Machine Learning for Bitcoin Price Prediction

Team: Karan Jaisingh (419, Shubham's Cohort), Brendan McAffery (519, Yang's Cohort), Raphael Iskra (519, Yang's Cohort).
Project Mentor TA: Yang.

Abstract

In the modern financial world, cryptocurrency has come to the forefront of everyone's focus - however, given the recency of the field's popularity, limited research has been conducted into quantifying approaches to estimating future price movements. This project concerns itself with predicting the next-day price movement of Bitcoin (BTC) using historical financial data, as Bitcoin is arguably the most well-known cryptocurrency today. It focuses on two approaches - using naive models that do not consider a degree of sequence, as well as advanced sequential models, in order to predict both directional and numerical price movement. Our project concludes that both novel methods and advanced methods can perform slightly better than analyst market predictions and random forecasts, though struggles to obtain a high accuracy suitable for steady implementation in a market setting.
Github Link to Code: https://github.com/kjaisingh/519-Project

Introduction

In 2020, Bitcoin led a massive revolution in the growth of cryptocurrency, with its market cap surpassing $1 Trillion for the first time in history. This revolution was met with institutional adoption: JP Morgan and Goldman Sachs reversed its previous stance which declared Bitcoin/crypto "is not an asset class," and Tesla became the first company in the S&P 500 to convert a portion of its balance sheet to Bitcoin. Most importantly, the events of 2020 have forced individuals across the globe to deeply investigate the question - "What is money?".

The members of this project group understand that a decentralized network of proof-of-work-based value recognition is far superior to the synthetic solutions (targeted inflation, TBill Rates, etc.) offered by centralized financial institutions. Brendan has experience trading crypto currencies with algorithms that employ technical analysis strategies. Raph, who boasts an Bitcoin portfolio with a Cost Basis of $14/BTC, is also an active crypto-trader. Karan has an interest in using quantitative/ML methods to predict market behavior, and is intrigued to investigate ML capabilities on blockchain networks. Thus, the members of this group are interested in using Machine Learning methods to predict short-term price fluctuations of Bitcoin and other crypto currencies.

Related Prior Work

Time-Series Forecasting of Bitcoin Prices using High-Dimensional Features

Mudassir, M., Bennbaia, S., Unal, D. et al. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. Neural Comput & Applic (2020). https://doi.org/10.1007/s00521-020-05129-6

In this paper, the authors explore the use of machine learning-based models for one, seven, thirty and ninety days time frames for Bitcoin price classification and price prediction. The classification model achieved up to 65% accuracy for next-day forecast and scored from 62 to 64% accuracy for seventh to ninetieth-day forecast. For daily price forecast, the error percentage is as low as 1.44%, while it varies from 2.88 to 4.10% for horizons of seven to ninety days. For 7th-day forecasts, SVM performs the best with accuracy of 62% and F1 of 0.60. ANN performs the poorest with accuracy of 51%. LSTM performs better than ANN with 55% accuracy but similar F1. For 90th-day forecasts, the LSTM model reports the highest accuracy of 60% with F1 of 0.66. The ANN model improves to 61% accuracy, and the SANN comes in third with 60% accuracy.

A Deep-Learning Based Cryptocurrency Price Prediction Scheme

Mohil Maheshkumar Patel, Sudeep Tanwar, Rajesh Gupta, Neeraj Kumar, A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions, Journal of Information Security and Applications, Volume 55, 2020, 102583, ISSN 2214-2126, https://doi.org/10.1016/j.jisa.2020.102583.

In this paper, a LSTM and GRU-based hybrid cryptocurrency prediction scheme is proposed, which focuses on only two cryptocurrencies, namely Litecoin and Monero. The results depict that the proposed scheme accurately predicts the prices with a reasonably strong accuracy, revealing that the scheme can potentially be applicable in various cryptocurrencies price predictions.

Anticipating Cryptocurrency Prices Using Machine Learning.

Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, Andrea Baronchelli, "Anticipating Cryptocurrency Prices Using Machine Learning", Complexity, vol. 2018, Article ID 8983590, 16 pages, 2018. https://doi.org/10.1155/2018/8983590.

This study analyzed daily data for cryptocurrencies for the period between Nov. 2015 and Apr. 2018. We show that simple trading strategies assisted by state-of-the-art machine learning algorithms outperform standard benchmarks. Both LSTM and GRU-based approaches have shown to be effective. Additionally, regression trees built by the XGBoost Algorithm used in the latter study showed to be effective in short term price prediction. The XGBoost Algorithm built different regression trees for each cryptocurrency, suggesting that it is effective across various currencies.

