Forcasting Frenzy: Machine Learning's Wild Ride on Wall Street

Zach Joachim, Taylor Jones, Trevor Wai, Carson Watkin

April 13, 2024

Abstract

The goal of this paper is to build a model to forcast stock prices and profits of the Coco-Cola Company. Using an Autoregressive Integrated Moving Average (ARIMA) model to create this forcast. Stock prices tend to follow trends during certain periods of time. We attempt to use this data to demonstrate that this data can be used to create reasonable predictions of the behavior of stock prices.. . .

1 Problem Statement and Motivation

Stock forecasting is vital for a company that wants to be successful and profitable. It helps companies with financial planning, allowing them to allocate resources efficiently and make informed decisions about investments and expenditures. It allows companies to identify and mitigate risks associated with a volatile market. Finally, along with many other things, forecasting provides a benchmark for the company's actual performance to be measured against. By comparing the forecasted stock prices with the actual prices, companies can evaluate the effectiveness of their strategies and operations.

Clearly, the ability to forecast stock prices and other important quantities is of great use and interest to companies everywhere. We will attempt to find the best method of forecasting stock prices that doesn't suffer from under- or over-fitting. We will try to predict the trendline the stock prices will follow over one year after training on the previous 10 years of data. This provides better information for companies that wish to analyze potential for growth rather than day traders who are trying to predict prices within a shorter window.

With the amount of money in the stock market, there have been many attempts to forecast stock prices. There are two main methods for stock

analysis: fundamental analysis and technical analysis[citation]. Fundamental analysis is attempting to measure stock prices based on a company's intrinsic value. Technical analysis is attempting to measure stock prices based on stock market activity. Our project will focus on a technical analysis approach. In recent years, many have used neural networks; however, neural networks do not forecast prices; rather, they "uncover opportunities" [add citation here]. Neural networks are currently being used to help find good investment opportunities. Our project, not only does not use a neural network but also attempts to predict the growth of a company's stock value, price, rather than investment opportunities.

Ideally, our project will be able to predict a projected trendline for a company. Thus, they can see how they are projected to perform based on previous performance, so they will know if they have changes that need to be made.

2 Data

The Coca-Cola stock price history data came from a dataset on Kaggle's website. The dataset contains 7 quantities indexed by the day of those quantities. These values are the stock's opening price, highesst price point in the day, lowest price point in the day, the closing price of the stock, the volume, dividends, and stock slpits of the stock for the day. Finding data like this was difficult to find for quantites like profits, margins, or quantities for specific drinks that fall under the Coca-Cola umbrella. However, the stock price for a company is a good indicator as to what the values are and where they are heading.

The stock price data reuqired a little cleaning. For about 50% the data, the index or date for the stock info comes in the form of YYYY-MM-DD, while the other 50% has the time next to the date, in the same format, YYYY-MM-DD HH:MM:SS-UTC. For the purposes of our analysis, we only needed the date, and the format it was in was acceptable, so what we needed to do was delete part of the index that gave the time. To do this, we stripped the index into its list of strings and only kept the part that gave the date, we made this into a new column called `df['date']`, and deleted the original column.

(Come back to this and talk about initial vs data we used...)

3 Methods

3.1 ARIMA Model

We decided to use an ARIMA model becasue we wanted to make future predictions based on past values. These models are great at making future trend predictions from time series data because they are able to take the moving average of the previous data points and use an autoregressive model to predict what the next data point. We chose an ARIMA model over an ARMA model because ARIMA models can take nonstationary time series and make the data stationary by differencing. Stock prices are nonstationary because they fluctuate with time; they reveal trends, seasonality, and grow at increasing rates over time.

3.2 Random Forest

In researching this project, we stumbled on a paper where Chinese machine learning engineers were able to use a random forest to make effective prediction on Chinese stock market prices [citation]. We decided that this would be a fun experiment to see if their results were replicable on the United States stock market. Random forests make sense ...

(Feature Engineering from Zach?)

