The Effect of Type of Machine on Accuracy

Akhil Ganesan & Prabhas Adivi

Centreville High School

Table of Contents

Abstract

The main problem this experiment was posed to solve was whether the type of artificial prediction engine (machine; statistical analysis, support vector regression, or neural network) affected the accuracy of the predictions. This experiment involved collecting/formatting public domain data sets (stock data from MSFT, AAPL, & NFLX), creating/obtaining many different artificial intelligences (using software on an electronic device), and testing the machines yielding the overall results (requiring the testing software, devices, and materials).

The results supported the hypothesis stating that the neural network would be the most accurate, then the statistical analysis, and finally the support vector regression. It also concluded that the MSFT dataset was the most predictable, then AAPL, then NFLX. The NFLX data set was most likely the toughest to predict due to its broad range of output values. In addition, the AAPL data set was more advantageous in its structure to the statistical analysis over the neural network. These were the main conclusions drawn upon completing this experiment.

This project provided a better insight of how certain machines work to predict data, as well as what type of data is the most predictable.

Introduction

This experiment is designed to provide a solution for the query, “which type of intelligent machine tends to be the most accurate in the environment of the stock market?”

Rationale

Advanced computer systems - including artificial intelligence (AI) - are predicted to be the next technological breakthrough. Although they do currently exist, they aren’t popularized among the general public. For those that do know, the stock market was their first response to a testing environment due to its rich data available and inability for most humans to accurately predict trends consistently (signifying an advanced task). But there is a catch, there isn’t only one testable AI; a multitude of machines exist in the modern world. That was the primary motive behind this experiment: use an advanced concept (AI) and test its variations (types of machines) in a controlled environment (stock market).

Experimental Design

This project plans to have 3 independent variable levels - statistical analysis (control), support vector regression, and a neural network. The dependent variable for this project will be the accuracy. The constants include the data, environment, format, builder, type of output function (linear), etc.

Expected Outcomes

After concluding some research on the topic of artificial machines in the stock market, here are a couple conclusions concerning each machine:

Machine	Advantages	Disadvantages
Statistical Analysis (SA)	- Recognizes overall pattern, may be practical in long-term data - Doesn't take long to compile	- Wouldn't work too accurately when considering small trends
Support Vector Regression (SVR)	- Analyzes training data intensely, results in very accurate training predictions	- May overfit data and not do as well on new test data
Neural Network (NN)	- Recognizes small patterns, puts together larger pieces - Takes a little longer to compile	- May not be accurate among large datasets where small trends throw the machine off

Hypothesis/Engineering Goal

If three different types of machines (statistical analysis, support vector regression, and a neural network) are tested for accuracy in the stock market (AAPL, MSFT, & NFLX), then the neural network will be the most accurate, then the statistical analysis, then the support vector regression because the neural network's complexity when compared to the other machines is far more, this means it is able to recognize more advanced, subtle trends; the statistical analysis will be next due to its ability to summarize trends; the support vector regression will be last due to its overfitting techniques. The goal of this experiment is to generate a machine with an accuracy greater than 90%.

Materials & Methods

The experiment was conducted in four primary recursive steps. The first step was to obtain the respective machines (whether they are hand-programmed or obtained from online sources). The next was to obtain the data set (typically from a public domain) and format it to match the machines (preprocessing, data entries and edits, etc). The following step was to run a pretest of the experiment; this means visualizing the data, testing the machines for programming errors, etc. The final step is to actually run the test command; upon doing so, each learning cycle will complete and yield results; the results are then categorized into a specific data table used for the final graphics. This step of the experiment is repeated for each machine per data set. By the end, there should be 9 significant figures in which the data is based from

Materials

This experiment isn’t too heavy on materials due to the fact that is is virtually stimulated. The primary materials list has a variety of variations as well as other alternatives to complete the same experiment. Here is the primary materials list:

- Essential data collecting materials (table for recording, basic supplies, etc)
- Computer (Dell Latitude 3300)
- Programming Software/IDLE (Spyder3)
- Dataset (Obtained from yahoo finance)
- Machines (premade)

Procedure

The procedure for completing this experiment consists of 4 primary steps. The final step will be repeated, in this case, 9 times.