Formal Problem Setup

T: For each day in our time frame, BTC's state is on day t is denoted as $c_{t+1} = [s_{t-89}, s_{t-89}, \dots, s_{t-1}, s_t]$, where each $s_i = [o_i, h_i, l_i, c_i, v_i, m_i]$, corresponding to Open, High, Low, Close, Volume, and Market Cap respectively on day s_i. Each s_i also includes several other extracted features which are calculated values from aggregation and numerical techniques commonly associated with equities. Our task is to classify BTC as a buy/sell (whether the price c_{t+1} will increase/decrease) each day based on the state of the coin each morning through a classification approach, as well as predict the next day's closing price of BTC by using a regression approach.

E: We explored a historical BTC dataset which contained raw information dating back to 2013 up to the early parts of 2021. This dataset had all of the defined attributes of the coinstate above, and was used to iteratively calculate others which we incorporated into our machine learning experiment in the next phase. We propose to use data from 90 previous days of BTC trading, and test our model's performance on the next-day BTC pricing. We partition our dataset with a 75-25 split, wherein 75% of days (and their performance over the previous 90 days) are utilised for training, whereas the other 25% are used for cross-validation and testing.

P: In our classification models, we define accuracy as follows: on a given day t, if our model classifies the coin as a buy/sell and the BTC price does in fact increase/decrease respectively, this will be a success, though if our model classifies the coin state as a buy/sell and the BTC price does in fact decrease/increase respectively, this will be a failure. The accuracy is hence the percentage of days which the model correctly classified the close price, c_{i+1}, higher or lower than the previous close (corresponding to buy/sell). We used the *F1 Score* metric to further evaluate the distribution of false positives and false negatives produced by our model's classification. In our regression models, we used Mean Squared Error (MSE) to measure the error of our models prediction y_{i+1} from the actual closing price c_{i+1}. We also utilise the R2 coefficient as $R^2 = 1 - \frac{\Sigma(y_{i+1} - c_{i+1})^2}{\Sigma(y_{i+1} - \overline{c_{i+1}})^2}$ to evaluate the fit of our model.

Methods

Dataset Curation and Processing: Our approach to this problem began by curating a processed dataset based on a Bitcoin dataset which contained daily financial data about the cryptocurrency stretching back to 2013 - this data involved basic features such as opening, closing, high and low prices, as well as market cap. Using these basic features, further features were extracted using numerous aggregation and numerical techniques - these involved establishing moving and exponential averages over varying time periods prior to the current day data, as well as creating additional metrics such as relative high, relative low and high-low ratio scores. This allowed us to generate a dataset with 26 features, 7 of which were initially present and 19 of which were extracted.

After this, a complete three-dimensional dataset was established, wherein rows corresponded to each instance, columns corresponded to the state for each of the previous 90 days and each element corresponded to a vector of metadata for that day. Two datasets were then extracted from this for the naive models - once which used the vector of the previous day's state to predict next-day closing price (metadata dataset), and another which used closing price from the previous 90 days to predict next-day closing price (historical dataset).

Baseline Approaches: Our baseline algorithms were split into four distinct categories and evaluated in different manners. Firstly, a series of models were trained on the metadata of BTC from the previous day, and used this to predict an upward or downward movement in BTC price on the next day. The second set of models approached a similar problem, but instead aimed to predict the exact next-day price of the cryptocurrency, making it a regression model. The third suite of models utilised the BTC closing prices from the previous 90 days to predict the next-day upward or downward movement, while the final set utilised identical data though to predict the next-day price instead. All models utilised processed and standardised data in their training.

Each set of models consisted of three distinct Machine Learning methods - Support Vector Machines (implemented with scikit-learn), Random Forest (implemented with scikit-learn) and Fully-Connected Neural Networks (implemented with Keras) - with the type of each method varying based on whether the problem at hand required a categorical or numerical output. Each baseline model was trained with the default parameters supplied by each library as these are known to be easily-generalisable. Following the training and evaluation of the baseline models, the top-performing model in each category was optimised with GridSearch and cross-validation to create even higher-performing baseline models.

