Proposal For Ethereum (ETH) Next Day Price Prediction

Siarhei Staravoitau
School of Computing
National College of Ireland
Dublin, Ireland
x18162070@student.ncirl.ie

I. Introduction

This project proposal describes motivation, research question, theoretical literature review, data source and statistics, ML approaches and evaluation while performing ETH next-day price prediction with a focus on BTC price as market sentiment used as a predictor. The project goal is to build a number of SciKit-learn-based (also, it could be considered other libraries during research) models and evaluate their accuracy while predicting ETH next-day price based on historical data of five years retrieved from Yahoo Finance. This should offer valuable insights for financial analytics, investors and other relevant stakeholders.

II. Motivation

In the last decade, cryptocurrency price fluctuation could cause significant implications in various economic sectors like finance, investment and economic research. This research study predicts next-day Ethereum (ETH) prices through Scikit-learn and other libraries and uses BTC trading volume and prices as market sentiment, which could give a better insight into a better understanding of modern economic trends. The following rationales make this project essential:

Investment Strategy:

Cryptocurrency price analysis and prediction using advanced ML techniques and frameworks could help investors and various financial institutions get a better picture, aggregate useful cryptocurrency insights and trend analysis, and optimise investors' decisions to maximise returns and minimise risks in the volatile and fragile cryptocurrency market.

Financial Products Development:

Advanced ML techniques in cryptocurrency price analysis and prediction help design new financial instruments like options, futures, and ETFs tailored for cryptocurrency markets. This gives more options for optimising risks and providing hedging opportunities and attracts more players to cryptocurrency markets.

Economic Research:

Using other cryptocurrencies as additional market sentiment for predicting target cryptocurrency behaviour could bring better insights into understanding cryptocurrency price interdependencies between cryptocurrency markets and contribute to scholarly research, market dynamics, and the overall impact on a global economy.

Regulatory and Policy Framework:

Using effective cryptocurrency market predictive models could help policymakers better understand market trends and their implications for the global financial system. This helps to improve market stability, predictability, market protection, fraud detection and prevention.

Technology and Innovation:

Using advanced ML libraries in financial research stimulates further using applied ML methods and pushes forward innovations, development, and AI applications while analysing highly volatile markets and data environments.

The project aims to amalgamate the usage of ML tools and cryptocurrency price data insights to predict ETH next-day prices, calculate Scikit-learn-based and potentially other ML libraries models accurately, compare them, and deliver them to the stakeholders, getting them better informed while dealing with cryptocurrencies.

III. Research Question

RQ 1: How do fluctuations in BTC Volume, Open, Close, Adjusted Close, High, and Low daily prices influence on and correlate with next-day ETH price prediction over the past five years? Which predictor features most significantly impact ETH price predictions?

RQ 2: Develop, implement and evaluate the ETH next-day price prediction ML pipeline, including data retrieving, visualising, cleaning, predictors selection, model fitting, predicting ETH next-day price, visualising results and aggregating accuracy measures.

IV. Initial Review

For conducting research and answering RQs, the following papers were elected to contribute to the theoretical base and insights into predicting Ethereum (ETH) prices with BTC prices as market sentiments using ML techniques covering the end-to-end data workflow pipeline, starting from data acquisition, preprocessing, modelling, predicting, results evaluation, and visualisation.

- S Semerikov in ”Financial Time Series: Short-term Forecasting Using Machine Learning Techniques” [1] discusses the usage of machine learning approaches using cryptocurrency datasets and applying various ML methods, including Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forests (RF), and Stochastic Gradient Boosting Machine (SGBM). The author stated that SGBM outperformed other approaches, and RF showed the worst results. To improve the model's accuracy performance, the author recommends the usage of additional indicators and oscillators, such as price rate of change and relative strength index. [1]

