Stock Prices Predictive System

Final Paper for University of Virginia

SYS 6014 Decision Analysis Spring 2020

Lin Wang
lw7kv

Apr 27, 2020

Abstract

As long as capital markets have existed, investors and aspiring arbitrageurs alike have strived to gain edges in predicting stock prices. In particular, by using of machine-learning techniques and quantitative analysis to make stock price predictions has become increasingly popular with time. In this project, firstly crawling google historical stock price data from kaggle.

We analyzed the multiple scenarios like predicting the closing price for day, given opening price and predicting the all **open, close, high, low prices and volume of the day** based on historical data since 03/17/2017 to 02/29/2020.

In this project, we applied the Support Vector Machines (SVM) model to achieve the final prediction result **，**When applying the model, we chose different algorithms to get the results separately In this process, we used Polynomial and Radial Basis Function (RBF) methods as SVR kernels to get different prediction results

After that, we use Mean Absolute Error (MAE) and Mean Squared Error (MSE) to evaluate the performance effect for different models

Finally, we summarize the performance of different models and use R-squared (R2) to discuss the payoffs of the stock recommendation system. According this project, users can figure out the trend of stock prices through historical data in some degree. It is helpful for users to avoid the failure, decrease risk and gain edges.

Keywords : Stock prices; Support Vector Regression; Deep Learning

1. Introduction

Stocks are called "money cubes", and the key to economic freedom is stock investment. Although this is the truth of the booming market, amateur trading stocks is still an attractive option today. The question is: which stocks? How to analyze stocks? How to determine the buying and selling of stocks

The efficient market hypothesis states that the factors that determine the price of stocks in the future are in the future thus making the future prices of stocks random andunpredictable. However, using the stock market prediction methods outlined in this program simplifies the process of prediction by removing the random element out of stock market price futures. The use of support vector machines for classification and regression analyses gives a scientific element to the prediction of stock market prices rather than relying on hunches and intuition. A combination of the thus computed results and sound investment planning will, therefore, raise an investor’s chance of stock market success.

In this article, we first grab the historical stock price data of Google from 03/17/2017 to 02/29/2020. You can adjust the start and end dates as needed. In the next analysis process, we will use the closing price, which is the final price of the stock at the end of the day of trading, as a reference for data analysis.

First, we use the support vector machine (SVM) model to obtain the final prediction results. In this process, we used the polynomial and radial basis function (RBF) method as the SVR kernel to obtain different prediction results.

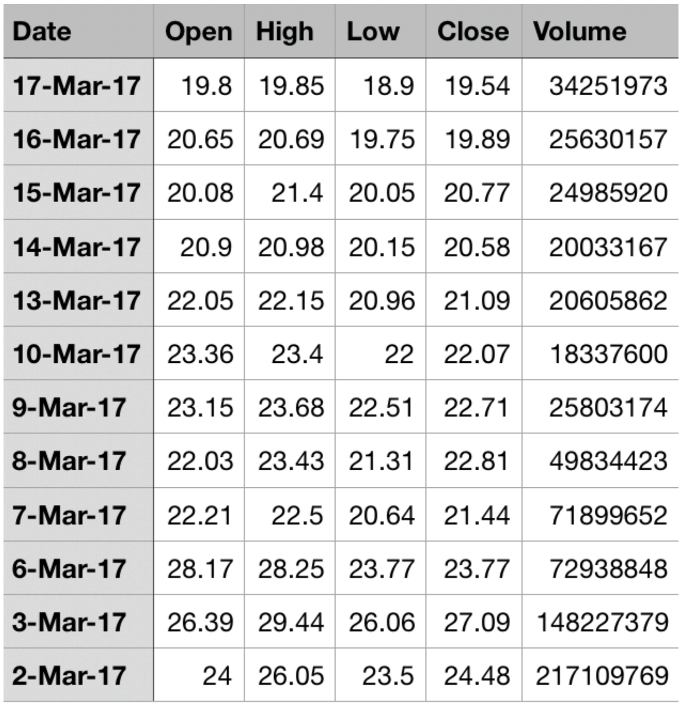
After that, we use mean absolute error (MAE) and mean square error (MSE) to evaluate the performance effects of different models

Finally, we summarize the performance of different models and discuss the returns of the stock recommendation system using R-squared (R2). Through this project, users can understand the stock price trend through historical data to a certain extent. This is very helpful for users to avoid failures, reduce risks and gain advantages

2. Data

2.1 Data Collection

The original data is crawled from kaggle, crawling all the stock records of Google from 03/17/2017 to 02/29/2020.