The optimal model that utilised historical closing prices to predict next-day price was then extended into newer contexts, and their generalisability was tested on similar yet different problems. Firstly, the architectures of optimal models were tested on shorter timelines than 90 days - this timeline was reduced to spans of 3, 10 and 30 days. Next, a downscaled version of the model which utilised Principal Component Analysis (PCA) was tested (with an optimal number of components found through plotting a scree diagram), as the results from the shorter timelines provided a hint that much of the data variation may be attributable to a certain few columns. Finally, both the model and the model architecture was tested on alternate cryptocurrencies to test its generalisability.

Implementation Details: We then experimented with more advanced models, namely Long-Short Term Memory (LSTM) models and 1-D Convolutional Neural Network (CNN) models. These models utilised all metadata (open, close, high, low, high-low ratio, etc) of BTC from the previous 90 days to predict upward or downward movement of the coin on the next-day. While the LSTM model utilised sequential data recognition to make next-day predictions, the 1-D CNN utilised techniques such as Pooling and Convolutions.

Once a baseline LSTM model was created, an optimised LSTM model was developed using a GridSearch-based hyperparameter selection process, and was found to benefit from recurrent dropout, inter-layer dropout and using the hyperbolic tangent as its activation functions. The 1-D CNN model that was found to work optimally was based on the WaveNet structure, a generative model found to outperform LSTMs for use on raw audio - as such raw audio appears in a similar sequential form, it was applied to this problem. Both such models were found to benefit from an Adam-based optimization technique, which utilises an exponential decaying average of previous gradients in a manner similar to momentum.

Experimental Results

We aim to answer the following questions: (1) Can our system predict upwards/downwards next-day price movements accurately? (2) How does it extend to predicting next-day numerical prices? (3) How do our models generalise to alternate settings - minimised time windows, downscaled feature set and alternate cryptocurrencies? (4) Does harnessing the sequential nature of the price data alter or improve results? (5) Is our model ample for real-world implementation, and if so, in what cases?

We first compare the results from our baseline models, which were trained in four different problem settings. This informed our decisions when creating optimised models.

	Classification: Metadata-Based (Accuracy / F1)	**Classification: Historical-Based (Accuracy / F1)**	**Regression: Metadata-Based (MAE / R2)**	**Regression: Historical-Based (MAE / R2)**
SVM	0.513 / 0.678	0.518 / 0.667	287.2 / 0.988	521.7 / 0.965
Random Forest	0.543 / 0.583	0.509 / 0.555	211.9 / 0.993	179.8 / 0.990
Neural Network	0.543 / 0.610	0.490 / 0.616	229.4 / 0.993	234.5 / 0.994

Following this, we conducted optimisations on such baseline models. Most baseline models demonstrated some degree of competency, though the classification models that utilised historical closing price data failed to perform well in any case, so we did not proceed with optimising it. We also abstained from displaying the optimisationg of the regression models trained on the metadata dataset, as it did not provide any significant improvement in performance. Hence, after conducting GridSearch and Cross-Validation based optimisations, the following results were obtained for the Metadata-Based Classifier and Historical-Based Regressor respectively.

	Accuracy	**F1**
Optimised Neural Network	0.567	0.725

	MAE	**R2**
Optimised Random Forest	160.9	0.996

The optimised historical-based regressor model was then extended to alternate contexts as described above - the results for each of these were as follows.

	MAE	**R2**
3-Day Historical	177.9	0.995
10-Day Historical	181.9	0.995
30-Day Historical	176.4	0.996
Downscaled Model (PCA)	165.74	0.996
Ethereum Model	11.4	0.993

Next, we can explore the results obtained from the optimised versions of our advanced methods, namely LSTMs and 1-D CNNs. The results for each one of these models on both classification and regression tasks can be seen below.

	Classification (Accuracy / F1)	**Regression (MAE / R2)**
LSTM	0.547 / 0.707	5120.2 / 0.564
1-D CNN	0.555 / 0.714	5434.3 / 0.533

Similar to the baseline models, further experiments were done with the optimised classification model to analyse how well it generalised to other similar problems, with the results seen below.

