Applied Machine Learning Intensive (AMLI), Summer 2021
Capstone Report

STOCK PRICE PREDICTION

July 28, 2021

Elayne Blancas, ECE, eb3770@nyu.edu
Gregory Perez, ECE, gap85@cornell.edu
Jonathan Zamudio, CSCE, jz034@uark.edu

1 Abstract

Our Stock Market Predictor dictates whether an individual stock will either over-perform or under-perform, calculate the long and short gains of the portfolios stocks, and obtain the portfolios stock average return and Sharpe Ratio. We used yfinance API to obtain a decade of daily historical stock data from Yahoo Finance for the top 30 Standard and Poor (S&P) 500 companies. Our data consisted of non-negative stock values with no upper bound for weekdays and non holidays. Our Stock Market Predictor uses Random Forest as a baseline for our LSTM model. To prevent over-fitting, we dropped 20 percent of our data, a common practice used in machine learning. Additionally, in an effort to obtain a higher accurate score, we use a decade of the stockâs history to predict its future trends. In making this model pipeline, we would use the 3 prior years as our training data. This pipeline would take the years of data and go over it in a series of sequences, each sequence containing 240 days. This sequence of data would then give us the value of the stock. Our model has 60% to 70% training accuracy. Our predictor would advice for a long investment on either Comcast Company or Adobe, and a short investment on either Protector Gamble Company or Salesforce. Although this type of predictor is difficult to generalize, our future work would be to include sentimental analysis, web scraping, and Beautiful soup for real time data to mitigate for its volatility. Our model does not account for anything external and should, therefore, not be used as a main investment tool.

2 Introduction

The most important applications of our Stock Market Predictor are the following:

1. To predict whether an individual stock will either over-perform or under-perform.
2. To calculate the long and short gains of our portfolioâs stocks.
3. To obtain the stock portfolioâs average return and Sharpe Ratio.

Any scale of investors gain the most benefit from the application of predicting the marketâs behavior. This application can be used as a tool to help investors decide when and which stocks to purchase for the growth of their investment. The nation, as a whole, also greatly benefits from a healthy stock market since the stock market affects the US economy in the following manners:

1. Stocks allow the individual investor to own part of a successful company. Without stock markets, only large private equity investors and financial institutions could profit from America's free market economy.
2. Stocks allow savers to overcome inflation. There is an approximately 7% annual increase on stock prices, after taking into account inflation, which generously com-

pensates most investors for the additional risk of owning stocks rather than bonds or keeping the money in a savings account.

3. Growing, successful businesses need capital to fund growth and the stock market is a key source. In order to raise money this way, owners must sell part of the company, and to do so they "take the company public" through an initial public offering (IPO) of the company's shares. An IPO raises a lot of cash. It also signals that the firm is successful enough to afford the IPO process. The drawback is that the founders no longer own the company; the stockholders do. Founders can retain a controlling interest in the company if they own 51% of the shares.

Typically, the stock market and economic performance are aligned. Therefore, when the stock market is performing well, it is usually a sign of a growing economy, vice versa. The most prominent measurement of economic growth is the Gross Domestic Product (GDP). GDP is the total monetary or market value of all the finished goods and services produced within a countryâs borders in a specific time period. As a broad measure of overall domestic production, the GDP functions as a comprehensive scorecard of a given countryâs economic health. As a result, when the GDP is increasing that is an indication that individual businesses are expanding. Expanding business activity usually increases valuations and leads to stock market gains. A stock market predictor could assist with early detection of threatening trends allowing us to uncover hidden nuances about the economy and gain a greater understanding of such a volatile industry.

Despite its critical role in the economy, the stock market is not the same as the economy. The stock market is driven by the emotions of investors. They can exhibit irrational exuberance. It occurs during an asset bubble and the peak of the business cycle. They become overly optimistic even though there is no hard data to support it. Our mission is to provide stability in the emotions of the investors by predicting if a stock is either over or under fit, and the stock marketâs trends. Our goal is to be utilized as an additional tool/resource in their decision-making process.