Date	Open	High	Low	Close	Volume
17-Mar-17	19.8	19.85	18.9	19.54	34251973
16-Mar-17	20.65	20.69	19.75	19.89	25630157
15-Mar-17	20.08	21.4	20.05	20.77	24985920
14-Mar-17	20.9	20.98	20.15	20.58	20033167
13-Mar-17	22.05	22.15	20.96	21.09	20605862
10-Mar-17	23.36	23.4	22	22.07	18337600
9-Mar-17	23.15	23.68	22.51	22.71	25803174
8-Mar-17	22.03	23.43	21.31	22.81	49834423
7-Mar-17	22.21	22.5	20.64	21.44	71899652
6-Mar-17	28.17	28.25	23.77	23.77	72938848
3-Mar-17	26.39	29.44	26.06	27.09	148227379
2-Mar-17	24	26.05	23.5	24.48	217109769

(Table 1 Stock prices from Google from 03/02/2017-03/17/2017)

In this project, we used 5 features that are more useful in stock trading. The opening price of the stock, the closing price, the highest price of day, the lowest price of the day and volume of the day.The whoel dataset is too large. It is difficult for individuals to calculate and process the data. Thus, we took one month of data as a sample to evaluate the entire system as a demo.In the next analysis, we will use the closing price, which is the final price of the stock at the end of the day's trading.

3. Prediction Model

3.1 Support Vector Regression

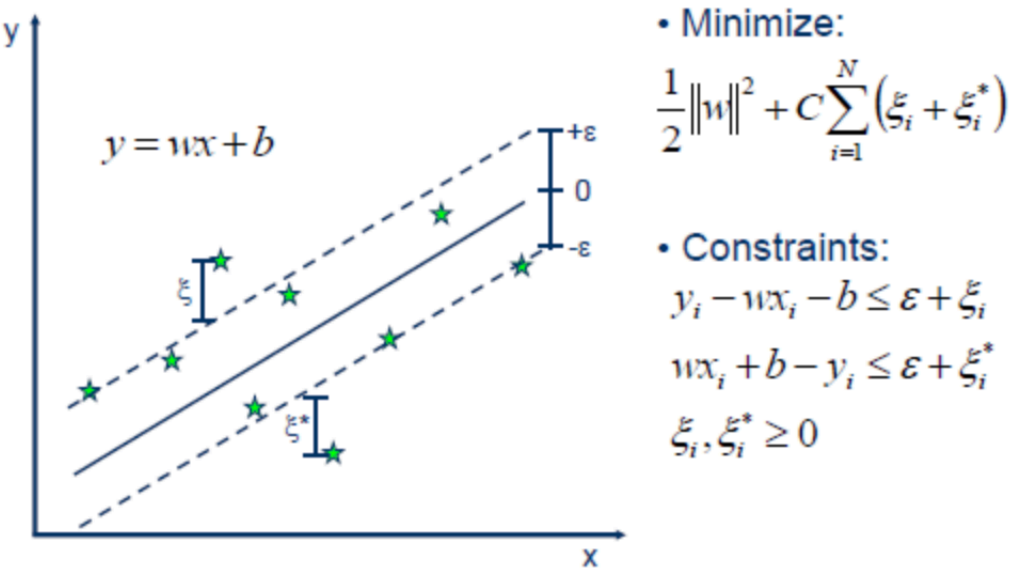
Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification.

For the SVR, the prediciton function is

$$Min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}|\xi_i| \tag{1}$$

and the constraints

$$\begin{aligned} y_i - wx_i - b &\le \epsilon + \varepsilon_i \\ wx_i + b - y_i &\le \epsilon + \varepsilon_i \end{aligned} \tag{2}$$


y
x
y = wx + b
ξ
ξ*
+ε
0
-ε
• Minimize:
$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$
• Constraints:
$y_i - wx_i - b \leq \varepsilon + \xi_i$
$wx_i + b - y_i \leq \varepsilon + \xi_i^*$
$\xi_i, \xi_i^* \geq 0$

(Figure 1. Principle diagram of SVM)

This method builds the predictive model and graphs it. It takes three parameters: dates, prices(the prices of train set), and x (price of target date). This function creates 2 models, each of them will be a type of support vector machine with different kernel.

As Figure .1 shown, tt will be such that the distances between the closest points in each of the two groups are farthest away. When we add a new data point in our graph depending on which side of the line it is, we could classify it accordingly with the label.
The support vector regression uses the space between data points as a margin of error and predicts the most likely next point in a dataset.

