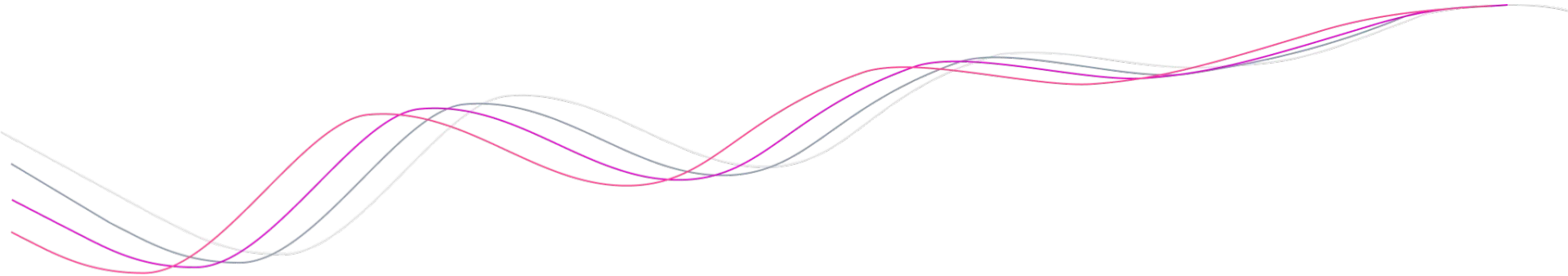
Predictive Model for Cryptocurrency Prices

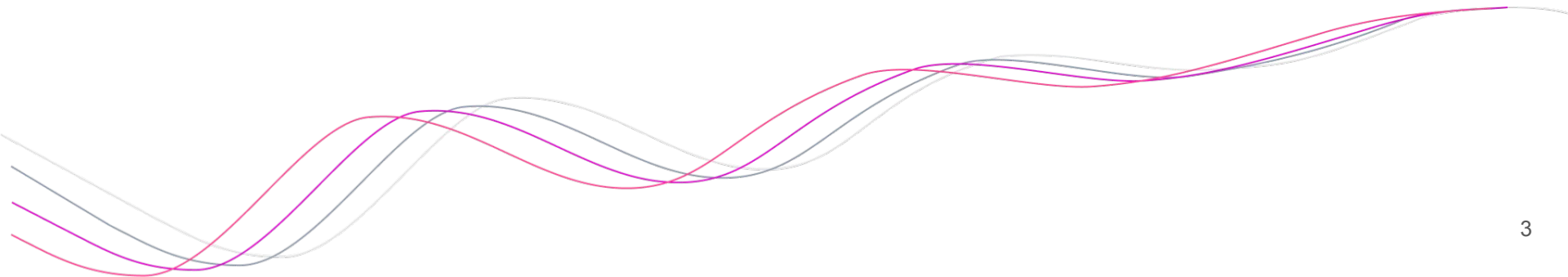


Anamaria Loznianu

Overview

1. Data Collection
    a. OHLCV Data
    b. Fear & Greed Index
    c. Coin Fundamentals and Market Metrics
    d. Google Trends
    e. Macro Economic Factors
2. Data Cleaning
3. Features
4. Model Development