The contributions and role of each members:

- Team Manager: Elayne
- Program Manager: Greg
- Resource Manager: Jonathan

1. **Phase 1:**

   Timeline: June 25, 2021 - July 2, 2021

   Goals: Our team is composed of Elayne Blancas, Gregory Perez, Jonathan Zamudio. We will build a model that predicts future stock prices of a particular index. (i) Form the team; (ii) Define the capstone topic; (iii) Make a plan on who does what for the coming phases; (iv) Estimate the task and goal for the next Phases: fill out the content for Phase 2, Phase 3, Phase 4.

2. **Phase 2:** Timeline: July 2, 2021 - July 09, 2021

   Goals:

   - Collect the data and annotate
     - **Date: July 6, 2021**
     - Gregory
   - Reproduce the code
     - **Date: July 7, 2021**
     - Jonathan
   - Understand the input data so that you can run the code on your custom data
     - **Date: July 8, 2021**
     - Elayne
   - Investigate ethical implications and impacts of our model.
     - **Date: July 8, 2021**
     - Gregory

3. **Phase 3:** Timeline: From July 9, 2021 to July 19, 2021

   Goals:

   - Build the model
     - **Date: July 11, 2021**
     - Jonathan
   - Consider Results and Determine if more data is needed
     - **Date: July 14, 2021**
     - Elayne
   - Iterate through the model and optimize
     - **Date: July 18, 2021**
     - Jonathan

   - Collect and implement feedback from peers and instructors
     - **Date: July 18, 2021**
     - Greg
   - Prepare for mid-project presentation
     - **Date: July 18, 2021**
     - Elayne

4. **Phase 4:**

   <u>Timeline:</u> July 19, 2021 - July 28, 2021

   <u>Goals:</u> Please fill with your plan including timeline and who will take care the task.

   - Final review of the model
     - **Date: July 20, 2021**
     - Jonathan
   - Prepare data and ethics analyses for presentation
     - **Date: July 23, 2021**
     - Elayne
   - Edit and review Capstone Report
     - **Date: July 26, 2021**
     - Gregory
   - Prepare results analysis for presentation
     - **Date: July 27, 2021**
     - Jonathan

3 Data Analysis

Our Stock Price Prediction model will analyze data pertaining to the opening and closing stock prices for a portfolio of the top 30 S&P 500 companies from the year 2010 to present day. Our model takes in sequences of standardized one-day returns as inputs and will output a prediction about the performance of stocks in our portfolio per day. We group this data in overlapping periods of 4 years, which defines our study period for the training and testing sets. These 30 companies are:

Table 1: **Table of Stocks**.

Top 30 Stocks		
Index	**Company**	**Service**
1. AAPL	Apple Inc.	It is an American multinational technology company that specializes in consumer electronics, computer software, and online services.
2. MSFT	Microsoft Corporation	It is an American multinational technology company which produces computer software, consumer electronics, personal computers, and related services.
3. AMZN	Amazon.com, Inc.	It is an American multinational technology company which focuses on e-commerce, cloud computing, digital streaming, and artificial intelligence.
4. FB	Facebook, Inc.	It offers other products and services beyond its social networking platform, including Facebook Messenger, Facebook Watch, and Facebook Portal. It also has acquired Instagram, WhatsApp, Oculus, Giphy and Mapillary, and has a 9.99% stake in Jio Platforms.
5. GOOGL	Alphabet Inc.	It is an American multinational technology company that specializes in Internet-related services and products, which include online advertising technologies, a search engine, cloud computing, software, and hardware.
6. GOOG	Alphabet Inc.	This is a Class C share that does gives the investor ownership stake but not voting rights. This share is normally given to employees and Class A holder.
7. BRK.B	Berkshire Hathaway Inc.	This is a Class B share gives 10 times the voting power and is usually held by the founders of the company.

