Table 1. Input variables used for each cryptocurrency.

Category	Variable
Price Trends	Close price (USD)
Price Trends	Log returns
Price Trends	7-day reversal
Price Trends	30-day reversal
Price Trends	6-month momentum
Price Trends	1-year momentum
Liquidity	24-hour volume
Liquidity	24-hour market cap
Volatility	30-day volatility
Volatility	90-day volatility
Volatility	180-day volatility
Volatility	90-day beta
Volatility	90-day beta squared
Network	Median transaction fee (USD)
Network	Active address count
Network	Transaction count
Network	Transfer count
Investor attention	Number of wallet users
Investor attention	Google search trend queries
Production	Avg. retail price of electricity in the United States
Production	Retail sales of electricity in the United States
Production	Net generation of electricity in the United States

Collection

We collect data on close prices, volume, and market capitalization from coinmarketcap.com. Coinmarketcap.com notes that the close price of a cryptocurrency refers to the latest updated price for a given day. The data from coinmarketcap is utilized to construct all of the price trend and liquidity factors. Volatility variables are computed using close prices for BTC, ETH, and XRP, as well as cryptocurrency market close prices. We choose the Crescent Crypto Market Index (CCMIX) to represent cryptocurrency market returns. The CCMIX employs a market capitalization based weighted average to construct their index. We collect production data from the U.S. Energy Information Administration (EIA), and network data is collected from coinmetrics.io. Data on the number of wallet users is accessed from blockchain.com, and Google search trend queries are obtained from trends.google.com. We employ search data for the term “bitcoin”, since it is larger, more liquid, and more widely recognized than any other cryptocurrency. Moreover, the term “bitcoin” is often used synonymously with the term “cryptocurrency” by those unfamiliar with blockchain technology. Overall, we visit trends.google.com, and download the historical data for the term “bitcoin” in the United States. The Google search trends are the historical datapoints of the weekly number of searches for the term “bitcoin”.

Variable Construction

While a plethora of the research’s independent variables are already processed and ready for use at the time of collection (e.g. production, investor attention, and network factors), we compute a handful of features manually. These

include metrics for returns, volatility, beta, beta squared, momentum, and reversal. Additionally, our target variable, the sign of next-day returns, is constructed from our calculated log returns.

Equation 1: We construct the daily log return, R_d, at day d, where P_d is the close price at day d and P_{d-1} is the close price at the day prior to day d:

$$R_d = \ln\left(\frac{P_d}{P_{d-1}}\right)$$

Equation 2: We construct beta on a 90-day rolling window, where c_{90} are the 90 daily log returns for cryptocurrency c. Conversely, m_{90} are the 90 daily log returns for the cryptocurrency market, which we base on the CCMIX:

$$\beta = \frac{\text{cov}(c_{90}, m_{90})}{var(m_{90})}$$

Equation 3: Beta squared is constructed by squaring the respective beta calculation at a given day:

$$(\beta)^2$$

Equation 4: 30-day volatility, V_{30}, is constructed as the standard deviation of returns on a 30-day rolling window, where c_{30} are the 30 daily log returns for cryptocurrency c:

$$V_{30} = \sqrt{var(c_{30})}$$

Equation 5: 90-day volatility, V_{90}, is constructed as the standard deviation of returns on a 90-day rolling window, where c_{90} are the 90 daily log returns for cryptocurrency c:

$$V_{90} = \sqrt{var(c_{90})}$$

Equation 6: 180-day volatility, V_{180}, is constructed as the standard deviation of returns on a 180-day rolling window, where c_{180} are the 180 daily log returns for cryptocurrency c:

$$V_{180} = \sqrt{var(c_{180})}$$

Equation 7: We calculate weekly reversal, Rev_7, as the 7-day holding period log return, where P_d is the close price at current day d and P_{d-7} is the close price 7 days prior to day d:

$$Rev_7 = \ln\left(\frac{P_d}{P_{d-7}}\right)$$

Equation 8: We calculate monthly reversal, Rev_{30}, as the 30-day holding period log return, where P_d is the close price at current day d and P_{d-30} is the close price 30 days prior to day d:

$$Rev_{30} = \ln\left(\frac{P_d}{P_{d-30}}\right)$$

Equation 9: We calculate 6-month momentum, M_{6m}, where P_{d-30} is the close price 30 days prior to current day d and P_{d-210} is the close price 210 days prior to current day d:

$$M_{6m} = \ln\left(\frac{P_{d-30}}{P_{d-210}}\right)$$

Equation 10: We calculate 1-year momentum, M_{1y}, where P_{d-30} is the close price 30 days prior to current day d and P_{d-395} is the close price 395 days prior to current day d:

$$M_{1y} = \ln\left(\frac{P_{d-30}}{P_{d-395}}\right)$$

Equation 11: Our target variable is the sign of daily log returns. We transform the returns into a binary classification problem, where 1 signifies a positive return and 0 represents a negative return. As such, the sign of daily returns at day d, S_d, is computed as follows:

$$S_d = \begin{cases} 1, & if\ R_{d+1} > 0 \\ 0, & otherwise \end{cases}$$

Note, since we are forecasting the sign of next-day returns, we shift our dataset such that the according sign value for day d is the original sign of returns for day $d+1$.

Cleaning Data

Our independent variables exist in varying levels of time-series frequency. Data like close-price, market capitalization, and liquidity are available on a daily frequency. However, larger macroeconomic variables like the net generation, average retail price, and total retail sales of electricity are only available on a monthly frequency. Moreover, Google search trend data is available on a weekly frequency. Since our target variable is the next-day sign of returns, we transform all data into a daily frequency. To be more specific, if the data is in a monthly frequency, the single monthly value is matched for every day of the according month. In the scenario where data is in a weekly frequency, the single week value is matched for every value of that respective week. We begin our research with a time-sample from September 5th, 2015 to January 1st, 2021. However, since the construction of yearly momentum requires that data from the previous year be available, momentum data for the first year is unable to be calculated. Thus, we eliminate all data from 9/5/15 to 9/4/16. Our final dataset consists of a time-sample from September 5th, 2016 to January 1st, 2021.