Data Collection

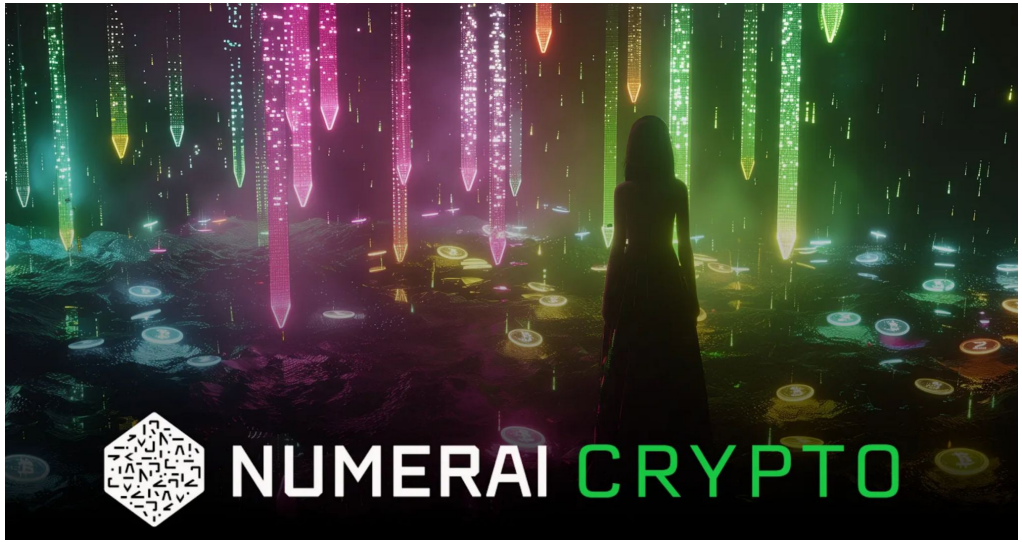


Data Collection

Target: Identifying Cryptocurrencies for Data Collection

The cryptocurrency datasets were sourced from Numerai.
Files:

- crypto/v1.0/historical_meta_models.csv
- crypto/v1.0/historical_meta_models.parquet
- crypto/v1.0/live_universe.parquet
- crypto/v1.0/meta_model.csv
- crypto/v1.0/meta_model.parquet
- crypto/v1.0/train_targets.parquet



Symbol Identification:
train_targets.parquet was used to identify the unique symbols for training the machine learning models.

The dataset provided **1439** unique cryptocurrency symbols.

ℹ️ **'RNDR' is being migrated and rebranded to RENDER**
ℹ️ **ONIT has been rebranded to LWA**
These 2 coins could be updated in the dataset.

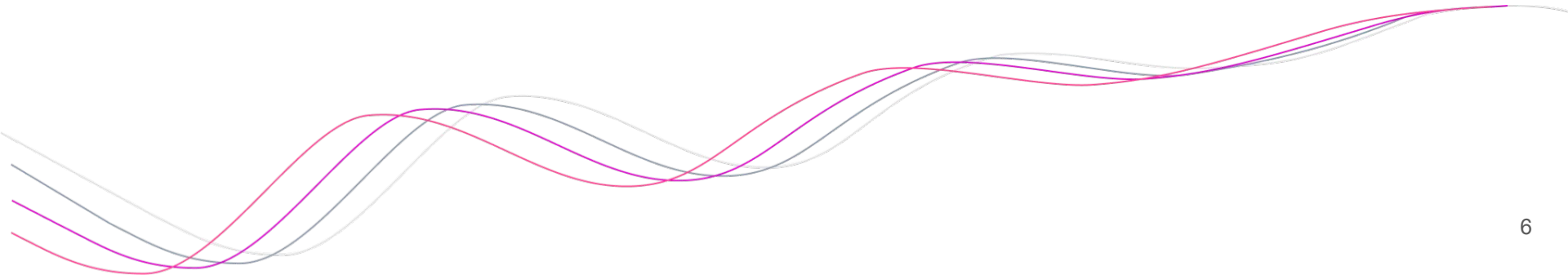
Data Collection

For all symbols, the following data was fetched, cleaned, and included in the training and validation datasets.

Data was collected from multiple sources covering various areas. The principal data sources include:

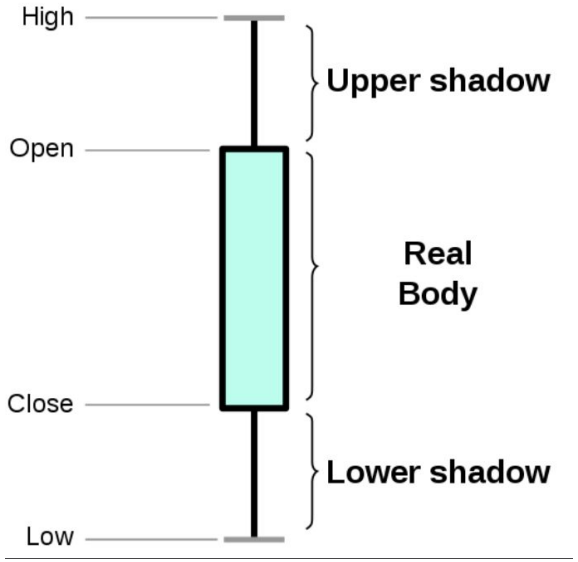
- OHLCV Data: Binance and Yahoo Finance
- Market Sentiment Data: Fear and Greed Index from Alternative.me
- Coin Fundamentals and Market Metrics: CoinMarketCap
- Search Sentiment Data: Google Trends
- Macro Economic Factors: World Bank, FRED (Federal Reserve Economic Data) and Trading Economics

OHLCV Data



Data Collection

Target: OHLCV Data and Its Impact on Predicting Crypto Prices



OHLCV Data Sources:

- Binance(via ccxt) and Yahoo Finance(via yfinance):
  - For the 1439 symbols used for training, OHLCV daily data was downloaded from Binance.
  - For coins not available on Binance, Yahoo Finance was used as the data source.

