Short-Term Cryptocurrency Price Prediction

Zijin Hu, Chenye Yuan, Hongyi Zheng

August 23, 2022

1 Abstract

Cryptocurrencies, a hotly discussed investment option this year, are being added to more and more people's investment portfolios due to their wild volatility, free trading environment, and correlation with global assets. As a new type of asset, his pricing strategy is still being explored. This paper used market data and public sentiment generated form Twitter posts and news headlines using finBERT as data sources, and tries a variety of models, including linear regression, decision tree, and XGBoost, to try to make bidirectional predictions of cryptocurrency price trends, and achieve stable and profitable results in backtesting.

2 Introduction

Cryptocurrency is a kind of digital currency enabled and secured by blockchain cryptography technology. As an emerging type of asset, cryptocurrency is less similar to equities or commodities which, more or less, have their intrinsic values, but much more like gold or even fiat money, despite that its technical decentralized features back the credits of cryptocurrency. Empirical asset pricing is a major branch of finance, but evaluating or predicting the price of cryptocurrency is very difficult as it is overly based on market consensus and less on fundamental or even macro information.

In this project, we study the short-term predictability of various cryptocurrencies. We choose to study crypto because of its unorthodox characteristics: it is less related to fundamentals and thus highly volatile, which also means more market signals. We selected and tested a comprehensive set of market signals and information, both traditional and nontraditional in the financial sense. We tried various models, through which we hope to achieve a balance of both predictivity and interpretability.

3 Business understanding

By Efficient Market Hypothesis, all publicly available information will be reflected in the price of an asset. On the one hand, EMH means all short-term fluctuations are simply random walks, and all effort trying to predict the market will end in vain.

On the other hand, we know that there is no perfect efficient market in the real world. Investors still gain edges by analyzing fundamentals, while short-term trading is more of a game between market agents. The price of an asset in the short term depends more on the market sentiments, instead of fundamental or macro information. Thus, the signal offered by the market could in return influence the market, which blurs the boundary of leading and lagging indicators, offering opportunities to predict the market.

Cryptocurrency, as an asset whose value is determined by consensus, its short-term fluctuation may be led by the momentum that could be uncovered by mining market indicators. Meanwhile, the high volatility of cryptocurrency provides more signals and trading opportunities.

Therefore, we choose some indicators (which we will explain in the later parts) that are traditionally or logically correlated with price movement and try to predict the price directional movement of cryptocurrency in the short term (5Min). Some of the indicators do not come from the cryptocurrency itself, but instead, come from a special kind of derivative for cryptocurrency, which is called a perpetual swap. The funding mechanism of perpetual swap ensures that its value is almost equivalent to its underlying cryptocurrency. Therefore, indicators that are able to predict the price movement of the perpetual swaps could also predict the price movement of the cryptocurrency.

4 Data Understanding

The data we used could be categorized into two types: market data and alternative data (sentiment score). Here is every feature under each category:

- Market Data
  - Open Interest: The total number of perpetual swap contracts held by market participants during a fixed period. It is used as an indicator to determine market sentiment and the strength behind price trends.
  - Top Trader Long/Short Ratio: The ratio between long and short orders made by traders with the largest trading volume in the past few days. Reflect on the opinion of these top traders regarding future price movements.
  - Global Long/Short Ratio: The ratio between long and short orders made by every single user of the platform. This could differ a lot from the top trader long/short ratio since the previous metric is determined by a few traders.
  - Taker Buy/Sell Volume: Different from long/short volume, this feature measures the amount of orders that are actually being executed. 99% of the orders on an exchange eventually ends up being canceled (since most market participants are high-frequency market makers). Instead of order imbalance, this featuers measures the trade imbalance.
  - Past Price Change: If we believe that there are momentum/reversion effects in the price movement (e.g. the price change is autocorrelated), then the past price change would be a useful indicator.
- Sentiment Data
  - Sentiment Score: The relative frequency of occurrence of positive tone in public opinion or news over a range of time which should reflect the level of public optimism about cryptocurrency.

5 Data Preparation

5.1 Market Signal

We collected market data mostly through the Binance Futures API and Yahoo Finance API. Since there are limits on the number of JSON objects returned over a single API call, we iterate through 30 day period day by day so that we could collect all available data and concatenate them into one large dataframe. The formats of the data from Binance Futures and yahoo finance are quite different, so we applied timestamp conversion and reindexing to ensure that our final dataframe is indexed by timestamp in chronological order.