3.3 Kalman Filtering*

3.4 SVM

[Look up what this is]

3.5 KNN

4 Results

4.1 ARIMA Forecast

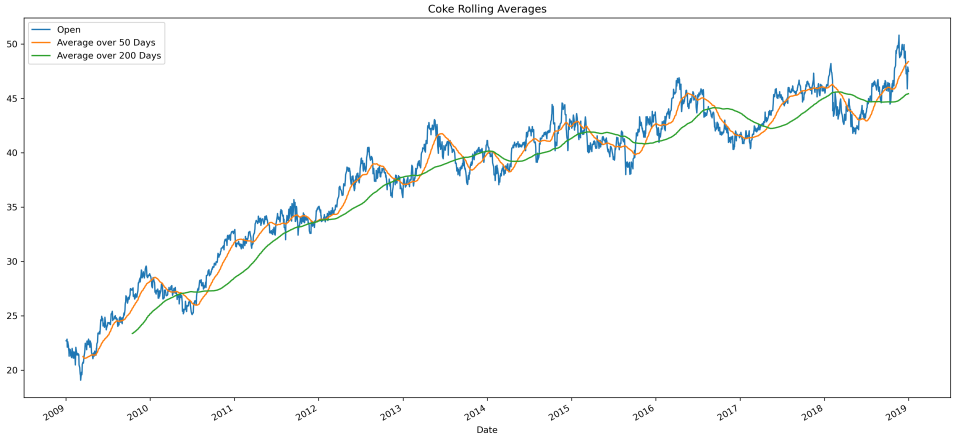
We wanted to see if an ARIMA model could accurately predict the next year (365 days) into the future. Due to the complexity of the Stock Market and the different parameters of a stock, such as its value at Open, Close, Daily High, Daily Low, and the Volume sold, we decided to just look at the value of Coca-Cola stock at Open. Due to the large amount of data from the Coca-Cola stock, we decided to just analyze the stock from 2009 to 2019 and then forecast into 2020.

Like most stocks, the Coca-Cola stock has a lot of noise. To try and compensate for this, we decided to take the rolling average over 50 and 200 days to get a smoother trendline. We then perform time series forecasting using the ARIMA model on the 50 day rolling average. After splitting our data, we then use the `auto_arima` class from the `pmdarima` python package. This allowed us to find the best parameters for the ARIMA model based on the training data. The best parameters are then used to train an ARIMA model on the dataset. After training the model, we then forecast the next year into the future. The resulting plot shows the historical data, the true future data, and the forecasted stock prices.

Here we see the different ARIMA models that were tested and the best model that was then selected to predict forecasting.

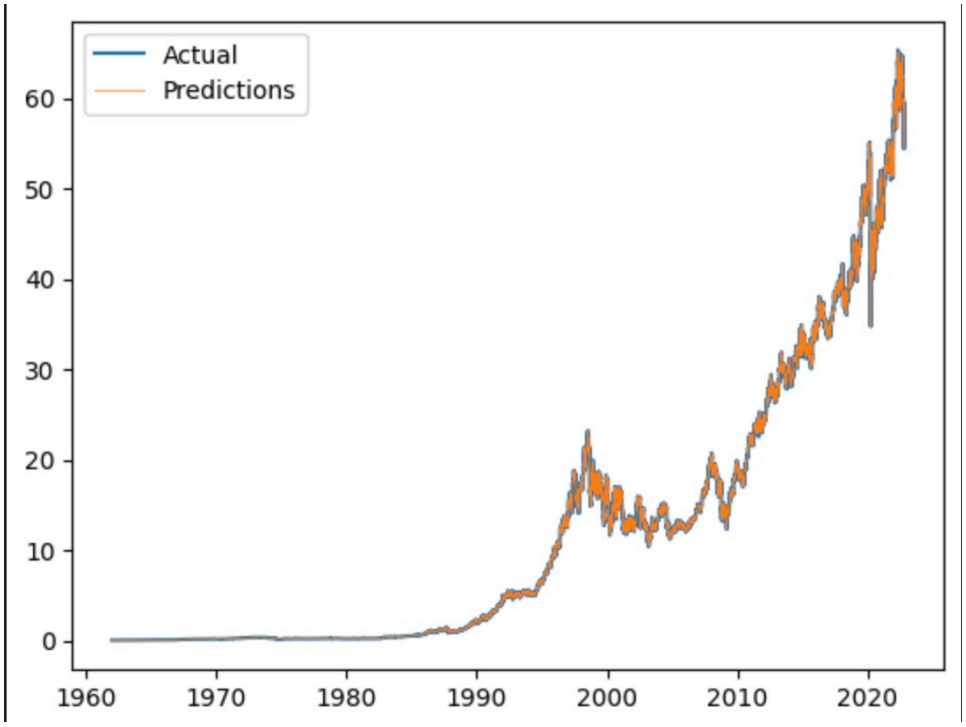
```
Performing stepwise search to minimize aic
 ARIMA(2,2,2)(0,0,0)[0] intercept   : AIC=-14614.840, Time=1.17 sec
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=-14614.126, Time=0.51 sec
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=-14619.359, Time=0.34 sec
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=-14619.441, Time=0.35 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=-14616.124, Time=0.24 sec
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=-14617.515, Time=0.66 sec
 ARIMA(0,2,2)(0,0,0)[0] intercept   : AIC=-14617.494, Time=0.61 sec
 ARIMA(1,2,2)(0,0,0)[0] intercept   : AIC=-14615.679, Time=0.87 sec
 ARIMA(0,2,1)(0,0,0)[0]             : AIC=-14621.441, Time=0.44 sec
 ARIMA(1,2,1)(0,0,0)[0]             : AIC=-14619.515, Time=0.41 sec
 ARIMA(0,2,2)(0,0,0)[0]             : AIC=-14619.495, Time=0.39 sec
 ARIMA(1,2,0)(0,0,0)[0]             : AIC=-14621.358, Time=0.12 sec
 ARIMA(1,2,2)(0,0,0)[0]             : AIC=-14617.680, Time=0.37 sec

Best model:  ARIMA(0,2,1)(0,0,0)[0]
Total fit time: 6.508 seconds
```