	Accuracy	**F1**
Downscaled Model (PCA)	0.509	0.647
Ethereum Model	0.544	0.705

Conclusions and Future Work

The results from the various models trained and optimised revealed that a Fully-Connected Neural Network (FCNN) model, which does not take the sequential nature of input data into account, slightly outperforms more advanced or sequential-based models such as an LSTM and 1-D CNN when predicting upwards or downwards next-day price movement. On the other hand, our optimal regression model involved a Random Forest system that obtained a MAE of $160.90, which equates to a deviation of about 0.93% relative to the weighted average historical price of Bitcoin. This was a significant improvement over advanced models such as LSTM and 1-D CNN, which had a maximum R2 correlation coefficient of 0.564 in comparison to to 0.996 with the Random Forest model - this can likely be attributed to the excessive complexity associated with the more complex models, generating a degree of over-fitting. While the regression-focused Random Forest model was able to generalise well to other contexts - obtaining similar MAE and R2 scores for much shorter time periods, performing similarly with a condensed dataset and operating well on an alternate cryptocurrency - we see that there was a reasonably significant drop in performance for the more advanced 1-D CNN method in similar yet new contexts.

Through these results, we are driven to the conclusion that while our model is able to predict cryptocurrency movement and pricing with a reasonable degree of accuracy, it is not ready for consistent real-world implementation. Our classification slightly under-performs the models in industry in terms of accuracy, though does obtain a better F1 score. By analysing the precision and recall of our FCNN model, we see that this is driven by a better recognition of downwards price movements relative to models in industry - it could be argued that this is more ideal, as in the financial world, shorting an asset (predicting that it will reduce in price) can lead to infinite losses, whereas this model does not incorrectly predict downwards price movements as often. Given that none of the models utilised had a particularly great performance for our problem though, we are led to the conclusion that for accurate financial predictions in the cryptocurrency domain, we may require the creation of more targeted algorithms for the financial world, or potentially even the novel aggregation of various existing models.

Ethical Considerations and Broader Impacts:

The rapid advancement and implementation of algorithmic trading, especially that which involves automated machine learning techniques, requires consideration and classification of ethical and legal boundaries. With recent estimates attributing about half of all trading volume in US equities to be algorithmic, the most common ethical consideration is to what extent autonomy of algorithmic financial systems will develop, and how an entity will recover assets in the case of a system or software failure. In a similar vein, as these autonomous financial tools continue to dictate an increasing amount of equity trading, it is uncertain how algorithmic systems will respond in unanticipated circumstances and what the potential consequences could be. Lastly, algorithmic trading requires access to big data, complex software and hardware infrastructure, and human capital which can lead to drastically different levels of access to individuals and investing entities of varying financial and contextual backgrounds.

References

1. Mohil Maheshkumar Patel, Sudeep Tanwar, Rajesh Gupta, Neeraj Kumar, A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions, Journal of Information Security and Applications, Volume 55, 2020, 102583, ISSN 2214-2126, https://doi.org/10.1016/j.jisa.2020.102583.
2. Laura Alessandretti, Abeer ElBahrawy, Luca Maria Aiello, Andrea Baronchelli, "Anticipating Cryptocurrency Prices Using Machine Learning", Complexity, vol. 2018, Article ID 8983590, 16 pages, 2018. https://doi.org/10.1155/2018/8983590.
3. https://www.kaggle.com/kaushiksuresh147/bitcoin-prices-eda-and-prediction-r2-0-99. A simple implementation of an ML algorithm on a BTC dataset
4. https://www.kaggle.com/jeongbinpark/lstm-can-we-predict-the-bitcoin-price. A good source of initial intuition for recurrent neural networks to find long-term dependencies in our dataset.
5. *Mudassir, M., Bennbaia, S., Unal, D. et al. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. Neural Comput & Applic (2020). https://doi.org/10.1007/s00521-020-05129-6*
6. Aggarwal, A., Gupta, I., Garg, N., & Goel, A. (2019, August). Deep learning approach to determine the impact of socio economic factors on bitcoin price prediction. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-5). IEEE.
7. Felizardo, L., Oliveira, R., Del-Moral-Hernandez, E., & Cozman, F. (2019, October). Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods. In 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC) (pp. 1-6). IEEE.
8. Samaddar, M., Roy, R., De, S., & Karmakar, R. (2021, February). A Comparative Study of Different Machine Learning Algorithms on Bitcoin Value Prediction. In 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) (pp. 1-7). IEEE.
9. Alonso-Monsalve, S., Suárez-Cetrulo, A. L., Cervantes, A., & Quintana, D. (2020). Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. Expert Systems with Applications, 149, 113250.
10. AYGUN, B., & GUNAY, E. K. Comparison of Statistical and Machine Learning Algorithms for Forecasting Daily Bitcoin Returns. Avrupa Bilim ve Teknoloji Dergisi, (21), 444-454.
11. Cavalli, S., & Amoretti, M. (2021). CNN-based multivariate data analysis for bitcoin trend prediction. Applied Soft Computing, 101, 107065.
12. Livieris, I. E., Kiriakidou, N., Stavroyiannis, S., & Pintelas, P. (2021). An Advanced CNN-LSTM Model for Cryptocurrency Forecasting. Electronics 2021, 10, 287.
13. Kumar, D., & Rath, S. K. (2020). Predicting the trends of price for ethereum using deep learning techniques. In Artificial Intelligence and Evolutionary Computations in Engineering Systems (pp. 103-114). Springer, Singapore.
14. Zhengyang, W., Xingzhou, L., Jinjin, R., & Jiaqing, K. (2019, June). Prediction of cryptocurrency price dynamics with multiple machine learning techniques. In Proceedings of the 2019 4th International Conference on Machine Learning Technologies (pp. 15-19).
15. Nazemi, A., Jakubik, J., Geyer-Schulz, A., & Fabozzi, F. J. (2020). Incorporating Financial News for Forecasting Bitcoin Prices Based on Long Short-Term Memory Networks. Available at SSRN 3733398.
16. Yasir, M., Attique, M., Latif, K., Chaudhary, G. M., Afzal, S., Ahmed, K., & Shahzad, F. (2020). Deep-learning-assisted business intelligence model for cryptocurrency forecasting using social media sentiment. Journal of Enterprise Information Management.
17. Cocco, L., Tonelli, R., & Marchesi, M. (2021). Predictions of bitcoin prices through machine learning based frameworks. PeerJ Computer Science, 7, e413.
18. Wellman, M.P. & Rajan, U. (2017). Ethical Issues for Autonomous Trading Agents. Minds and Machines, 27, 609–624.