Continuation of Table 1

Index	**Company**	**Service**
8. TSLA	Tesla Inc.	Their current products include electric cars, battery energy storage from home to grid-scale, solar panels and solar roof tiles, as well as other related products and services.
9. NVDA	NVIDIA Corporation	It designs graphics processing units for the gaming and professional markets, as well as system on a chip units for the mobile computing and automotive market.
10. JPM	JPMorgan Chase & Co.	It is a global leader in financial services, offering solutions to the world's most important corporations, governments and institutions in more than 100 countries. As announced in early 2018, JPMorgan Chase will deploy $1.75 billion in philanthropic capital around the world by 2023.
11. JNJ	Johnson & Johnson	It is an American multinational corporation founded in 1886 that develops medical devices, pharmaceutical, and consumer packaged goods.
12. V	Visa Inc.	It facilitates electronic funds transfers throughout the world, most commonly through Visa-branded credit cards, debit cards and prepaid cards.
13. UNH	UnitedHealth Group Inc.	It offers health care products and insurance services. It is the second-largest healthcare company by revenue with $400.1 billion, and the largest insurance company by net premiums.
14. PYPL	Paypal Holdings Inc.	It is an American company operating an online payments system in the majority of countries that support online money transfers, and serves as an electronic alternative to traditional paper methods such as checks and money orders.
15. HD	Home Depot Inc.	It is the largest home improvement retailer in the United States, supplying tools, construction products, and services.

Continuation of Table 1		
Index	**Company**	**Service**
16. MA	Mastercard Inc.	It is a technology company that connects consumers, financial institutions, merchants, governments and businesses across the world, enabling them to use electronic forms of payment.
17. PG	Proctor & Gamble Co	It is a multinational consumer goods corporation; specializing in a wide range of personal health/consumer health, personal care ,and hygiene products.
18. DIS	Walt Disney Co	It is a multinational mass media and entertainment conglomerate. It is the leader in American animation, and diversified into live-action film, television, and theme parks.
19. ADBE	Adobe Inc	It is an American multinational computer software company, with software specialized for creation and publication of content. This includes graphics, photography, animation, motion pictures and has expanded into digital marketing management software.
20. BAC	Bank of America Corp	It is an American multinational investment bank and financial services holding company. It is the second largest banking institute in America, falling behind JPMorgan Chase. Its primary services revolve around commercial banking, wealth management, and investment banking.
21. CMCSA	Comcast Corporation	It is an American telecommunications conglomerate. It is the second-largest broadcasting and cable television company in the world, the largest home Internet service provider in the United States, and the nation's third largest home telephone service provider.

Continuation of Table 1

Index	**Company**	**Service**
22. XOM	Exxon Mobil	It is an American multinational oil and gas corporation. With 37 oil refineries to make a combined daily refining capacity of 6.3 million barrels, Exxon is the seventh largest refiner in the world.
23. CRM	Salesforce.com, Inc	It is an American cloud-based software company. It provides customer relationship management (CRM) service along with a complementary suite of enterprise applications focused on customer service, marketing automation, analytics, and application development.
24. PFE	Pfizer Inc.	It is an American multinational pharmaceutical and biotechnology corporation. It develops and produces medicines and vaccines for immunology, oncology, cardiology, endocrinology, and neurology.
25. CSCO	Cisco Systems, Inc.	It is an American multinational technology conglomerate. It develops, manufactures and sells networking hardware, software, telecommunications equipment, and other high-technology services and products.
26. VZ	Verizon Communications Inc.	It is an American multinational telecommunications conglomerate. It is the second-largest wireless carrier in the US, and the second-largest telecommunications company by revenue after AT&T.
27. NFLX	Netflix, Inc.	It is an American over-the-top content platform and productions company. Its primary business is a subscription-based streaming service, also producing and distributing content.
28. KO	The Coca-Cola Company	It is an American multinational beverage corporation. It has interests in the manufacturing, retailing, and marketing of nonalcoholic beverage concentrates and syrups.

Continuation of Table 1		
Index	**Company**	**Service**
29. INTC	Intel Corporation	It is an American multinational corporation and technology company. It is the world's largest semiconductor chip manufacturer, supplies microprocessors for computer system manufacturers, and also manufactures motherboard chipsets, network interface controllers and integrated circuits, flash memory, graphics chips, embedded processors and other devices related to communications and computing.
30. PEP	PepsiCo	It is an American based multinational food, snack, and beverage corporation. Its encompasses all aspects of the food and beverage market, overseeing the manufacturing, distribution, and marketing of its own products.

