Equity Portfolio Construction, Optimization, and Performance Analysis using Machine Learning

Alex Cole[a], Chris Cline[a], Will Bailey[b], Rachel Piraino[b], Hum Nath Bhandari[c,*]

[a] *Department of Business, Roger Williams University, Bristol, RI, USA*
[b] *Department of Computer Science, Roger Williams University, Bristol, RI, USA*
[c] *Department of Mathematics, Roger Williams University, Bristol, RI, USA*

---

*Corresponding author

Email addresses: `acole246@g.rwu.edu` (Alex Cole), `ccline616@g.rwu.edu` (Chris Cline), `wbailey758@g.rwu.edu` (Will Bailey), `rpiraino489@g.rwu.edu` (Rachel Piraino), `hbhandari@rwu.edu` (Hum Nath Bhandari)

Abstract

In this study we implement multiple machine learning techniques as they apply to the financial sector in order to create a comprehensive predictive deep learning model which we use to construct an optimization equity portfolio based on a return prediction. The overall goal is to create a portfolio which can generate an excess risk-adjusted return relative to industry benchmarks. In the first stage, we collect and prepare various relevant data including social media sentiment, technical indicators, and fundamental metrics for the stocks contained within the S&P Top 50 Index. In the second stage we use an optimized LSTM model to predict daily returns for the stocks. In the third stage, we construct an optimized portfolio daily, and use this portfolio to test performance against our benchmarks. Our model proved to have significant results and was able to optimize a portfolio that significantly out performed our benchmarks on a risk-adjusted basis over our testing period. While price predictions were not perfect, our model showed a significant ability to predict large price movements which allowed our optimization algorithm to adjust its weighting scheme to minimize excess risk while preserving returns. Overall our model showed great promise and our results were great given our limited time and resources for this project. Further expansion of this project could yield more accurate predictions, and by extension better performance.

Keywords : Data Science, S&P 500, text mining, financial literacy, financial sector, stock prices, optimized portfolio

1. Introduction

Modern Portfolio Theory

In 1952 Harry Markowitz introduced his process for constructing and optimizing an efficient portfolio which laid the framework for the Modern Portfolio Theory (MPT). The MPT states that investors can optimize their risk-adjusted return by constructing a diversified portfolio with respect to the assets' variance and correlation. The expected return of a portfolio can be assumed to be the weighted expected return of each asset. The total portfolio, risk which is measured by the standard deviation of returns, will actually be lower than a weighted sum of the asset risk due to asset correlations. The expected return and risk of all possible portfolios for all possible combinations of assets can be graphed, an upward sloping curve can be drawn to connect the most efficient portfolios; this curve is called the efficient frontier. The point along this curve which maximizes expected return for a given level of risk defines the optimally constructed portfolio in which an investor should invest in for the best risk-adjusted return. This return is usually measured using the Sharpe Ratio which compares a portfolio's return excess of the risk-free rate of return divided by the standard deviation of the portfolio's excess return. This process introduces a number of problems for the modern investor, most importantly the MPT considers both upside and downside variance when evaluating portfolios; this is the main problem that the Post Modern Portfolio Theory (PMPT) attempts to solve.

Post Modern Portfolio Theory

In 1991, Brian Rom and Kathleen Ferguson created the Post Modern Portfolio Theory (PMPT) which considers the standard deviation of only negative returns as the measure of a portfolio's risk. While the MPT assumes symmetrical risk, PMPT asserts that the returns of portfolios and assets can not be accurately represented by a joint elliptical distribution, such as a normal distribution. PMTP also introduces the Sortino ratio to replace MPT's Sharpe ratio as a measure of a portfolio's risk-adjusted return. It compares a portfolio's return excess of the risk-free rate of return divided by the standard deviation of only the portfolio's negative returns. PMPT also considers volatility skewness, which compares the distribution percentage of the total variance of an asset's returns above and below the mean; that is, they quantify asymmetric returns. For a modern, active investor this new approach to portfolio construction and optimization relies on the idea that upside variance is not something to be avoided, and said investor should seek to maximize positive returns while minimizing downside variance. While this framework works well to define how an active investor should approach risk, another problem arises in the PMPT which has not yet been addressed; this problem being that historic returns cannot accurately predict short-term future returns. The expected return of an asset and, by extension, of a theoretical portfolio is usually calculated as a function of historic return which raises an important problem for an active investor: historic returns cannot accurately predict short-term future returns.

Applications of Machine Learning in Portfolio Construction/Optimization

In an attempt to solve this problem, this study sets out to implement a comprehensive deep learning model by implementing a number of various modern machine learning techniques in order to provide much more accurate short-term expected return values. We will use the PMPT, using our one day return prediction as the expected return input to construct and optimize an equity portfolio to maximize short-term returns while minimizing downside risk. A time series prediction will be used taking in multiple categories of inputs, including technical, sentiment, fundamental, and macroeconomic data. Machine Learning has been used in prior studies to analyze sentiment, expected return, risk, and future stock prices. We look to build on this prior research by implementing our own set of inputs to optimize our model for the best return possible.

Machine Learning

Machine learning(ML) stems from the artificial intelligence(AI) sector of computer science. It focuses on using data and algorithms to imitate human behavior and learned abilities. It can be applied to a wide range of fields like marine biology, medicine, social media, etc. As previously mentioned, the team will implement machine learning techniques and apply it to the financial sector by creating a logistic regression model using data from the S&P 500 Index. There are multiple categories of machine learning models, including, but not limited to, supervised learning, unsupervised learning, and semi-supervised learning, with each having distinct characteristics that differentiate them from each other.

Importance of Text-Mining and Sentiment Analysis

Text mining uses Natural Language Processing (NLP) to create structure for sample text while organizing it from unstructured data to structured. Unstructured data involved more complex search, for example images and videos or text from emails. Structured data is easy to search and use such as database tables. Sentiment analysis is the use of algorithms to categorize sample text into whether they are positive, negative, or neutral. Sentiment analysis is also able to take emoticons into account. Implementing texting-mining and sentiment analysis allows the system to use ML and NLP to analyze the authors emotions and feelings which allows for a more accurate sentiment score. Sentiment analysis will allow the team to have a strong understanding of the environment's attitude towards particular stocks. This will aid the team in understanding the investors' perspective on the S&P 500 Top 50 Index.

