I. Definition

a. Project Overview

i. Predicting the stock market has been the Holy Grail of investors ever since it was established in 1817. Each new generation comes up with a new and fantastic way to predict prices and of course the hot topic right now is machine learning. Can a machine be trained to predict the stock market? This is a question that is being actively researched and tested every day around the world, specifically using deep learning. One such technique that has shown promise is recurrent neural networks (RNN). Networks that have a "memory" of what has been seen in the past. Unfortunately, RNNs do not have a long memory and thus long short-term memory (LSTM) was born, which has allowed RNNs to have a sort of perpetual memory in regards to weights and bias'. The goal of this project was to predict stock prices using RNN-LSTM to get state of the art results. The result was definitely less than stellar and shows why predicting the stock market remains an unattainable goal.

ii. Stock market prediction has been researched for many years and the driving force is financial gain [4]. A quick Google search delivers a plethora of different research papers devoted to predicting the stock market. There are many different approaches to predicting the stock market, from the traditional moving averages, candle charts, and technical analysis to the cutting edge machine learning algorithms. One such analysis is Particle swarm optimization (PSO) and least square support vector machine (LS-SVM), this is a combination of two different algorithms to tune and adjust to different weaknesses in the models [5].

b. Problem Statement

i. In this project I will strive to predict stock prices. The problem is it is very difficult to predict future prices of stocks. Stocks are very tangible things. You have prices that go up and down, and this data is available everywhere but the problem and solution seem to be very difficult to define and it is definitely not an easy prediction to make. Before I have even begun the project, I have an intuition that the Volume will play a big role in the data and the ending result. The method with which I will structure this project is regression. The inputs for this will be the normal stock market features, opening price, closing price, highest price during the day, lowest price during the day, volume, in addition to some derived features such as the Sharpe ratio and possibly Bollinger bands to determine boundaries for the model.

ii. The nature of the stock market is not altogether rational or logical. From observations it is based upon emotion. It involves a huge human element in which there can develop a herd mentality. It is near impossible to predict where this herd will take the market and your model could be correct 80% of

the time with the same features, but the market can take a turn the model wasn't prepared for.

iii. As a result of the analysis I expect my result to not be very accurate. The plan is to attempt to predict stock prices for the next day both for an individual stock on the stock market as a whole.

iv. This report uses two different models that compose the data pipeline for predicting stock prices. The first is a preliminary model in order to get a sense of the data and how a model would choose to go about "learning" how the data is structured. Before the model can be used The preliminary model is a Support Vector Regression (SVR), which is similar to a Support Vector Machine (SVM), but used for regression analysis instead of classification. It will be used to determine the feasibility of the models and to get a baseline for further analysis.

c. Metrics

i. The metric chosen to evaluate the model is the Root Mean Squared Error (RMSE). It is chosen because it is a common evaluation metric for complicated analysis, which predicting stock market prices is a very complex process. It allows for a regularization of the errors that allow for flexibility in the analysis.

II. Analysis

a. Data Exploration

i. The dataset I used was one from a competitor on Kaggle and the code can be found on this site: https://github.com/CNuge/kaggle_code. The dataset consists of 504 different stocks and their prices collected from August 13, 2012 to August 11, 2017. The data represents the companies in the S&P 500. It consists of the following features: Date, Open, High, Low, Close, Volume, and Name. In addition I have added a few different features in order to predict stock price, Daily Returns, Rate of Change, and the Sharpe Ratio. The main one of these is the Sharpe Ratio, which is, "The Sharpe ratio is the average return earned in *excess* of the risk-free rate per unit of volatility or total risk." [1]

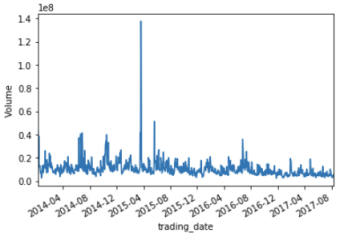
ii. The feature to be predicted is the Closing Price of the stock. The Closing price is an important element in the stock market. If an investor is able to predict if the closing price is higher the next day he/she can purchase the stock today and sell it tomorrow for a gain.

iii. Here is a sample of the dataset for the stock American Airlines (AAL)

