

Predicting Stock Market Prices

Chandler Baggett

chandler.baggett@colorado.edu

Jacob Toomey

Jacob.Toomey@colorado.edu

Dmytro Ryzhkov

Dmytro.Ryzhkov@colorado.edu



Figure 1: An image of S&P 500 prices since inception. X-axis is year, y-axis is price

PROBLEM STATEMENT/MOTIVATION

In this paper, we describe a project concerning mining the stock market for data in order to search for discernible statistical trends. Many statistical methods have been used to mine the stock market for information in order to predict which direction a given stock will move. By definition, a stock market prediction is “the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange.” The counter to the idea that stock market prices can be predicted is based on the efficient-market hypothesis. This hypothesis suggests “that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable.” Prediction methods of the stock market consist of fundamental analysis, technical analysis, and data mining technologies. In this paper, we are focusing solely on the latter.

- Is it possible to find patterns in previous stock prices and the history of the stock market in order to predict its future performance?

- Can we utilize data mining and machine learning to recognize discernible patterns in stock prices?

CCS CONCEPTS

- Information Systems -> **Data Mining**

KEYWORDS

Finance, Economics, Stocks, Stock market, Trends, Prediction, Prices, Patterns, Future performance, NASDAQ, Machine learning, Python, Data mining

1 INTRODUCTION

Since the stock market was created, investors have taken numerous measures in order to try and increase their income from the stock market. While there have been very few successful investors, the stock market has proven to be very difficult to predict. Burton Malkiel, a well-known name in the world of finance, attributes this fact to what he calls the efficient-market hypothesis. The theory states that “stock prices are a function of information and rational expectations, and that newly revealed information about a company’s prospects is almost immediately reflected in the current stock price. This

would imply that all publicly known information about a company, which obviously includes its prices history, would already be reflected in the current price of the stock. Accordingly, changes in the stock price reflect release of new information, changes in the market generally, or random movements around the value that reflects the existing information set.” Malkiel addressed this theory in his book [A Random Walk Down Wall Street](#), arguing that stock prices could therefore “not be accurately predicted by looking at prices history.” In this paper, we plan to challenge the theory that stock prices cannot be predicted using data mining methods found in Python’s scikit-learn.

2 PREVIOUS WORK

- Genetic algorithms have been used to try and predict the stock market.
- Banks and high frequency trading firms hire analysts to predict the stock market and have talented developers writing algorithms and utilizing machine learning to try and predict the stock market.
- There are many papers written on the subject and a lot of academic research going into utilizing machine learning in order to predict the stock market.

2.1 Application of data mining techniques in stock markets; A survey by Ehsan Hajizadeh, Hamed Davari Aradakani, and Jamal Shahrabi ("Applying Data Mining Techniques." *Data Mining for Managers* (n.d.): n. pag. Web.)

- Keywords: Stock market, data mining, decision tree, neural network, clustering, association rules, factor analysis, time series
- “In this paper, an overview of application of data mining techniques such as decision tree, neural network, association rules, factor analysis, and etc in stock markets is provided.”

2.2 Financial Stock Market Forecast using Data Mining Techniques by K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik and P. Arumugam ("Financial Forecasting Problem and Data Mining Techniques." *Ordinary Shares, Exotic Methods* (2003): 1-4. Web.)

- Keywords: Data mining, Time series Analysis, Binomial test, Typical Price, Bollinger Bands, Relative Strength Index and Moving Average
- “This paper attempts to determine if it is possible to predict if the closing price of stocks will increase or decrease on the following day. The approach taken in this paper was to combine six methods of analyzing stocks and use them to automatically generate a prediction of whether or not stock prices will go up or down. After the predictions were made they were tested with the following day’s closing price. If the following day’s closing price can be predicted to

increase or decrease 70% of the time at the .07 confidence level, then this analysis would be an easy and useful aid in financial investing. Furthermore, the results would show that the results are better than random at a reasonable level of significance.”

2.3 Stock market time series forecasting with data mining methods by Milan Csaba Badics (Milan Csaba Badics, Awar, Kochmeister. *Stock Market Time Series Forecasting with Data Mining* (n.d.): n. pag. Web.)

- Keywords: Stock time series forecasting, Trading strategy, Neural networks, ICA, EMD, Data mining methods
- “This paper focuses on the best-known data mining methods suitable for active portfolio management, as well as their advantages and disadvantages, discussing which should be used when and how, and also touching upon the most important current research trends. The objective is to present the entire process, from the selection of stock prices to forecastic (OTP and MOL daily closing prices in this paper), through the definition of the necessary input variables and available data mining methods right up to the execution of the trade, essentially giving the reader a roadmap for forecast-based active portfolio management.