Technical Analysis

Technical Analysis is a trading discipline that evaluates investments along with identifying trading trends and opportunities. At the core, technical analysis is the study of supply and demand and how it affects the market overall. By looking at the past stock trends, technical analysts' believe that it can be very telling for future stock prices and trends. Implementing machine learning along with our previous knowledge from technical analysis will aid in creating a more efficient and accurate model. Technical indicators can give portfolio managers clear buy or sell signals based on the movement and velocity of stock prices. This allows for optimal entry and exit points for any equity position which will aid our model in optimizing an efficient portfolio.

Our contribution

In this study will be utilizing machine learning to create a logistic regression model for each of the stocks contained in the S&P 500 Top 50 Index. The study will use various inputs including sentiment analysis of financial news and Twitter activity, fundamental data, and derivative trading activity to statistically quantify predicted one day forward return. The tools that will be utilized are text scraping, sentiment analysis, machine learning, etc.

2. Research Framework

Throughout this paper we will discuss other relevant publications that supports our ideologies and research. The model will be presented along with the various factors and justification of each contribution. We will discuss our overall steps for data collection and integrating the model. Next, we will show our results and discuss our conclusion and potential next steps.

Our research was set up in three distinct stages, data collection & preparation, machine learning implementation, and portfolio optimization. During data collection, we scraped technical, text, and fundamental data from many existing Python libraries, including SNScrape and yfinance. This data acted as our input features to our model. Once the data collection was successfully completed, we used our optimized LSTM model with the goal of accurately predicting the following day's closing price. With those prediction, in stage three we used linear programming to construct 10,000 potential portfolios, and the one with the best Sharpe ratio (risk to return ratio) was selected to be used for that day's weighting scheme. With a weighting scheme for each day, we simulated our portfolio over a nine month period and compared our results against benchmark Index funds.

3. Literature Review

Theme 1: Investment portfolio construction and optimization

Dziwok Ewa 2014: Presented and compared asset allocation methods used in modern investing. Differing portfolio construction methods were compared on the basis of efficiency, diversification, and limitations of each methodology. The study tested six differing portfolio construction methods, those being: mean-variance optimization, Black-Litterman model, naive diversification, global minimum variance approach, most diversified portfolio, and equal risk contribution. A comparative analysis method was used to distinguish the strengths and weaknesses of each method. One overarching method was used in all methods however, financial statement analysis. The company selection was not tested in this study, rather the weighting of pre-selected companies in a mock portfolio. In conclusion, the study found distinct pros and cons with each methodology and could not identify one superior weighting method. Each method has benefits given a certain market environment.

Harry Markowitz 1952: This paper details portfolio selection in two stages. The first stage starts with observation and experience which is translated into beliefs about the future performance of securities. The second stage, which this paper is concerned with, has to do with relevant beliefs about future performances which are followed by portfolio choice. The author then details the E-V rule and tries to find efficient combinations of both E and V. He concludes that on a large scale diversification with the E-V rule leads to efficient portfolios.

Theme 2: Applications of Machine Learning in Finance: Portfolio Construction and Optimization

Yilin Ma, et. al. 2020: Used time series prediction models to improve the performance of pre-existing portfolio optimization techniques. The study compared these time series predictions to another model which gave return predictions using Random Forest, Support Vector Regression models, LSTM, DMLP, and CNN models as well. The study uses each of these return predictions to construct a mean-variance (MV) portfolio optimization model. This study shows the superiority of return prediction models when compared to a time series forecast for portfolio optimization. However, the performance of these models were hindered by the high turnover ratio associated with using return prediction models in portfolio optimization, as transaction fees diminished any additional unexpected return, or alpha. Turnover was responsible for lowering the models returns by nearly half, and highlights a key issue that comes with re-balancing a portfolio using predictive models. The LSTM Model was tested with three different optimizers, SGD, RMSprop, and Adam. The study finds LSTM models gave the highest excess return based on Jensens Alpha, calculated using the CAPM method. The study did daily re-balancing with each of the models to compare the performance of each.

Gupta, A. et. al. 2020: There has been an increased availability for the opportunity to utilize text mining in the finance field. Texting mining in the financial sector deals with three main financial categories, forecasting, banking, and corporate finance. Financial forecasting deals with the stock market and foreign exchange market predictions. Banking looks at money laundering detection, risk management, and customer relations. While corporate finance analyzes reports and fraud detection. All three categories of finance use sentiment analysis, text clustering, and text classification techniques. There has been recent growth in digitization of the banking sector. Sentiment analysis uses opinion mining techniques and therefore allows us to predict future stock market trends and prices from the analysis of financial news articles. Textual data is generated through the process of acquiring information about customers. Text mining has challenged and changed the outlook of whether or not financial markets are predictable.

Wei Chen et. al. 2021: In order to successfully predict the future performance of the stock market, the study implemented the firefly algorithm and tailored it to the financial sector. The firefly algorithm (IFA) has been improved to select optimal parameters of the extreme Gradient Boosting (XGBoost). The firefly algorithm allows them to predict the following stock prices. The algorithm dynamically divides the group into subgroups and search based strategies are designed accordingly. The predictability of stock prices is directly correlated with the volatility, or rapid change, over time. Stable stocks are easier to predict than relatively noisy or unstable ones. The training set is used to train the model and adjust the parameters. The test set is used to evaluate the performance of the final model. The study improves the prediction accuracy and avoids the negative influence of parameter selection. Overall, the firefly algorithm improved stock forecasting.

