Predicting Most Profitable Altcoin Based on Bitcoin and Crypto Market Data

Maryam Anvaripoor

Abstract

In this analysis, we used cryptocurrency market data to predict which altcoin (from a set of six) would deliver the highest return over the next 12-hour interval. We tested simple models, ensemble approaches, and recurrent neural networks (RNNs).The Decision Tree Classifier achieved an average profit of 13% over two weeks. By combining models into an ensemble, we improved performance to an average profit of 20% in the same period. We also implemented RNNs, which produced an average profit of 6%. Our findings also showed that changes in Bitcoin's price, the number of consecutive positive days for Bitcoin, and Google Trends data are valuable predictors for identifying profitable altcoins.

Introduction

Machine learning is important for crypto market prediction because it can uncover hidden patterns in highly volatile price data that are difficult for humans to detect. By leveraging historical trends, technical indicators, and sentiment data, ML models can enhance decision-making and improve the timing of trades in the unpredictable crypto landscape.

In the first part of this analysis, we focused on collecting relevant features for our task. We gathered OHLC data for each symbol, Google search trends, market sentiment, gold prices, and the U.S. dollar index. We then derived additional financial and temporal features to enrich the dataset.

Next, we experimented with different classifiers to predict the most profitable altcoin over 12-hour intervals. This led to the development of two sets of models. The first set, which included decision trees and logistic regression, delivered high potential profits but also carried significant risk. By combining these models in an ensemble, we were able to reduce risk while improving profitability. The best high-risk model achieved an average profit of 20% over two weeks, with a maximum drawdown of -9%.

To balance risk and return, we also tested recurrent neural networks (RNNs). This model had an average profit of 6%. However to lower the maximum loss, further experimentation is required.

Finally, we examined feature importance to validate whether the factors behaved as expected. SHAP analysis revealed that features such as Bitcoin's closing price, the number of consecutive positive days for Bitcoin, Google Trends data for altcoins, and technical indicators like moving averages and RSI were highly influential in driving model predictions.

Data

To build the dataset, we collected 12-hour OHLC cryptocurrency data from Binance for each symbol: Bitcoin, Bonk, Hamster Kombat Coin, Ripple, ADA, and Floki Inu. We obtained daily market sentiment scores from *alternative.me*, weekly search interest for each symbol from Google Trends, and daily dollar index and gold prices from Yahoo Finance. The dataset covers the period from 2021 to August 2025, resulting in approximately 3,000 data points. While our aim was to obtain precise 12-hour interval data, rate limits prevented us from achieving full granularity.

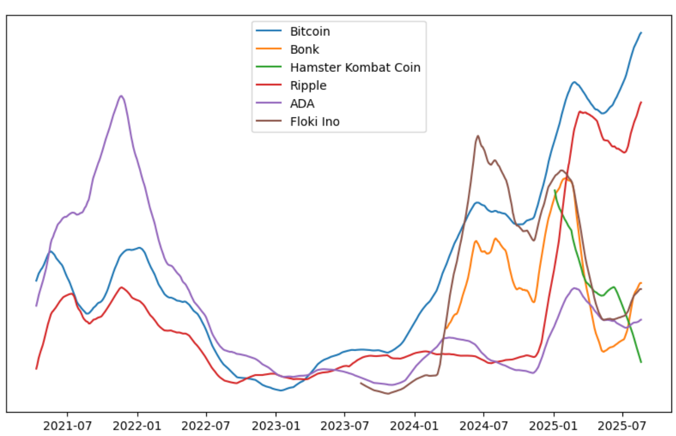
Features and preprocessing

The first challenge was handling data frames with different time intervals, ranging from weekly to 12 hours. We resampled all data to 12-hour intervals using forward fill. Since some altcoins were not available as early as 2021, the dataset contained many missing values. To prevent these from influencing the labels, we replaced them with 0.

For each symbol, we created a new feature called *percent change*. The final dataset included the percent change for each cryptocurrency or market symbol, along with Google Trends data and sentiment indicators. Labels were generated using the following rule: the label corresponds to the cryptocurrency with the highest positive percent change. If Bitcoin was the most profitable asset, or if all symbols were negative during a given interval, the label was set to “don’t.”

We also engineered additional features. For example, *Days Positive* counted the number of consecutive days in which Bitcoin closed with a positive return. Financial indicators such as moving averages and the Relative Strength Index (RSI) were added, as well as temporal features including hour, day of the week, day of the month, and whether the interval fell on a weekend. Labels were encoded with a label encoder. To ensure that predictions were based on past data, we shifted all features by -1 (except temporal features), so that each sample used the previous 12 hours to predict the next 12 hours.

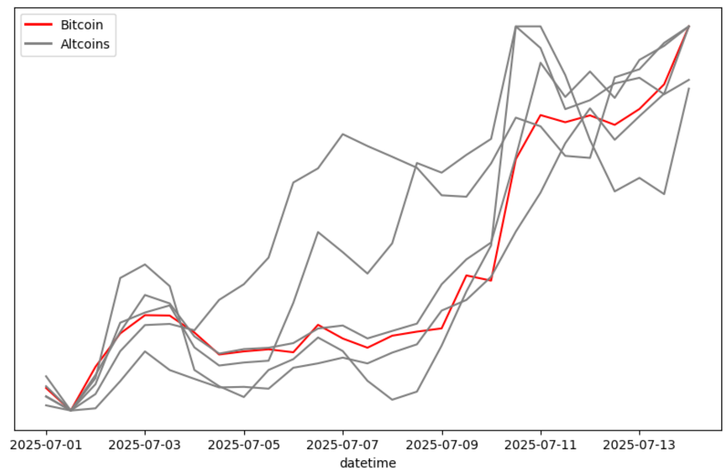
Further preprocessing included imputing missing values, feature scaling, and applying *SelectKBest* for feature selection.