1. Attain the respective machines (from online or build from scratch)
2. Obtain the necessary data and format it correctly (stock data)
    a. This includes dividing the dataset into training and testing sets
3. Run a pretest of the experiment to ensure all is in order (visualize the data)
4. Reset the program, begin testing each machine for each dataset
    a. At the end of each test, the program will report results; record and repeat for the next machine

Data Collection

The data collection will consist of two primary phases: the experimental phase (in which uses a data table) and the post-experimental phase (in which incorporates a graphic representation of the data). For the experimental phase, the data table should resemble this:

Acc.									
Data	Aapl	Nflx	Msft	Aapl	Nflx	Msft	Aapl	Nflx	Msft
Type	Statistical Analysis			Support Vector Regression			Neural Network		

For the post-experimental data, the plan is to use a triple bar graph with each dataset on the X-axis, accuracy on the Y-axis, and a key for three bars per level of the independent variable.

Risks & Safety

Due to the fact that the testing environment is the void of digital space on a computing device, there are very few to no risks or safety concerns associated with this experiment. Most of these possible risks include technological errors such as computer overheating, out-of-control hardware, etc. Other than that, there are very few plausible risks posed with this experiment.

Results

The results of the experiment supported the hypothesis in its assertion of the neural network having the highest accuracy, then the statistical analysis, and finally, the support vector regression. This section will cover the data & experimental results leading into the discussion and analysis.

Data Tables

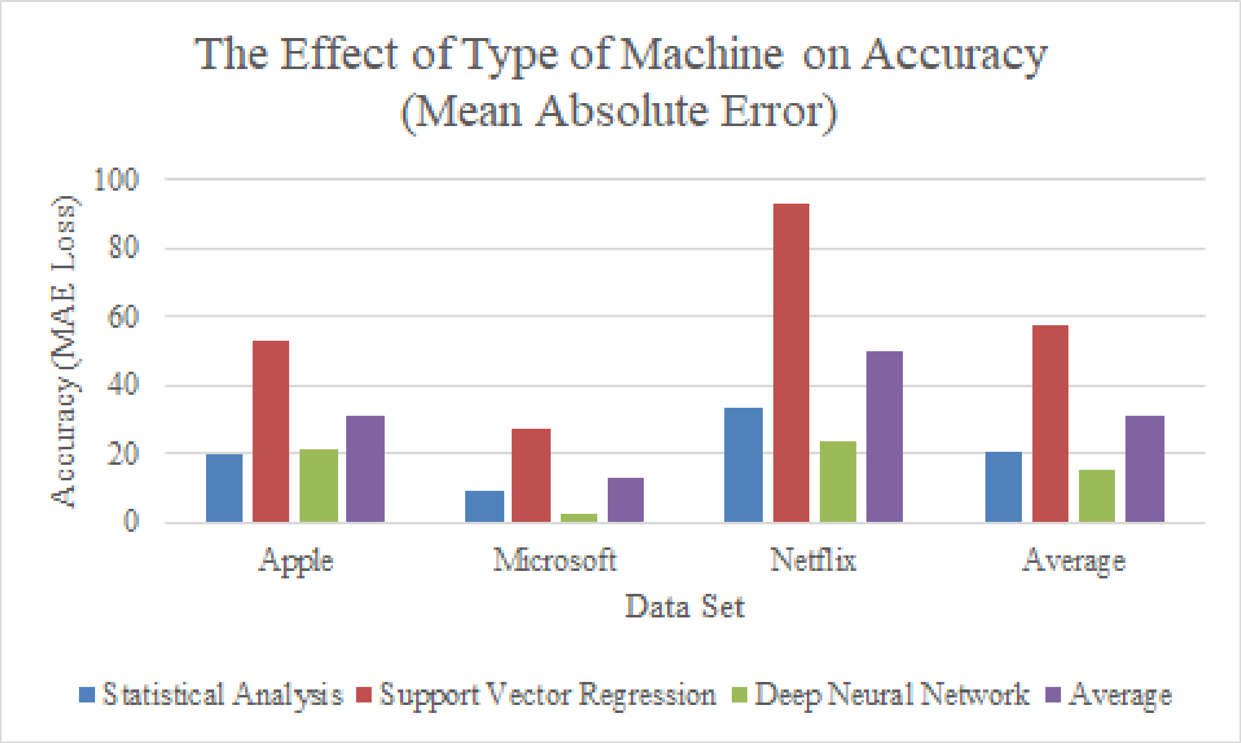
Acc	19.6	9.05	33.24	20.6	53.3	27.49	92.65	57.81	21.45	2.285	23.51	15.75	31.45	12.94	50.16	31.51
Data	Appl	Msft	Nflx	Avr.	Appl	Msft	Nflx	Avr.	Appl	Msft	Nflx	Avr.	Appl	Msft	Nflx	Avr.
Type	Statistical Analysis				Support Vector Regression				Neural Network				Average			