Abe, M., Nakayama, H. 2018: This paper investigates the performance of deep learning models in predicting one-month-ahead stock returns in the cross-section of the Japanese stock market. The study compares the performance of deep neural networks with conventional three-layer neural networks, support vector regression, and random forests as representative machine learning models. The results show that deep neural networks generally outperform shallow neural networks and the representative machine learning models in predicting future stock returns. The study suggests that deep learning has potential as a skillful machine learning method to predict stock returns in this cross-section. The most effective model was able to achieve a 53.48% directional accuracy with p 0.001.

Theme 3: Text Scraping and Text Mining Strategies

Richardson Leonard 2019: This article provided the documentation for the Beautiful Soup 4 library which we used to scrape data from HTML files. The Beautiful Soup library is a Python library used for pulling data from HTML or XML files. The documentation serves to illustrate all of the major features of the library, showing how the library works, what is can be used for, and how to navigate every function. The documentation listed is for the Beautiful Soup version 4.8.1, which we used in collaboration with Python 3 to scrape news headlines from MarketWatch.com. This library is widely used by Python programmers and saves programmers vast amounts of time when scraping data.

Hongkee Sul 2014: This research proposes to analyze data collected from Twitter from March to October of 2011 and links it to the average daily return of the S&P 500. Twitter is a social media platform that allows users to post messages of up to 140 characters. Twitter has been proven as a tool to determine the sentiments of many domains media and education. Tweets with a dollar sign indicate the tweet involves investments. There were 2.5 million of these tweets collected and analyzed. It was found that the sentiment of tweets was significantly related to stock returns on subsequent days. Furthermore, that sentiment on Twitter was often associated with same-day abnormal returns. Specifically, tweets from those with fewer followers had a more substantial impact on future returns, while those with many followers had a more substantial impact on same-day returns.

Theme 4: Importance of Text mining and sentiment analysis in Portfolio construction, prediction, and optimization

Thomas Renault 2019: The study used a dataset of one million messages from the platform StockTwits to perform sentiment analysis on specific stocks. It found that the use of emojis and/or biagrams significantly improved the performance of the model, giving the model clearer indication of positive or negative sentiment as opposed to specific keywords, which may be used in differing contexts. The study performed daily sentiment analysis and found the correlation between investor sentiment and stock returns is high. Additionally, Renault found evidence that the method of preprocessing and the size of the dataset have significant impacts on the correlation found between investor sentiment and stock returns. The study found the highest classification accuracy score using a dataset of one million messaged, but suggested the accuracy would increase with a larger dataset. The study concluded that using sentiment data does not help in forecasting large market capitalization stock returns on a daily frequency, but suggested the preprocessing methods used to derive investor sentiment impacts the correlation coefficient.

Yi Yang 2020: This paper details the financial-specific BERT model, FinBERT which was created using a large scale of financial communication corpora of 4.9 billion tokens. This data included corporate reports, earnings conferences call transcripts, and analyst reports. There is growing interest in using NLP (Neuro Linguistic Programming) techniques to monitor market sentiment. The idea is that if the sentiment for a stock is good then the price will increase. The model is initialized from the original BERT model, this model was then used to create four different variants of FinBERT. The models are cased or uncased and BaseVocab or FinVocab. It has been shown that training the BERT model in the financial domain results in FinBert outperforming generic BERT models. Specifically, the uncased FineBERT-FindVocab model performed the best. Furthermore, the creators hope that researchers and analysts can use FinBert to detect sentiment without having access to the significant computational power that was used to create the model.

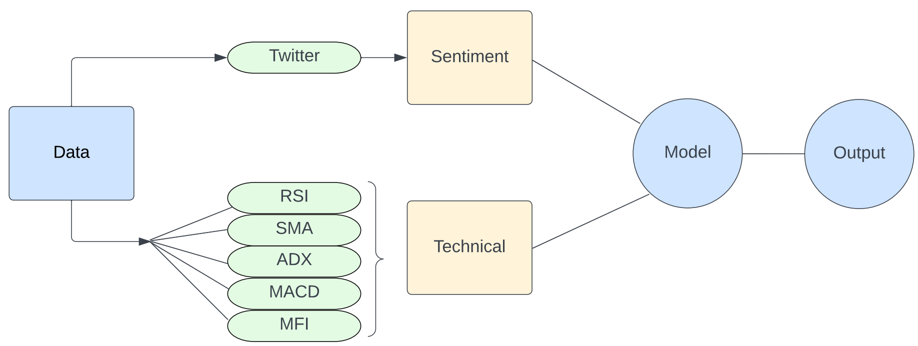
Malandri, L., Xing, F.Z., Orsenigo, C. et al 2018: Shows the relationship between emotions and their impact on stocks. Having become a technology driven society, the internet allows to curate public opinion, and we are able to use the public opinions to determine stock portfolios. Efficient-Market Hypothesis, EMH, states the current stock prices reflect all past information and prices react to new information. There is a need to adapt to society and develop tools to determine financial mood, therefore public opinion gets analyzed through text mining. The paper took financial data over 5 years and used the Quandl API, the 3 different learning algorithms for portfolio allocation were LSTM, MLP, RFC. The LSTM algorithm produced the best results; LSTM is in the RNN family and best used with data over long periods of time. Portfolio optimization is possible with an algorithm that uses public mood data results optimal allocation. Lastly, using LSTM networks and collective mood data positively improves portfolio management.

Paul Tetlock 2007: This paper examines the relationship between media content and daily stock market activity, using the Wall Street Journal's "Abreast of the Market" column from 1984 to 1999. The study finds that high levels of media pessimism predict downward pressure on market prices followed by a reversion to fundamentals, and unusually high or low pessimism predicts high market trading volume. These results are consistent with theoretical models of noise and liquidity traders and are inconsistent with theories of media content as a proxy for new information about fundamental asset values, as a proxy for market volatility, or as a sideshow with no relationship to asset markets. The paper concludes that measures of media content serve as a proxy for investor sentiment or non informational trading. The study introduces the method of quantitative content analysis as it is employed in this study for analyzing daily variation in the WSJ "Abreast of the Market" column.