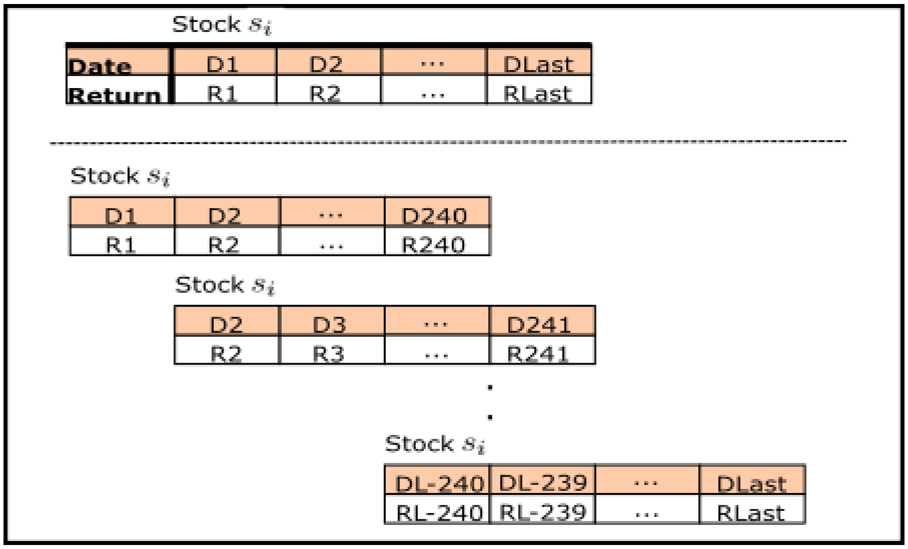
To accomplish this goal, historical day-by-day stock data is collected from Yahoo Finance using the `yfinance` API. The daily return for each stock is calculated using the open and closing prices and is then standardized. Then, data is split into 240 sequences of consecutive points in time. The first 3 years of data sequences makes up the training set, and the last year of the period will be the test set. While training, 20% of the data is set aside for validation.

4 Project Implementation

Our project implementation includes 5 steps. First, we process the raw data to fit the feature space and target for making predictions and training. Second, our processed data is split into training and testing sets. Third, we provide a discussion on the architecture and implementation of LSTM. Fourth, we describe random forests and its use as a benchmark model. Lastly, we present the trading strategy from our models predictions. At the end, a discussion about the challenges that we encountered during the implementation of these models is included.

4.1 Generating feature and target variables

Our feature space includes the daily and next day returns for each company, calculated by:



Stock s_i

Date	D1	D2	···	DLast
Return	R1	R2	···	RLast

Stock s_i

D1	D2	···	D240
R1	R2	···	R240

Stock s_i

D2	D3	···	D241
R2	R3	···	R241

Stock s_i

DL-240	DL-239	···	DLast
RL-240	RL-239	···	RLast

Figure 1: Construction of our sequence vectors from each stock vector s_i (All vectors are shown transposed).

$$R_t^{m,s} = \frac{P_t^s}{P_{t-m}^s} - 1$$

where the simple return (R) at a time (t) for a stock (s) in our portfolio over m periods, is calculated for a price process (P). Additionally, the percent change of the closing price from a prior day $(t-1)$ to the current day (t) is included in our feature space. These features are then standardized by subtracting the median $(\tilde{x}_{train})$ and dividing them by the Interquartile Range:

$$\tilde{R}_t^{m,s} = \frac{R_t^{m,s} - \tilde{x}_{train}}{Q_3 - Q_1}.$$

Because LSTM networks work particularly well with time-series data, the inputs to our model will be sequences of standardized one-day returns $(R_t^{1,s})$. Our model requires sequences of length 240 days. We generate overlapping sequences of 240 day periods for each stock, which includes approximately a year of trading history. For each stock, the data is sorted by date in ascending order. Each sequence is constructed as $\{\tilde{R}_{t-239}^{1,s}, \tilde{R}_{t-238}^{1,s}, ..., \tilde{R}_t^{1,s}\}$ for each stock s and each $t \geq 240$ within the study period. Thus, the first sequence would consist of the first 240 one-day returns $\{R_1^{1,s_i}, R_2^{1,s_i}, ..., R_{240}^{1,s_i}\}$. The second sequence contains $\{R_2^{1,s_i}, R_3^{1,s_i}, ..., R_{241}^{1,s_i}\}$ and so on [1]. This is further illustrated in **Figure 1**. In total, each study period consists of approximately 22000 sequences. This results in approximate data splits of 15000 and 7000 sequences for training and testing sets.