This table was from the original format; the layout for easier reading.

Data set	*Statistical Analysis*	*Support Vector Regression*	*Neural Network*	*Average*
Apple	19.6	53.3	21.45	31.45
Microsoft	9.05	27.49	2.285	12.94
Netflix	33.24	92.65	23.5	50.16
Average	20.6	57.81	15.75	31.51

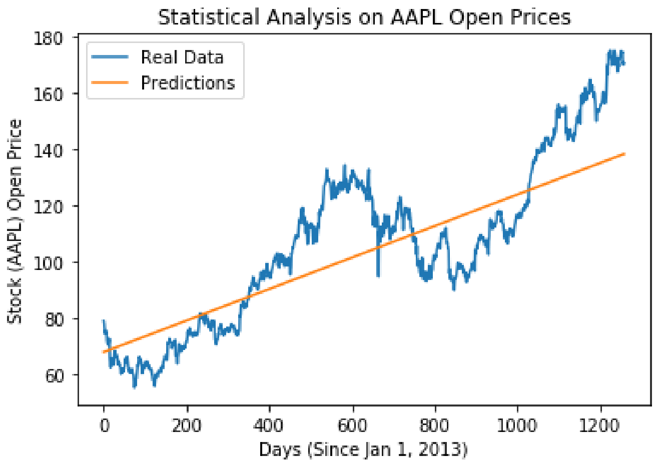
This was the final data table used for creating the visuals. It consists of the specific performances for each machine on each data set, the average per machine/data set, and the overall machine performance average. As seen, data table structure was an adjustment from the original plan (see “Discussion/Changes from Original Research Plan”) as well as the graphic structure in turn.

Graphs

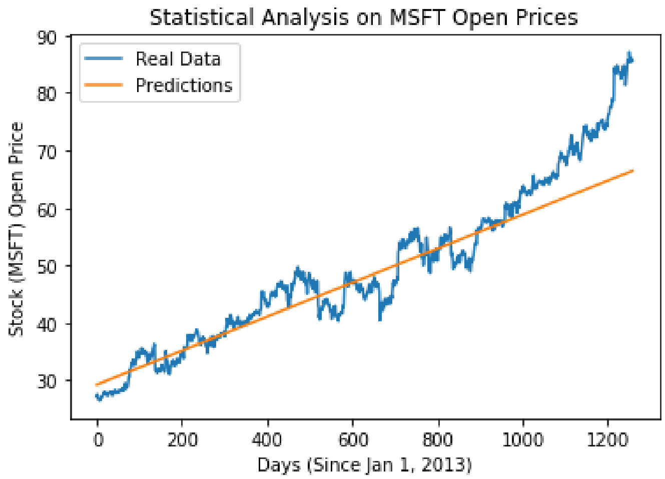


This graph was the main graph representing the final collected data. It is grouped by data set due to the fact that it allowed for better comprehension of the machine performance while also representing the average "toughness" of the data set. As shown, this data arises many conclusions, which are addressed in the next section (see "Discussion/Conclusions"). This was the most comprehensive graph of the experiment; in addition to this, we also created graphs showing each machine's predictions.