Most Relevant Publication

From the publications above, the most relevant publication is entitled 'Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages' by Thomas Renault. The publication uses sentiment analysis on specific stocks and tracked their progress over time. It found that the use of emojis positively impacted the model. The publication also provides useful guidelines that apply to sentiment analysis specifically in the financial sector; for example, the study utilized a large data set and found it to be a key factor in their success. This publication relates directly back to this study as the team is going more in-depth on the topics of sentiment analysis and how to predict a stock portfolio based off of public opinion that is taken from Twitter.

4. Data Preparation



- https://www.sciencedirect.com/science/article/pii/S1568494620308814?fr=RR-2&ref=pdf_download&rr=79c8e86af9c2176c
- https://link.springer.com/article/10.1186/s40854-020-00205-1#citeas

5. Modelling Approach

Our model approach can be seen below. First we will input our data, after we will visualize it. Then we will prepare our data to be entered into the constructed LSTM and GRU models. We will run the LSTM and GRU models with various numbers of neurons per layer and determine which is the best model to then utilize. During the model implementation, we will split the data; 80% of the data will be used as training data and the remaining 20% is test data. From there, we will run hyper-parameter tuning on the model. Once that is done, we will create our output visualization and statistical analysis and come to our conclusion on which is the best model. Lastly, we will support our conclusion and discuss future work.

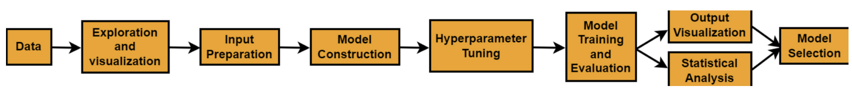


Figure 1: **Deep Learning approach for LSTM and GRU Architecture.https://www.sciencedirect.com/science/article/pii/S2665963822000902**

5.1. Data Input and Visualization

Below is a table of the dataset, this includes the date, close price, simple moving averages, MACD, etc., as well as our correlation heat map, model features, and a snapshot of the sentiment analysis data.

Date	Close	SMA	MACD	MACD$_{Signal}$	RSI	ADX	MFI	sentiment$_{v}alue$	
2016-02-11	43.00	45.51900005	-1.361498879	-0.645677321	36.57533552	29.84265138	37.22797229	0	
2016-02-12	44.56000137	45.36050014	-1.251041264	-0.766937751	43.2885457	28.91943144	42.63440179	0	
2016-02-16	46.72999954	45.40600014	-0.977138249	-0.809029958	51.05001122	27.06012399	43.82767598	0	
2016-02-17	48.40999985	45.52200012	-0.617389273	-0.770663825	56.06342784	26.23748358	50.8332948	0	
2016-02-18	49.08000183	45.6795002	-0.275051436	-0.671462752	57.91467047	25.83852612	61.56995094	0	
...	...	...	...	...	...	...	...	...	...
2021-12-30	182.73	181.865	3.540829906	4.326096739	55.97080302	39.84097954	48.72366825	0	
2021-12-31	182.87	181.865	3.54083	4.326097	55.97080302	39.84097954	48.72366825	0	

Table 1: **Snapshot of the dataset.**

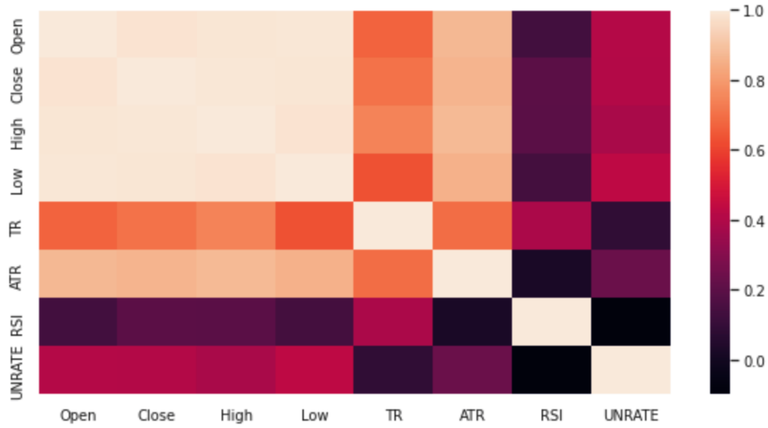


Figure 2: **Correlation heat map among the attributable variables.**

Data	Source	Library
Technical Indicator		
20 Day Simple Moving Average	Yahoo	stock-indicators
Moving Average Convergence Divergence	Yahoo	stock-indicators
Stochastic	Yahoo	stock-indicators
Relative Strength Index	Yahoo	stock-indicators
Average Directional Index	Yahoo	stock-indicators
Money Flow Index	Yahoo	stock-indicators
Average True Range	...	...
Sentiment		
Investor Sentiment	Twitter	Snscrape

Table 2: **Model Features**

Date	sentiment_value
2021-12-03	0.007692
2021-12-06	0.052941
2021-12-07	0.103175
2021-12-08	0.160000
2021-12-09	0.008547
2021-12-10	0.106796
2021-12-13	0.104478
2021-12-14	0.029630
2021-12-15	0.061856
2021-12-16	0.016260
2021-12-17	0.084746
2021-12-20	0.066667

Figure 3: **Sentiment Analysis**

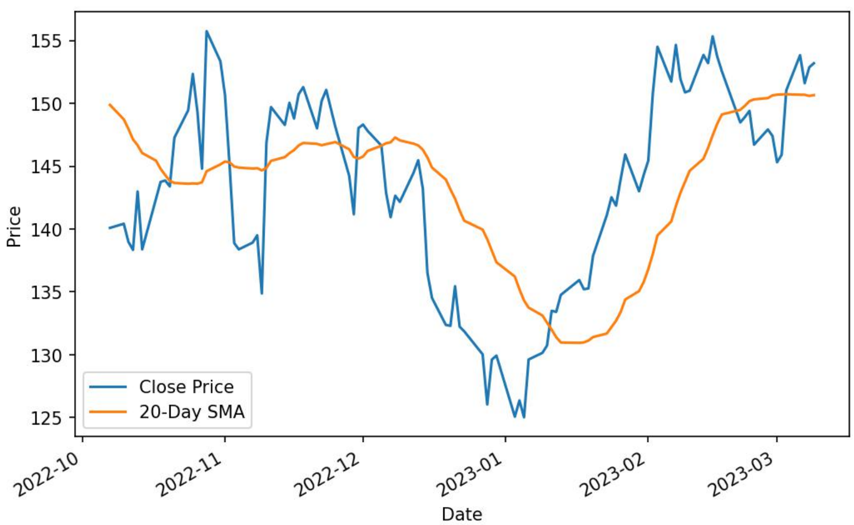
5.2. Input Definitions

Close Price: The last transaction price of a stock at the end of a day's trading session. To calculate the close price, divide the total number of shares / total number of shares within the last 30 minutes. The close price helps those who invest in the stock market to understand the market sentiment surrounding the particular stock over a period of time.