Variable Correlation

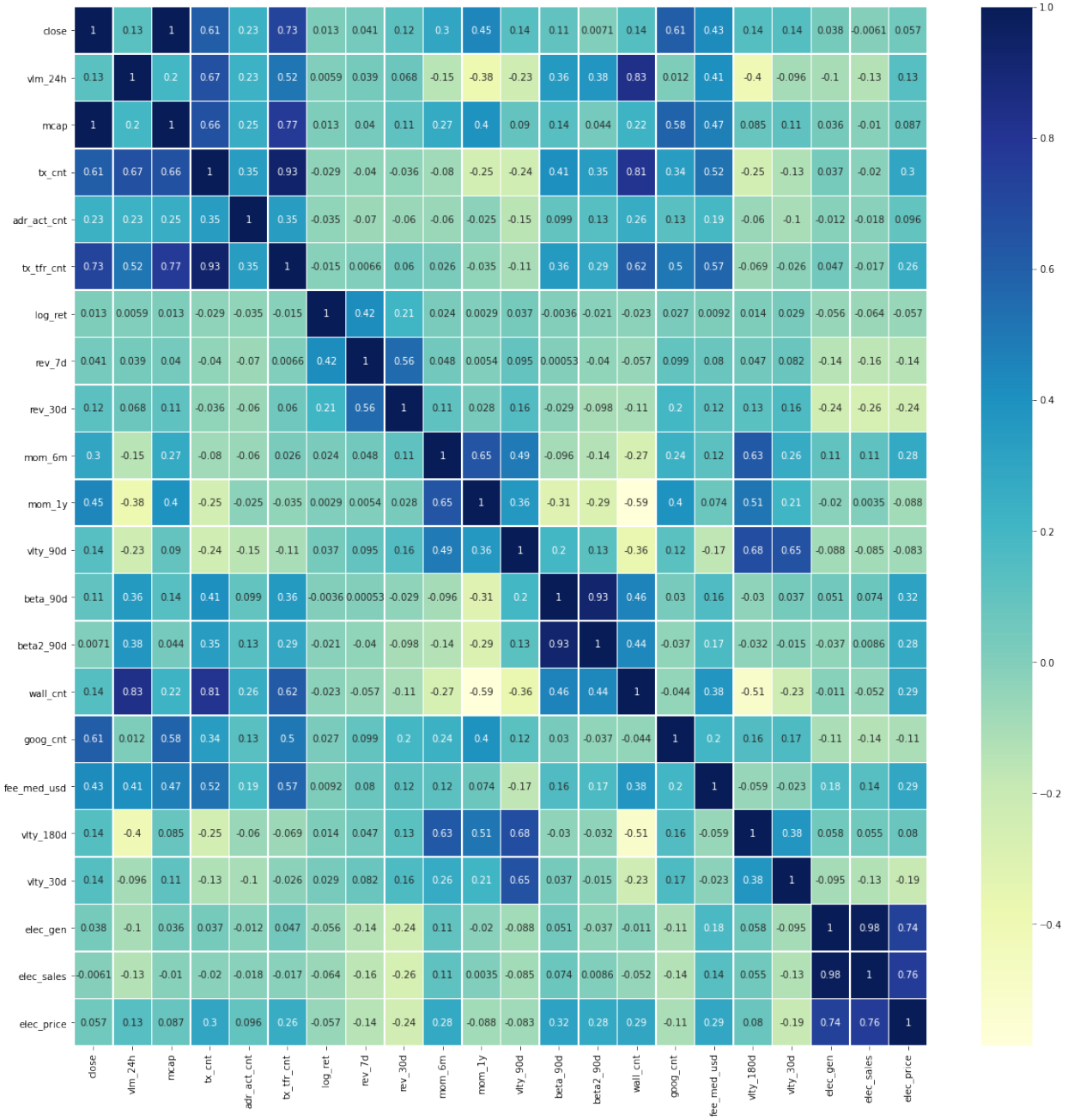

Figure 1. Heat map of all feature variable correlations. The displayed correlations are computed as the mean correlation value across BTC, ETH, and XRP for a given cell.

We reinforce the conclusion made by Liu and Tsyvinski (2020), substantiating that "there is a strong time-series momentum effect…". Additionally, of all variables analyzed within the research, the 7-day and 30-day reversals consistently maintain the strongest correlation towards returns. For all volatility variables, the 90-day volatility showed the most significant correlation with returns. Regarding investor attention, Google search trend queries showed a stronger correlation than wallet count. Moreover, we find that production variables demonstrated a powerful, although negative, correlation with returns. Similarly, three out of the four examined network factors also showed a strong but negative correlation. Factors for liquidity, namely the 24-hour volume, were insignificant.

Methodology

Software Materials

All data processing and analysis is done via Python 3.7. The dataset is cleaned and processed with Pandas and Numpy. Scikit-learn is utilized to implement the logistic regression, support vector machines, K-nearest neighbors, Gaussian process, decision tree, random forest, AdaBoost, and multilayer perceptron classifiers. The XGBoost is implemented via the native XGBoost Python library. Parameter optimization is done via autosklearn's Python package.

Preprocessing

We prepare our data for the machine learning models by using 80% training data and 20% testing data. Due to the time-series nature of our dataset, we avoid the use of cross-validation when training and testing data. Our feature variables are scaled to values between 0 and 1 using scikitlearn's MinMaxScaler. Overall, our dataset contained a target class imbalance, with the majority of all next-day targets having a positive sign. We combat the imbalanced data by applying a penalization score to the classes, utilizing scikitlearn's "class_weight" parameter to balance each class.

Models

We utilize a myriad of machine learning classification models to predict the sign of next-day returns. In all, our models included logistic regression, support vector machines, K-nearest neighbors, Gaussian process classifier, decision tree, random forest, AdaBoost, XGBoost, and multilayer perceptron. We apply the listed models to all three of our compiled cryptocurrency datasets: BTC, ETH, and XRP.

Trading Strategy

We develop a trading strategy that secures either a long or short position based not solely on the classification prediction, but also on the model's measured confidence that the classification is correct. In essence, we transform a mere binary classification problem into a probability prediction problem, where the model returns a probability for the sign of next day returns for the three cryptocurrencies. By optimizing our trading strategy to account for probability, we limit the positions that the model takes to solely those with the highest likelihood of classification success. Overall, we construct a simple probability-based trading strategy. First, for the three analyzed cryptocurrencies in this research (BTC, ETH, XRP), we predict the sign of next-day returns using what we determine to be the most accurate classification model for our dataset. Second, we take a single position each day on one cryptocurrency based on the model's prediction. The cryptocurrency we make a position on as well as the type of position (long or short) is the one with the model's highest probability of being correct. To be more specific, if the model determines that there is 55% probability for the next day sign of BTC to be positive, a 70% probability for ETH to be negative, and a 60% probability for XRP to be positive, we take a short position on ETH. Positions are liquidated every 24 hours and a new position is taken based on the model's predictions for the next day's sign. Probabilities are determined using scikitlearn's "predict_proba" method.