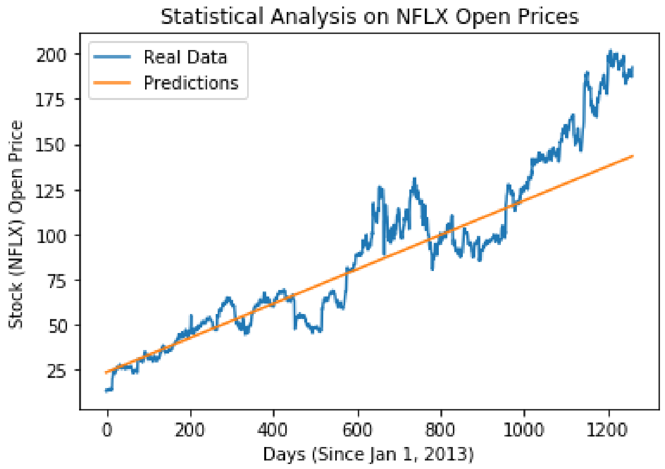
Data in Depth



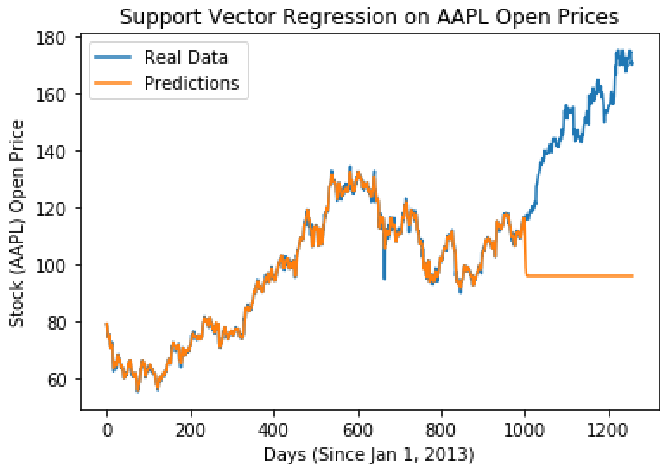
This graph represents the statistical analysis's trend line and its predictions on the AAPL dataset. This machine had a loss of about 19 and was the most effective for this dataset. It was trained till day 1,000 and tested from then on.



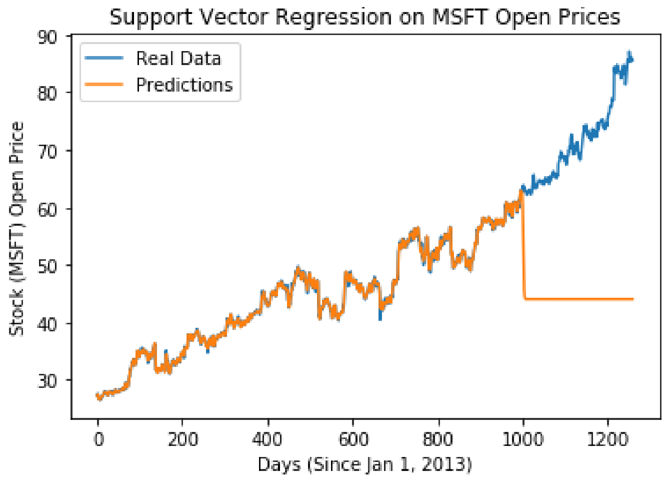
This graph represents the statistical analysis's trend line and its predictions on the MSFT dataset. This machine had a loss of about 9 and was the second most effective for this dataset. It was trained till day 1,000 and tested from then on. It satisfied the engineering goal.



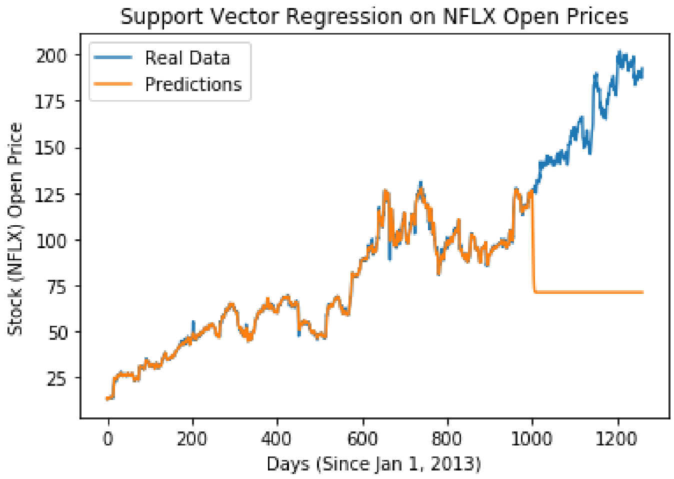
This graph represents the statistical analysis's trend line and its predictions on the NFLX dataset. This machine had a loss of about 33 and was the second most effective for this dataset.



