Can qualitative and quantitative data increase accuracy of AMZN stock price prediction in a RRN?

By Anthony Gonzalez

Github: https://github.com/AMGonz96/RNNForStockPrediction

Highlights

- Adding increased qualitative or quantitative information to the RNN has close to the same effect in the neural networks performance.
- Using both qualitative and quantitative data in the RNN has the greatest increase in price accuracy
- the model using both qualitative and quantitative data was able to achieve profitability over a period of 250 days

Background

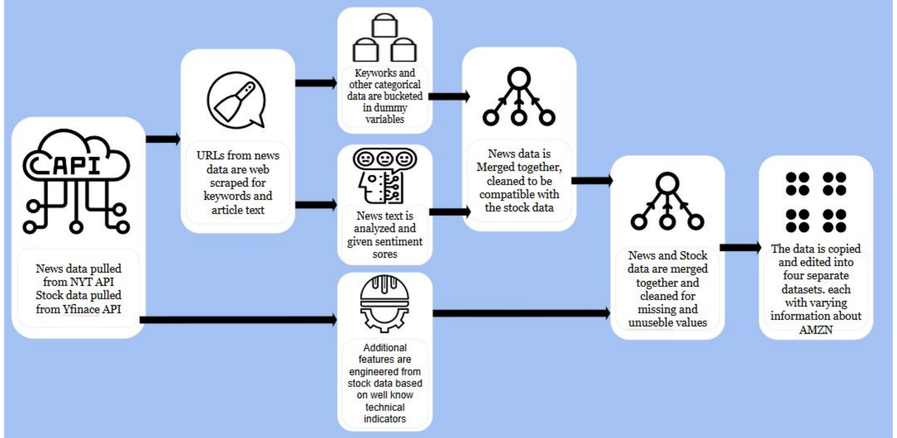
The stock market has shown to be one of the most sensitive areas of economics where numorouse outside factors can cause large swings in prices overnight. While some indicators for these factors may be able to be found within the numbers of market data other qualitative indicators need to be found else where.By even capturing a fraction of these indicators could help investors increase profits or minimize their losses. This project aimed to discover if both qualitative and quantitative data could be used to increase accuracy predictions of the next day opening price of Amazon Inc. stock.

Data

Data was pulled via an API from two different databases, pulling only relevant information about Amazon Inc. from 2010/6/1 - 2020/3/31. This data was then used to extract a total of 352 features. 227 from the qualitative news data and 125 from the quantitative stock data

Once all the features were extracted and the datasets merged the data was copied and turned into four additional data sets

1. Having only basic stock market data, 7 columns total
2. Having basic stock data & the additional engineered stock data, 127 columns total
3. Having basic stock data & the extracted News data, 233 columns total
4. Having all the news and stock data generated, 352 columns total



Model

A recurrent neural network or RNN was used due to the size and format of the data being used. Since upwards of 350 features would be used an artificial neural network was needed and due to the data being a time series a recurrent network would provide better results the a convolutional network. Long short term memory cells were used in the first to layers due to their ability to store relevant information for longer, then dense layers were used for the back half of the neural network to get the result to a linear output. The model was trained with data from 2010/6/1 - 2019/04/05 and tested with 2019/04/08- 2020/3/31.

Results

Dataset 1, Basic stock data

Dataset 2, Basic stock data & engineered stock data

Dataset 3, Basic stock data & extracted news data

Dataset 4, All news and stock data data

With the use of qualitative and quantitative data the RNN is able to more accurately predict the next day opening price of AMZN.

The graphs to the left shows how each dataset's predictions from the RNN compares to the real next day opening value for each day in the test set.

- Dataset 1 performed the worst with an average difference of $70 between the predicted and real price, equating to about a 3.80% difference.
- Datasets 2 and 3 come in a close 3rd and 2nd with a 1.3% and 1.01% average difference respectively, equating to about $24 and $18.
- Dataset 4 had the best performance with the model with only an average .51% difference in the real vs. predicted price. Thus showing how using qualitative and quantitative data a RNN can improve accuracy

Market Performance

To test how effective the model would be at making money a script was developed to simulated a bot trading AMZN stock based off of the results of dataset 4.

The plot to the right shows the value of the bot's profolio against the real and predicted prices of AMZN. If price was predicted to be higher at opening it would buy if lower it would sell. These buy and sell points are dotted in the figure

When simulated a bot was able to achieve over a 20% increase in revenue using the RNN with both qualitative and quantitative data