3 PROPOSED WORK

Our goal is to feed preexisting datasets that contain stock prices over a period of time into machine learning algorithms in order to make more educated predictions about the future prices of individual stocks. Our methodology will include code written in Python and consistent usage of the Scikit-Learn Library. After we have used machine learning tools to learn from past data, we will then interpolate our results to predict what the future price of a stock may be. Our goal is to make a hypothetical “profit” based on our statistical findings by running our algorithms on a past data set and checking to see if there is any correlation between how accurate our predictions are of different stocks. Currently, we have developed a program that places our stock data into a machine learning algorithm using Python library Scikit-Learn. After the data is inputted, the system uses Support Vector Regression in order to predict the stock price of a stock on a certain day. The original data will be plotted, with the model over it, and the prediction will be printed to the terminal. The graphs of these results can be seen below.

4 DATA SET

Our data set consists of .CSV files of past stock performance of the following: AAPL, AXP, BA, CAT, CSCO, CVX, DD, DIS, GE, GS, HD, IBM, INTC, JNJ, JPM, KO, MCD, MMM, MRK, MSFT, NKE, PFE, PG, TRV, UNH, UTX, V, VZ, WMT, and XOM. Each individual CSV file has the following attributes:

Volume, Symbol, Adjusted Close, High, Low, Date, Close, Open. The volume is defined as the number of shares that changed hands during a given day. The ticker symbol or stock symbol is an abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market. A stock symbol may consist of letters, numbers or a combination of both. The adjusted close is defined as a stock's closing price on any given day of trading that has been amended to include any distributions and corporate actions that occurred at any time prior to the next day's open. The high is defined as a security's intraday high trading price. Low is defined as a security's intraday low trading price. Open is defined as the price of the stock as the market opens and close is defined as the price of a stock as the market closes for the business day.

5 EVALUATION METHODS

We will be comparing our results to the previous work we have cited:

- Application of data mining techniques in stock markets; A survey by Ehsan Hajizadeh, Hamed Davari Aradakani, and Jamal Shahrabi
- Financial Stock Market Forecast using Data Mining Techniques by K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik and P. Arumugam
- Stock market time series forecasting with data mining methods by Milan Csaba Badics

6 TOOLS

We will be using Python in order to write code that will use machine learning in order to make predictions. We will be using the Scikit-Learn library in Python in order to help us make calculations. Scikit-Learn provides more than one method of mining for data so this will suffice as our tool for the duration of the project.

7 MILESTONES

7.1 What We Have Achieved

1. Data sets have been downloaded.
2. Python Library Scikit-Learn has been researched and is decidedly our main tool for conducting machine learning and data mining of the stock market.
3. We have met and decided to test our machine learning algorithms on a specific bundle of stocks.
4. We have run data sets to make a prediction of stock prices using a Support Vector Regression method.

7.2 What Remains

1. Our plan is to do more research on methods of mining for data in the stock market i.e. clustering
2. Using these new methods, we will attempt to make more predictions of how stocks might behave in the future. Scikit-Learn provides more than one method of mining for data so this will suffice as our tool.

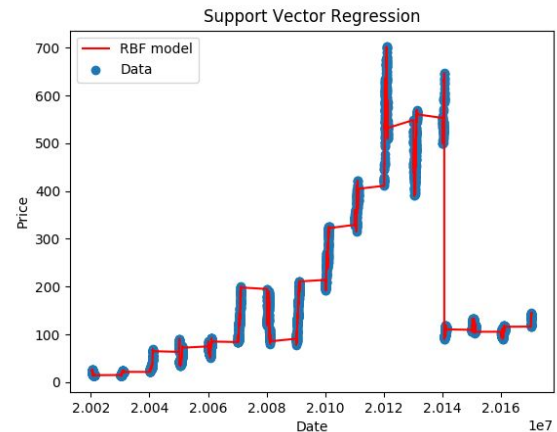
8 RESULTS

Predictions of AAPL, AXP, BA, CAT, CSCO for 5/1/17:

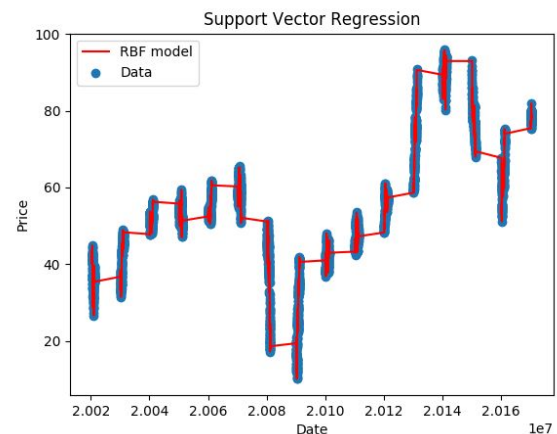
Example of code to run program (we will not be posting all commands in order to produce all the graphs as it becomes redundant):

AAPL:

```
[Chandler@MacBook-Pro:scripts Chandler]$ python predict.py AAPL.csv 20170501
Creating RBF model... (This takes a while)
The predicted price on 2017-05-01 is 188.19568899
```

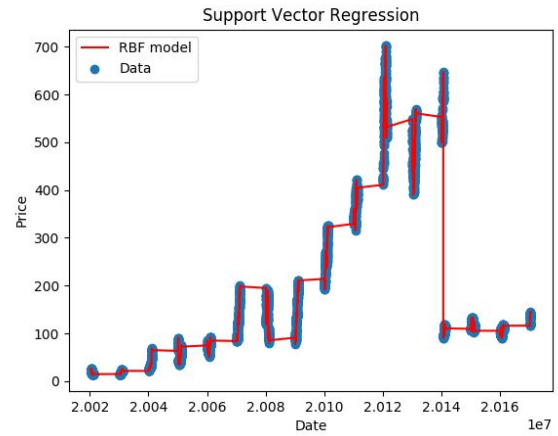
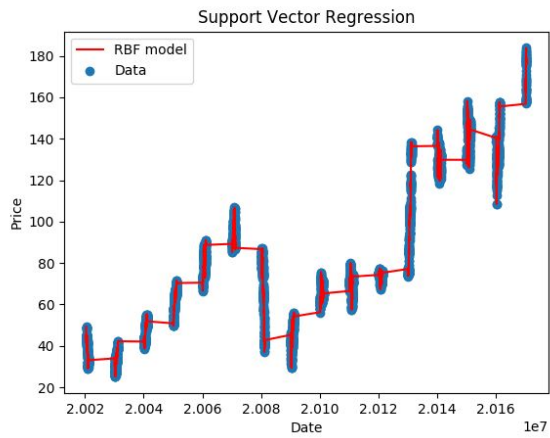


AXP:

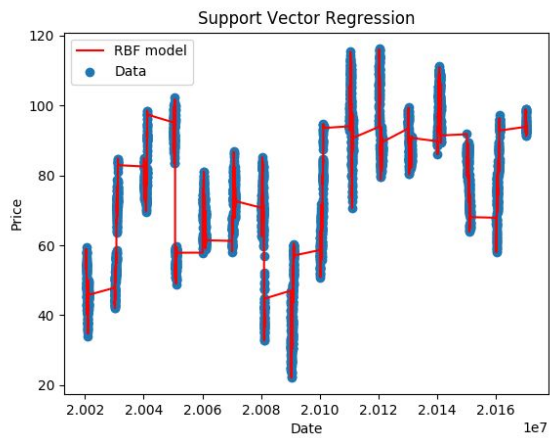


BA:

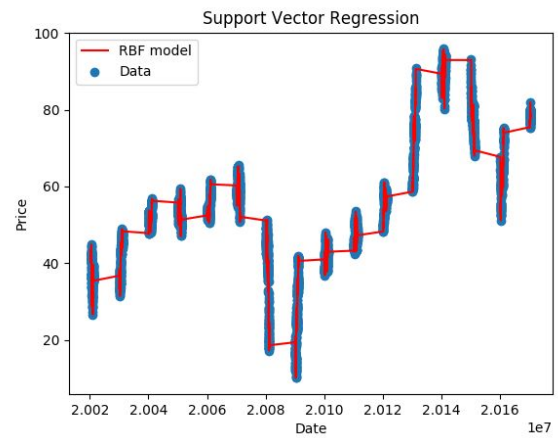
```
[Chandlers-MacBook-Pro:scripts Chandler$ python predict.py AAPL.csv 20170601
Creating RBF model... (This takes a while)
The predicted price on 2017-06-01 is 188.19568899]
```