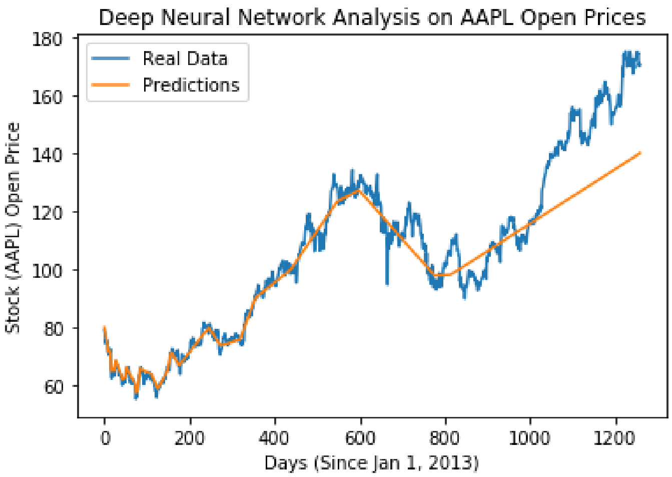
This graph represents the support vector regression's trend line and its predictions on the AAPL dataset. This machine had a loss of about 53 and was the least effective for this dataset. It was trained till day 1,000 and tested from then on.



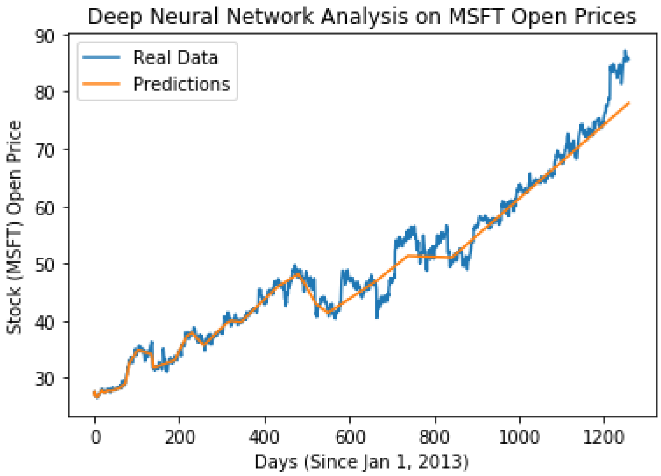
This graph represents the support vector regression's trend line and its predictions on the MSFT dataset. This machine had a loss of about 27 and was the least effective for this dataset. It was trained till day 1,000 and tested from then on.



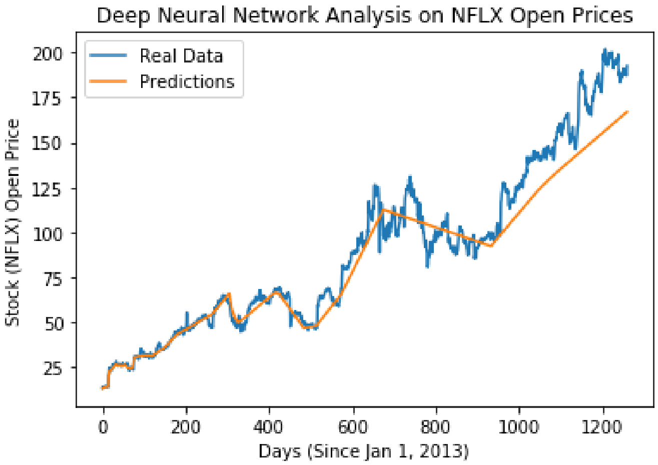
This graph represents the support vector regression's trend line and its predictions on the NFLX dataset. This machine had a loss of about 92 and was the least effective for this dataset. It was trained till day 1,000 and tested from then on.



This graph represents the deep neural network's trend line and its predictions on the AAPL dataset. This machine had a loss of about 21 and was the second most effective for this dataset. It represents the importance of big data for a more accurate trend summary line.