After we successfully converts the data format and put them together, we started to apply transformations to the features to facilitate the model construction. The first transformation we applied is the log transformation. Since many features are ratio data, their distributions are highly skewed, which makes our model less robust to outliers. Log transformation effectively converts the skewed distributions to bell-shaped distributions and reduces the magnitude of the outliers.

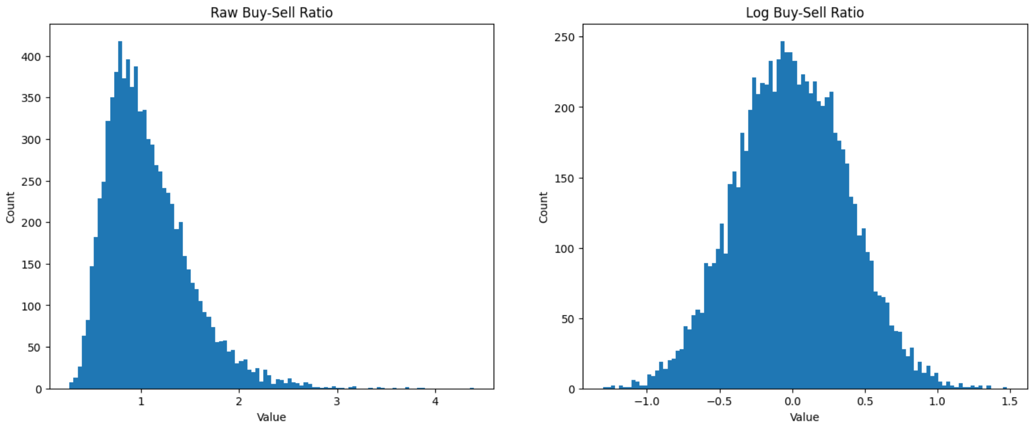

Figure 1: Buy-sell ratio distribution before and after log transformation

Then, we convert the open interest value and price data to their percentage change compared with the previous value. This conversion is very intuitive since for us the absolute value of the cryptocurrency price and the amount of the open interest do not matter at all. What provides us information is the change in those values, which reflects the market sentiment. Also, if in the future the open interest or price value move out of their range in the dataset, our model would absolutely be broken by the price movement, which is what we do not want to see.

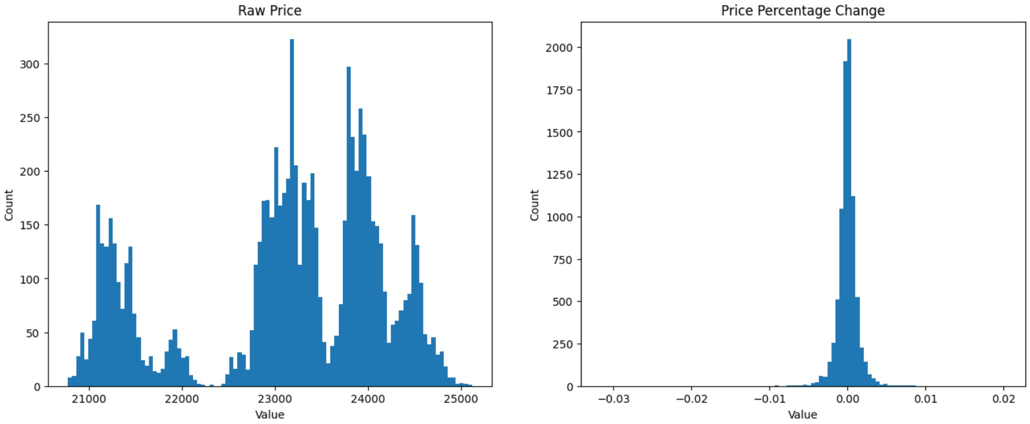

Figure 2: Price distribution before and after percentage change transformation

Lastly, some correlation analysis and binned heatmap visualization is conducted to check whether there exists significant relationship between each feature and the target variable (5-min return). The correlation analysis indicates that there are some degree of correlation between some of the features, highlighting the need of regularization in our model.