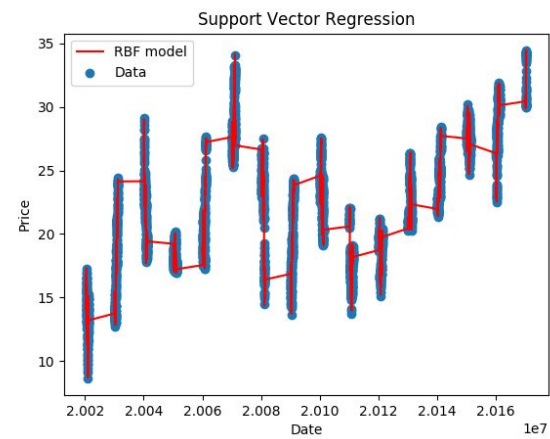
CAT:



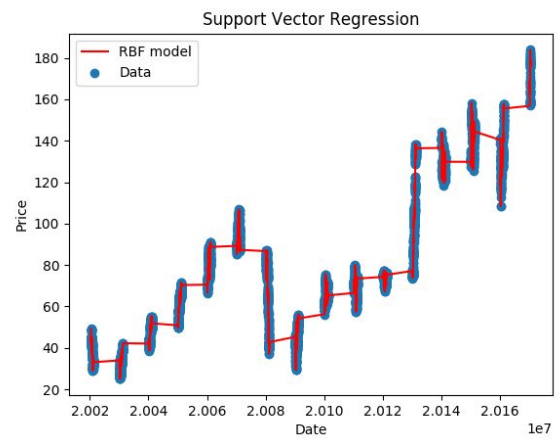
AXP:



CSCO:



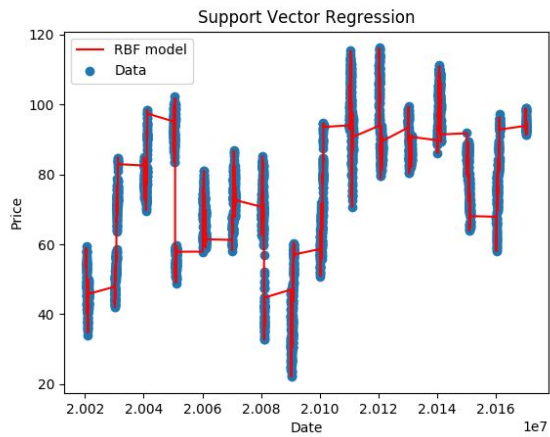
BA:



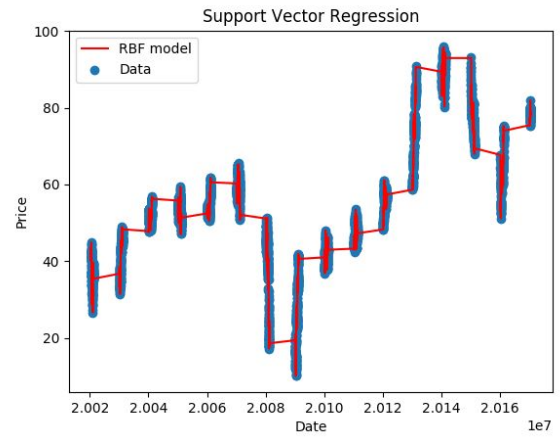
Predictions of AAPL, AXP, BA, CAT, CSCO for 6/1/17:

AAPL:

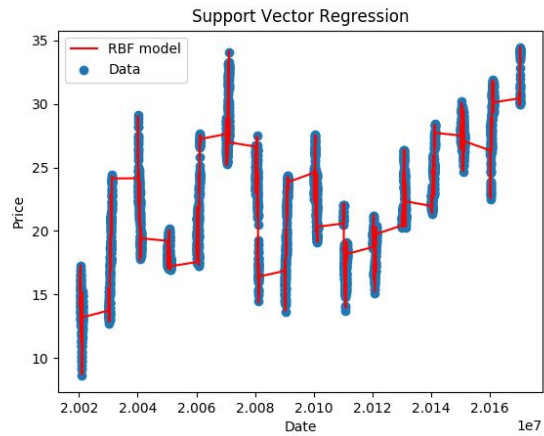
CAT:



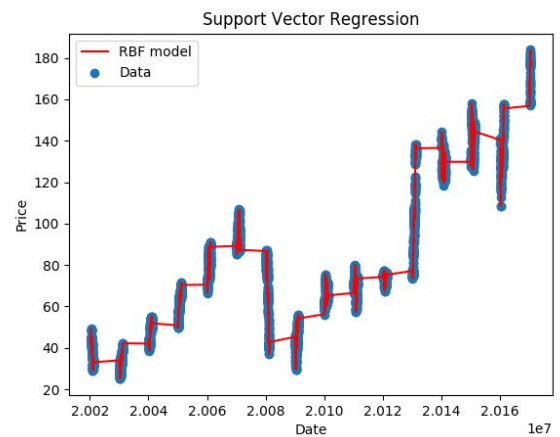
AXP:



CSCO:



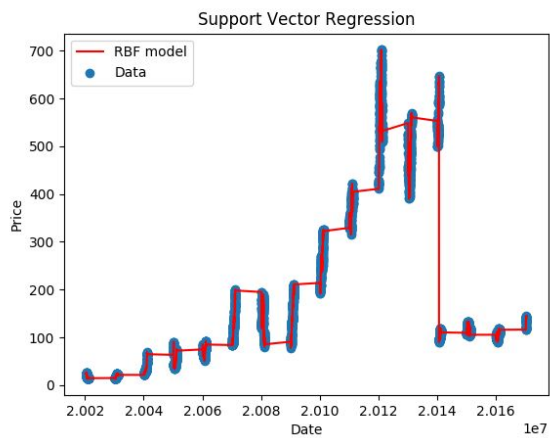
BA:



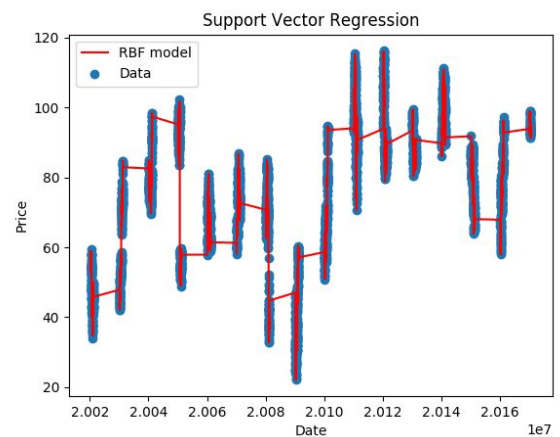
Predictions of AAPL, AXP, BA, CAT, CSCO for 6/1/18

AAPL:

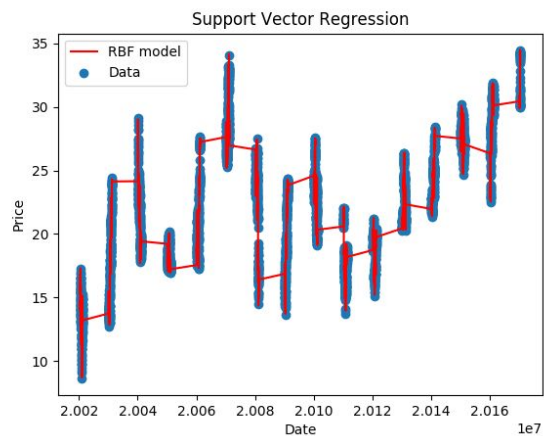
```
Chandlers-MacBook-Pro:scripts Chandlers$ python predict.py AAPL.csv 20180601
Creating RBF model... (This takes a while)
The predicted price on 2018-06-01 is 188.19568899
```



CAT:



CSCO:



A.1 Problem Statement/Motivation

A.2 Previous Work

A.2.1 Application of data mining techniques in stock markets; A survey by Ehsan Hajizadeh, Hamed Davari Aradakani, and Jamal Shahrabi

A.2.2 Financial Stock Market Forecast using Data Mining Techniques by K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik and P. Arumugam

A.2.3 Stock market time series forecasting with data mining methods by Milan Csaba Badics

A.3 Proposed Work

A.4 Data Set

A.5 Evaluation Methods

A.6 Tools

A.7 Milestones

A.7.1 What We Have Achieved

A.7.2 What Remains

A.8 Results

A.9 References

9 REFERENCES

- K. Senthamarai Kannan, P. Sailapathi Sekar, M. Mohamed Sathik and P. Arumugam, "Financial Forecasting Problem and Data Mining Techniques." *Ordinary Shares, Exotic Methods* (2003): 1-4. Web.
- Ehsan Hajizadeh, Hamed Davari Aradakani, and Jamal Shahrabi, "Applying Data Mining Techniques." *Data Mining for Managers* (n.d.): n. pag. Web.
- Milan Csaba Badics, Awar, Kochmeister. *Stock Market Time Series Forecasting with Data Mining* (n.d.): n. pag. Web.
- "Stock Market Prediction." *Wikipedia*. Wikimedia Foundation, 19 Jan. 2017. Web. 11 Apr. 2017.