Ensemble Model Crypto Stock Prediction

Eshan Kaul, Stephen Goodwin and Milen John
Virginia Tech

Abstract

- Cryptocurrency price prediction has been a popular topic of research in recent years due to the soaring prices., However, the volatility of cryptocurrencies had made this task challenging.
- Using an ensemble learning approach, you can build a robust model that maximizes the strengths of multiple models, and minimizes their weaknesses. This model can then be used in a pairs-trading strategy and can see improved performance upon optimization.

Introduction

- The recent surge in cryptocurrency popularity has led to increased interest in strategies to capitalize on market inefficiencies and volatility.
- Our project focuses on implementing a pairs trading strategy, which is a market-neutral approach that leverages mean reversion in correlated assets.
- To maximize the effectiveness of this strategy, we will employ a Regression Voting Ensemble with soft voting for accurate future price predictions of cointegrated assets.
- Our robust crypto forecasting model will utilize a combination of Random Forests, Orthogonal Matching Pursuit, Gradient Boosting Regressor, and many other parametric and non parametric regression models within the Ensemble Voting framework.

Datasets

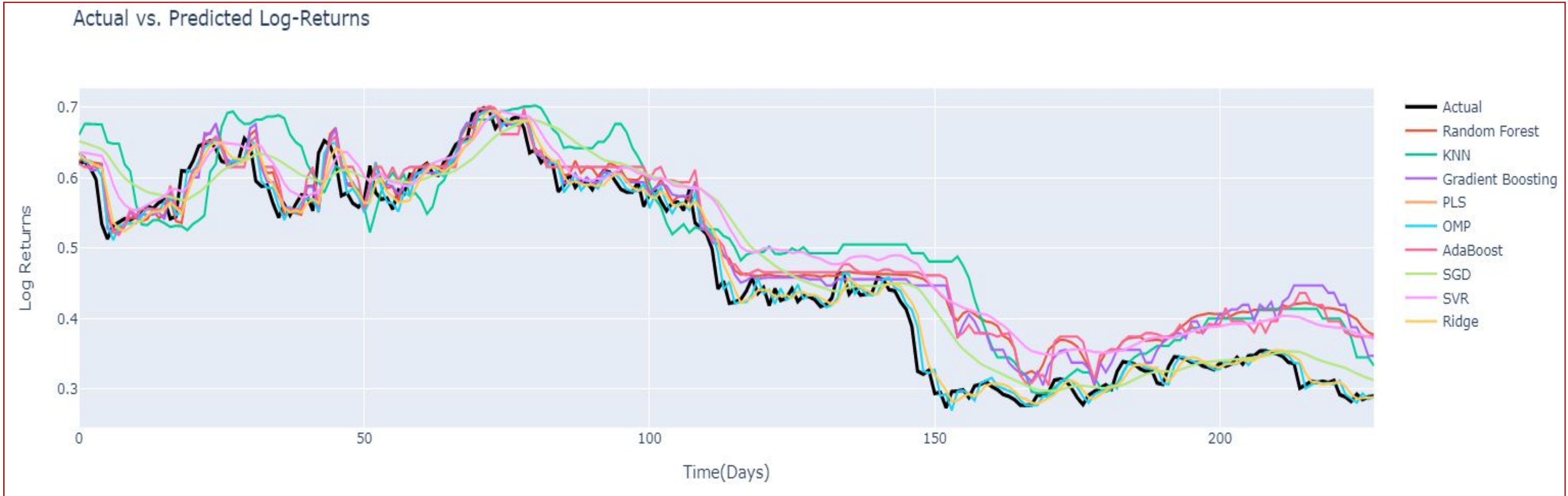
- The dataset is crypto data that is scraped from Yahoo Finance utilizing a custom python script that takes in inputs names of the tickers for various crypto/stocks and dates specified by the user. The script scrapes several attributes in daily increments including: Date, Open Price, High Price, Low Price, Close Price, Volume etc. The final data set included twenty three different assets with daily observations from Jan. 1 2017 to present day (April 2023).

