2023

A Machine Learning Approach to Sector Based Market Efficiency

Angus Zuklie
Bowdoin College

A Machine Learning Approach to Sector Based Market Efficiency

An Honors Paper for the Department of Computer Science

By Angus Zuklie

Bowdoin College, 2023

I believe the market accurately reflects not the truth, which is what the efficient market hypothesis says, but it accurately and efficiently reflects everybody's opinion as to what's true.

Howard Marks

Institute, Investments & Wealth, Capturing Inefficiencies: The Rare Insight of Howard Marks (April 1, 2016). Journal of Investment Consulting, Vol. 17, no. 1, 4-10, 2016, Available at SSRN: https://ssrn.com/abstract=2792274

Dedicated to all the staff of Bowdoin College, from the health services folks and the café baristas to the library staff and landscapers. Bowdoin is a complex machine that runs so smoothly due to the care and positive energy of many committed employees. My time at Bowdoin has been enriched by everyday interactions with this wonderful community.

ACKNOWLEDGMENTS

I would like to thank the members of my thesis committee for their help in preparation of this work – Karen Jung, who personally sat down to inform all honors students the submission procedures and guidelines, Carmen Greenlee, for her work organizing copyright guidance, Meredith McCarroll, for informing the requested writing rhetoric and presentation of project.

This project would not have been possible without the tremendous support and guidance from Professor David Byrd. As we met throughout the year and he not only patiently answered my questions but also pushed me to think deeply about this topic of research from many different angles.

Special thanks are due to the friends and colleagues who made this work possible. Josh-Pablo Patel, Tara Mullen, and J.T. Wooley were all invaluable as both friends and classmates, helping me find time for my work on this project by joining me as group mates and lab partners for the rest of my course load. I am grateful to my partner Jenny for her constant support and willingness to listen to my neural network ramblings at all hours of the day.

Finally I would like to thank the facilities of Bowdoin, without Bowdoin's Slurm HPC Cluster this project would not have been possible for such a short time frame.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AI Artificial Intelligence

CNN Convolutional Neural Network

DVN Devon Energy Corp

EMH Efficient Market Hypothesis

ETF Exchange Traded Fund

LOBSTER Limit Order Book Reconstruction System

LSTM Long Short Term Memory

ReLU Rectified Linear Activation Unit

RNN Recurent Neural Network

SPDR Standard and Poor's Depositary Receipts

UAT Universal Approximation Theorem

ABSTRACT

In economic circles, there is an idea that the increasing prevalence of algorithmic trading is improving the information efficiency of electronic stock markets. This project sought to test the above theory computationally. If an algorithm can accurately forecast near-term equity prices using historical data, there must be predictive information present in the data. Changes in the predictive accuracy of such algorithms should correlate with increasing or decreasing market efficiency.

By using advanced machine learning approaches, including dense neural networks, Long Short Term Memory (LSTM), and Convolutional Neural Network (CNN) models, I modified intra day predictive precision to act as a proxy for market efficiency. Allowing for the basic comparisons of the weak form efficiency of four sectors over the same time period: utilities, healthcare, technology and energy. Finally, Within these sectors, I was able to detect inefficiencies in the stock market up to four years closer to modern day than previous studies.

CHAPTER 1
INTRODUCTION AND RELATED WORK

1.1 Research Question

The weak form of the long-standing Efficient Market Hypothesis (EMH) asserts that the value of assets cannot be predicted from past sequences of price and volume. This work analyses how our understanding of stock market efficiency through the application of machine learning tools. The EMH is a widely used and accepted theory in the economic field; however, we do not know if it exists as a spectrum or a blanket statement that can be applied accurately over the entirety of a market. This research builds off the work produced by Byrd et al. in August of 2019[1], which found convincing evidence that market efficiency may not be consistent over time. The authors however, only analyzed the S&P 500 monolithically, and did not address any potential disparities across further sub divisions of the market. The focus of this work is whether markets have inconsistent efficiencies across the dimension of sectors when measured within the same time periods.

1.2 Introduction

The United States stock market is a vast and complex system, composed of various sectors, each with its own unique characteristics and behaviors. Understanding these sectoral differences is crucial for investors looking to build a diversified portfolio and manage risk effectively. The different sectors analyzed in this project are summarized below to illustrate the ways people interact with these complex systems and how behavioral variation may affect sectoral efficiency.

The first sector, utilities, consists of stocks that tend to be stable and predictable, with relatively low volatility and a steady stream of income. Companies in the utilities sector typically operate in regulated and mature markets, making them less sensitive to outside influences[2]. For this reason they are often sought out by investors with low risk tolerance.

