MSc Project - Reflective Essay

Project Title:	Bitcoin Stock Prediction using SARIMA
Student Name:	Harnoor Singh Oberai
Student Number:	190753898
Supervisor Name:	Dr Luk Arnaut
Programme of Study:	School of Electronic Engineering and Computer Science

The technology likely to have the most significant impact in the next few decades has arrived; and it is not social media, robotics, and not even AI. It is the underlying condition of digital currency like bitcoin called the blockchain. My inspiration to work on stock prediction began when I was working at TiVo Incorporation in India. In 2017, My colleagues made huge profits on cryptocurrencies, and I had no idea about this new financial currency. I was astonished by the fact; 100$ invested in my high school will be worth $7588953.009 in 2017 (DQYDJ, 2020).

The co-founder of PayPal, Peter Theil testimonies that Bitcoin is the first [encrypted money] having a potential to do something that will change the world. I became curious and wanted to know about bitcoin and see if I can make a prediction based on the data. My MSc gave me a first-hand experience researching bitcoin and Time Series Data. I did make mistakes during the process. However, this allowed me to learn some valuable lessons as a result of this.

In my research, I took the historical bitcoin market data at a 1-min time interval. The data multiple attributes. Some of them are Open, High, Low, Close, Weighted Average, Date. The records were updated on the conditions: If Open, High, Low & Close value were Null; the record is removed else the attributed was updated with preceding value. As a result, the data modified 28% of instances and pre-processed it. The cleaned data set contains 3126480 transactions.

Extensive comprehensive data analysis is conducted in regards to trading in Bitcoin. Some of the highlights were: bitcoin started trading at $13.5. The highest value ever recorded by bitcoin was $19,783. The high-value graph shows a meteoric rise and fall of Bitcoin at different time intervals.

The paper focuses on building a SARIMA model as Time forecasting model. The reason to choose this model over ARIMA is the fact it takes account seasonal component in the series. It combines three models. The Auto Regression(AR) calculates the attribute based on the linear combination of past its values. The Moving Average (MA) component that forecasts based on error accounted. The Integrated(I) component to make the series stationary. The disadvantage of the SARIMA model is that it can only extract linear relationships within the time series data. Artificial neural networks are useful for mapping when a non-linear relationship exists within the time series data. The disadvantage of ANN is a specific non-linear function within time series may be difficult to explain in practice. The ANN also had limitation learning patterns as recorded by Kim and Hann (2000). The stock prices are equipped with multiple dimensions having enormous noise. The quantity of stock data will eventually conflict with learning patterns. As a result, I went to build a SARIMA model for forecasting using Weighted Average attribute.

The most challenging part was making the data stationary and testing against Augment Dicky Fuller test. The weighted average is grouped by month. It is then made stationary and normally distributed using Box-Cox transformation. The ADF test rejected the null hypothesis; suggesting that the data is stationary. After, the SARIMA model is built using various combination of different parameter. The total number of combination used is 81 to yield the lowest AIC value of 175.917.

After building the model, it is validated via Out-Of-Time cross-validation. We use such a matrix because the SARIMA model is heavily dependent on previous value and shuffling the data will not yield an accurate result. I observed choosing the ratio of training and testing data is a massive factor for forecasting accuracy. Multiple ratios of train and test data are considered to yield an average accuracy of 64%. The model is successful in predicting the last 15 months, with 73% accuracy.

The only condition that is satisfied is the quantity of data. There is a limited understanding of the factors that can affect the cryptocurrency exchange rate. Factors like an epidemic, change in government policy, demonetization, COVID-19 may directly affect the exchange rate, which has not been taken to account, as the data is not available. Political events can have a significant influence on stock prices (Kuo, Chen and Hwang, 2001). Sometimes well-published articles can also alter the hypothesis. If a not so well-published article forecast that the prices are going to drop, people will rush to sell so the forecast will become self-fulling.

Predicting monthly prices on Weighted Price can give an understanding of change in price, but the accuracy is not what I expected. The range is quite broad (31-73%). Train: Test ratio of 92:9 and 93:8 had the lowest accuracy. The model performs better when predicting the immediate next five months weighted price (Accuracy 76%).

For future work, I would want to implement ML-Ops pipeline (Google Cloud. 2020) of Continues Integration and Development. Implementing not only SARIMA model, but deploying the ML pipeline to automate the retraining an deployment of updated new SARIMA model. Since the time series is heavily dependent on previous values, setting up a CI/CD system will enable me to test the new historical data and update the forecasted prices. This paper focuses on forecasting monthly prices and not daily change. My next task will be every day forecast prices on bitcoin and do extensive research of different factors affecting bitcoin indirectly. I want to explore this area to find if the accuracy can be increased and the range reduced.

Overall, the process was challenging, that introduced me to the iterative nature of the model building. It became obvious that stock prediction is not an easy task. I struggled with increasing accuracy. The market is extremely volatile, making it difficult to forecast.

Looking back upon this past semester, I have opened my eyes to the financial ecosystem in a manner I had never expected. I gain new insights and perspective about the cryptocurrency exchange and have started investing in bitcoin. I developed the skills of python, pandas, time series and started looking at the data at a more scientific level. This was my first thesis ever written and the feedback provided by Dr Luk Arnaut was helpful. The MSc project was an informative learning experience, and I thoroughly enjoyed it.

References

DQYDJ – Don't Quit Your Day Job... 2020. Bitcoin Return Calculator - Investment On Any Date And Inflation - DQYDJ. [online] Available at: <https://dqydj.com/bitcoin-return-calculator-inflation-adjusted/> [Accessed 11 August 2020].

Google Cloud. 2020. Mlops: Continuous Delivery And Automation Pipelines In Machine Learning. [online] Available at: <https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning> [Accessed 13 August 2020].

Kuo, R., Chen, C. and Hwang, Y., 2001. An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. Fuzzy Sets and Systems, 118(1), pp.21-45.

Kim, K. and Han, I., 2000. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert Systems with Applications, 19(2), pp.125-132.