Volume: The number of total shares of a stock traded during a day. An increase in trading volume is often correlated with investors reacting to a change in the underlying stock, like from breaking news, and usually results in a larger than normal price action of the stock.

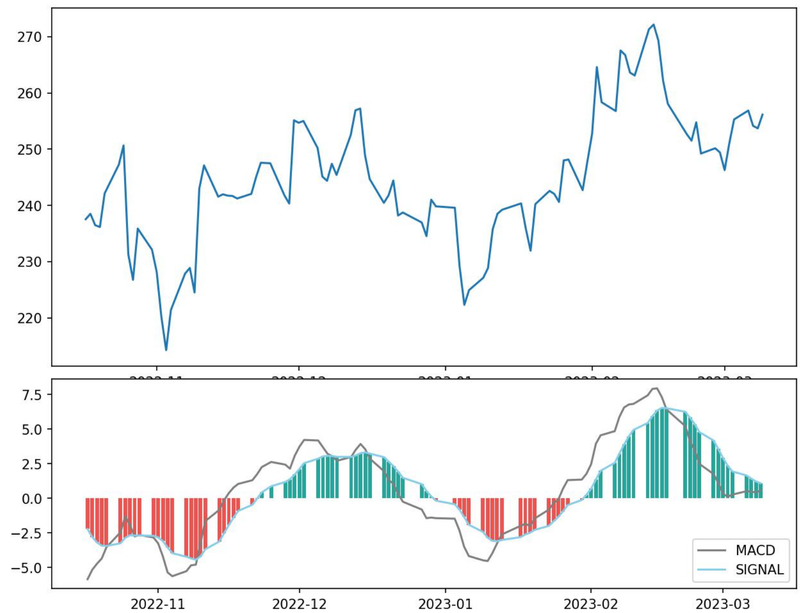
Short Interest Ratio: The ratio of the number of a stock's shares held short to the stock's average daily trading volume. Investors sell a stock short to make a bet on downward price movement, or to hedge their investment against price depreciation. A stock with a higher short interest ratio indicates investors see a lot of risk associated with the stock and are hedging their investments to the downside. Likewise, a stock with a low short interest ratio is seen by investors as safer, and less likely to depreciate in value.

N-Day Simple Moving Average (SMA): Calculated as the average closing price of the last N days, the simple moving average provides a rolling value of the stock's average price looking back. This simple technical indicator is used to visualize smoothed price movement, and help filter out extreme short-term price fluctuations; for this reason, it can be used to visualize broader directional trends within a stock over longer periods of time. For this project, we used a 20-day moving average to provide a medium-term look at a stock's general trend.



Moving Average Convergence Divergence (MACD): Calculated by subtracting the 26-day exponential moving average (EMA) from the 12-day EMA (EMA is similar to the SMA however it exponentially weights recent returns over distant ones). The result of this calculation is the MACD line; because the shorter EMA is constantly converging towards, or diverging from, the longer EMA, the MACD line oscillates around 0. To analyze this indicator, the MACD line is plotted against a 9-day EMA of the MACD line itself, which is called the "signal line." This allows investors to visualize the momentum of a stock's price.

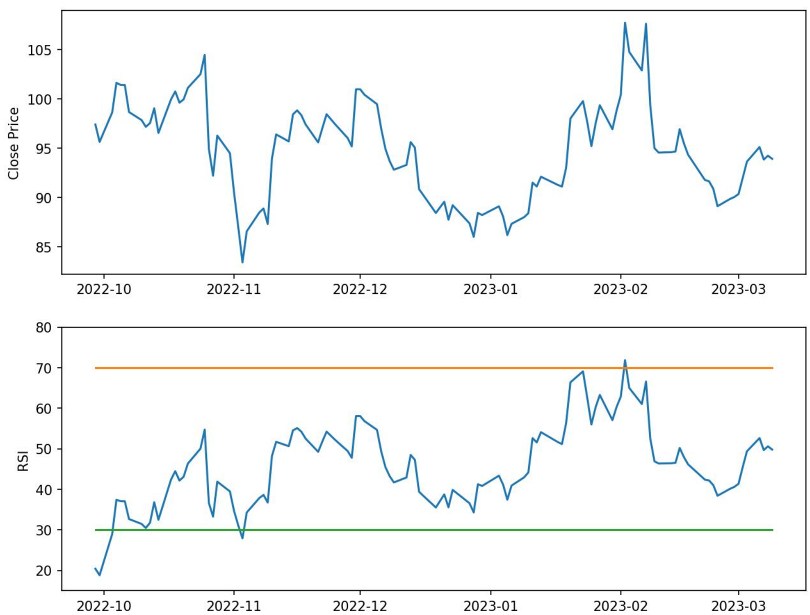
Most commonly a buy signal is formed when the MACD line crossed above the signal line, signifying that the momentum of the stock is starting to increase. Conversely, a sell signal is formed when the MACD line crosses below the signal line, indicating that the stock's momentum has stopped and started to reverse. The MACD is useful for identifying the strength of a directional move, and indicating a reversal in momentum.