- Troy J. Strader et al. in ”Machine Learning Stock Market Prediction Studies: Review and Research Directions” [2] review an application of ML approaches like Linear regression, Ridge regression, Lasso regression, Elastic Net regression, and Random Forest to predict the next day stock price, how to perform features selection, build model, evaluate and visualise results. Research suggests that for numeric analysis and predictions, the best results are demonstrated by ANN, the SVM approach demonstrates better results in classification tasks, and the Genetic approach shows good results in identifying high-quality system inputs like composing stock portfolios.
- Han-Min Kim, et al. in ”Predicting Ethereum prices with machine learning based on Blockchain information” [3] reviews influence factors on ETH price by other cryptocurrencies like BTC, Litecoin, etc. Apart from BTC price, as an additional feature could be used blockchain information like ”Ethereum-specific Blockchain information variables such as gas used, gas limits, gas price, uncle block, Blockchain information of other coins, macroeconomic factors, and generic Blockchain information” [3].
- Haoyu Jiang in ”Comment on papers using machine learning for significant wave height time series prediction: Complex models do not outperform auto-regression” [4] reviews the performance of various simple regression models vs complex models and concludes that in certain market conditions, complex models do not outperform the simple models.
- N. Khemani in ”Predicting Ethereum Prices Based on Blockchain Data” [5] performs research based on the LSTM model and suggests using blockchain-related metrics to enhance the prediction accuracy of cryptocurrency prices. A similar could be applied to non-ANN prediction models.
- H. Sebastião and P. Godinho in ”Forecasting and trading cryptocurrencies with machine learning under changing market conditions” [6] discusses the ML usage for developing trading strategies for BTC, LTC and ETH using linear models, RF and SVM. It was noted poor performance of ARIMA models due to price volatility for all instruments; all models demonstrated inconsistent accuracy, and ensemble models outperformed individual models.
- Jethin Abraham et al. in ”Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis” [7] investigate the application of sentiment analysis from Twitter, Google Trends and news sources for predicting the ETH prices as one of the predictors in a multiple linear regression model, which delivered consistent accuracy results.
- Alice Zheng and Amanda Casari in ”Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists” [8] highlight the importance of advanced feature engineering techniques, such as lag features and rolling averages, in enhancing the performance of predictive models. This could be used in project research during data preprocessing and features engineering.

V. Data Source & Statistics

Data Source: The proposed research project will use publicly available price data retrieved from Yahoo Finance using the Yahoo Finance API. The dataset contains ETH and BTC daily prices, including Open, High, Low, Close, Adjusted Close, and Volume data retrieved over five years.

Data Statistics: The data set will be divided into training and test datasets with ratios of 80% and 20%. This will allow us to effectively train and evaluate ML models and compare them.

Features Selection: To select predictor features will be used following approaches [9]:

- Univariate Feature Selection - based on univariate statistical tests like ANOVA, chi-square, and F test - what allows establishing a relationship between predictors and the target variable
- Tree-Based Feature Selection - uses Random Forest to evaluate the relationship between features and can capture the non-linear relationship between features
- Greedy Backward Feature Selection - this method uses iterative exclusion of features from the list based on criteria like AUC, accuracy and number of features to be used in the model and stops when criteria are met (Fig. 1).

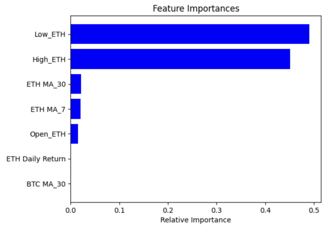

Fig. 1: Features selection

VI. Machine Learning Methods

For conducting a research project to predict Ethereum (ETH) prices, incorporating Bitcoin (BTC) prices as market sentiment, it was selected following SciKit Learn ML models.

- **Linear Regression** will be used as a starting point to create a basic model and evaluate its performance. [10] [11] [12].
- **Ridge Regression** extends linear regression, introducing L2 regularization to avoid overfitting. It enhances generalizability in multicollinear scenarios [13].