The next sector, healthcare, is driven by emerging technologies, making it more volatile than utilities as well as more lucrative[2]. However, due to regulatory processes and required testing, the healthcare market has a natural limit to its volatility.

The technology sector is often associated with rapid growth and disruption, resulting in high volatility and the potential for significant gains or losses. While technology stocks have been among the top-performing stocks in recent years, they are also subject to rapid changes obsolescence and shifting consumer preferences, making them susceptible to boom-and-bust cycles[3]. Stocks in this sector are preferred by most noise traders.

Energy is a sector that is sensitive to global economic conditions such as oil prices. While energy stocks can offer high returns, they can also be vulnerable to market downturns tied to geopolitical risk[4].

We hypothesise that the differing interactions between people and various market sectors effect their volatility and liquidity, possibly leading to varying efficiency.

In recent years, Artificial Intelligence (AI) and machine learning have radically changed many fields including medicine and education. Applying technology and AI to novel spaces can have significant effects on how we live, work, and interact with our environment. By leveraging the power of these tools in novel ways, we can find new solutions to old problems and unlock opportunities for innovation and growth. The research underlying this paper uses the sequence predicting ability of AI to create a novel quantification of efficiency that can be used to compare a market across sectors and time.

1.3 Background and Related work

The EMH is a theory that states that financial markets are ”informationally efficient,” [5] meaning that asset prices already reflect all available information about the underlying asset. According to the EMH, it is impossible to consistently achieve above average returns in the stock market through active trading or stock picking, because all available information is already priced into the stock.

There are three forms of the EMH: weak, semi-strong, and strong. The weak form of the EMH states that current stock prices already reflect all past market data and prices, including past stock prices and trading volume. The semi-strong form of the EMH states that stock prices already reflect all publicly available information, including financial statements, news releases, and other data that is widely disseminated. The strong form of the EMH states that stock prices already reflect all information, whether public or private. This paper will focus solely on the weak form of the EMH, and the notion that "Future equity prices cannot be predicted from past sequences of price and volume."[5]

The Universal Approximation Theorem (UAT) is a mathematical theory that states that a neural network with a single hidden layer can approximate any continuous function, given enough hidden units. In other words, a neural network with just one hidden layer and a sufficient number of hidden units can be used to model any arbitrary function, to any desired degree of accuracy. The UAT has significant implications for the design and training of neural networks, as it suggests that a simple neural network architecture is capable of recreating any function. However, it is important to note that in practice a neural network is not guaranteed to learn a given function due to programmer-based user error. The success of a neural network depends on a variety of factors, including the quality and quantity of training data, the choice of activation functions, and the optimization algorithm used during training[6].

With the combination of these two ideas we have a novel way to validate the EMH. In other words, if there is any predictive information in past equity prices of a stock, a neural network should be able to recreate the function to predict the unknown future price as per the UAT.

Byrd et al. applied this ideas in their work, *Intra-day Equity Price Prediction using Deep Learning as a Measure of Market Efficiency*, which found that a neural network could predict intraday equity prices at a rate significantly better than random chance until 2008 when the predictions became too noisy to rely on[1]. This paper builds off the work of Byrd et al. aby looking across sectors to further refine the axes of the market and illuminates further inconsistencies through time.

CHAPTER 2

METHODOLOGY

2.1 Data Sourcing and Collection

The process of data collection began with an analysis of stock market delineators. This paper will create resolution within time frames by dividing the market into its trading sectors. Significant sectoral delineations were selected by following the Standard and Poor's Depositary Receipts (SPDR) Exchange Traded Fund (ETF) portfolios. These portfolios are subsets of the S&P500. The S&P500 is a rolling list of the top 500 companies in United States stock exchanges ranked by Market Capitalization which is calculated with the following equation:

$$M_x = P_x * N_x \tag{2.1}$$

Where M = market capitalization, P = price of each share, N = number of publicly traded shares, and x = the stock being observed. The selected sectors were represented by the ten stocks with the highest market capitalization. From the most recent publishing of these portfolios to create each data set. The following sector ETFs were used to represent each sector.

Table 2.1: Selected Sectors for Comparison and Sourced Sector Representatives.

Sector	Representative ETF
Utilities	XLU Select Sector SPDR Fund
Technology	XLK Select Sector SPDR Fund
Healthcare	XLV Select Sector SPDR Fund
Energy	XLE Select Sector SPDR Fund

In the interest of computational constraints the selected stocks were then received as limit order book snapshots from Limit Order Book Reconstruction System (LOBSTER). Data used for the research conducted in this paper was compressed to a frequency of one data point per minute before being passed to the training models to reduce noise and focus on daily trends.