Supplementary Material

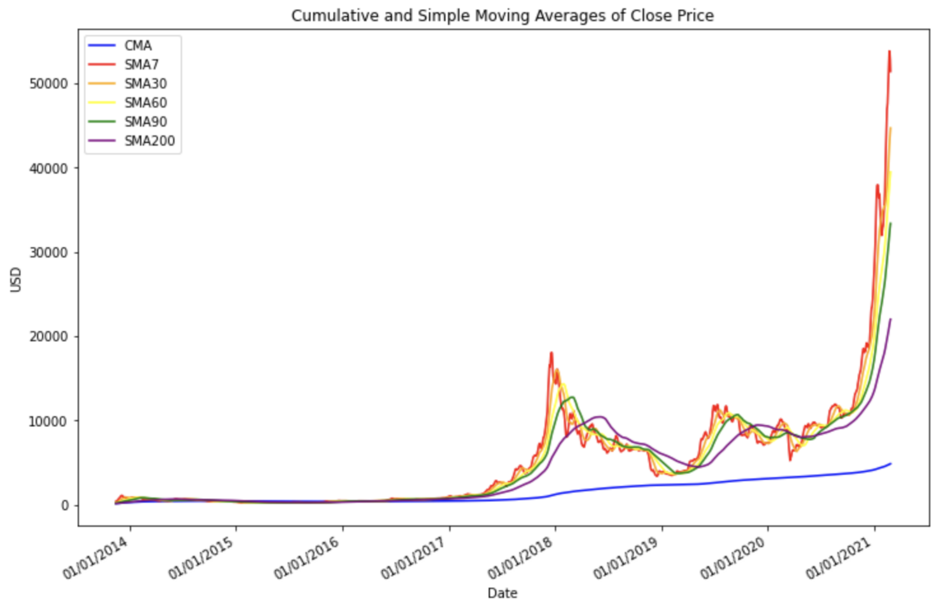

Figure 1. Graph of cumulative and simple moving averages over our chosen time periods as calculated from the raw BTC dataset.