For our model, we define a binary classification problem. All one-period returns $R^{1,s}_{t+1}$ of all stocks s are ranked in ascending order. Class 0 is achieved when the one period return of a stock s is lower than the cross-sectional median return of all stocks in the period. Class 1 is achieved if the the one period return of a stock s is equal or greater than the cross-sectional median return of all stocks [1]. As a result, both classes are of equal size.

4.2 Producing training and testing sets

The sequence vectors for each stock s in our portfolio are stacked into one single feature vector. From the study period ($T = 4$ years), the first 3 years of historical stock data constitutes the training set and the last year defines the testing set. Due to companies exhibiting a subset of historical data during the study period, the amount of input data fluctuates slightly.

4.3 LSTM architecture

Long Short Term Memory networks are a class of RNN, designed to avoid the long-term dependency problem. Essentially, LSTM adds or removes information to the cell state through structures called gates. These gates are usually labeled as the "forget gate", "input gate", and "output gate" [3]. In **Figure 2**, an illustration of the LSTM cell structure is shown. In this figure, the "forget gate" is governed by the equation $f_t = \sigma(W_f x_t + R_f h_{t-1} + b_f)$. The "input gate" is governed by the equation $i_t = \sigma(W_i x_t + R_i h_{t-1} + b_i)$. Then the "output gate" is governed by the equation $o_t = \sigma(W_o x_t + R_o h_{t-1} + b_o)$. These gates regulate the cell state to maintain "long-term memory" throughout the network. The cell state c_t governed by the equation $c_t = f_t \circ c_{t-1} + i_t \circ c'_t$. Here we can see that the previous cell state is multiplied with the forget gate to determine what earlier information to preserve and what current information to add into the cell state. The output is governed by the equation $h_t = o_t \circ tanh(c_t)$, which determines what information is output from the cell [2].

Our model architecture starts with an input layer that accepts inputs with shape (240,3). This corresponds to the sequence length and the size of our feature space. This layer feeds into our single layer many-to-one LSTM network of 25 cells. Then, a dropout layer will drop the result 10% of the time. Finally, the output layer (dense) has 2 hidden neurons with softmax activation function.

4.4 Random forest

For our benchmark model, we implement a random forest model. Random Forest is a "state-of-the-art machine learning model" that delivers good results with virtually no tuning. Additionally, it is a popular choice for large-scale machine learning applications on stock market data [1]. For benchmarking, our model uses cumulative returns $R_{t,m}s$ as

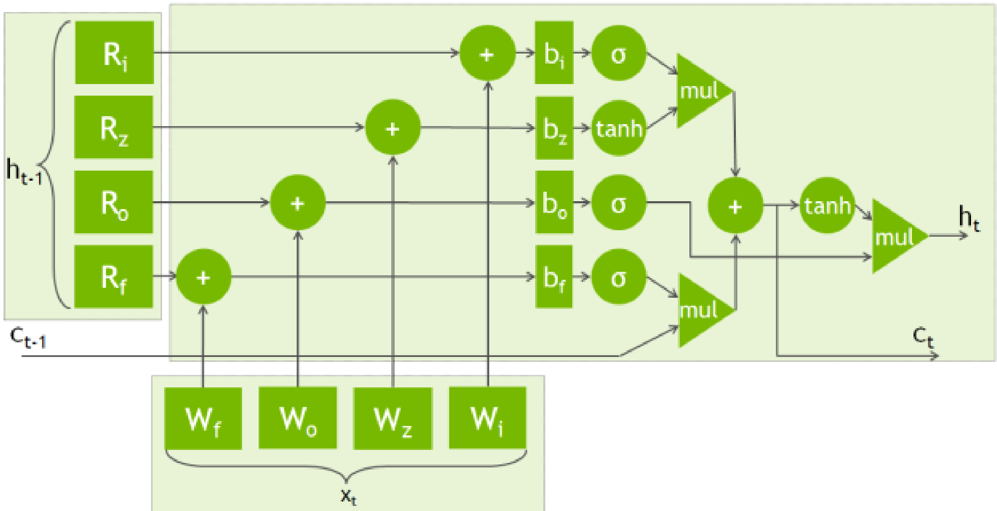


Figure 2: Cell structure of the LSTM cell used in our model [2].

features with $m \in \{\{1, ..., 20\} \cup \{40, 60, ..., 240\}\}$ and has the same target as described in our LSTM model. Our RF model contains 1000 tress and has a max depth of 10.