2.2 Data Prepossessing

LOBSTER delivers data in a limit order book format. The Research conducted in this Thesis required the use of appropriately 120GB of raw CVS data starting in 2007 and running until 2021 for each requested stock in the following format:

Table 2.2: Raw LOBSTER Data Sample.

Level 1 Ask Price	Level 1 Ask Volume	Level 1 Bid Price	Level 1 Bid Volume
2239500	100	2231800	100
2239500	100	2238100	21
2239500	100	2238100	21
...	...	...	...

At our chosen order book depth of 1 layer, only the most competitive bid and ask prices are reflected. We estimate the value of an asset at each timestamp and frequency by taking the midpoint of the lowest and highest buy and sell offers respectively. We compiled the sectors into separate files. As shown below, each column now represents one of the ten stocks in the requested sector at the desired time interval.

Table 2.3: Pre-Processed LOBSTER Data Sample.

TIME	COP	CVX	DVN	EOG	OXY	...
2007-06-27 09:30	754050.0	824550.0	757550.0	717350.0	551000.0	...
2007-06-27 09:31	748200.0	822150.0	759450.0	717350.0	551700.0	...
2007-06-27 09:32	747500.0	818100.0	758750.0	719400.0	551450.0	...
2007-06-27 09:33	748250.0	817750.0	758950.0	718900.0	550950.0	...
...	...	...	...	...	...	...

In order to improve generalization, each day was independently standardized to converge at zero for the moment of prediction. This was achieved by dividing the price calculated at each time step by the price of the asset at the time of the prediction then subtracting one from all values using the following equation:

$$\forall t_i \in [t_1, t_2, ..., t_{predict}, ..., t_{close}] : t_i = (t_i/t_{predict}) - 1 \tag{2.2}$$

In this way, the agent can use the intraday history of relative price changes up to the prediction to predict the relative change between the time of prediction and closing time of the day.

2.3 Approach

The experiment used three learning models: Multi Layer Dense, LSTM, and CNN. Each model was given the opportunity to make predictions across the four selected sectors, with two controllable hyper parameters: the time of day at which the model will make a prediction, and the minimum return required to consider a prediction significant.

Multi Layer Dense: In this experiment, a Multi Layer Dense model will begin by feeding the known price history into the first layer of the neural network. This layer is called the input layer, and it consists of a series of neurons that are fully connected to the input data. The input layer will pass its output into a series of hidden layers. Each hidden layer consists of densely connected neurons that process the data from the previous layer. Each neuron applies a non-linear Rectified Linear Activation Unit (ReLU) activation function to the weighted sum of its inputs to generate an output. A ReLU function is a function with a non-negative cut off. For example:

$$f(x) = max(0, x) \tag{2.3}$$

After passing through all of the hidden layers, the forecast will be returned in the output layer[7].

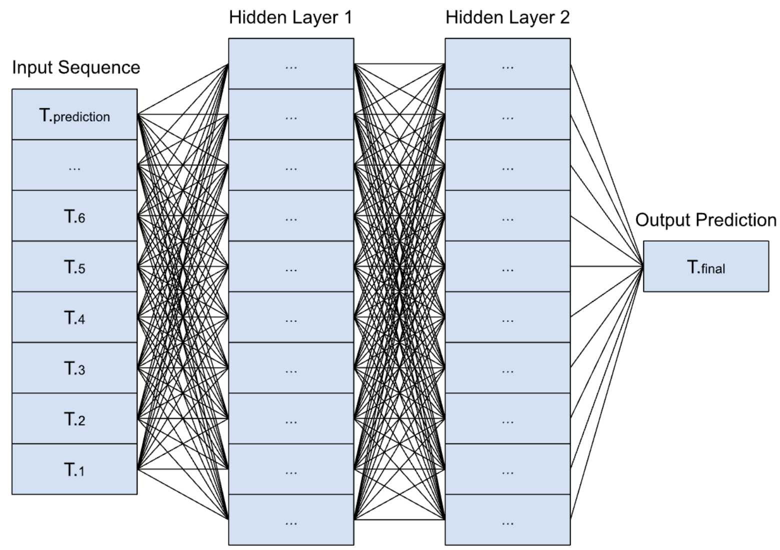

Figure 2.1: Depicting the dense nature of the hidden layers, each layer has output to every layer in the next hidden layer.