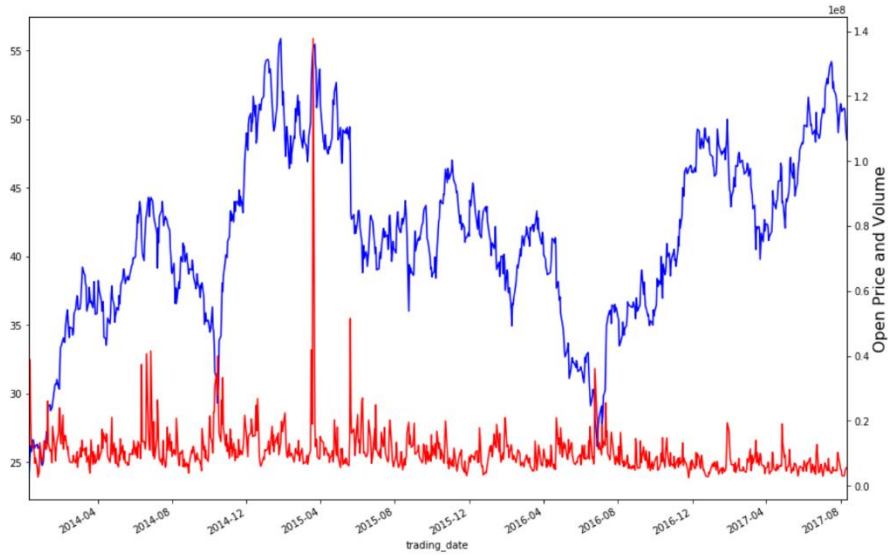
Date	Open	High	Low	Close	Volume	Daily_Returns	Sharpe_Ratio	Rate_of_Change
12/9/2013	23.85	25.44	23.45	24.6	43197268	-3.04878	-0.490958	3.144654
12/10/2013	24.5	25.17	24.41	24.88	18660625	-1.527331	-0.249594	1.55102
12/11/2013	25.48	27.2	25.37	25.99	38843371	-1.962293	-0.318597	2.00157
12/12/2013	26.2	26.71	25.45	25.45	19981824	2.946955	0.460209	-2.862595
12/13/2013	25.75	26.3	25.52	26.23	12192421	-1.829966	-0.297604	1.864078
12/16/2013	26.63	26.77	26.35	26.61	13190945	0.07516	0.004626	-0.075103
12/17/2013	26.48	26.59	25.95	26.1	11413199	1.455939	0.223674	-1.435045
12/18/2013	25.99	26.23	25.55	26.23	9994162	-0.914983	-0.152451	0.923432
12/19/2013	26.12	26.49	25.82	26.12	6916497	0	-0.007297	0
12/20/2013	26.18	26.49	26.14	26.33	8530924	-0.569692	-0.097674	0.572956

b. Exploratory Visualization

i. These two graphs represent my initial graphs regarding the data for one particular stock. These graphs were beginning of the data preprocessing/analysis step. The graph on the left signifies the ups and downs of the stock prices for the given years. The graph on the right signifies the volume of trades for the given years. This lead to a question how does the data on the graphs relate to each other when graphed on top of the volume. In other words how does the volume either influence or is influenced by the stock prices. Is there any correlation?



ii. The below graph was a result of that question. Since the stock prices on the combined graph follow a very similar path, I chose one feature to graph against the volume. Below is the graph of 'Open' vs 'Volume.' From the graph I gathered that there was some type of pattern between the two features but it was not immediately known how they relate, but from the graph there are obvious points of spikes in the 'Volume', but how does that tell what's happening between the features. More exploration is needed.



c. Algorithms and Techniques

i. The algorithms being utilized are Support Vector Machines (SVM), Recurrent Neural Network (RNN), and Long short-term memory (LSTM).

ii. SVM is being used as a preliminary model to test out the likelihood and the feasibility of predicting stocks. The SVM technique has been a technique used for predicting stock prices. However, it is not the best way to necessarily predict stock prices. "We find little predictive ability in the short-run but definite predictive ability in the long-run." [2]

iii. The first approach was to run the data through the SVM to observe the results and develop a better understanding of the data. The data needed to be preprocessed so through research with the sklearn library and StackOverflow, the *LabelEncoder* class was used. This allowed the data to be normalized for analysis. Next, was using the *LabelEncoder's fit_transform* method in order to get the encoded labels to use. The next step was to split the data into the training and testing sets. I used sklearn's *train_test_split* method to accomplish this. After the split the *fit* method was called to develop the model, than the *predict* and finally an accuracy score was obtained.

iv. After the SVM was established and the model is in place, it was time to develop the RNN-LSTM model for the data. RNN-LSTM is a particularly difficult model to implement. Much research beyond the primary resource Udacity was utilized in order to develop the model as is noted in the Resources section. It begin with preprocessing the data in order to get a proper form for the analysis. The next step was to split the data so it could be used in the model (the *train_test_split* did not work to give adequate data). Because the model is particular there were many attempts at trying

to reshape the data to a form the model could use. Unfortunately, many different approaches were taken, but all failed to produce results.

d. Benchmark
    i. The clear benchmark for this problem is the market itself. But as the analysis didn't yield fitting results the ability of the model to predict would not be worth the time to run it. The SVM was a particularly low accuracy, and the RNN-LSTM was not able to be ran due to misconfigured data and multiple attempts to rectify.