This graph represents the deep neural network's trend line and its predictions on the MSFT dataset. This machine had a loss of about 2 and was the most effective for this dataset. This machine was the most accurate overall and shows the best data/machine combination possible.



This graph represents the deep neural network's trend line and its predictions on the NFLX dataset. This machine had a loss of about 24 and was the most effective for this dataset. It represents the importance of big data for a more accurate final slope.

Discussion

This section will review over the plan changes, conclusions, trends, data, adjustments, discussions, and real world applications.

Changes from Original Research Plan

The experiment was quite static throughout and didn't go through many different changes. For one, the original plan was to measure accuracy as a percent; this was then changed to another method called a loss (how many dollars off the machine was on average) due to the fact that it summarized the data in a more comprehensive manner (using a loss called mean absolute error). In addition to this, we decided we wanted a machine who had a loss lower than 10 (the top 10% is 90 and above assuming the max loss is 100). In response to this, an average factor was added to the tables, graphs, and charts. The experiment also was intended to incorporate a learning curve; this was soon dropped as it skews the overall meaning of the data and causes confusion with other parameters. Overall though, the experiment endured through subtle changes.

Conclusions

Upon resolving the data from the experiment, a couple conclusions arose. First and foremost, the neural network was the most accurate in predictions, then the statistical analysis, and finally the support vector regression (as previously predicted). The neural network was most likely effective due to its far higher complexity in comparison to the other machines (in which it could recognize smaller trends while still being able to interpret the big picture through its weighted sum infrastructure). The statistical analysis was probably the next most effective because it summarized the data to the most simplest form possible (a line with a variable degree of 1) while taking into account the trends of a continuous variable. The support vector regression was most likely last due to the fact that it summarized the data into a line too generally (in the sense that it wasn't expecting a continuous variable, rather, one that came previously). It also was a classical example of overfitting data. Another important trend to notice was the difference in data sets. As shown by the average column, Microsoft data tended to be the most predictable, then Apple, then Netflix (probably due to its high range of y values resembling a scaling). Almost all of the machines followed these trends except for one instance in which the statistical analysis outperformed the neural network on the Apple data set. This creates some generalizations about the apple data set and its graphic in comparison to other data sets.

Sources of Error

While the experiment yielded useful results and fulfilled the engineering goal of a loss less than 10, the machines used in this experiment are far from ready for practical usage. This is mostly due to the multitude of sources of errors possible. For example, one major confirmed source or error was the fact that all of the datasets were positively changing (in the sense that they were all rising). This would lead to all of the machines being tried only on a positively changing dataset (which isn't realistic in the real world). Another error was the use of support vector regression. The main problem with this is that it wasn't expecting a continuous value (one of which is out of the train range); instead, it expected a value provided and averaged to achieve that. Another primary error which could have occured is a programming error. Whilst this is unlikely due to the results and alignment with other sources, it is a possibility for overall improvement. In addition to these, there were many other sources of error in this experiment, but this doesn't mean the data is useless. Rather, the data helps imply which machines to use where, which are the overall most effective, what is the average loss possibility, etc. All in all, this experiment had major flaws; despite this, it still proves useful in conveying an analysis.

References

Artificial Intelligence. (2019). In Gale In Context Online Collection. Detroit, MI: Gale. Retrieved from https://link.gale.com/apps/doc/VOJSVI051385296/SUIC?u=fairfaxcps&sid=SUIC&xid=7cc5c749.

Denoeux, T. (2019, February 8). PDF. Compiegne, France. Retrieved October 24, 2019, from https://arxiv.org/pdf/1807.01846.pdf.

Denœux, T. (2019, March 27). Logistic regression, neural networks and Dempster–Shafer theory: A new perspective. Retrieved October 24, 2019, from https://www.sciencedirect.com/science/article/abs/pii/S0950705119301558.

Ghosh, P. (2019). How Is Machine Learning Used in the Stock Market? PC Quest, 32(2), 28. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=f6h&AN=134620092&site=ehost-live

Linear Regression. (2019). Retrieved from Yale.edu website: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

O'Shea, A. (2017, June 19). Stock Market Basics: What Beginner Investors Should Know - NerdWallet. Retrieved October 26, 2019, from NerdWallet website: https://www.nerdwallet.com/blog/investing/stock-market-basics-everything-beginner-investors-know/