LSTM: LSTM is a type of recurrent neural network that is commonly used for sequence-to-self learning tasks. The main idea behind LSTM is to address the vanishing gradient problem that can occur in traditional Recurent Neural Network (RNN), which can make it difficult for them to learn long-term dependencies in sequential data. At a higher level, an LSTM model consists of a sequence of LSTM cells that are connected in a chain. Each LSTM cell has three gates (input, forget, and output gates) and a memory cell, which allow the cell to selectively read, write, and forget information from the input sequence as it passes through the network[8].

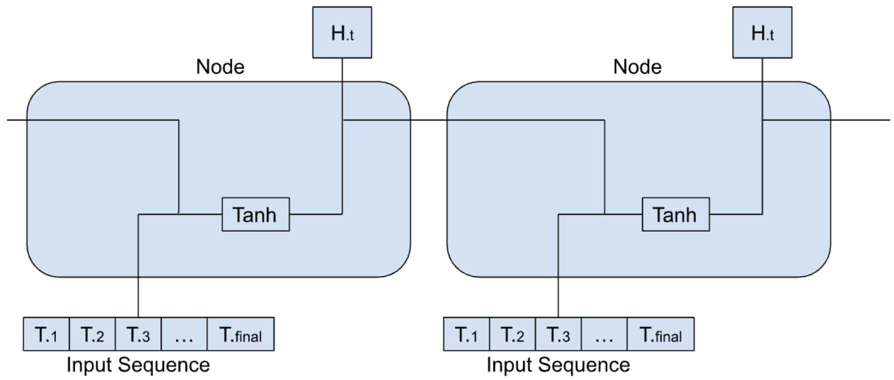

Figure 2.2: Depicting how each node of the LSTM model has a forget gate in addition to its input and outputs.

Convolutional: A CNN is a type of neural network often used for image processing. However, by using a 1D convolutional filter and sliding it over the input sequence to compute the weighted sum of the temporal feature values in a small window, and then applying a ReLU activation function to the result, the CNN can be adapted to our data set nicely.

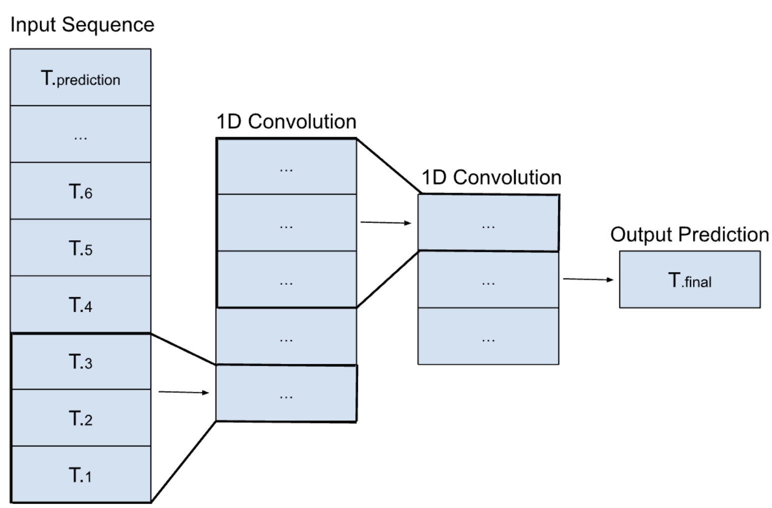

Figure 2.3: Depicts how each window of the CNN model is able to learn from not only the inputs surrounding it but also the surrounding convolutional layers.

The same trained model was used to make all final predictions for one sector for one month. For month m, four models would be trained on intraday data for months m-13 to m-

2, one model per daily prediction time. The four models would be evaluated on validation month m-1 and only the best-performing model used to predict results during month m. After all predictions are made, m slides forward one month and the entire process repeats. This allows the algorithm to select the best prediction time for each month using recent data, without leakage of future data[9].

Figure 2.4: Displaying the price of the asset Devon Energy Corp (DVN). Here the green lines delineate the sections in which our agents switch modes; the first section is training, the second section is validation and the final section is testing. After completing the test time section, the agent will restart the whole process from the red dotted lines. This is the same pattern slid forward in time one month. (In this example time period, the model is set up to make predictions just at the end of the 2008 financial crisis).

CHAPTER 3

RESULTS

3.1 Models

In the interest of time, the different architectures were evaluated on a temporal subset of the data, only the highest preforming architecture was used to run the final experiments of the work. Each architecture was run over the first 12 test months, the losses were averaged to select the highest preforming architecture.

Table 3.1: Selected Model Loss Comparisons.