Methodology

- The first step of the project is to identify cointegrated assets because pairs trading is only effective on highly cointegrated assets. Examples of assets pairs that exhibit the mean reverting behavior include Bitcoin and Ethereum. The Engle-Granger and Johansen tests are employed to determine these relationships.
- The next step is to build and train each individual model to be used in the ensemble voting regressor using a 60-20-20 training, test, validation set.
- This is followed by stacking the individual regression models into the ensemble, and utilizing a bootstrap aggregation (bagging) method to determine the weights of each regressor inside the ensemble in an effort to minimize the variance of the total error term.
- Finally the results of the ensemble voting regression model is compared to the results of an ARIMA model which is a popular time series model for forecasting financial assets price movements.

Results

- In an ensemble machine learning approach, multiple models are combined to achieve improved overall performance. In this case, the 9 models - Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, Partial Least Squares (PLS), and Orthogonal Matching Pursuit (OMP), Ada-Boost, Ridge, Support Vector Regression (SVR), Stochastic Gradient Descent (SGD) - will work together to produce a more accurate and robust prediction. The performance metrics provided for each model will be used to analyze their individual contributions to the ensemble.



	KNN	PLS	OMP	Gradient Boost	RF	Ada-boost	SGD	SVR	Ridge
R-squared	0.726	0.863	0.982	0.849	0.837	0.839	0.895	0.811	0.975
RMSE	0.070	0.049	0.017	0.052	0.054	0.053	0.043	0.058	0.0211

- The OMP model exhibits the best performance among the five models, with the lowest values for MSE, RMSE, and MAE. Consequently, it will likely have a strong positive influence on the ensemble's overall performance. The PLS model follows closely with the second-best performance, making it another valuable contributor to the ensemble. These two models can potentially compensate for the weaker performance of the other three models, especially in cases where the weaker models might overfit or fail to capture specific patterns in the data.
- The Gradient Boosting model, which presents a middle ground between the OMP and PLS models and the Random Forest and KNN models, could still contribute valuable information to the ensemble. While it may not be as accurate as the OMP and PLS models, its distinct learning approach can help capture patterns missed by the other models, thus enhancing the ensemble's performance.
- The Random Forest and KNN models, despite having relatively higher error values, can still play a role in the ensemble approach. Their unique learning mechanisms can capture different aspects of the data, complementing the other models and potentially improving the overall performance. However, it is crucial to carefully manage their weights in the ensemble to avoid negative impacts on the final predictions.

Conclusion

In summary, the ensemble voting mechanism that stacks several models can lead to improved performance by leveraging the strengths of each model while mitigating their weaknesses. The OMP and PLS models will likely have a significant positive impact on the ensemble's performance, with Gradient Boosting, Random Forest, and KNN models providing additional complementary information. To fully benefit from this ensemble approach, it is crucial to optimize the combination of models and their respective weights to achieve the best possible predictions.

References

- Junwei Chen. 2023. Analysis of Bitcoin Price Prediction Using Machine Learning.Journal of Risk and Financial Management 16, no. 1: 51.
- Miroslav Fil and Ladislav Kristoufek. 2020. Pairs Trading in Cryptocurrency Markets. IEEE Access 8 (2020), 172644–172651. https://doi.org/10.1109/ACCESS. 2020.3024619
- David Hitchcock and Shan Zhong. 2021. S&P 500 Stock Price Prediction Using Technical, Fundamental, and Text Data. In Statistics, Optimization Information Computing, 9(4), 769-788.
- Y Huang, L. F. Capretz, and D Ho. 2021. Machine Learning for Stock Prediction Based on Fundamental Analysis. In 2021 IEEE Symposium Series on Computational Intelligence (SSCI).
- Phaladisailoed, Therasak, and Thanisa Numnonda. 2018. Machine learning models comparison for bitcoin price prediction. In 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE).
- Selmi, Refk, Walid Mensi, Shawkat Hammoudeh, and Jamal Bouoiyour. 2018. Is Bitcoin a hedge, a safe haven or a diversifier for oil price movements? A comparison with gold. In Energy Economics 74.
- Gu JW. Yu FH. et al. Zhu, DM. 2021. Optimal pairs trading with dynamic mean- variance objective. Mathmatical Methods of Operations Research 94 (2021), 145–168. https://doi.org/10.1007/s00186-021-00751-z