3.1.1 RBF

Gaussian **RBF**(Radial Basis Function) is another popular Kernel method used in **SVM** models for more. **RBF** kernel is a function whose value depends on the distance from the origin or from some point.

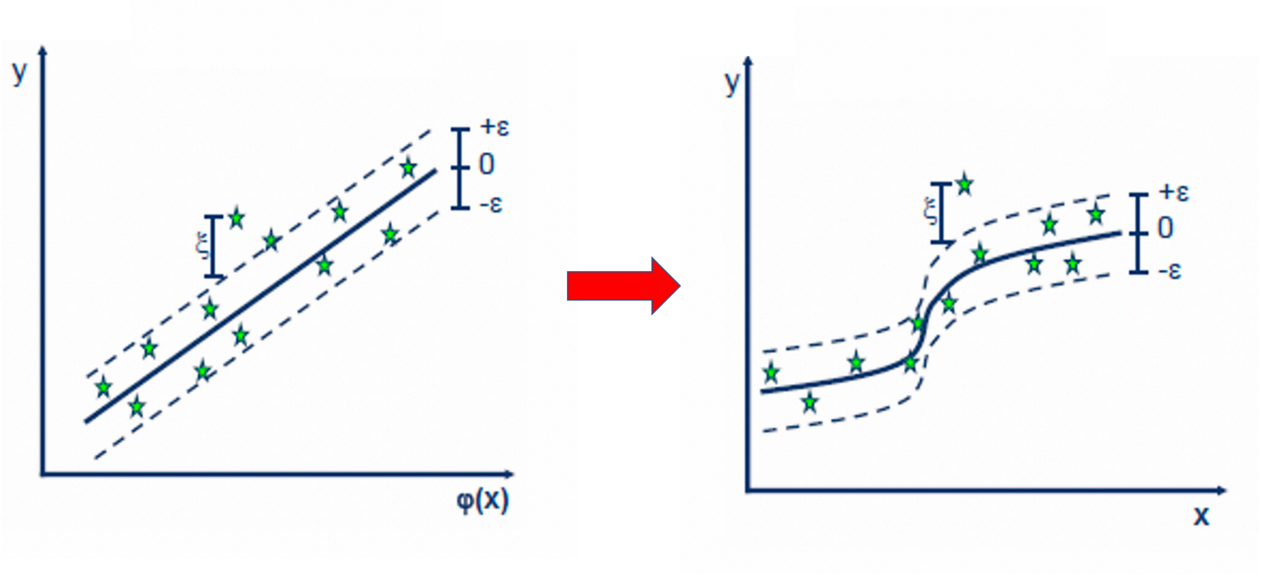
And in this project, we use kernel functions transform the data into a higher dimensional feature space to make it possible to perform the linear separation.

The prediction function is

$$y = \sum_{i=1}^{N}(a_i - a_i^*).\, K(x_i, x) + b \tag{3}$$

where for the RBF the kernel function is

$$K(x_i, x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \tag{4}$$



(Figure 2 Transformation of from liner to non-linear SVR when apply different kernel)

3.1.2 Polynomial

Intuitively, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the context of regression analysis, such combinations are known as interaction features. The (implicit) feature space of a polynomial kernel is equivalent to that of polynomial regression, but without the combinatorial blowup in the number of parameters to be learned.

In this porjcet, similarly, we have same prediction function as rdf, but with different kernel function:

$$K(x_i, x_j) = (x_i, x_j)^d \tag{5}$$

3.2 Evaluation index

3.2.1 Mean Squared Error

the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss.

$$\text{MSE} = \frac{1}{|n|}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_2\right)^2 \tag{6}$$

3.2.2 Mean Absolute Error

In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{7}$$

3.3 Evaluation of Payoffs

3.3.1 R-squared

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. In this project, we use R-squared to evaluate the payoffs of using this predictive model .

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_{pred} - y_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} \left(\bar{y}_i - y_i\right)^2} \tag{8}$$

If R2 is laydown between 0 ~ 1, the closer to 1, the better the regression fitting effect. Generally, the model with a value of more than 0.8 is considered to have a good fit.