4.5 Trading Strategy

Now, our models predictions are used to rank the performance of the 30 stocks within our portfolio in ascending order. Our trading strategy is to go long the top 10, or 10 most undervalued, companies and go short the worst 10, or flop 10, companies.

4.6 Challenges

A large portion of our time was spent finding a method to collect and process our data for use in our model. Many of the APIs for retrieving stock data are freemium or deprecated, so a bit of research went into finding an API to download the data. Initially, we manually downloaded our datasets, but had later found the yfinance API to automate our data collection. Combining the stock data and splitting it into study periods was also a challenge. We solved this by parsing through each stock and writing the data to a `.csv` file, as seen in our Google Colab notebook.

Another major challenge was understanding our models inputs and outputs. Reading through the literature related to our GitHub repository, we were able to understand the details of our model architecture.

Finally, we encountered issues with saving, loading, and making predictions with our model. Similar to understanding our model, we wanted to extract only the predictions from our model from the present day. This took a little more research and understanding the code, but eventually reached a solution to getting graphs of our predictions.

5 Experimental Results

5.1 Testing

To test our model performance we use the simulate function, which takes predictions from our model to determine long and short investments. The top ten trending stocks are classified as long investments, while the the flop ten trending are the short investments. Once we obtain which companies to invest in for long/short, we calculate the average daily return on the portfolio. The average returns are plotted with the benchmark RF model to determine performance. The Sharpe ratio is also calculated with the Sharpe Ratio formula:

$$S_a = \frac{E[R_a - R_b]}{\sigma_a}$$

The Sharpe ratio (S_a) is a measure that indicates the average return (R_a) minus the risk-free return (R_b) divided by the standard deviation (σ_a) of return on an investment. This metric helps define the overall risk to return of investing in a given stock.

5.2 Evaluation

Our model's accuracy and loss values were really great when testing. The earlier years we ran the model for gave us much smoother curves for our accuracy/loss per epoch. When it came to our later models we saw more oscillations at the end of our curves. This means there could be overfitting, or could come from the volatility of stock data during 2020 given the pandemic.