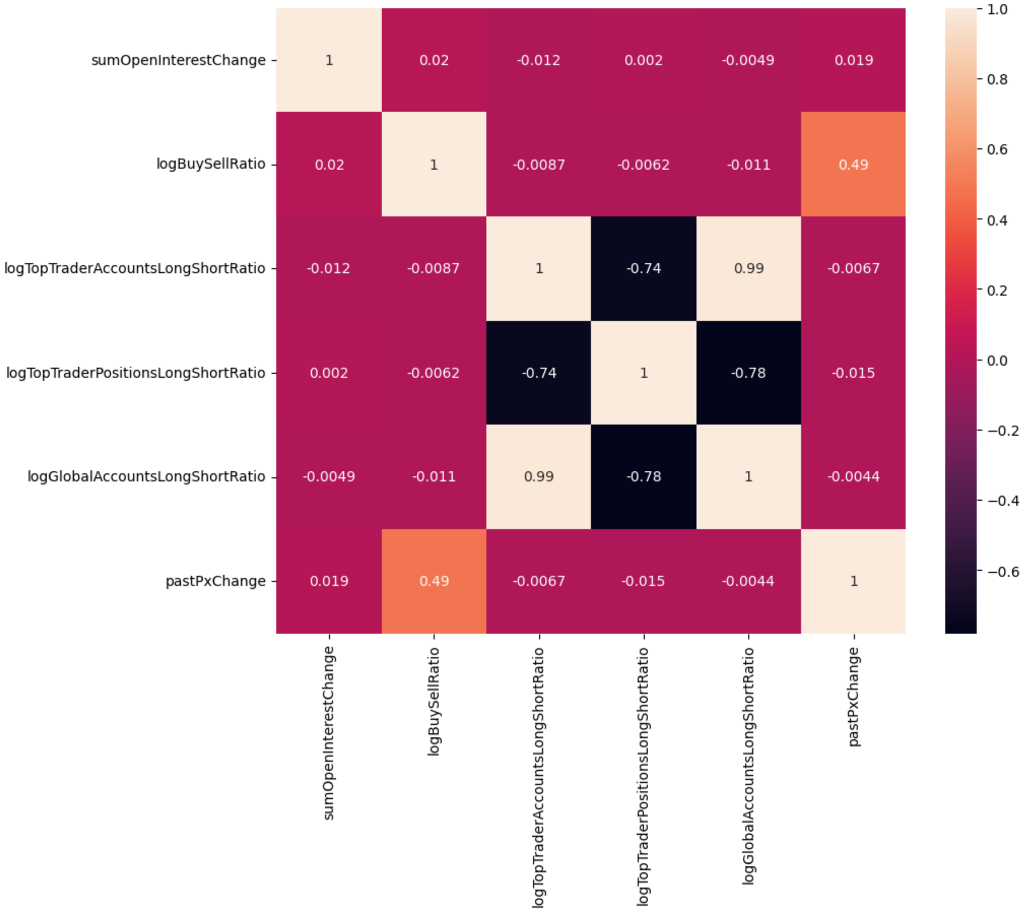

Figure 3: The correlation matrix

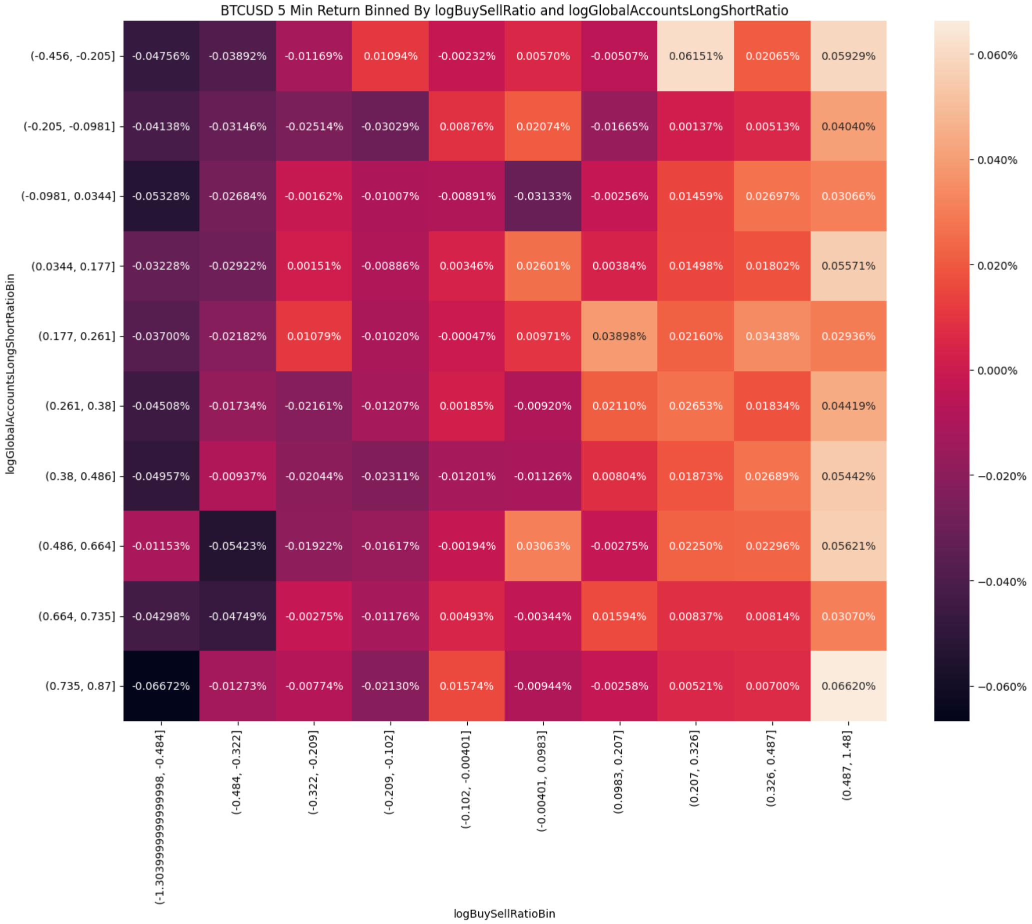

Figure 4: The heatmap showing the relationship between buy-sell ratio, long-short ratio, and 5-minutes return

5.2 Sentiment Score

We obtained a collection of public tweets under 4 major cryptocurrencies (BTC, ETH, ADA, UNI) from August 16 to August 23. For each tweet post, it contains the time of post, context annotations to indicate the possible field of tweets, and the text content of the Tweet, which is less than 140 characters. The amount of data is limited by our Twitter API access level to a maximum of 400 Tweet posts. To address the problem of a small data set, we will aggregate all tweets together to calculate an average overall sentiment, rather than separate sentiments by category. To remove the influence of ad bots, we observe some representative advertisement tweets and summarize a set of words that appear frequently in ads as filters in our query request, such as ”give away”, ”giving”, or ”free BTC”. We also filter out Tweets that contain URLs or retweets.

As shown in Figure 3, another data source we used is the news from google news under the search keyword ”crypto”. Unlike tweets, news headlines are a good summary of the content of the article, so we only use news headlines and publication times to put into our database. Although more data can be obtained from google news, 400 pieces of news from August 16 to August 23 were collected to balance with the data volume of Twitter. Next, we used finBERT, a domain-specific pre-trained NLP model, to classify each piece of text as one of: ”Positive”, ”Neutral”, and ”Negative”.