III. Methodology

a. Data Preprocessing
    i. The preprocessing that was done for the SVM preliminary model was straightforward and routine using the sklearn library. It utilized a few methods from the library to also split the training and test data.
    ii. The RNN-LSTM model preprocessing was very cumbersome, difficult, and ultimately did not work. Many different approaches were attempted but did not succeed. The data was not in a correct format for the LSTM layer to do an adequate analysis.

b. Implementation
    i. The SVM was a simple implementation model. It required very little data intervention from the researcher, and it still yielded results. If it was the main focus of the analysis the parameters would have been adjusted to a more refined degree.
    ii. The main focus of the implementation was the RNN-LSTM model. This model required much more intervention and adjustment than the SVM model. During the analysis data needed to be pre-processed to a high degree and was not achieved in the implementation.

c. Refinement
    i. The data, because it is stock prices, by definition is a time series. This means the data needs to stay in strict order, because predicting prices is not an exact science. Having the data in a very ridged format gives some type of consistency in the data so it can be used accurately.
    ii. As stated above the SVM implementation did not need very much refinement, because it was not the main focus of the investigation. I mainly chose to go with a vanilla implementation of the SVM so there was relatively little refinement that was needed.
    iii. The RNN-LSTM model needs much more refinement than the SVM. Refinements to the shape, data type, training & testing sets, and the model itself were made with not much success. There were refinements regarding the data in the pre-preprocessing, but nothing was found to get results. Much time was spent on research across the internet, including but not limited to StackOverflow.com, Keras.io, Udacity.com,

MachineLearningMastery.com, and others. Overall, there is much more need for refinement that seems to be beyond my current abilities using RNNs.

IV. Results

a. Model Evaluation and Validation

i. The final model was not very successful. In the end the SVM was wholly inaccurate and would not be able to predict any stock prices even if it was a few minutes later.

ii. The RNN-LSTM was a complete disaster with its difficult getting the data into a form it needed. Which was failure on the part of the researcher and lack of knowledge. The RNN-LSTM was a failure in predicting and couldn't get past that data, which would seem to show that it lacks the robustness needed for analysis. It was discovered that there is so much data pre-processing that is needed its ineffective at predicting stocks at a reasonable rate.

b. Justification

i. The SVM and the RNN-LSTM models developed in this report are not adequate to solve the problem. The accuracy of the SVM was very low to not even guess a stock price correctly, whereas the RNN-LSTM model was not sufficiently developed to provide any type of solution.

V. Conclusion

a. Reflection

i. The problem of stock market prediction is one which continues to perplex many researchers every day. The solution proposed in this report is not one that would solve this problem. From the beginning of the report it attempts to break down the problem and get the data into a usable format. That effort was not successful, and failed to provide an adequate solution or process. The SVM could be developed more to provide a better accuracy, but in the opinion of the researcher it would still not provide an adequate answer to the problem, and it would be just another failed experiment in which to learn from.

ii. The RNN-LSTM was an extremely difficult challenge, mainly due to lack of understanding and domain knowledge. It seemed like a great challenge but ultimately was not solvable by a novice in the field. Much more research will need to be done in order for a novice deep learning practitioner to advance. The most difficult portion of this entire project was the data pre-processing.

iii. The data pre-processing from what I have observed is the most difficult portion of any machine learning problem. It needs to be done in a consistent and logical manner and comes with experience in how to process

and divide the data. More work needs to be done on the pre-processing of the data, which will allow the models to move forward in the analysis.

b. Improvement
    i. As a model RNN-LSTM can be a great resource for more experienced researchers and those that are involved with investigating deep learning possibilities. However, it is not a practical solution.
    ii. Researching more data pre-processing solutions would be a possibility for improvement. The RNN-LSTM model needs to be given very clean, tightly organized data in order to form an answer to the problem set.
    iii. Also, picking a more developed dataset could have concluded with a better more robust result.

Resources:

[1] Investopedia. (2017). *Sharpe Ratio* [Online]. Available: https://www.investopedia.com/terms/s/sharperatio.asp

[2] S. Madge. (2015). *Predicting Stock Price Direction using Support Vector Machines* [Online]. Available: https://www.cs.princeton.edu/sites/default/files/uploads/saahil_madge.pdf

[3] J. Brownlee. (2016, July 21). *Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras* [Online]. Available: https://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/

[4] P. Pawar. (Accessed: 2017, December 4). *Machine Learning applications in financial markets* [Online]. Available: https://www.iith.ac.in/~saketha/research/ppBTP2010.pdf

[5] O. Hegazy. O. Soliman. M. Salam. (2013, December). *A Machine Learning Model for Stock Market Prediction* [Online]. Available: https://www.researchgate.net/publication/259240183_A_Machine_Learning_Model_for_Stock_Market_Prediction