Data Overview:

- Open: The price at which the cryptocurrency started trading at the beginning of the day.
- High: The highest price reached during the day.
- Low: The lowest price reached during the day.
- Close: The price at which the cryptocurrency ended trading at the end of the day.
- Volume: The total amount of the cryptocurrency traded during the day.

Dataset:

- Size: ~1.4 million rows of data.
- Quality: Cleaned OHLCV data, ensuring no missing parameters.
- Timeframe: Data is distributed from 2020-06-01 to the current date.

Importance in Predicting Crypto Prices:

- Price Trends: OHLCV data helps in identifying price trends and patterns. For instance, a consistent rise in the 'close' price over several days could indicate an upward trend.
- Volatility Analysis: The 'high' and 'low' prices help in understanding the volatility of the cryptocurrency, which is crucial for risk assessment.
- Market Sentiment: The 'volume' data indicates the level of interest and trading activity in the market. High volume often accompanies significant price movements.
- Technical Indicators: OHLCV data forms the basis for many technical indicators like moving averages, RSI, MACD, which are widely used in trading strategies.

Data Collection

Target: OHLCV Data and Its Impact on Predicting Crypto Prices

Correlation Between Trading Volume and Close Prices:

There is a **noticeable correlation** between the sum of close prices and trading volume over time. Both metrics tend to rise and fall together, particularly visible in the peaks around early 2021 and the significant drop around mid-2022. This suggests that higher trading volumes are often associated with higher closing prices, indicating strong market activity during those periods.

There is a notable decline in both the close prices and trading volume towards mid-2024. This recent drop suggests that the market may be experiencing lower interest or confidence, potentially due to macroeconomic factors, regulatory changes, or other external pressures affecting the cryptocurrency market.

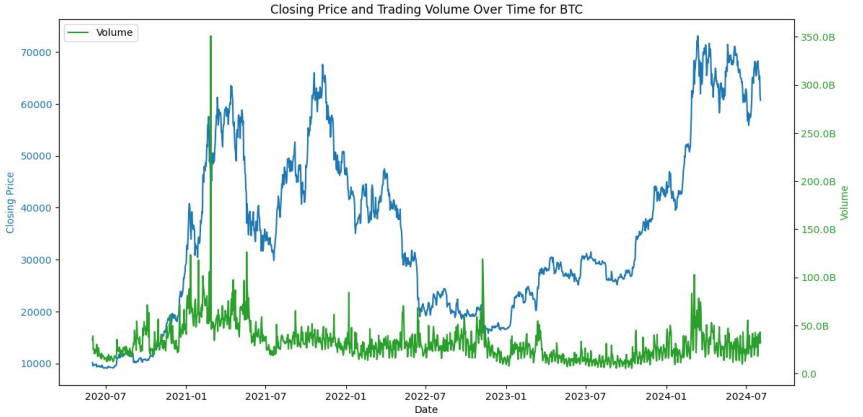
Data Collection

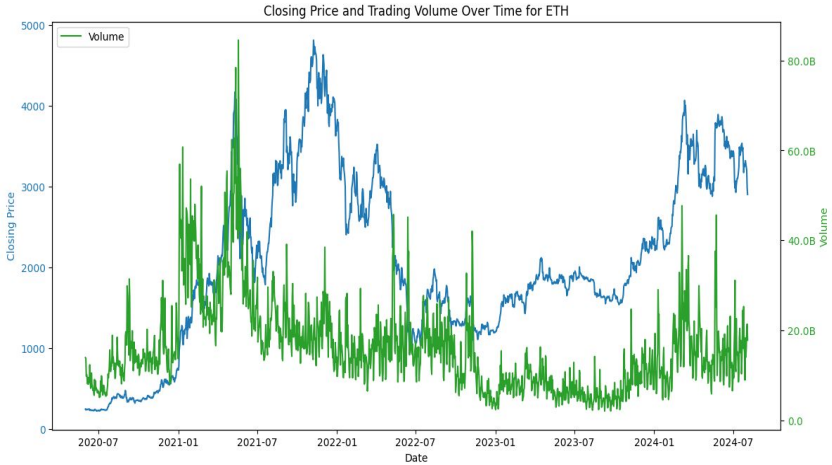
Target: OHLCV Data Analysis for Major Cryptocurrencies