Stochastic: Calculated as $\%K = \frac{C-L}{H-L} * 100$ where C is the most recent closing price, L is the lowest price traded of the last 14 trading sessions, H is the highest price of the last 14 trading sessions, and %K is the current value of the Stochastic indicator. The Stochastic Oscillator is a momentum oscillator comparing the price of a stock relative to its price range over the last 14 sessions and is measured from 0% - 100% where 0% means the stock is currently trading at its lowest price in 14 days, and 100% means its trading at its highest. This can be used by traders to identify overbought or oversold stocks, usually represented by values over 80 or under 20. However, a strong trend can sustain overbought or oversold levels for an extended period of time. Instead, investors can use changes in Stochastic to predict changes in momentum. Like the MACD, the

Stochastic line is plotted against a 3-Day SMA of itself where, again like the MACD, Stochastic readings above the signal line indicate bullish momentum, while readings below the signal line indicate bearish momentum.

Relative Strength Index (RSI): Similar to Stochastic, the RSI is a momentum oscillator used to measure the momentum of a stock's recent price movements. The indicator compares price movement on days where prices increase and decrease respectively to calculate a value between 0 and 1 (0% - 100%) where 1 indicates the strongest trend, and 0 the weakest. This indicator is the most widely used to identify overbought and oversold stocks, correlating with values over 70 or under 30 respectively.



Average Directional Trend (ADX): The ADX measures the strength of a trend, and is often used in conjunction with positive and negative directional indicators which measure the direction of a trend. On its own, an ADX value over 25 indicates a strong trend, while a value under 20 indicates a weak trend. Again, ADX only measures the strength of a trend regardless of direction so Investors can use this with directional indicators to spot strong upward/downward trends forming and enter a long/short position. The ADX is used to measure whether the market is fluctuating or trending. If the market is trending, it is used to determine the trend's strength. Lastly, it can be used to determine future changes and trends.

Money Flow Index (MFI): The MFI measures the flow of money in and out of a stock using price and volume data over a specific time period. The MFI is calculated by gathering money flow values and creating a ratio and then normalized and put into the money flow oscillator. The oscillator moves between 0 - 100. This allows traders to analyze the stock price and volume.

Average True Range (ATR): The ATR formula is as follows: Calculated as $ATR = \frac{PreviousART(n-1)+TR}{n}$, where n is the number of periods and TR represents the true range. Average true range was developed by J. Welles Wilder, Jr. and measures degree of price volatility. This takes gaps within the market price movement into account.

Investor Sentiment: Investors are able to see when the overall market is trending as a bull versus a bear market which then indicates that people will be willing to buy stocks at higher prices rather than be willing to sell at this time. Investor sentiment is a inkling that investor have knowing past stock trends.

5.3. Input Preparation and Model Construction

We used a basic train test-split with a test data size of 20 percent. Once the input data was ready, we started building our models. The first model we decided to use was a Long Short-Term Memory network(LSTM), and secondly gated recurrent unit (GRU).

Our data was collected from various Python libraries, including SNScrape, and yfinance. Additionally, we calculated our own Technical Indicator data as additional inputs for our model. Our data was normalized using a MixMax Scaler, to ensure the data was prepped for model implementation.

5.4. LSTM Architecture

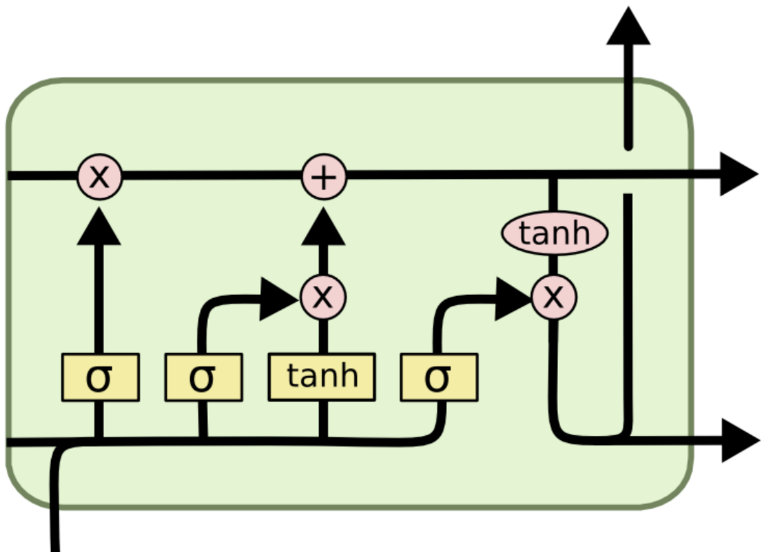
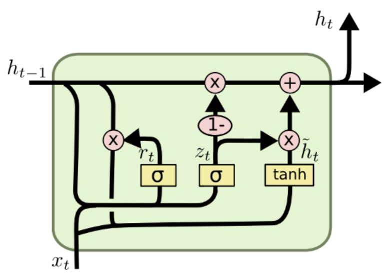


Figure 4: **Visualization of LSTM model. https://colah.github.io/posts/2015-08-Understanding-LSTMs/**

Figure three above is a visual representation of the LSTM model architecture. At the top of the diagram is what's called the cell state. Information moves along this line. The model adds or subtracts information based on the gates input. The gates are made of a sigmoid neural layer and a point wise multiplication operation. The sigmoid layer outputs a number between 0 and 1 so that the sell state knows how much information should be let through. If the cell layer is fed 0 it means to let nothing through and if its fed 1 it means let everything through. LSTM modeling uses three gates to control the cell state. The model first decides what information it doesn't want at the sigmoid layer or "forget layer". After this we go to the next layer called the input gate layer which decides which values to update. Then a tanh layer creates a vector of new values that may or may not go into the final state. Now that it has come up with all this information it just needs to create the new cell state by updating the old cell state. We first forget everything the new state wants to forget then add the input layer multiplied by the tang layer of new values. Thus creating our new cell.

