Survey of Predictive Models for Stocks

Braden Huffman

Theory of Predictive Modeling

CS 580/PHSCS 513R

Introduction

Predictive modeling as a field exists because of utility of predicting and understanding complex systems in the real world. Stock change price due to the buying and selling of their shares by humans who inherently make complex decisions. The goal of this project is to survey the utility of different types of buckets of models with regards to the stock NIO.

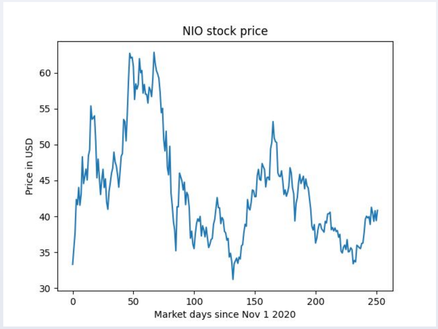

Fig 1: Nio price from 11/01/2020 to 11/01/2021

Data Collection

Data was collected from investing.com's historical data tab on the NIO page. Data was collected from 11/01/2020 to 12/06/2021.

For each day the market was open four points of data were collected: opening price, highest price, lowest price, closing price.

Assumptions

Let "the current day" be any arbitrary day the model is looking at.

The following assumptions were made all analysis methods:

- The opening, high, low and closing prices are the only things that affect stock price.

The following assumptions were made for Linear and Nonlinear Models:

- Prior day's stock action does not affect the current day's stock action.
- Closing price on a day can be predicted using that day's opening, high and low stock price.

The following assumptions were made for Dynamical Systems:

- Only the previous day's stock action directly affects the current day's stock action.
- Current day's opening, high and low prices do not affect the current day's closing price.
- Stock action happens in discrete time.

Notation

- o_i - The opening price on the ith market day since 11/01/2020.
- c_i - The closing price on the ith market day since 11/01/2020.
- h_i - The highest price on the ith market day since 11/01/2020.
- l_i - The lowest price on the ith market day since 11/01/2020.
- θ_i - The ith parameter in the column vector θ.

Linear Regression

To better understand stock price movement, 10 different buckets of linear models were analyzed. Opening, high and low prices on the current day were used to predict the current day's closing price. This was done using the following equation.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} ||A\theta - c||_2$$

Where A is a matrix representing a linear equation for all (o,h,l) in the training data set, θ is a column vector containing the parameters, and c is a column vector containing the closing prices for the training set. The parameters that reduce the 2-norm of Aθ-c are said to be "optimal" for a given bucket. The optimal parameter values were found using np.linalg.lstsq().

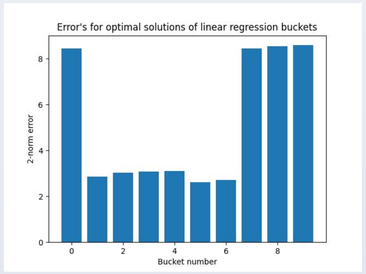

Fig 2: Validation set errors for different optimal linear buckets from 11/01/2020 to 11/01/2021.

The bucket that performed best on the validation set (10% of the entire data set) which it had not seen before was bucket 5. Bucket 5 can be represented by the following equation:

$$\begin{aligned} c_i = {} & \theta_1 + \theta_2 o_i + \theta_3 h_i + \theta_4 l_i + \theta_5 o_i h_i + \theta_6 o_i l_i + \\ & \theta_7 h_i l_i + \theta_8 o_i^2 + \theta_9 h_i^2 + \theta_{10} l_i^2 + \theta_{11} o_i^2 h_i + \theta_{12} o_i^2 l_i + \\ & \theta_{13} h_i^2 l_i \end{aligned}$$

This model had the most parameters out of all of the models tried. The average closing price was \$1.45 different than the opening price, \$1.29 off the highest price and \$1.42 off the lowest price. The model was off by \$0.52 on average. This shows the model was a more accurate predictor of stock closing price than the opening price, highest price or lowest price.