Model	Out of Sample Loss
Multi-Layer-Dense	0.009
Long Short Term Memory	0.011
Convolutional	0.015

This Table displays model loss compared across the various models. The values depict average mean squared error loss over the 12 month test period using the highest preforming hyper parameters each month. Below x and y are D dimensional vectors, and x_i denotes the value on the ith dimension of x used to calculated the mean squared error for the table above.

$$\sum_{i=1}^{D}(x_i - y_i)^2 \tag{3.1}$$

Because of the structure of the CNN the model was making predictions through tout the day where as the Multi-Layer Dense and LSTM models only making predictions at the end of the day. For this reason, it seems likely that predictions closer to the end of the day are easier to make than those near the beginning and middle when using intra day data sets. Between Multi-Layer Dense and LSTM, Multi-Layer Dense slightly out performed LSTM and will be the model used for the remainder of the experiments in this paper.

3.2 Hyper Parameters

This section will summarize the effects of applying hyper parameters to the Multi-Layer Dense model making predictions in the four sectors: Healthcare, Energy, Technology, and Utilities.

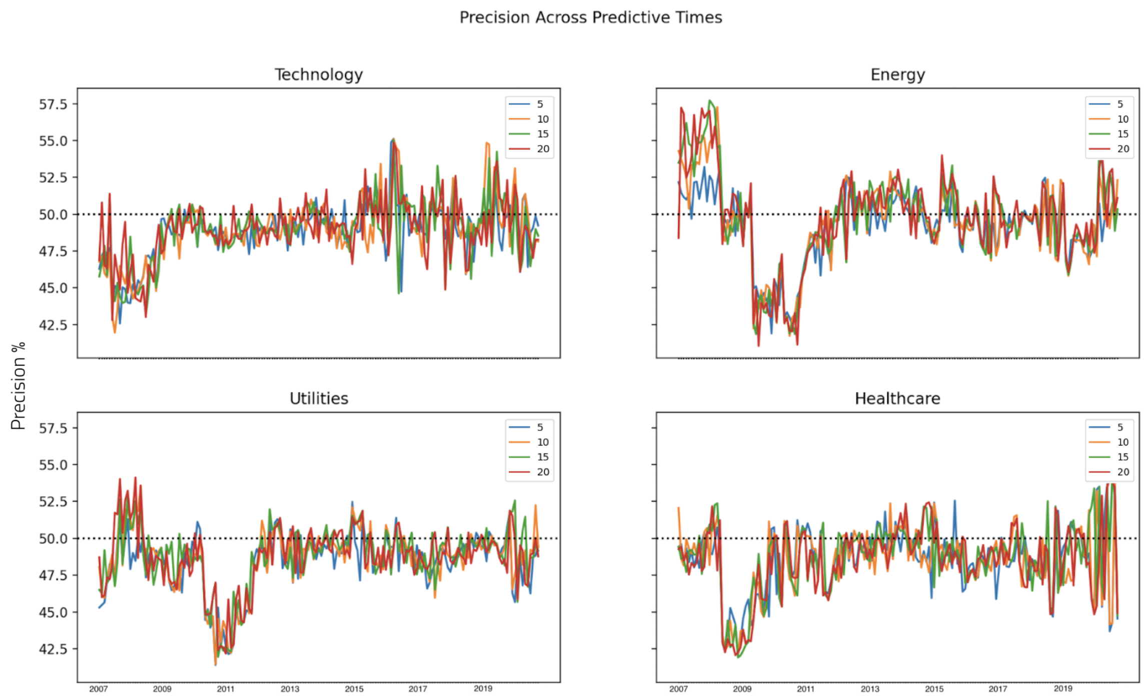

Figure 3.1: Displaying the precision of the most precise model: Multi-Layer Dense. (The first-degree hyper parameters are the prediction times before market close: [5 min, 10 min, 15 min, 20 min]

Figure 3.1 displays all prediction time hyper parameters applied to each stock. While the sectors have the most impact on the overall shape of the lines, the 20-min and 15-min prediction times consistently out perform the 5-min and 10-min prediction times. This may be caused by the increased noise present when predicting a final outcomes relative position closer to one's own. When predicting from further away more of the general trend may be present in the close prices relative change.

The prediction significance thresholds were also assessed, however were not plotted as that hyper parameter had no significant effect on the predictive accuracy of the models.

3.3 Results

This section analyses the ability of the most precise model to search for meaningful differences across the dimension of sectors. Each month used in these charts was tested using the hyper parameter which performed best in its validation period.