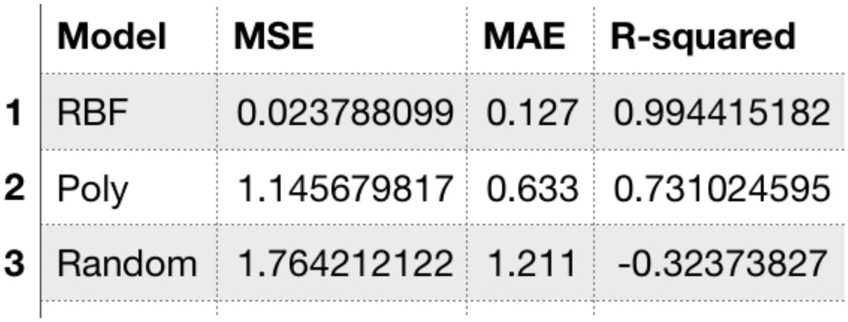
3.3.2 Payoffs Calculation

The original probability of retaining a user successfully is assumed as 0.5. Based on the result of different model, this probability would be calculated and renewed as follows:

$$P = 0.5 \times \left(1 + R^2\right) \tag{9}$$

4. Result Analysis

4.1 Performance of different model

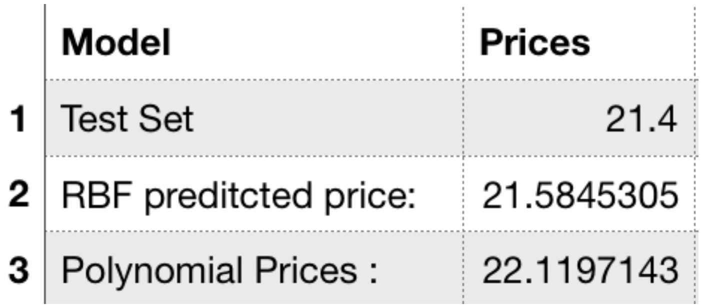


	Model	MSE	MAE	R-squared
1	RBF	0.023788099	0.127	0.994415182
2	Poly	1.145679817	0.633	0.731024595
3	Random	1.764212122	1.211	-0.32373827

(Table 2 performamce of different model)

From the Table 2, we can figure out that when apply RBF model, we get the best performance. RBF has less errors than Poly in MAE and MSE, also has better scores in R2.Thus, we choose RBF as kernel of SVR in this prices prediction project.

4.2 Prices compare

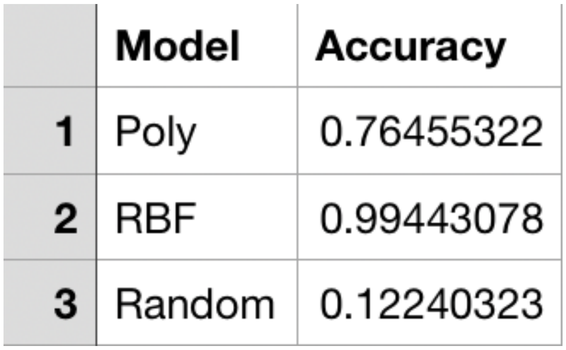


	Model	Prices
1	Test Set	21.4
2	RBF preditcted price:	21.5845305
3	Polynomial Prices :	22.1197143

(Table 3 The prices compare with testset)

From Table 3, In this process, we randomly selected a point to test the accuracy of our model stock prediction results. We took March 7 as the object. It was found that the value of RBF(21.5845305) is closer to the real value(21.4) than Poly(22.1197143) compared with the actual value.

