scope of this project and our objectives, but future work could be done to see if these indicators are consistent across other modeling tasks, and to understand exactly why they may be important.

Another very curious result is the feature selection of ZL, ZC, and ZW TRIX in the target-pair plus additional futures modeling tasks. In the last section, we determined that the addition of additional futures does not increase accuracy, so one would presume features generated from these additional futures are not important in the models. However, these results show that they are quite important and often some of the top selected features. This leads to the understanding that while these features can be of importance in prediction, they do not help increase prediction accuracy. This is a very interesting and unusual result, and future work should be done to understand why this is the case.

8.3 LASSO

LASSO (least absolute shrinkage and selection operator) [19] is a regression analysis model for variable selection and regularization. It allows for better feature selection, understanding of the underlying model, and overall performance. By using LASSO with our Hidden Markov Model [20], we can generate a more useful to predict correlations and prices as well as understand what features impact the model most. This will give Belvedere a more useful overall end-product, and drive final modeling insights for us data scientists.

8.3.1 LASSO with Classification Models

8.3.2 LASSO with Hidden Markov Model

9 Bibliography

References

[1] Belvedere trading.

[2] James Chen. Futures contract definition, Jan 2020.

[3] A. Gupta and B. Dhingra. Stock market prediction using hidden markov models. In *2012 Students Conference on Engineering and Systems*, pages 1–4, March 2012.

[4] Md. Rafiul Hassan, Baikunth Nath, and Michael Kirley. A fusion model of hmm, ann and ga for stock market forecasting. *Expert Systems with Applications*, 33(1):171 – 180, 2007.

[5] Md. Rafiul Hassan, Baikunth Nath, Michael Kirley, and Joarder Kamruzzaman. A hybrid of multiobjective evolutionary algorithm and hmm-fuzzy model for time series prediction. *Neurocomputing*, 81:1 – 11, 2012.

[6] Sonam Srivastava and Ritabrata Bhattacharyya. Evaluating the building blocks of a dynamically adaptive systematic trading strategy. *SSRN Electronic Journal*, 2018.

[7] James D. Hamilton. Regime switching models. *The New Palgrave Dictionary of Economics, 2012 Version*.

[8] Yung-Keun Kwon, Sung-Soon Choi, and Byung-Ro Moon. Stock prediction based on financial correlation. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, GECCO '05, page 2061–2066, New York, NY, USA, 2005. Association for Computing Machinery.

[9] John Y. Campbell, Sanford J. Grossman, and Jiang Wang. Trading Volume and Serial Correlation in Stock Returns*. *The Quarterly Journal of Economics*, 108(4):905–939, 11 1993.

[10] Gunter Meissner. Correlation trading strategies: Opportunities and limitations. *The Journal of Trading*, 11(4):14–32, 2016.

[11] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo. Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3):1014–1020, Aug 2003.

[12] G.Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159 – 175, 2003.

[13] Ping-Feng Pai and Chih-Sheng Lin. A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497 – 505, 2005.

[14] Nguyen Hoang Hung and Yang Zhaojun. Profitability of applying simple moving average trading rules for the vietnamese stock market. *J. Bus. Manag*, 2(3):22–31, 2013.

[15] Julián Andrada-Félix and Fernando Fernández-Rodríguez. Improving moving average trading rules with boosting and statistical learning methods. *Journal of Forecasting*, 27(5):433–449, 2008.

[16] Machine learning and ai via brain simulations. *Stanford University*, Aug 2019.

[17] Ta-lib.

[18] Stock market forecasting using hidden markov model: a new approach. *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pages 192–196, 2019.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[20] A regularized hidden markov model for analyzing the "hot shoe" in football. pages 192–196, 2019.