Key Insights:

- **Bitcoin (BTC):**
  - The **closing price** of Bitcoin shows significant peaks around late 2020 and early 2021, indicating major market rallies.
  - **High trading volumes** coincide with price surges, suggesting increased market activity during bullish trends.
  - **Long-Term Trends**: Despite short-term volatility, Bitcoin's long-term trend has been generally upward, reflecting its increasing adoption and recognition as a store of value.
  - **Market Sentiment**: Price spikes often follow positive news events or market sentiment, such as institutional adoption or regulatory approvals.

- **Ethereum (ETH):**
  - Ethereum's price **trends mirror** those of Bitcoin, with notable peaks in early 2021 and a sustained upward trend in 2021.
  - **Volume** spikes often precede or coincide with price peaks, indicating trading momentum and market interest.
  - **Network Upgrades**: Significant price increases often align with major Ethereum network upgrades or improvements (e.g., Ethereum 2.0).
  - **DeFi and NFTs**: The rise of decentralized finance (DeFi) and non-fungible tokens (NFTs) on the Ethereum network has driven substantial interest and investment, reflected in the price and volume spikes.





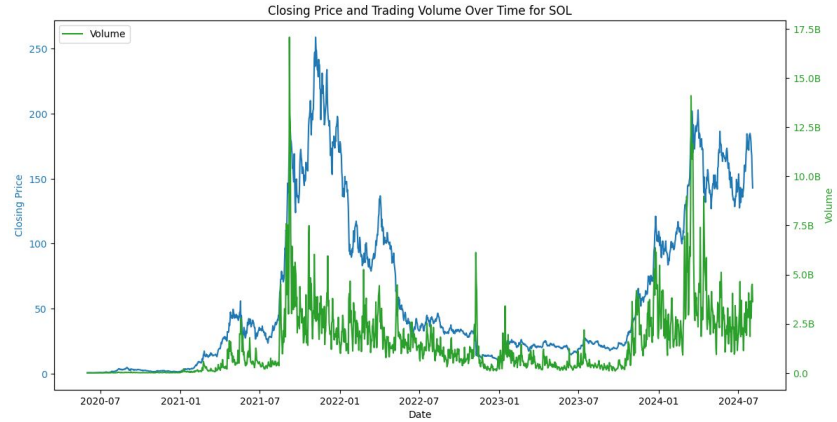
Data Collection

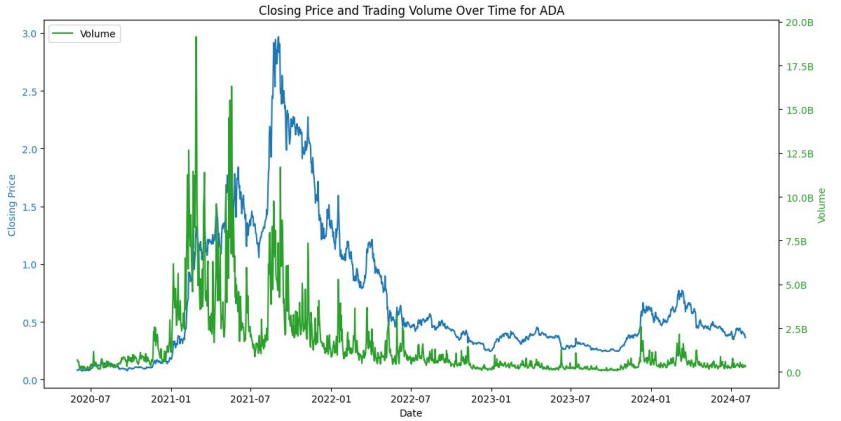
Target: OHLCV Data Analysis for Major Cryptocurrencies

Key Insights:

- **Solana (SOL):**
  - Solana experienced a rapid price increase in mid-2021, followed by periods of high volatility.
  - Trading **volumes** increased significantly during price surges, reflecting strong market participation.
  - **Ecosystem Growth**: Solana's growth was fueled by the expansion of its ecosystem, including decentralized applications (dApps) and DeFi projects.
  - **Performance and Scalability**: Solana's focus on high performance and scalability attracted developers and investors, contributing to its price rise.