Nonlinear Regression

The nonlinear regression models aimed to solve the same problem as described in the linear regression section above; however this time nonlinear sets of equations were used to find the optimal parameters to minimize the following 2-norm:

$$\hat{\theta} = \underset{x}{\operatorname{argmin}} ||f_i(x) - c||_2$$

Where f_i is the ith nonlinear bucket. The optimal parameter values were found using the scipy.optimize.minimize() SLSQP method.

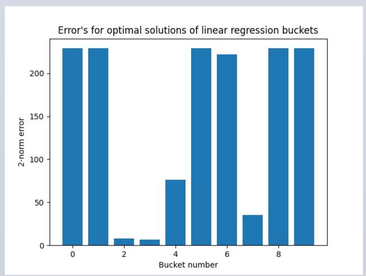

Fig 3: Validation set errors for different optimal nonlinear buckets from 11/01/2020 to 11/01/2021.

Nonlinear Regression cont.

There errors in (fig 3) are the errors of the models on the same validation set that the linear models were tested on. The best performing nonlinear model (model 3) was off on average by \$1.37. Model 3 can be represented by the following equation:

$$c_i = \theta_1 \log(|\theta_2 o_i h_i|)$$

This bucket of models did not contain the most parameters of the nonlinear buckets, and it also did not use the lowest price of the day. While more testing needs to be done, preliminary results suggest it is easier to find a better fit for the problem describe in this and the previous section using linear regression.

Dynamical Systems

Dynamical systems were used to predict stock price action from one day to the next. The 10 models that tested were intended to predict the stocks action for a year given only the opening, highest, lowest and closing prices on 11/01/2020.

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmin}} \| \widehat{c}(o_0, h_0, l_0, c_0) - c \|$$

where optimal solution was of the form:

$$\begin{aligned} \hat{o}_{i+1} &= \theta_1 o_i + \theta_2 h_i + \theta_3 l_i \\ \hat{h}_{i+1} &= \theta_4 o_i + \theta_5 h_i + \theta_6 l_i \\ \hat{l}_{i+1} &= \theta_7 o_i + \theta_8 h_i + \theta_9 l_i \\ \hat{c}_{i+1} &= \theta_{10} o_i + \theta_{11} h_i + \theta_{12} l_i \end{aligned}$$

This model outperformed a similar model that also included closing price of the previous day. The following graph shows the model's prediction given the initial conditions on 11/01/2020. The optimal parameter values were found using scipy.optimize.minimize() SLSQP method.

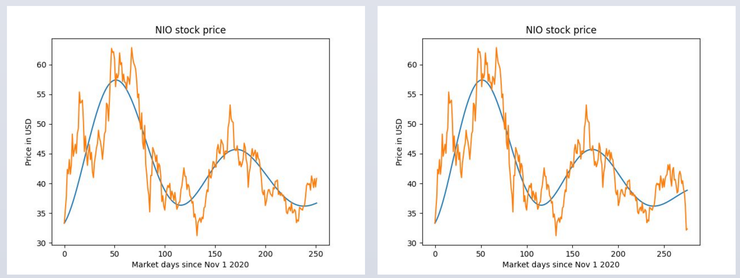

Fig 4: Predictions from 11/01/2020 to 11/01/2021 on training data (left), Predictions from 11/01/2020 to 12/06/2021 including data not trained on (right).

In reality, humans (and stocks by extension) are complex. The optimal model did not contain enough parameters to adjust to every sharp stock movement. The model follows the overall trend of the graph, and demonstrates dynamical systems can be used to represent stock action.

Conclusion and Future Work

This survey of models in regards to NIO's stock price demonstrates that modeling can be a useful tool in market prediction assuming a useful bucket of models is chosen.

Future work involves creating more complex linear models in order to ascertain more information on how stock prices vary within a day to use for swing trading. Furthermore, the the bucket of the dynamical system shown above will be introduced to a larger time frame of NIO data, as well as other stocks and ETFs.