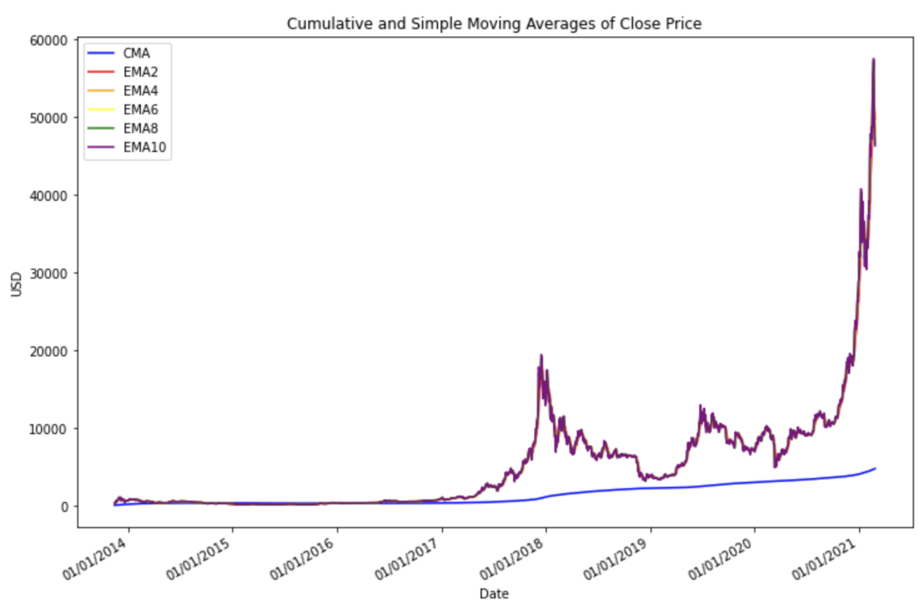

Figure 2. Graph of cumulative and exponential moving averages over with chosen alpha values as calculated from the raw BTC dataset.

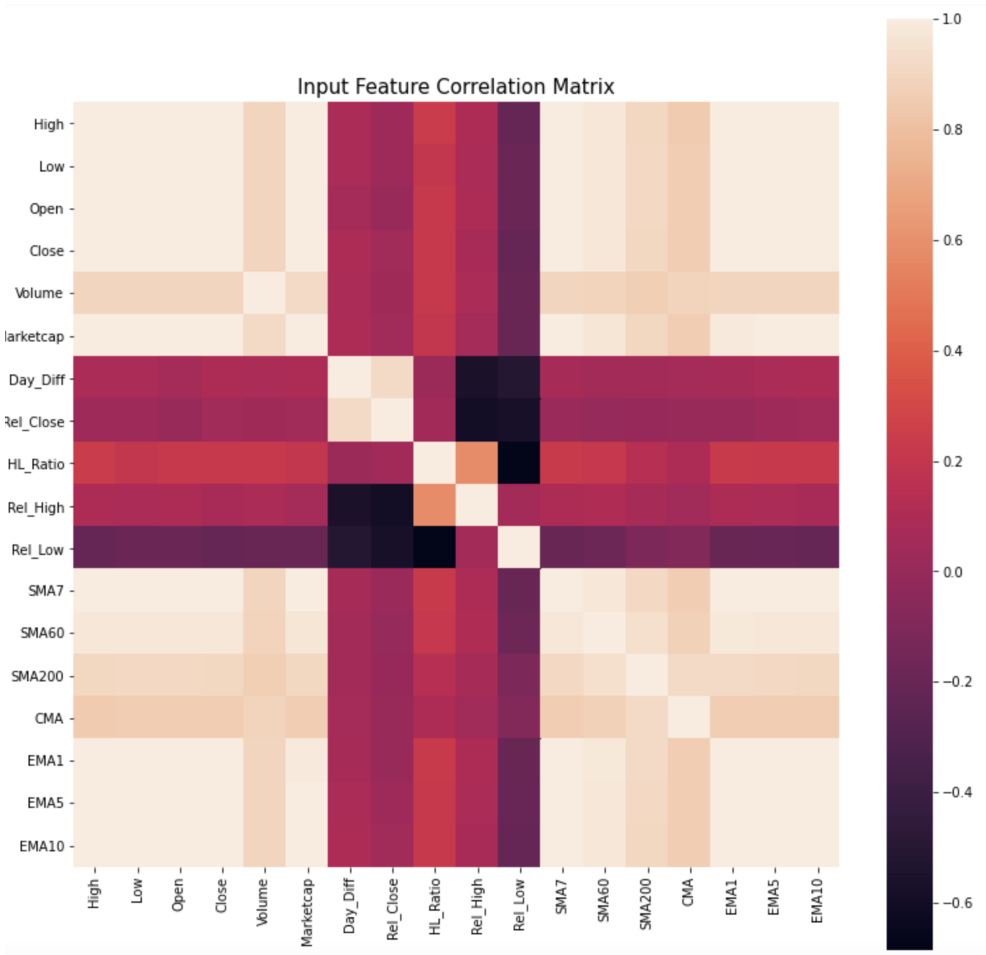

Figure 3. Correlation matrix of chosen input features using a simple Spearman Rank Correlation.