To convert the tone label into numerical data that best represent the sentiment within a certain period, the sentiment score was calculated by the relative frequency of occurrence of ”Positive” to reflect the level of optimism toward cryptocurrency.

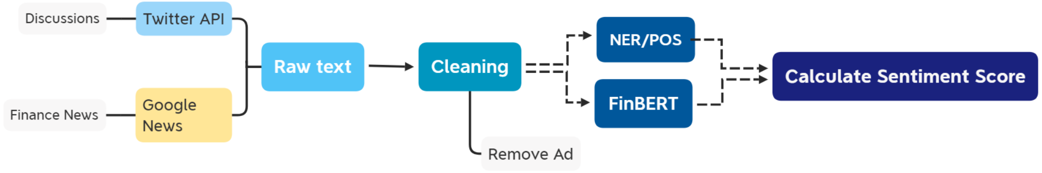

Figure 5: Sentiment score

6 Modeling

6.1 Linear Regression (With Regularization)

The first model we tried is the linear regression model with regularization. One may question why the linear regression model is chosen over other fancy models such as neural networks or reinforcement learning agents, and there are multiple reasons to justify this choice. First of all, we do not have a sufficient amount of data: the dataset we have consists of less than 10000 lines, which is suitable for linear regression but definitely inappropriate for neural networks. A complex model will be deemed to overfit the dataset and produce bad predictions. Secondly, when it comes to trading, rolling validation is extremely important: for each date d, one needs to fit the model on the train set between date $d-n$ and $d-1$ and validate the model performance on date d's data. This approach requires the model to run multiple times in a short period of time. If we use models that are too complex, rolling validation combined with hyperparameter tuning would take days to complete, which is not a realistic choice. Moreover, linear regression is compatible with one extremely feature selection method: LASSO. In a LASSO regression, redundant and unuseful data would be automatically discarded. Compared with forward or backward selection, LASSO is extremely efficient since it finishes the feature selection in one run.