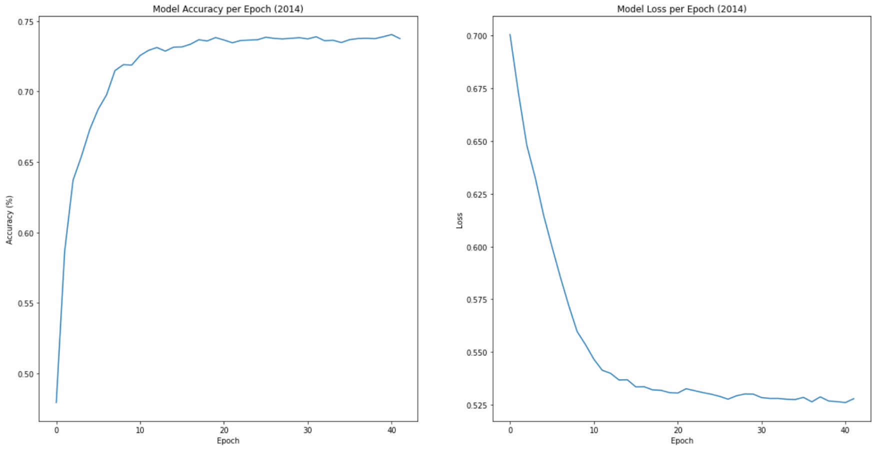


Figure 3: Accuracy and Loss for 2014 LSTM model

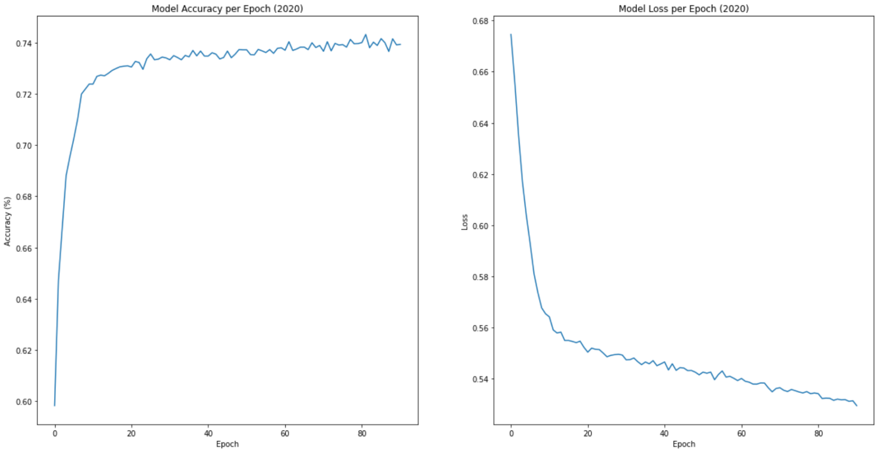


Figure 4: Accuracy and Loss for 2020 LSTM model

5.3 Results

The main outputs we got from our model were Average Daily Returns and the Sharpe ratio. Comparing our LSTM model to the benchmark RF model, we can see that the LSTM model outperforms the RF model. LSTM is more accurate and has higher percentage returns.

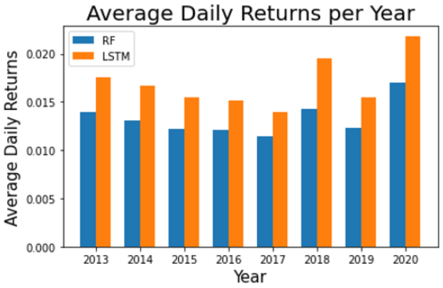


Figure 5: The average returns of all stocks in the portfolio for each day, grouped by year.

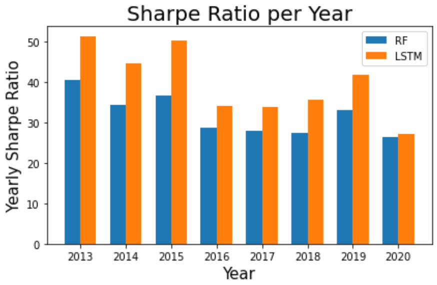


Figure 6: The yearly Sharpe ratio for the stocks in the portfolio.

6 Conclusion

We created a stock market predictor using an LSTM model. This model took in historical stock data from the last decade and created predictions on whether or not a stock would over or underperform. The model computed the average daily return and Sharpe ratio to determine which companies to invest in or sell. The LSTM model we chose does better in all years compared to our benchmark RF model.

Overall, our model worked well, ending with a 70% accuracy rating. This model does decent in showing which stocks would be good to take long or short positions in, however it is not the most accurate given the nature of stocks. The stock market has always been very quick to change given outside factors that a prediction model cannot always take into account. In order to further improve our model, we would have to do some sentiment analysis. This would allow our model to account for outside factors that commonly affect stock prices, like the general populations feelings towards a stock taken from news and social media. We can also use web scrapping and Beautiful Soup to take in more real time and accurate stock information, along with news information on these stocks.

7 Acknowledgment

We would like to thank our advisors for their guidance, the financial support from our families and sponsors: NACME and GOOGLE, as well as the TA's and our colleagues.

References

[1] Fischer, Thomas, and Christopher Krauss. ”Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions.” Econstor, 2017, www.econstor.eu/bitstream/10419/157808/1/886576210.pdf.

[2] ”Optimizing Recurrent Neural Networks in CuDNN 5.” NVIDIA Developer Blog, 25 Aug. 2020, developer.nvidia.com/blog/optimizing-recurrent-neural-networks-cudnn-5/.

[3] ”Understanding Lstm Networks.” Understanding LSTM Networks – Colah's Blog, colah.github.io/posts/2015-08-Understanding-LSTMs/.