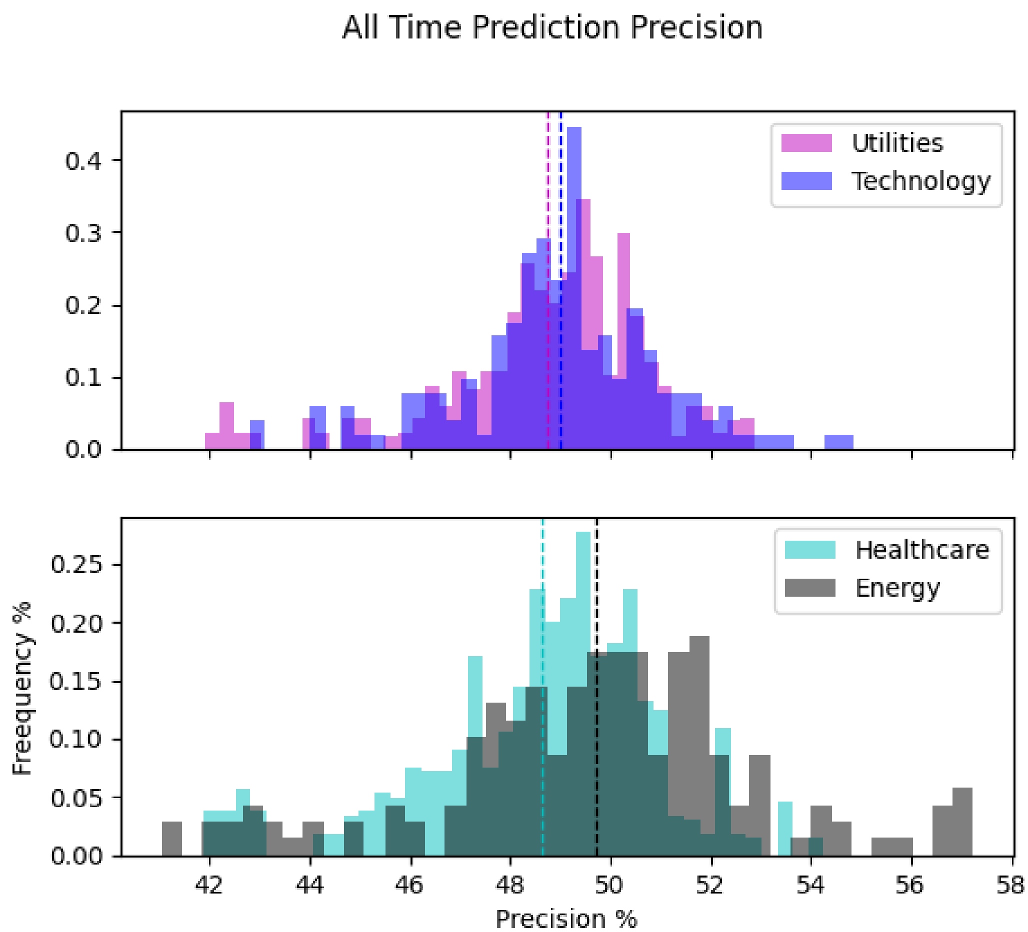

Figure 3.2: Displaying the precision of the most precise model, Multi-Layer Dense across all sectors. The dashed line represents their respective means.

In the above display, we can identify that the predictions of Utilities, Healthcare, and Technology have a closer distribution to their respective means and are heavily left skewed. Energy has a smaller skew running slightly right, and a further average distribution from its mean.

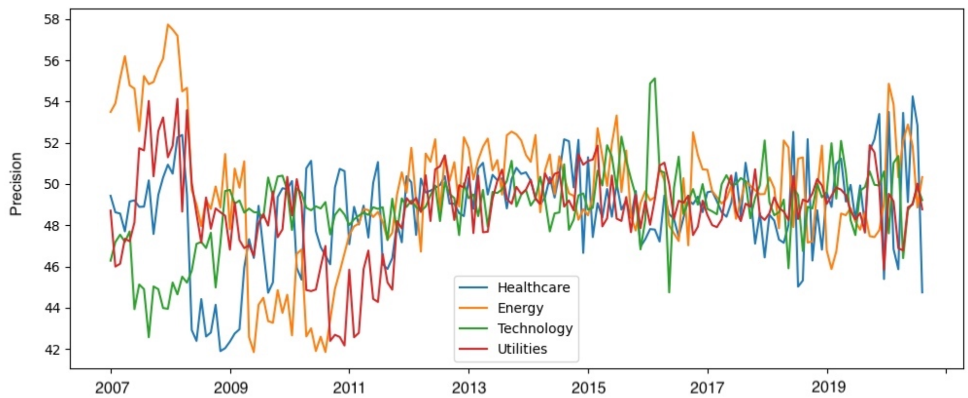

Figure 3.3: Displaying the precision of the most precise model, Multi-Layer Dense predicting the closing price of sectors Healthcare, Energy, Technology, and Utilities from 2007 until 2021.