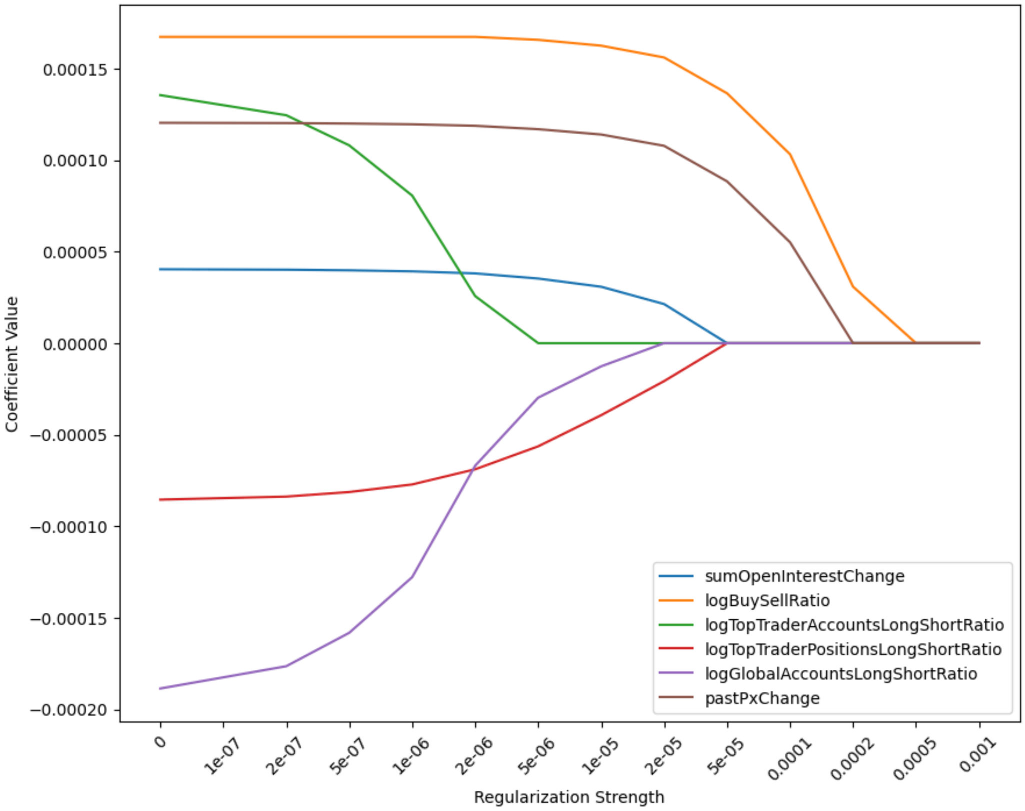

Figure 6: Relationship of parameter coefficient and LASSO coefficient magnitude

Therefore our model mainly relies on linear regression. We built a powerful pipeline to help us customize and speed up the hyperparameter search and evaluation process which will be introduced in the section 7.

That being said, we still made some attempts and explorations on relatively more complex models, which are introduced below.

6.2 Decision Tree & XGBoost

The relationship between market signals and the price might be highly non-linear, as different forces might act against each other. In order to capture this character and at the same time also preserve the interpretability of the model, we used simple decision tree model and XGBoost to predict the directional change of the price of cryptocurrency. XGBoost is a type of GBDT(Gradient Boosting Decision Tree), basically, the model predicts the real value by summing the predicted results of all "sub-trees", and then the next sub-tree fits the residual of the error function to the real value, and the process goes on until the predicted value equals to the real value. In our model, to use decision tree, we first convert the price change to directional binary value, and then applied the model using python library sklearn and xgboost.