- **Cardano (ADA):**
  - Cardano's price showed a significant upward trend in early 2021, **similar** to other **major** cryptocurrencies.
  - **High trading volumes** during price increases indicate robust market activity and interest.
  - **Smart Contract Launch**: Anticipation and subsequent launch of Cardano's smart contract capability (Alonzo upgrade) led to price surges.
  - **Research-Driven Development**: Cardano's unique approach of peer-reviewed research and development added to investor confidence and interest.





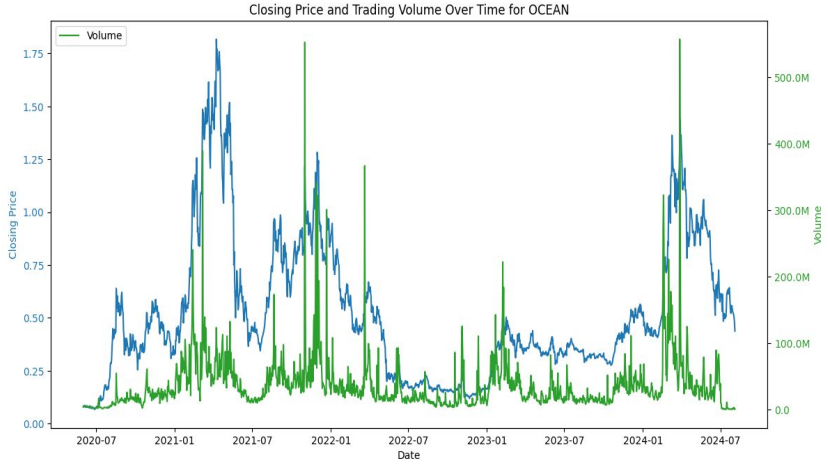
Data Collection

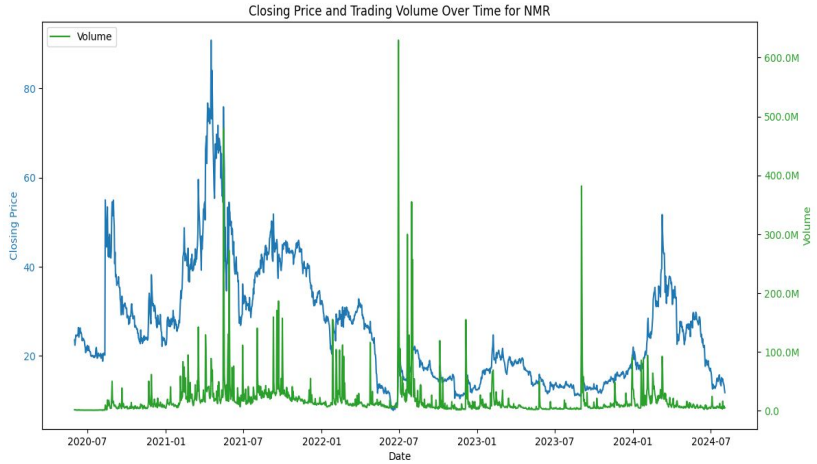
Target: OHLCV Data Analysis for Selected Projects

Key Insights:

- **Ocean Protocol (OCEAN):**
  - Ocean Protocol's price has shown high volatility, with several distinct peaks and troughs.
  - Trading volume patterns align with price changes, reflecting market reactions to price movements.
  - Data Economy: Ocean Protocol focuses on the decentralized data economy, and its price movements often correlate with developments and partnerships in this space.
  - DeFi Integration: Integrations with DeFi platforms and protocols have influenced its market performance and trading activity.

- **Numeraire (NMR):**
  - Numeraire's price trends reveal periodic peaks, with significant price increases observed in mid-2021.
  - Volume spikes are evident during these price peaks, suggesting increased market engagement.
  - Unique Token Model: Numeraire is used in the Numerai hedge fund, where data scientists stake NMR tokens on their models' performance, influencing its demand and price.
  - Predictive Modeling: The focus on AI and predictive modeling has driven interest and investment, particularly among data scientists and quants.