While the sector lines in the plot above quickly converge to the expected value of randomly guessing up or down for the end of the day, the first four years on the graph look very different. By regenerating the histogram shown in Figure 3.2 using only the earliest 24 months of data we derive the Figures below.

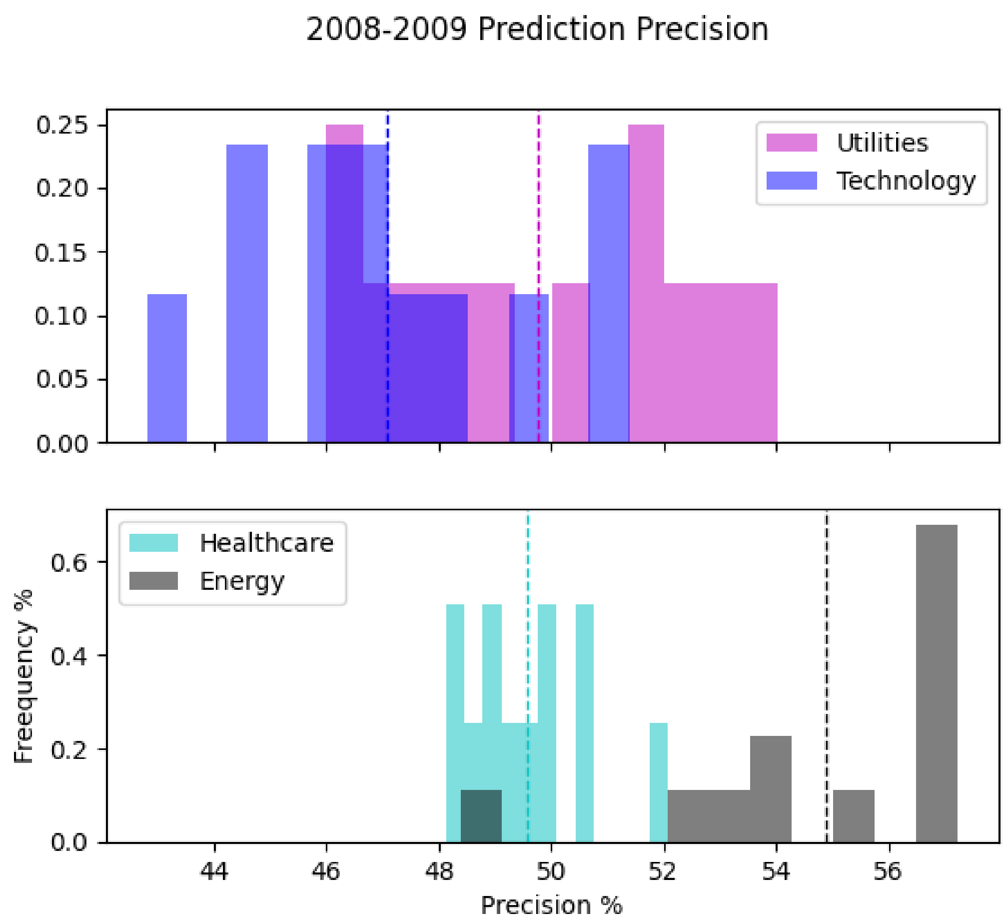

Figure 3.4: Displaying the precision of the most precise model, Multi-Layer Dense, predicting the different sectors using only data from 2008 - 2009. Hashed lines represent respective means.

While there remains room for interpretation, given the limited number of data points available before the precisions shown in Figure 3.3 converge, the distributions shown in Figure 3.4 of the energy sector in particular appear less normal than those in Figure 3.2. This, along with the further spread of the means from the two data sets, seems indicative of a larger difference between the efficiency of the four sectors.

3.4 Interpretation

A significant limitation in this research is lack of access to older data through LOBSTER, particularly during the time periods when the models should have preformed best.

However, preliminary data suggests that the window seen from 2008 to 2012 in Figure 3.4 allows each of the sectors selected for analysis to tell a unique story and indicates that further subdivision of the market could allow for further resolution when assessing market efficiency. The limitations regarding data may mean that the data gathered in this research may not be sufficient for creating a general use market efficiency quantification tool, but comparisons of the discussed sectors are still informative.