5.5. *GRU Architecture*



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Figure 5: **Visualization of GRU model. https://colah.github.io/posts/2015-08-Understanding-LSTMs/**

The second model we decided to use was a Gated Recurrent Unit or GRU model. The GRU model is shown above. As you can notice it is very similar to the LSTM model. The biggest difference between LSTM and GRU is that the GRU model combines the forget layer and input layer into one layer called the "update gate". The update gate is equivalent to the forget gate and input gate of LSTM. It is responsible for long-term dependencies. The reset gate is responsible for the short-term dependencies. This model is considered less complex then the LSTM model.

5.6. *Hyper Parameter Tuning*

When tuning the models there were three main factors we looked at to determine which was performing best. First, we had to look at the performance of LSTM and GRU architectures. Second the performance of each individual neuron layer. Then finally the hyper parameters of these models. These hyper parameters include optimizer, learning rate, batch size, and time step. We created average score graphs to observe these factors and how they behaved in the models we created. These graphs contained the model's average RMSE, MAPE, and R scores. When selecting the best model, RMSE was our primary selection criterion.

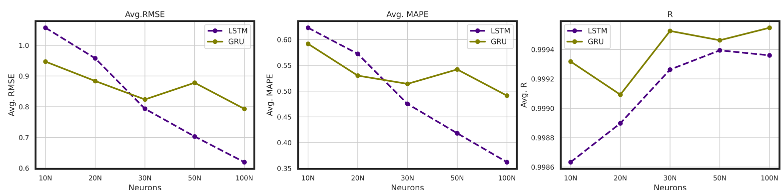


Figure 6: **Average Model Scores**

We looked at the loss plot graphs to determine how many epochs to use. After looking at the graphs, we observed that at around 4-5 epochs the loss plateaued. We ultimately choose to use 15 epochs as our constant for the model. We did this to account for any outliers in the data. This was also done because the time taken to run the full 15 epochs did not affect the timeline of this project.

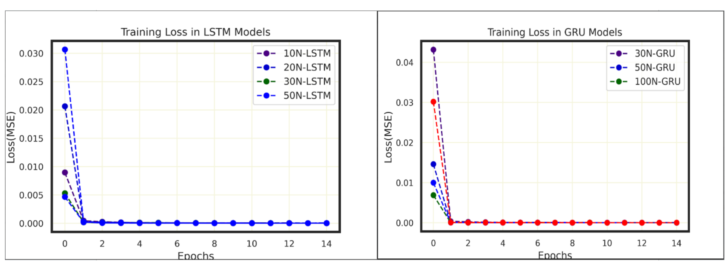


Figure 7: **Loss Plotsl**

6. Experimental results

After tuning our model, we were able to determine that the LSTM model with the hyper parameters listed below consistently gave us the strongest RMSE, MAPE, and R values. Below are those values as well as the average scores for the best model we produced.

Hyper Parameters	Value
Optimizer	Adam
Learning Rate	.0075
Batch Size	8
Time Step	10

Strongest Hyperparameters

RMSE	MAPE	R
.62	.36	.99

Strongest Model Output

With our model and optimizers chosen we created 50 different models for each stock. Figure 8 below is a graph of the true vs. predicted values for Apple. This graph shows how our model was trained and what it predicted. The blue line is the true value of the stock. The red line shows the predicted values of the train set which was the first eighty percent of our data. The grey line shows what the model predicted in the test set. By looking at the graph we can see that predicted values in the test set look strong and trail the true values accurately.

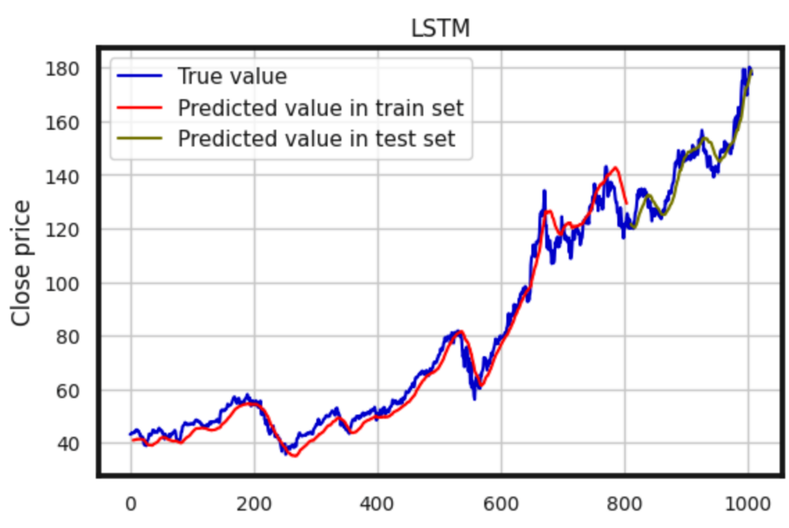
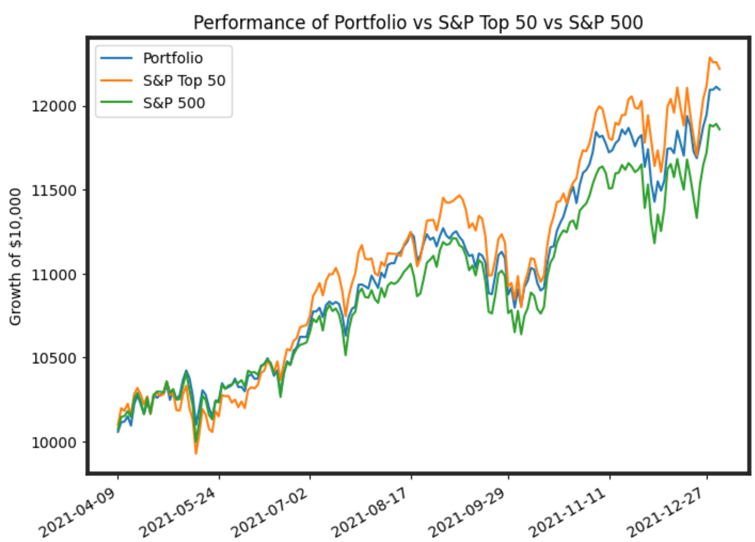