Data Collection

Target: Importance of OHLCV Data in Predicting Crypto Prices

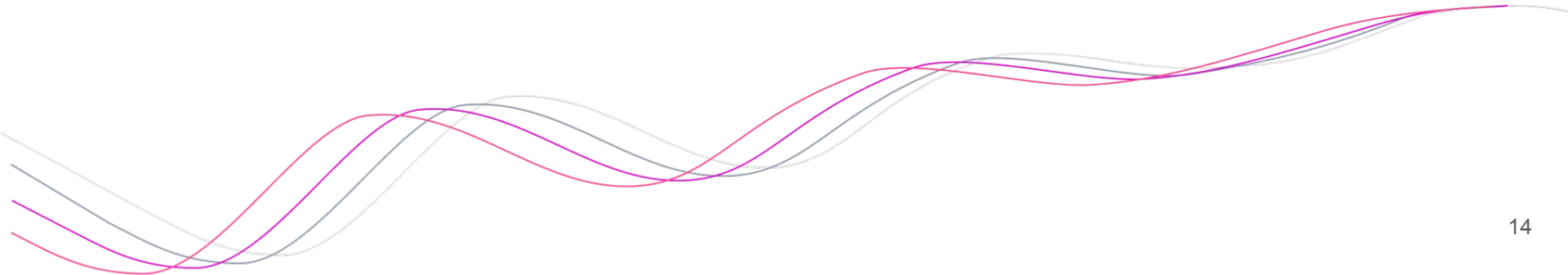
- **Historical Patterns**: OHLCV data helps in identifying historical price trends and patterns, providing a basis for forecasting future price movements.
- **Trend Analysis**: Enables the detection of uptrends, downtrends, and sideways trends over various time frames.
- **Risk Assessment:** Analyzing the high and low prices helps in understanding the volatility of the cryptocurrency, which is crucial for risk assessment.
- **Market Stability:** Provides insights into the stability of the market by examining the range between high and low prices over time.
- **Trading Activity:** The volume data indicates the level of trading activity and market interest, which can signal bullish or bearish sentiment.
- **Investor Behavior:** Helps in understanding investor behavior and market dynamics based on trading volumes and price changes.
- **Indicator Calculation:** Forms the basis for calculating many technical indicators like moving averages (MA), Relative Strength Index (RSI), and Moving Average Convergence Divergence (MACD).
- **Trading Strategies:** These indicators are essential tools in developing and implementing trading strategies.
- **Liquidity Measurement**: High trading volumes indicate liquidity, which is important for entering and exiting positions without significant price impact.
- **Volume Spikes**: Volume spikes often correlate with significant price movements, suggesting strong market interest and potential trend reversals.
- **Inter-asset Correlation**: Examines the correlation between different cryptocurrencies, aiding in portfolio diversification and risk management.
- **Market Movements**: Identifying correlated movements helps in predicting how one cryptocurrency might affect another.

Data Collection

Target: Impact on Prediction

- **Direction Prediction:**
  - Trend Forecasting: By analyzing OHLCV data, models can predict the likely direction of future price movements, helping traders to make buy or sell decisions.
  - Pattern Recognition: Recognizes bullish and bearish patterns, such as head and shoulders or bullish engulfing, to forecast future price directions.
- **Price Prediction:**
  - Future Prices: Makes informed predictions about future prices based on historical OHLCV data and technical analysis.
  - Market Timing: Assists in determining the best times to enter or exit positions.
- **Risk Management:**
  - Volatility Insights: Provides insights into the potential risk by analyzing price fluctuations and trading volumes.
  - Position Sizing: Helps in determining appropriate position sizes and setting stop-loss levels based on historical volatility and support/resistance levels.

Fear & Greed Index



Data Collection

Target: Fear & Greed Index Analysis and Its Impact on Predicting Crypto Prices

- **The Fear & Greed Index**, downloaded from Alternative.me, measures the **market sentiment** of cryptocurrencies.

- The index ranges from 0 (Extreme Fear) to 100 (Extreme Greed), providing insights into the emotional state of the market.