Assuming a perfectly efficient market, it should be just as hard to be consistently incorrect when trying predicting the market as consistently correct. This is because if ones predictions were consistently wrong over a long period of time this too would indicate the existence of predictive information in the market. Because of this, the equation formulated to represent market inefficacy for the use of sector comparison is double sided. In the following equation P_s^e represents the adjusted comparable precision and P_s represents the raw predictive precision of a sector s.

$$P_s^e = (50 - P_s * 100)^2 \tag{3.2}$$

This non-linear scaling function captures deviations of the model's precision from fifty percent, or the expected performance of randomly guessing. By squaring this difference, the formula not only captures both sides of inefficiency, but this also allows for a higher weight to be applied to a sector with a higher precision. Note that the adjusted precision P_e now represents inefficiency. Therefore zero represents a perfectly efficient market.

Table 3.2: Adjusted Predictive Precision as a Proxy for Efficiency.

Sector	2008-2021	2008-2009
Utilities	1.58	6.72
Technology	1.32	6.50
Healthcare	1.78	0.24
Energy	0.13	13.15
Market	1.07	0.25

Looking across the table in the 2008-2009 column, the data suggests that healthcare is the most efficient sector, and is only slightly more efficient than the market overall.

CHAPTER 4
CONCLUSION

According to the EMH a trade should only occur on the stock market in one of two scenarios. The first is in economic surplus, when both parties (the buyer and seller) believe that they are profiting from the trade due to their differing valuations of the asset they are obtaining. The second is when the market lacks informational efficiency, meaning the value of an asset is mispriced in the market. Under this model it should not only be impossible for an individual to outperform the market via actively trading, but it should also be impossible to under perform. However, this project was able to look at the technical analysis of a stock (the study of a stocks trade history) and both over and under perform when making informed predictions at the sectoral level until 2012. This is 4 years later than Byrd et al. were able to predict when following only the market as a whole[1]. It is already known and implemented by market analysts and brokers that sectors are often seen as having separate uses depending on one's investment strategy and risk tolerance. This new result suggests it may also be reasonable to quantify investment opportunities based on the informational efficiency on the input sectors. By dividing the market into sectors we were able increase our ability to detect and quantify market inefficiencies. Due to limited sample size, further experimentation is necessary. However, by further subdividing the market it may be possible to extend the date which models can accurately predict the market further forward into the present.

CHAPTER 5

DISCUSSION AND FURTHER WORK

While working on this project, two additional questions became apparent. The first involves broadening research dimensions to the market capitalization of the stocks being predicted. While I assembled lists of stocks that I felt represented the high, medium, and low market capitalization and liquidity respectively, when requesting the limit order book snapshots from LOBSTER large sections of the data had gaps and were ultimately unusable. Second, the implementation of a transformer based models for prediction of the stock market. While I only used three models in my final data collections, over the course of the project I created logarithmic, linear, random, multiple baseline models, and an attempt at a transformer based approach. The transformer, which is a generative sequencing model, offers intriguing possibilities for forecasting high-frequency stocks. However, due to time constraints, repurposing a large language model into a stock prediction model presented technical challenges. Despite this, I believe these two questions represent promising avenues for future research.

REFERENCES

[1] D. Byrd and T. H. Balch, *Intra-day equity price prediction using deep learning as a measure of market efficiency*, 2019. arXiv: 1908.08168 [q-fin.TR].

[2] K. J. Chang, D. C. Chichernea, and H. R. HassabElnaby, "On the dupont analysis in the health care industry," *Journal of Accounting and Public Policy*, vol. 33, no. 1, pp. 83–103, 2014.

[3] W. Tarczyński, S. Majewski, M. Tarczyńska-Łuniewska, A. Majewska, and G. Mentel, "The impact of weather factors on quotations of energy sector companies on warsaw stock exchange," *Energies*, vol. 14, no. 6, p. 1536, 2021.

[4] O.-I. Bunea, R.-A. Corbos, and R.-I. Popescu, "Influence of some financial indicators on return on equity ratio in the romanian energy sector - a competitive approach using a dupont-based analysis," *Energy*, vol. 189, p. 116 251, 2019.

[5] B. G. Malkiel, "The efficient market hypothesis and its critics," *Journal of Economic Perspectives*, vol. 17, no. 1, pp. 59–82, Mar. 2003.

[6] M. R. Baker, "Universal approximation theorem for interval neural networks," *Reliable Computing*, vol. 4, no. 10, pp. 235–239, Apr. 1998.

[7] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks," *Chemometrics and Intelligent Laboratory Systems*, vol. 39, no. 1, pp. 43–62, 1997.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] S.-C. Huang and T.-H. Le, "Chapter 8 - convolutional neural network architectures," in *Principles and Labs for Deep Learning*, S.-C. Huang and T.-H. Le, Eds., Academic Press, 2021, pp. 201–217, ISBN: 978-0-323-90198-7.