Trend of Altcoins and Bitcoin

Finally, we analyzed patterns between changes in Bitcoin and other altcoins. While the general trends were highly correlated, their magnitudes differed. Bitcoin tended to move more steadily, whereas altcoins such as Floki Inu and Bonk exhibited sharper, more exaggerated spikes. This suggests that altcoins often act as “leveraged bets” on Bitcoin — rising more aggressively when Bitcoin increases, and crashing harder when Bitcoin declines.

To examine the lagged correlation between Bitcoin and altcoins, we analyzed shorter time intervals. While changes typically occurred simultaneously, there were instances where altcoins reacted to Bitcoin with a noticeable delay.



A Closer View of Trend of Altcoins and Bitcoin

Methods

Decision Tree Classifier:

A decision tree classifier is a supervised learning algorithm used for classification tasks. It works by splitting the data into branches based on feature values, creating a tree-like structure where each internal node represents a decision rule and each leaf node represents a class label. The model aims to partition the data so that observations within each final group are as homogeneous as possible with respect to the target class.

Ensembling Models:

Ensembling models combine predictions from multiple individual models to improve overall performance and robustness. A voting classifier, for example, aggregates the outputs of several base classifiers by either taking the majority vote (hard voting) or averaging predicted probabilities (soft voting). This approach reduces the risk of relying on a single model’s weaknesses, often leading to better generalization on unseen data.

Recurrent Neural Networks (RNNs):

Recurrent neural networks are a type of deep learning model designed to handle sequential data, such as time series or natural language. Unlike traditional neural networks, RNNs include connections that form cycles, allowing information to persist across steps in the sequence. This makes them well-suited for tasks where context and order matter. For this task we used a combination of cross entropy (for classification) and expected profit (to maximize the profit) as loss function.

$$\text{Loss} = \alpha \times CrossEntropy(x) + (1 - \alpha) \times Mean(Sum(x \times Profits))$$

Results and Discussion

The details of the top three models are summarized in the table below. For evaluation, we considered both accuracy and profit. Profit was calculated under two assumptions: (1) the investment amount was fixed for every trade, and (2) transaction fees were set at 1%. Under these assumptions, overall profit was obtained by summing the profit/loss (%) and fees (%) across all trades.

For the first three models, missing values were imputed using the mean of each feature, followed by standard scaling. Feature selection was performed using *SelectKBest*.

To compare model performance, we trained them up to a specific threshold and then evaluated them over the subsequent 14 days, which included a total of 27 trades (since the dataset used 12-hour intervals and the first interval was excluded). This process was repeated 16 times over different periods to ensure robustness.

Performance of models

rank	Accuracy	Classifier	Parameter(s) Name	Parameter(s) Value	Profit	Min Profit/ Max Loss	Max Profit
1	0.2881	Voting Classifier of models 2 and 3	Voting	hard	20	-6	73
2	0.2857	Decision Tree	Max depth	5	14	-9	48
3	0.2405	Logistic Regression	-	-	8	-32	75
4	0.2000	RNN	Hidden size	32	6	-14	37
			Number of layers	6			

Models 1, 2, and 3 are considered high-risk but have the potential for higher profits. Model 3 on its own is not usable due to its large losses; however, when combined with Model 2, it produced a more balanced outcome with higher profits and reduced risk.

Model 4 was designed to lower the risk, which wasn't achieved so far. We experimented with different number of layers, hidden size, dropout and number of epochs for training. Further suggestions for future work are provided in conclusion.

Conclusion

In conclusion, our analysis demonstrated that different classifiers can be leveraged to predict the most profitable altcoin within 12-hour intervals, each offering distinct trade-offs between risk and return. While models such as the decision tree classifier and logistic regression achieved higher profit levels, they also carried a greater risk of loss, -8% and -32% respectively. By applying ensembling techniques, we were able to balance these outcomes, lowering the risk while still improving profitability. The best high-risk model achieved an average profit of 20% in two weeks, with a maximum loss of -9%.

We also created a RNN, which resulted in 6% profit on average. Finally, our feature importance analysis confirmed that key drivers such as Bitcoin's closing price, the number of consecutive positive Bitcoin days, interests over time from Google Trends for altcoins, and technical indicators like moving averages and RSI, all contributed significantly to classification performance. These findings emphasize the value of combining multiple approaches and feature sets to improve both accuracy and reliability in crypto market prediction.

We expect better results if all data were collected in 12-hour intervals, as this would improve data quality. Exploring even smaller intervals might further enhance model performance. For the RNN in this analysis, we only experimented with the number of layers and hidden states, but changing the architecture for this specific problem, could potentially yield a more effective model. Another avenue for improvement is selecting a different set of altcoins for prediction, which may significantly influence outcomes.