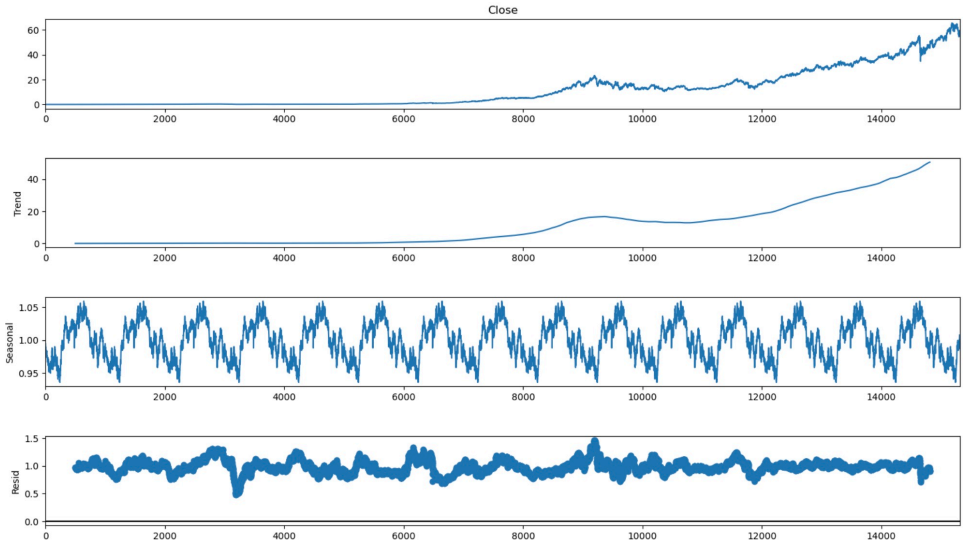
4.2 Random Forests

In Random Forest Regression, trees are built randomly, which helps prevent over-fitting. Random Forest Regressors are also well suited for noisy data, which is quite typical in stock price data. We trained our model using 100 trees on subsets of our data and got an out-of-bag score of 0.9999427215159137, which was surprising given the results of the ARIMA models. After training the model, we then predicted the stock price using the whole dataset and got the results shown below:



4.3 Time Series Decomposition of Closing Prices

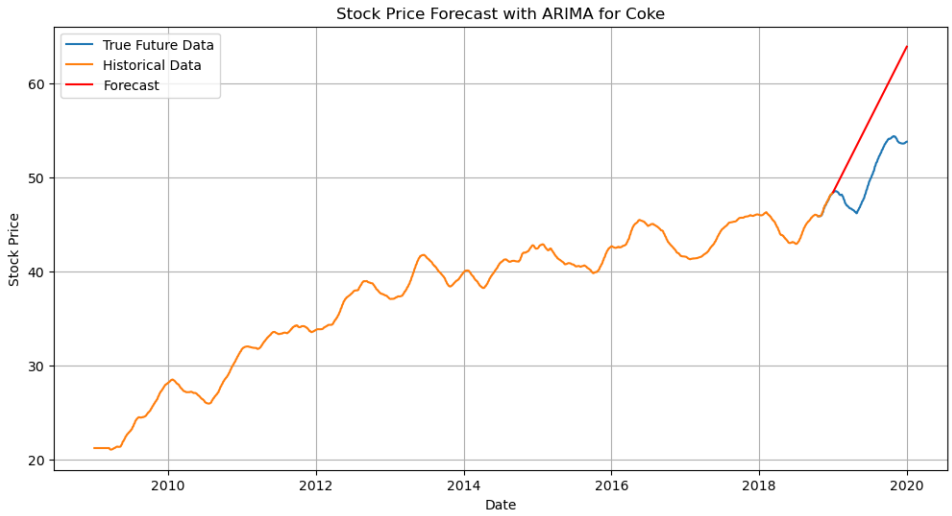
Because prices of stocks fluxuate so much we decided to use a seasonal of prices to break down the data. It revealed the trend line as well as the seasonal periodic graph as well as the residual noise. We found this to be a good method to break down the data because stock prices are slightly periodic through the day and are definetly not perfectly linear and so this allowed us to notice trends in our data and see how it changes over time.



5 Analysis

5.1 ARIMA

Our ARIMA model's performance on the stock price prediction indicates some shortcomings in accuracy. While our ARIMA model did predict that the Coca-Cola stock would increase (which overall it did), it failed to show the trends of how the stock will drop. It also predicted that the Coca-Cola to be $63.86 where it should have been at $53.78. This is a $10.08 difference between the actual cost of Coca-Cola and the forecasted cost. This discrepancy between the forecasted and actual values could be attributed to various factors. It's possible that the model failed to capture the full complexity of the underlying stock price dynamics, including non-linear trends, seasonality, or sudden shifts in market conditions. Additionally, external factors such as unexpected events or changes in economic conditions might not have been accounted for in the model, leading to inaccuracies in the predictions. Further analysis, including model refinement, incorporating additional data features, or exploring alternative forecasting methods, may be necessary to improve the accuracy of future predictions and better capture the dynamic nature of stock price movements.



5.2 Random Forests

Random Forest Regressors tend to perform well with noisy data such as ours, however, the model still had trouble predicting the larger spikes over time such as that through 1995–2005. Overall, this model performed better than the ARIMA model that we suspected would be good for this dataset, we had not thought of the impact that the noise in the data would have on the ARIMA model.

6 Ethical Implications

The primary ethical concern of our project is the application of our analysis. Individuals wanting to manipulate the stock market would be able to use our model as their evidence to artificially manipulate supply and demand. This interferes with the free and fair operation of the stock market and is illegal in the United States. An example of how manipulators can use our analysis to affect the market is through stock bashing. Our analysis could predict the fall of a stock price and then perpetrators can use this as evidence for stock bashing or convincing the public that a stock price is worth less. Therefore allowing them to purchase the stock at a cheaper price to try and later turn a profit.

Therefore we should add a disclaimer in our project that all of the analysis and predictions created are simply models and not prophecies of events to come. Therefore people understand that our analysis is simply trying to

predict how the stock prices are going to fluctuate and this can simply be a factor in their buying and selling of shares.

7 Conclusion

Overall, the traditional time-series model (ARIMA) that we expected to perform well with our data did not predict the future price of stocks accurately. The predictions seemed to be very sensitive to the behavior just before predicting. This could be due to the volatility of stocks and the lack of a clear period to base our models on.

On the other hand, the random forest regressor model tended to perform really well, with error in the estimates being really low. We don't know why this is.