7 Evaluation

One of the most important features of our project is the project's customizable hyperparameter tuning pipeline. According to the definitions in the config file, every possible set of hyperparameters is searched and the corresponding model is rolling validated. After we determine the hyperparameter combination for each trial, we will fit the data on the training data over day 1 to day 15, and validate on the day 16's data. Then we shift the training and validation data by one day and repeat until we reach the end of the dataset. The out-of-sample R^2 for each trial on each validation date are recorded.

After we get all out-of-sample R^2 metrics, we sort the result and get the hyperparameter combinations that yield the highest out-of-sample R^2, train the model with that set of hyperparameter again and get the corresponding PnL assuming \$10000 notional per trade.

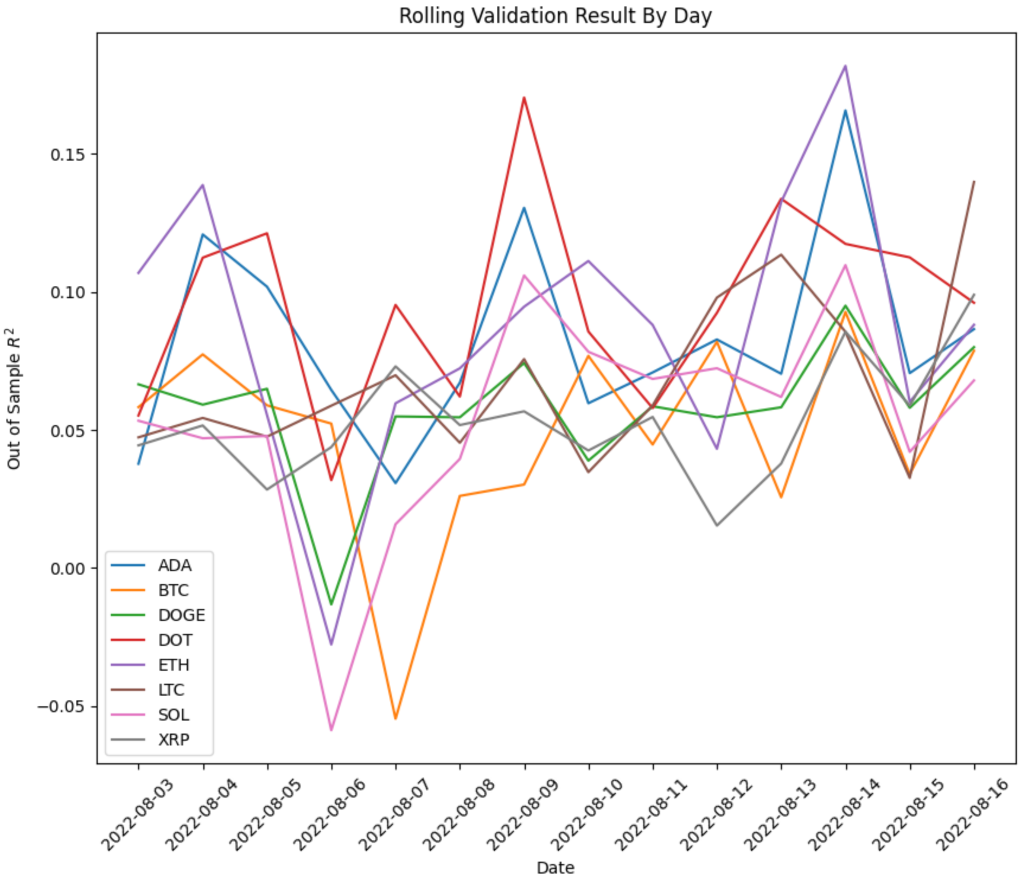

Figure 7: Best out-of-sample R^2 performance for each coin

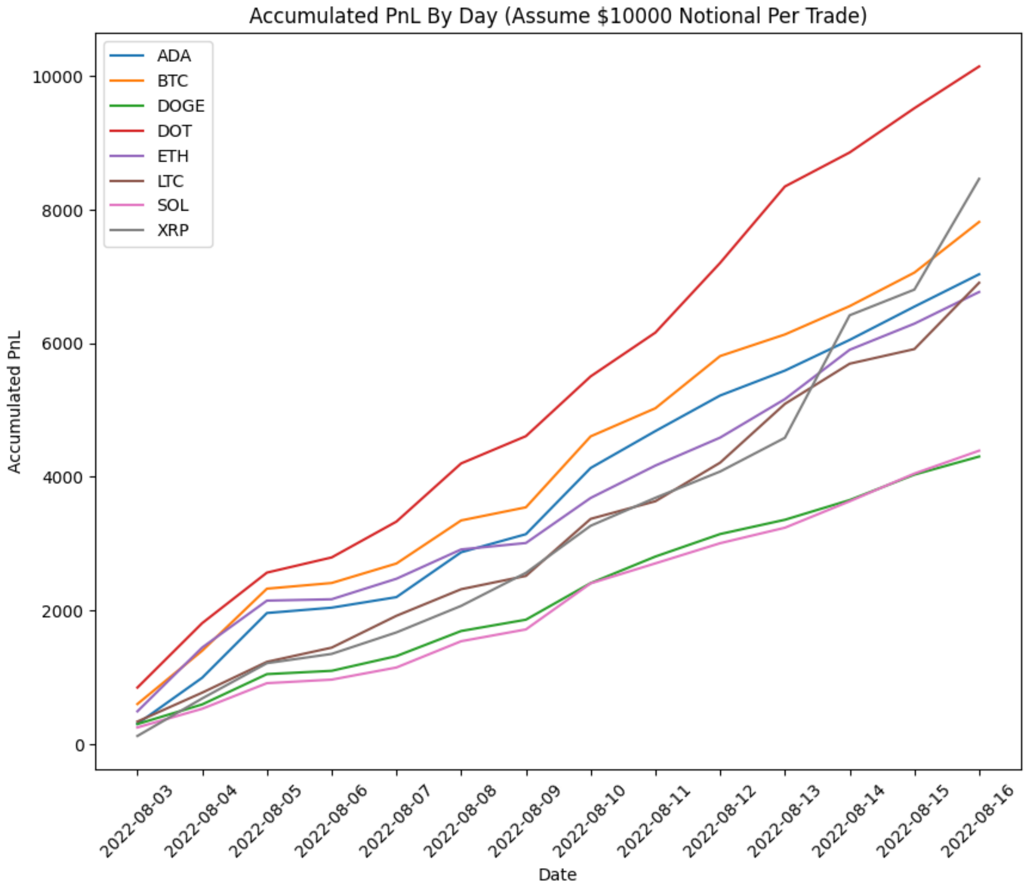

Figure 8: Best PnL performance for each coin

Our model achieves around 0.05 average out-of-sample R^2 for each coin during the rolling validation stage, which is satisfactory based on the noisy market conditions. Also, our model is steadily making money without considering the commission fee.

In terms of the decision tree model and XGBoost model, we only predict the direction of the price movement given the limited computing power. The accuracy of the Decision Tree model is about 60%, and the accuracy of the XGBoost model is 68%. It seems to be a very promising probability in the financial world. However, we notice that a naive model predicting price going up will have an accuracy of approximately 50%. Also, the market signals are always very noisy, which leads to questionable overfitting problems when training the model. In order to deploy the model, we also need to consider the TP/SL criteria instead of simple up/down direction. For future work, we may want to separate noise and train the model based on clearer signals.