Figure 8: **AAPL True Vs. Predicted Price**

Given the limited scope of this project, we saw very encouraging results. By using our model to predict returns we were able to significantly reduce our portfolio's risk while preserving returns similar to our benchmarks. Over our testing period of the nine months from April 1st through December 2021, our portfolio had a simulated return of 20.95%, compared to the S&P 500's 18.58% and the S&P Top 50's 22.18%. While we did slightly under perform our direct benchmark based on raw return, our portfolio had a simulated beta of 0.83 and an annualized standard deviation of 10.65% compared to the S&P Top 50's beta of 1.01 and standard deviation of 12.65%. With our lower risk, we would have been able to generate alpha, or risk-adjusted excess return, of 5.03% over the nine months of testing compared to our benchmark's 3.52%. For the risk-conscious investor, these results are extremely significant, rivaling professionally managed investment funds.



Measure	Portfolio	S&P Top 50	S&P 500
Holding Period Return	20.95%	22.18%	18.58%
Annualized Return	29.38%	31.17%	25.96%
Beta	0.83	1.01	1.00
Std Dev.	10.65%	12.65%	12.03%
Holding Period Alpha	5.03%	3.52%	-%
Annualized Alpha	7.15%	5.09%	-%
Holding Period Treynor Ratio	0.21	0.22	0.19
Annualized Treynor Ratio	0.31	0.31	0.26
Sharpe Ratio	2.43	2.19	1.87

7. Conclusion

A major limitation within this study is the short period of time to complete this project as well as our limited resources. Having said that, our model performed significantly well and there is room for future work. Despite our limitations, our model was able to predict stock market returns with enough accuracy to significantly reduce risk while preserving returns. Our portfolio has a significant advantage over our benchmark funds in that, while the weighting schemes and holdings of the Index funds are pre-determined, and all assets will be owned every day broadly speaking, our model can decide to overweight, underweight, or not own a particular asset at all on a given day. While our model might not be able to perfectly predict specific movements, it was specifically successful at predicting large movements which allowed our model to adjust its holdings to avoid the risk inherent in large price swings. Because of our limited resources we also had to limit our stock selection to just 50 stocks, while a larger pool of assets might have extended our model's success.

Another limiting factor to consider is transaction costs; while traditionally each transaction would come with a broker's and/or other fees, today almost all transactions are able to be made without any extra cost. For this reason, we choose to omit the analysis of transaction cost which might otherwise limit the profitability of this strategy.

Overall, our study has shown the potential that deep learning has within the financial market. Significant research has shown that Machine Learning models are able to predict stock performance to a degree where the implementation of such prediction can improve the optimization of a stock portfolio against comparable Index funds.

8. Future Work

Further improvements to this project could include the implementation of Transfer Learning to improve the performance of our modeling and prediction stage. With a more precise prediction of closing price, the model would be able to adjust fund weights more optimally leading to improved performance of the portfolio overall. The inclusion of a more diverse range of sentiment data from news headlines or social media such as Reddit could also help improve the accuracy of the model.

9. Ethics and implications

There can be many ethical implications of this study, while the end results were favorable this algorithm should not be used without performing the proper due diligence and research required when investing in stocks or other equities. Additionally, trading costs were not factored into the end performance of our portfolio, which proved to be a limitation on other trading algorithms that used similar strategies. This algorithm, when employed properly could serve to improve the performance of already existing equity portfolios, given they are well diversified across a variety of industries or sectors.

This algorithm could have implications of scale on the U.S. stock market. With enough assets under management, this algorithm could serve to move markets on its own, which would undoubtedly impact the performance of the model and the optimization techniques employed in this study. These factors were not taken into consideration when evaluating the performance of our model, as garnering the scale needed to have these impacts is unlikely.

10. Acknowledgment

Roger Williams University. The Math, Business, and Computer Science Department. Dr. Hum Nath Bhandari

References

Abe, M., & Nakayama, H. (2018). Deep Learning for Forecasting Stock Returns in the Cross-Section. Lecture Notes in Computer Science, 273–284. doi:10.1007/978-3-319-93034-3_22

Chen, Wei. "Mean–Variance Portfolio Optimization Using Machine Learning-Based Stock Price Prediction." ScienceDirect, https://doi.org/10.1016/j.asoc.2020.106943.

Dziwok, E. (2014). Asset Allocation Strategy in Investment Portfolio Construction–A Comparative Analysis. Journal of economics and management, (18), 124-132.

Markowitz, H. (1952), PORTFOLIO SELECTION*. The Journal of Finance, 7: 77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

H. Sul, A. R. Dennis and L. I. Yuan, "Trading on Twitter: The Financial Information Content of Emotion in Social Media," 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 2014, pp. 806-815, doi: 10.1109/HICSS.2014.107.

Gupta, A., Dengre, V., Kheruwala, H.A. et al. Comprehensive review of text-mining applications in finance. Financ Innov 6, 39 (2020). https://doi.org/10.1186/s40854-020-00205-1

Kliengchuay, W., Srimanus, R., Srimanus, W. et al. Particulate matter (PM10) prediction based on multiple linear regression: a case study in Chiang Rai Province, Thailand. BMC Public Health 21, 2149 (2021). doi: https://doi.org/10.1186/s12889-021-12217-2

Neubauer, Jiří, et al. "Impacts of Built-Up Area Geometry on PM10 Levels: A Case Study in Brno, Czech Republic." Atmosphere, vol. 11, no. 10, Sept. 2020, p. 1042. Crossref, https://doi.org/10.3390/atmos11101042.

Malandri, L., Xing, F.Z., Orsenigo, C. et al. Public Mood–Driven Asset Allocation: the Importance of Financial Sentiment in Portfolio Management. Cogn Comput 10, 1167–1176 (2018).

TETLOCK, P.C. (2007), Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance, 62: 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Yang, Yi and UY, Mark and Huang, Allen. (2020). FinBERT: A Pretrained Language Model for Financial Communications.