Capstone project report: Prediction of future security prices

Jayendra Vadrevu, Shweta Gaur, Nikhil Maddirala, & Vidit Sharma

1 Introduction

Is it possible to predict future security prices based on historical data? Since the advent of the stock market, this question has been hotly debated by economists, investors and business professionals of all stripes. Economists who subscribe to the efficient markets hypothesis argue that stock market prediction is by definition impossible, while most professional investors on the other hand stake their entire livelihood on the assumption that stock markets are predictable. Given that both sides continue to attract and retain strong proponents — the former camp features Nobel Laureate economists such as Eugene Fama, while the latter camp features multi-billionaire investors such as the Warren Buffet — it is unlikely that this debate will be settled anytime soon.

In recent times, this debate has been further intensified by the explosion of big data and the increasing ubiquity of analytics. Whereas traditionally stock market prediction was limited to two methodologies: (1) fundamental analysis and/or (2) technical analysis (a.k.a. charting), we are now seeing the emergence of a new methodology for stock market prediction which can be called (3) big data analytics. Although professional investors now have a powerful new tool in their arsenal for stock market prediction, there is no clear consensus on the kinds of data and the methods of analysis that can yield useful insights for the purposes of stock market prediction.

One class of data that has received significant attention from investors, analysts and academics alike in this context is sentiment data. While it is common knowledge that markets react sharply to news reports, it has been a challenge to quantify the impact of this reaction. The goal of this project was to build a stock market prediction model (for individual securities and for portfolios of securities) that takes sentiment data as a key input into the model.

2 Description of the data

There were three primary types of data we collected: (1) historical stock prices, (2) macroeconomic and financial indicators, and (3) sentiment score. Each of the three types of data had to be collected from various sources, and then appropriately cleansed / transformed, and then finally reconstructed into new variables (if necessary):

- Stock prices
  - Source: Yahoo finance
  - Cleansing / transformation: Making the data static by taking the differential (1st and 2nd differential)
- Macroeconomic and financial indicators (S&P, USD Crude, etc.)
  - Source: Multiple publicly available sources (CBOE, World Bank, etc.)
  - Cleansing / transformation: Dealing with missing values, standardization of dates, etc.
- Sentiment scores: individual stock sentiment as well as market sentiment
  - Source: Hedgechatter
  - Cleansing / transformation: Starting with the source data from Hedgechatter, which was represented by 5 variables (Strong buy %, Buy %, Hold %, Sell %, and Strong Sell %), we constructed three variables to summarize the sentiment: positive, neutral, and negative.

3 Methodology

ARIMA model. Our model of choice is the Autoregressive Integrated Moving Average (ARIMA) model, one of the most prominent methods in economic and financial forecasting, which has been explored extensively for

time series prediction. ARIMA models have been shown to generate efficient short-term financial forecasts, consistently outperforming complex structural models in short-term prediction.

Modeling technique. We built three models for the three individual stocks in the automobile industry (Tesla, GM, and Ford). The procedure is as follows:

1. First, we ensure that the time series stationary (by means of the first / second differential)
2. Next we attempt to fit the auto.arima model (i.e. ARIMA 0,0,0) and evaluate the quality of this estimate based on ACF, PACF and RMS error terms.
3. In case the auto.arima model is not a good fit, we try varying the p and r terms by a process of trial and error to see which yields the optimal model.
4. Choose the best ARIMA model based on the above process.
5. Add the appropriate external regressors (sentiment data and macroeconomic data) through a continuous and iterative process of model building and validation

4 Results and analysis

Results and analysis are model specific. As an illustration, we refer to the Tesla model's results:

- Macroeconomic and financial indicators are not significant
- Negative stock sentiment (sentiment about Tesla) is significant (but not positive or neutral)
- Positive market sentiment (sentiment about the total market) is significant (but not negative or neutral)
- Adding market and stock sentiment score improves the predicted return direction from 55% to 67%.

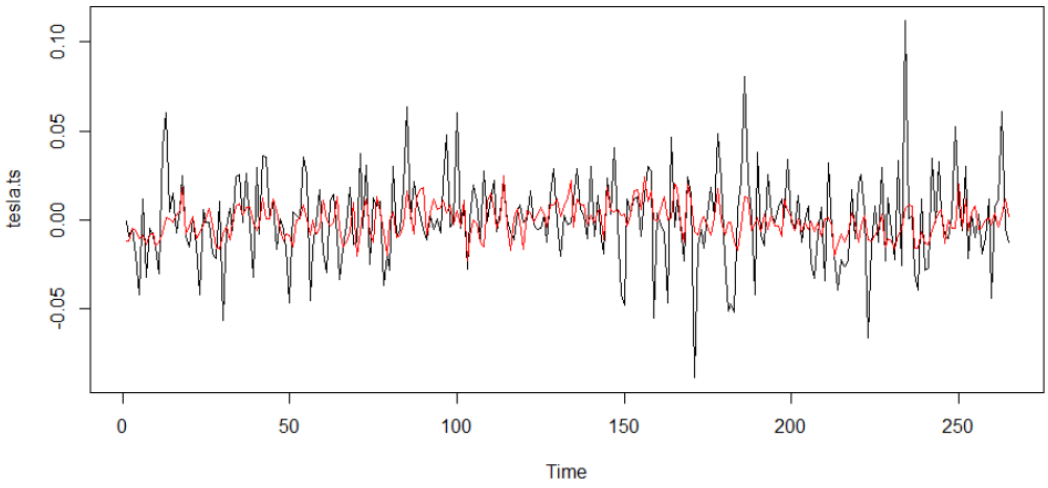

Figure 1: Tesla: Plot of actual vs. fitted.

It is surprising to note that while Tesla stock sentiment is significant only if it is negative, total stock sentiment is significant only if it is positive. This result does not hold good for Ford and GM (in fact, the result is inverted in the case of Ford). Therefore, the general result of this analysis is that there is no uniform pattern by which sentiment is related to security price movements. The variation is specific to particular securities at particular time periods — there is no general model / "one size fits all" model.

5 Conclusion

Let's come back to the original question: Is it possible to predict future security prices based on historical data? Whether it is predictable by professional firms with cutting edge technology is still an open question, but I think we can safely conclude that the stock market is certainly not predictable by four amateurs running R code on their consumer laptops! Although we weren't able to predict the stock market and get rich, we did learn a lot of interesting things on this journey:

Set up the problem right. Abraham Lincoln once remarked (perhaps apocryphally): "If I had six hours to chop down a tree, I'd spend the first four hours sharpening the axe." Sharpening the axe in this case involves understanding and clearly defining the problem statement, reaching consensus on a strategy and approach before diving into the analysis. Start with a holistic view of the problem before navigating through it.

Be an expert in your domain. No amount of data or modeling techniques can make up for a lack of domain knowledge. Invest into building knowledge in the domain of your choice, or work with domain experts who understand and can measure the causal factors involved.

Use a process-driven approach. Modeling is an art, a science, and a process (especially so when it is a team effort). The process needs to incorporate a clear breakdown of the work into tasks and milestones, and a systematic and periodic review/validation process.

Get creative. Don't be a slave to your processes. Be curious and open to thinking outside the box. Sometimes the most seemingly unrelated factors generate the most valuable insights.

Don't be a perfectionist. There is no perfect model. Ask and answer the question "how good is good enough?" Complex models are not always better than simple models. Until what point is it worth it to improve the accuracy at the cost of increasing complexity? Consider the costs of a model that is too complex to explain or to implement in real time scenarios.