- **Lasso Regression** uses L1 regularization for features selection and effectively reduces the number of variables in the model. [14].
- **SVR (Support Vector Regression)** uses kernel tricks for building non-linear complex models, typically in financial markets. ”Support vector machine (SVM) analysis is a popular machine learning tool for classification and regression, first identified by Vladimir Vapnik and his colleagues in 1992. SVM regression is considered a non-parametric technique because it relies on kernel functions” [15] [16] [17].
- **Decision Tree Regressor** applies a tree-like model to analyse non-linear data relationships without transformations. ”It is a tree-structured classifier with three types of nodes. ... This algorithm is very useful for solving decision-related problems.” [18] [19].
- **Random Forest Regressor** ”combines ensemble learning methods with the decision tree framework to create multiple randomly drawn decision trees from the data, averaging the results to output a new result that often leads to strong predictions/classifications.” [20] [21].
- **Gradient Boosting Regressor** ”builds an additive model in a forward stage-wise fashion; it allows for optimising arbitrary differentiable loss functions. In each stage, a regression tree is fit on the negative gradient of the given loss function.” [22] The model runs trees sequentially, correcting previous trees’ errors and improving the model’s accuracy. [23].

The overall rationale for selecting the above approaches is to build different models employing their features, test how they behave using time-series volatile data, and establish the best-performing model. The project aims to develop and establish the most comprehensive ETH next-day price predicting model and provide insights into the factors influencing these price changes.

VII. Evaluation Methods

For effective models’ evaluation predicting Ethereum (ETH) prices using Bitcoin (BTC) as market sentiment will be used following comprehensive evaluation methods, statistical analyses and visualizations:

- **Accuracy**: RMSE, MSE, MAPE - accuracy metrics helping to assess the models’ performance, providing an overall universal indicator for all models [24].
- **Statistical Analysis and Data Visualization**:
  - *Descriptive Statistics*: Calculate descriptive statistics for each variable, including mean, median, standard deviation and range. This is required for a better understanding of the central tendency and dispersion [25].
  - *Distribution Checks*: It will be analysed using histograms and density plots to better understand the variable’s distribution for planning possible data pre-processing [25].
  - *Boxplots*: It will be used for continuous variables distribution visualisation and outliers detection check [25].
  - *Time Series Visualization*: Prices Time series plots could help to identify patterns, trends, or cyclic behaviours while feature selection and engineering, along with possible model selection [25] [24] (Fig. 2).

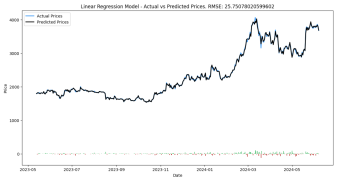

Fig. 2: Time Series ETH predicted vs actual prices

  - *Q-Q plot* Fig. 3 will be used for visualising and evaluation predicted prices vs actual prices.

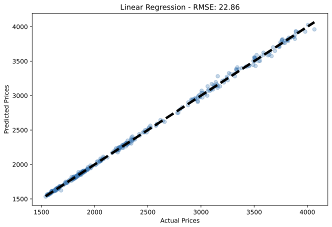

Fig. 3: Q-Q plot

References

[1] S. Semerikov, “Financial time series: Short-term forecasting using machine learning techniques,” *Journal of Computer Science and Technology*, no. 004.94, 11 2021, depositing User: . Last Modified: 25 Nov 2021 21:54. Institute for Digitalisation of Education ¿ Joint laboratory with SIHE ”Kryvyi Rih National University”. [Online]. Available: https://lib.iitta.gov.ua/id/eprint/727244

[2] T. J. Strader, J. J. Rozycki, T. H. Root, and Y.-H. Huang, “Machine learning stock market prediction studies: Review and research directions,” *Journal of International Technology and Information Management*, vol. 28, no. 4, 2020. [Online]. Available: https://scholarworks.lib.csusb.edu/jitim/vol28/iss4/3

[3] H.-M. Kim, G.-W. Bock, and G. Lee, “Predicting ethereum prices with machine learning based on blockchain information,” *Expert Systems with Applications*, vol. 184, p. 115480, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421008915