4.3 Payoffs Analysis



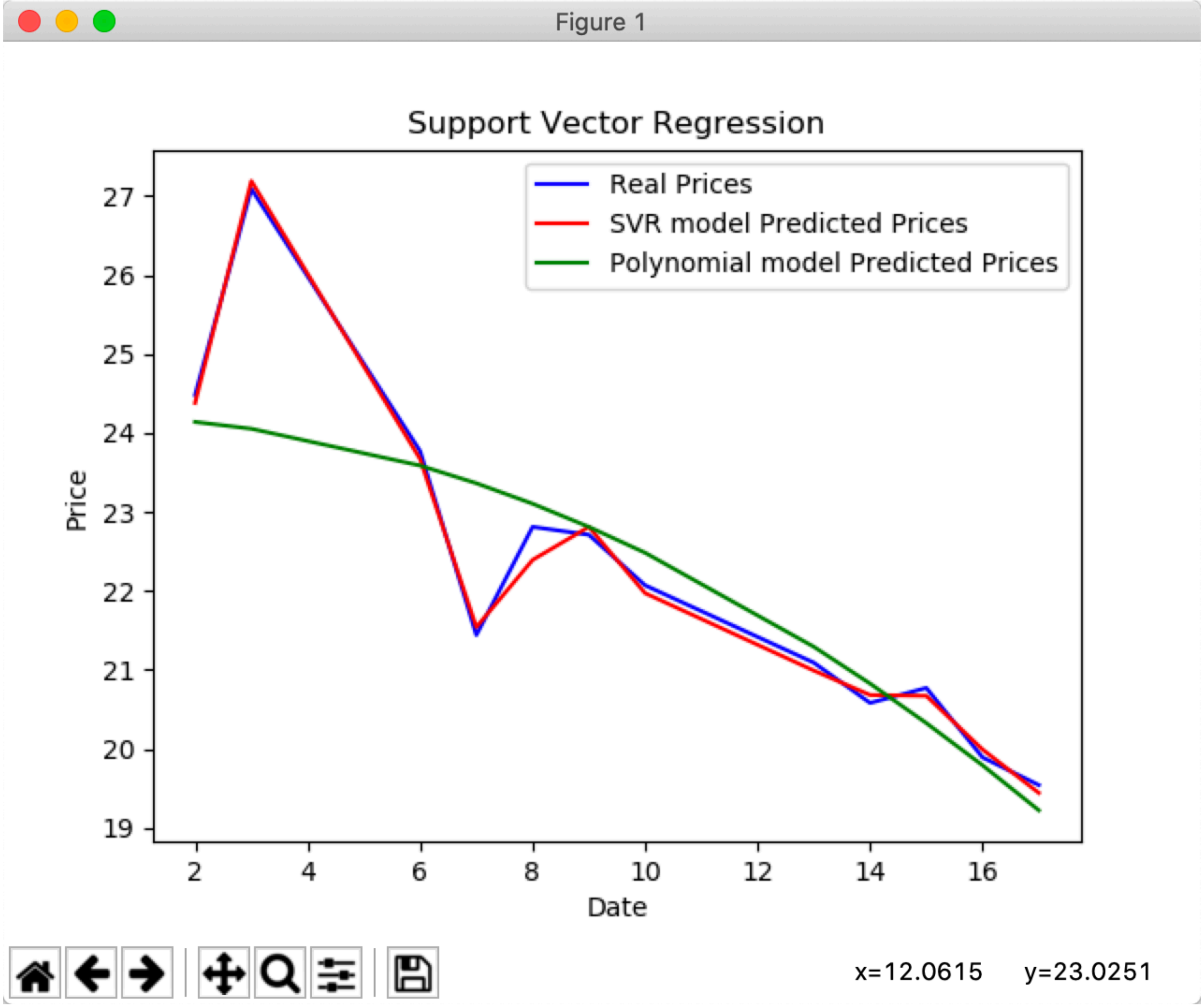
	Model	Accuracy
1	Poly	0.76455322
2	RBF	0.99443078
3	Random	0.12240323

(Table 4 Payoffs of different model)

According to equation(9), we can calculate the payoffs of each model. And we can easily to figure out the when apply RBF kernel we get the best performace. And result show It shows predictive model could get 87.44% improvement.

4.4 Poly Analysis

5The support vector regression estimates how each addition or modification of data affects the prediction and outlook on the future prices of stock. The support vector regression can be developed by using either the linear function model, the polynomial functions model or the ration basis model. The different results can then be plotted on one or different graphs for analysis). These graphs are then compared with the actual data from the company’s history and the model that matches the historical data and trends can then be used to predict how the figures will react to market stimulate.



(Figure 3 The polt of final result)

On analyzing the graph, we see that each of our models shows up in the graph and the RBF model seems to fit our data the best. From the Figure 3, We can find that the overall trends of Poly and RBF are consistent with the actual values, but compared to poly RBF, the actual data has a better fit and performs better in model prediction, but as the amount of data increases, we It can be found that the entire image has the same approach value in the future

5. Conclusion & Future Work

In this project, we have implemented a stock price prediction system based on the SVM model. The data comes from the trend of Google's stock changes from 2017 to 2020. In the process of establishing this price prediction model. Different kernels are used as algorithms to compare results such as RBF and poly. We conclude that RBF can achieve better performance and performance in this project. In addition, we also compared the real data and the predicted data and found that the error is within the acceptable range, so we verified the accuracy and reliability of the predicted results. Users can use this prediction system to avoid risks in stock trading and obtain greater benefits

But the prediction system is still relatively simple. First, because of personal computing power, we only sampled one month in the data. The so-called training object results are relatively simple. In addition, for the real world, we need to consider more parameters and variables, and the model is more complicated. And changeable. In the furture work, we can use more complex deep learning algorithms to provide better, more accurate and reliable stock price prediction models based on big data

Reference

1. Introduction to Support Vector Machines
2. Stock Market Prediction Using
3. Stock Price Prediction With Big Data and Machine Learning
4. How I made $500k with machine learning and HFT (high frequency trading)
5. Boston Data Festival Hackathon
6. Buffalo Capital Management