8 Future Works

From the data preparation perspective, Unlike news headlines, tweets posts' sentiment can be ambiguous or objectively neutral, resulting in poor prediction results, making the accuracy of the sentiment score very low. To solve this problem, we need to train the finBERT model optimized for social media text to make better predictions for this kind of text. Also, on smaller time scales, whether Twitter posts are sufficiently correlated with price movements remains to be further verified, so more alternative data sources need to be added, such as the trading community on Telegram, Reddit, or the price of relative assets.

From the prediction model perspective, the current model is applied to all data and time without selection. For deployment, besides directional prediction, we need better models for timing selection to reduce noise and increase accuracy. SL/TP strategy is also required. Also, to process more data in real-world application scenarios, the overall architecture needs to be optimized for parallel computing, including the use of input data streams caching, concurrent NLP pipelines, and re-implementing of models using the MapReduce paradigm.

9 Resources

- https://developer.twitter.com/en/docs/twitter-api
- https://huggingface.co/yiyanghkust/finbert-tone
- https://github.com/Iceloof/GoogleNews
- https://www.binance.com/en/binance-api
- https://finance.yahoo.com
- https://doi.org/10.48550/arXiv.1908.10063
- https://arxiv.org/abs/2107.08721
- https://www.sciencedirect.com/science/article/pii/S2405918821000027