[4] H. Jiang, Y. Zhang, C. Qian, and X. Wang, “Comment on papers using machine learning for significant wave height time series prediction: Complex models do not outperform auto-regression,” *Ocean Modelling*, vol. 189, p. 102364, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1463500324000519

[5] N. Khemani, “Predicting ethereum prices based on blockchain data,” 2022, accessed: 2024-06-16. [Online]. Available: http://cs230.stanford.edu/projects_fall_2022/reports/149.pdf

[6] H. Sebastião and P. Godinho, “Forecasting and trading cryptocurrencies with machine learning under changing market conditions,” *Financial Innovation*, vol. 7, no. 1, pp. 1–30, 2021. [Online]. Available: https://doi.org/10.1186/s40854-020-00217-x

[7] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, “Cryptocurrency price prediction using tweet volumes and sentiment analysis,” *SMU Data Science Review*, vol. 1, no. 3, p. Article 1, 2018. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss3/1

[8] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., March 2018. [Online]. Available: https://learning.oreilly.com/library/view/feature-engineering-for/9781491953235/

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Feature selection*, Scikit-learn, 2023, accessed: 2024-06-29. [Online]. Available: https://scikit-learn.org/stable/modules/feature_selection.html

[10] Scikit-Learn Developers, "Scikit-Learn: Machine Learning in Python," 2021, accessed: 2024-06-16. [Online]. Available: https://scikit-learn.org/stable/

[11] A. Jidge, "The complete guide to linear regression analysis," 2023, accessed: 2024-06-19. [Online]. Available: https://towardsdatascience.com/the-complete-guide-to-linear-regression-analysis-38a421a89dc2

[12] M. Huang, "Theory and implementation of linear regression," in *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, 2020, pp. 210–217.

[13] D. Kumar, "Ridge regression and lasso estimators for data analysis," MSU Graduate Theses, Missouri State University, 2019, accessed: 2024-06-19. [Online]. Available: https://bearworks.missouristate.edu/theses/3380

[14] L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga, "A critical review of lasso and its derivatives for variable selection under dependence among covariates," 2020. [Online]. Available: https://arxiv.org/abs/2012.11470

[15] "Understanding support vector machine regression," accessed: 2024-06-19. [Online]. Available: https://uk.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

[16] "An introduction to support vector regression (svr)," accessed: 2024-06-19. [Online]. Available: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[17] N. Rf, "Support vector regressor: Theory and coding exercise in python," accessed: 2024-06-19. [Online]. Available: https://medium.com/@niousha.rf/support-vector-regressor-theory-and-coding-exercise-in-python-ca6a7dfda927

[18] "Machine learning basics: Decision tree regression," accessed: 2024-06-19. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

[19] "The only guide you need to understand regression trees," accessed: 2024-06-19. [Online]. Available: https://towardsdatascience.com/the-only-guide-you-need-to-understand-regression-trees-4964992a07a8

[20] "Random forest regression," accessed: 2024-06-19. [Online]. Available: https://towardsdatascience.com/random-forest-regression-5f605132d19d

[21] "Machine learning basics: Random forest regression," accessed: 2024-06-19. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a

[22] "Gradientboostingregressor," accessed: 2024-06-19. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[23] "Gradient boosting with scikit-learn, xgboost, lightgbm, and catboost," accessed: 2024-06-19. [Online]. Available: https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/

[24] S. Staravoitau, "Efficiency of machine learning cloud-based services vs traditional methods in stock prices prediction," Master's thesis, Dublin, National College of Ireland, September 2022, submitted. [Online]. Available: https://norma.ncirl.ie/5972/

[25] G. Malato, "Statistical analysis of a stock price," Towards Data Science, 2023, accessed: 2024-06-20. [Online]. Available: https://towardsdatascience.com/statistical-analysis-of-a-stock-price-e6d6f84ac2cd