Results

Models

We find that the support vector machines give the most accurate classifications when forecasting the sign of next-day returns. On the contrary, we find that boosted methods like AdaBoost and XGBoost give the most inaccurate classifications. See Table 2 for a comprehensive overview of the various models' performance on each individual cryptocurrency.

Table 2. Classifiers and their accuracy for the next-day sign prediction of each cryptocurrency.

Model	BTC Acc.	ETH Acc.	XRP Acc.
Logistic regression	0.5233	0.5374	0.5196
Support vector machines	0.5698	0.5691	0.5572
K-nearest neighbors	0.5172	0.5001	0.4973
Gaussian process	0.5665	0.5686	0.5544
Decision tree	0.5601	0.5624	0.5499
Random forest	0.5665	0.5521	0.5501
AdaBoost	0.4778	0.4851	0.4713
XGBoost	0.5032	0.5005	0.4936
Multilayer perceptron	0.5506	0.5442	0.5415

Trading Strategy

We employ support vector machines for our long/short probability-based trading strategy, since our results show that it is the most accurate classifier for the sign prediction of next-day returns. See Table 3 for an overview of the cumulative log returns and Sharpe ratios for our trading strategy (SVM probability-based trading strategy), compared to the those from a standalone investment in BTC, ETH, or XRP over the same holding period.

Table 3. The research's trading strategy compared to a single long position in all three cryptocurrencies.

Model	Log returns	Sharpe
SVM probability-based trading strategy	3.72	2.8
BTC	1.11	1.49
ETH	1.01	1.06
XRP	-0.15	0.13

Overall, the SVM's probability-based trading strategy has a log return of 3.72. This correlates to a rate of return of approximately 41.3%. Thus, with a $100 investment, one would profit $41.3.

Discussion

This research substantiates the literature's general consensus that machine learning classifiers are capable of predicting the next-day sign of returns for cryptocurrencies. Moreover, we corroborate the notion that both the returns and prices

of cryptocurrencies maintain a strong relationship with time-series momentum. We find added significance in the use of additional price trend variables to forecast the next-day sign of returns, most notably with 7 and 30-day reversal. Regarding the comparative analysis of machine learning models for the research's task, our study encompasses a larger list of models than the vast majority of work in the literature, eight total models, to be exact. Despite our larger array of models, they do lack diversity. More specifically, we utilized a substantial number of tree-based methods, quite possibly more than necessary. Additionally, the literature has found great success in the use of recurrent neural networks (RNN's), to which we used none. Despite the fundamental similarity shared across many of our models, this did pose some added benefits. When assessing the accuracy of the models, for example, we discovered that both XGBoost and AdaBoost showed the weakest classification accuracies. This further corroborated our conclusion that boosted methods have the least promise for the research's given prediction task. Had we instead used solely one boosted method; our conclusion would have been nowhere near as significant. We ultimately find that support vector machines demonstrate the highest classification accuracy for all three cryptocurrencies. We attribute the success of the support vector machines towards the dataset's relatively short time-series horizon, spanning a little over four years. Another significant observation from the research regards the predictability of ripple (XRP). We find that all models could more accurately forecast BTC and ETH than XRP. We suspect that this phenomenon could be attributed to XRP's more volatile nature.

Aside from assessing the significance of cryptocurrency valuation factors and machine learning models, the remainder of our research focused on developing a profitable trading strategy, to which we found great success. The majority of the research that exists on the forecasting of cryptocurrency returns, devises a trading strategy from only one cryptocurrency. By constructing our models such that they forecast the next-day sign of returns for BTC, ETH, and XRP, we are granted significantly more flexibility regarding the means by which we can construct a trading strategy. As such, we formulated one not solely by a model's predicted classification, but also the probabilities of the classification being correct for each of the analyzed cryptocurrencies.

Our trading strategy subsequently saw results that significantly outperformed standalone long positions in all three cryptocurrencies over the same holding period. The average log return of an investment in BTC, ETH, or XRP was 0.66, whereas our support vector machines probability-based trading strategy showed a log return of 3.72. Additionally, the average Sharpe for the cryptocurrencies was 0.89. On the other hand, our strategy delivered a Sharpe of 2.8. Overall, the strategy formulated within this research outperformed log returns and Sharpe ratios by a factor of 5.64 and 3.15, respectively. Our strategy saw more success than studies focused on cryptocurrency predictability over short time horizons (1 and 60 minute). Taking a multitude of positions over small time periods makes a strategy more vulnerable to trading fees, which may seem marginal at first, but deteriorate returns over a longer period of time.

