An Exploratory Research into Stock Price Prediction

Opeyemi Openiyi, Francisco Baca

Abstract

This is exploratory research performed with the goal of making detailed comparisons between popular research and algorithms used to predict future values from existing time-series data. The specific goal of this paper is to see how well these machine learning algorithms and models work for making predictions from past stock market data. The machine learning models were trained on existing data and checked for accuracy with different model evaluation tests in order to determine how effective each model is at making the expected prediction. This research elaborates on the strengths and weaknesses of each model as they are applied to stock price prediction.

Introduction

Time series data is a collection of measurements made over time. These measurements are gathered so that meaningful inferences can deducted from them.

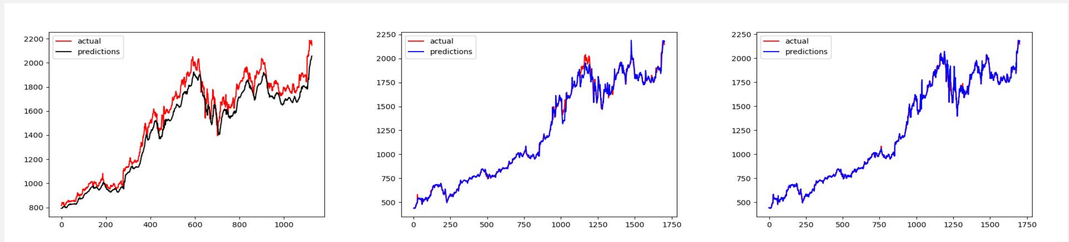
For many modern companies, the search for high-fidelity time-series data prediction methods has become analogous to that of the holy grail. Businesses often find themselves in the situation where they have collected market data (time series data) and they want to make certain inferences from it to minimize financial risk or maximize profit. Furthermore, the decisions of businesses and investors alike are heavily influenced by market forces, and they often look to stock market analysis before making such high-risk decisions. Hence the motivation for this research; an exploration of Machine Learning applied to stock market analysis.

As a part of this research, a Long Short-Term Memory (LSTM) model and Convolutional Neural Network (CNN) model were developed and tested through various model evaluation measures later introduced in the methodology.

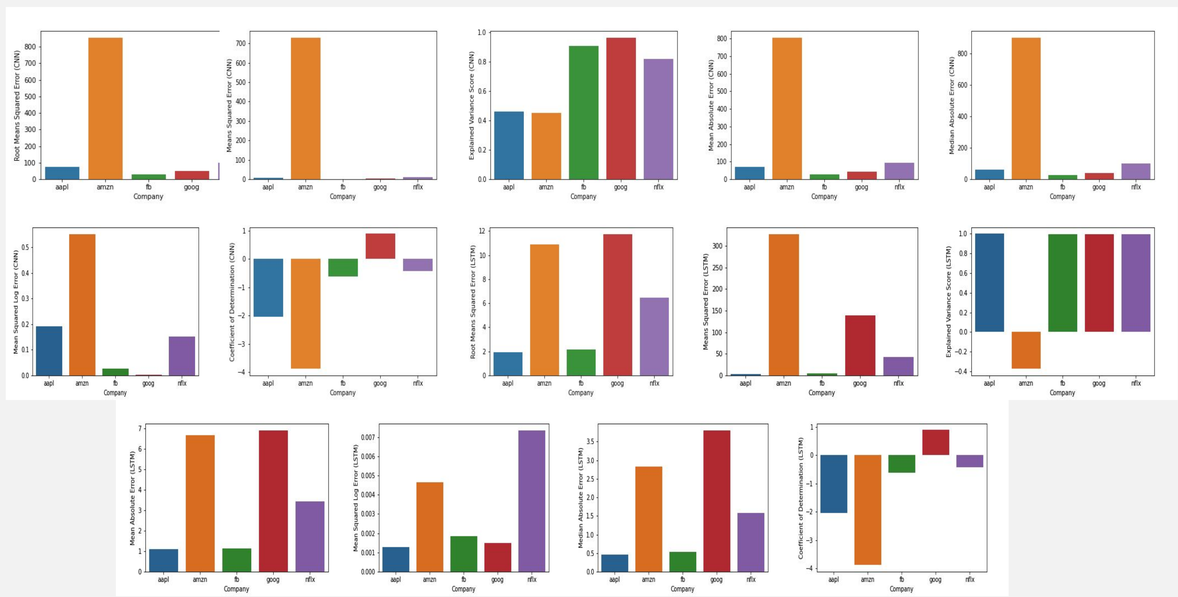
Methodology

Stock data from five major tech companies was collected *Yahoo Finance*. The companies used in this research are commonly known as *FAANG* (*Facebook*, *Amazon*, *Apple*, *Netflix*, and *Google*). For Facebook, the data collected ranged from 05/18/2012 to 02/21/2020, while for the rest of the companies, the range was from 08/19/2004 to 02/21/2020, this was based on the free availability of data. Due to the nature of both the stock market, certain dates (weekends and holidays) would show missing data for the data we collected. Both the CNN and LSTM models expect the time-series data to be continuous to work properly, the missing data points were therefore altered to contain the same values as the immediately preceding date to reflect no change in stock value. An alternative to this approach would have been to train the model on the data as is; in the results section you can see a direct comparison between this alternative as well as the former. The evaluation measures revealed that the first approach yielded the best results. For CNN, the data was partitioned at 60% for training, 20% for validation and 20% for testing, while for LSTM, the data was partitioned at 70% for training and 30% for testing. Both of the models were trained only on the 'high' attribute of the data. Various model evaluations were applied to the CNN and LSTM models; visualizations of the values for these measures can be seen in the results. The evaluations used were Root Means Squared Error, Mean Squared Error, Explained Variance Score, Mean Absolute Error, Mean Squared Log Error, Median Absolute Error and Coefficient of Determination.

Results



The graphs above show trendlines for actual vs predicted values for AMZN, from left to right: CNN predicted vs actual, LSTM predicted vs actual with data as is and finally LSTM with missing data replacement as discussed in the methodology.



The graphs above show the evaluation measurements for CNN and LSTM, more specifically, the measurement when the algorithms are applied to the stock records for each company. Some values had to be normalized in order for the barplot labels to be visible, specifically MSE for CNN and LSTM MSLE.

Conclusion

From the visualizations in the results section, it can be seen that the prediction trendlines for the LSTM graph more closely resemble (and almost completely overlap) the actual trendlines, so it is clear that LSTM did a better job predicting actual values than CNN. The visualizations also revealed that for most of the evaluation measures except for MSE, MAE and Coefficient of determination, the scores for LSTM were better than those for CNN. The scores for the aforementioned evaluations were at worst; the same for LSTM as for CNN. From the results we can conclude that LSTM performs far better than CNN on time-series (specifically stock) data.

Further Research

With more time and resources, further research could be conducted on hourly stock data rather than daily and tests could be conducted to find better predictors and an optimal lag time.

References

Brownlee, Jason. "Time Series Forecasting with the Long Short-Term Memory Network in Python." Machine Learning Mastery, 5 Aug. 2019, machinelearningmastery.com/time-series-forecasting-long-short-term-memory-network-python/.

Esling, Philippe, et al. "Time-Series Data Mining." ACM Computing Surveys (CSUR), 1 Dec. 2012, dl.acm.org/doi/10.1145/2379776.2379788.

Microsoft Azure. "Azure/DeepLearningForTimeSeriesForecasting." GitHub, Azure, 7 Aug. 2019, github.com/Azure/DeepLearningForTimeSeriesForecasting.

Link to source code