Conclusion

We capitalize on a wide array of machine learning models to continue the literature's study on the overall risks and returns of bitcoin (BTC), ethereum (ETH), and ripple (XRP). First, we begin our research by assessing the most significant valuation factors towards cryptocurrency prices and returns. Second, we execute a comparative analysis of machine learning models for the next-day sign prediction of the three cryptocurrencies. Third, we utilize our findings to develop a profitable, probability-based trading strategy.

We reinforce the conclusion made by Liu and Tsyvinski (2020), substantiating that "there is a strong time-series momentum effect...". We also find that for all variables analyzed in the research, the 7 and 30-day reversals show the strongest correlation towards returns. We conclude that support vector machines provide the highest classification accuracy when forecasting the sign of next day cryptocurrency returns. This is in contrast to boosted methods like AdaBoost and XGBoost, whose performance was the worst of the examined models. Ultimately, we employ our conclusions to construct a probability-based trading strategy that delivered a Sharpe of 2.8 and a cumulative log return of 3.72. In contrast, the according Sharpe ratios and returns were 1.11 and 1.49, 1.01 and 1.06, and -0.15 and -0.13 for standalone long positions in BTC, ETH, and XRP, respectively.

Limitations

Significant improvements could be made in future research to construct more accurate models and subsequently develop a more profitable trading strategy.

More specifically, a larger list of tested classifiers could potentially prove to be beneficial. The literature has shown success in cryptocurrency prediction via the use of recurrent neural networks, none of which were used in the research. Moreover, a more comprehensive parameter optimization for each of the tested models would likely result in positive contributions towards their accuracy. Building off the literature's conclusions that cryptocurrency returns have a significant relationship with momentum, our research could have also benefited from a deeper focus on varying frequencies of time-series data. Tests on different levels of time frequency like the previous 5 days, week, and month to predict the sign of next day returns may have improved the effectiveness of the models. A deeper emphasis on feature selection could have helped to eliminate unnecessary noise within the data, resulting in more accurate classifiers. In addition, the accuracy of the models could have been enhanced by compiling a dataset that consisted of more price trend related features, since our research shows that those variable categories posed the strongest predictive power. Regarding the trading strategy, we could have more accurately assessed its performance by testing the strategy in a wide array of changing market conditions, rather than the single market condition (bull) utilized within our testing data. Additionally, compiling a larger list of cryptocurrencies to the dataset may have resulted in both a more successful trading strategy, as well as a more comprehensive understanding of the variable correlations.

The most significant limitation we encountered throughout the research was in the data collection process. Due to the novelty of cryptocurrencies, limited time-series data is available for most variables. As a result, our training and testing data was of a considerably smaller size than datasets used for traditional assets like stocks and bonds. For the time being, however, little can be done to combat this issue. Over time, as cryptocurrency grows in attention as both a phenomenon and an investment, more data will become available, fostering the subsequent growth and advancement of literature.

Acknowledgments

We especially thank Tianshu Lyu from the Yale School of Management for his guidance, advice, and mentorship throughout the entirety of the research process. We also thank Lumiere Education for facilitating the relationship with Tianshu Lyu and for their general support in the formulation of the research paper.

References

Baur, D., Hong, K., & Lee, A. (2018). Bitcoin: medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money,* 54, 177-189. https://doi.org/10.1016/j.intfin.2017.12.004

Bouri. E., Molnar. P., Azzi, G., Roubaud, D., & Hagfors, L. (2017). On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier? *Finance Research Letters*, 20, 192–198. https://doi.org/10.1016/j.frl.2016.09.025

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies,* 33(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Jaquart, P., Dann, D., & Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning, *Journal of Finance and Data Science*, 7, 45-66. https://doi.org/10.1016/j.jfds.2021.03.001

Kristoufek, L. (2013). Bitcoin meets Google trends and Wikipedia: quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3, 3415. https://doi.org/10.1038/srep03415

Liu, Y., & Tsyvinski, A. (2021). Risks and returns of cryptocurrency, *Review of Financial Studies*, 34(6), 2689-2727. https://doi.org/10.1093/rfs/hhaa113

Panagiotidis, T., Stengos, T., & Vravosinos, O. (2019). The effects of markets, uncertainty and search intensity on bitcoin returns. *International Review of Financial Analysis*, 63, 220–242. https://doi.org/10.1016/j.irfa.2018.11.002

Sebastiao, H., & Godinho, P. (2021). Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financial Innovation*, 7, 3. https://doi.org/10.1186/s40854-020-00217-x