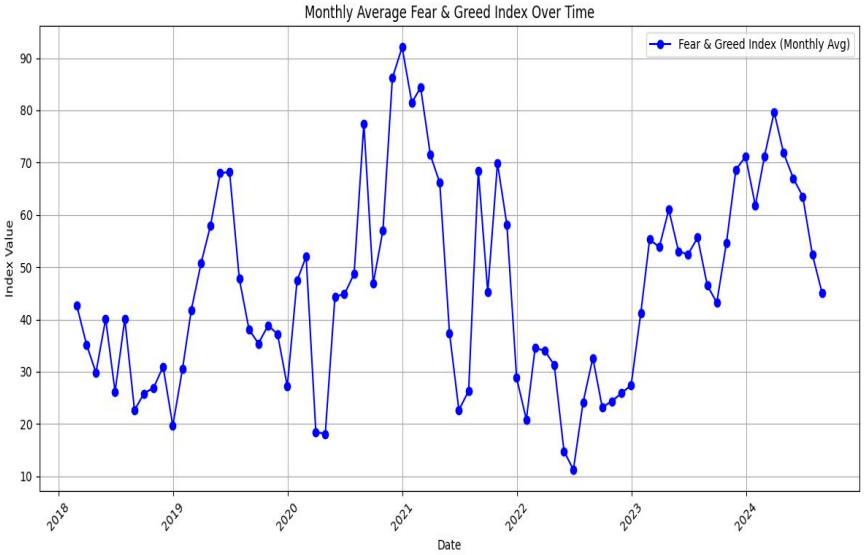
Impact and Importance:

- **Market Sentiment Indicator**: The Fear & Greed Index reflects the current sentiment of the market, helping traders gauge the emotional state of investors.
- **Contrarian Indicator**: Often used as a contrarian indicator, suggesting buying opportunities when the market is fearful and selling opportunities when the market is greedy.
- **Volatility Prediction**: High levels of fear or greed can indicate potential market volatility, which is crucial for risk management and trading strategy adjustments.
- **Timing Market Entries/Exits**: Helps in timing market entries and exits by understanding the prevailing sentiment, thereby improving the accuracy of price predictions.

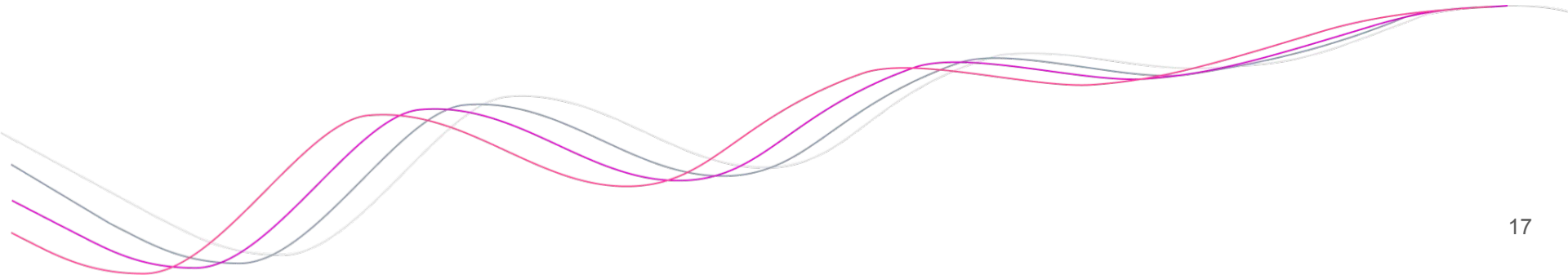
Data Collection

Target: Fear & Greed Index Analysis and Its Impact on Predicting Crypto Prices

- **Sentiment Fluctuations:** The index shows significant fluctuations over time, corresponding to market highs and lows.
- **Correlation with Price Movements:** Periods of extreme fear often precede market recoveries, while periods of extreme greed can precede market corrections.
- **Trend Analysis:** Long-term trends in the Fear & Greed Index can provide insights into the overall market cycle, helping in long-term investment decisions.
- **Behavioral Economics:** Reflects collective investor psychology, showing how fear and greed drive market behavior.
- **Predictive Power:** High predictive power for short-term market movements, as extreme sentiment often leads to corrective actions in the market.
- **Investment Strategy:** Incorporating the Fear & Greed Index into investment strategies can enhance decision-making processes by aligning with market sentiment trends.
- **Risk Management:** Assists in assessing risk levels, where high fear suggests potential low-risk buy opportunities and high greed indicates potential high-risk conditions.



Coin Fundamentals and Market Metrics



Data Collection

Target: Coin Fundamentals and Market Metrics

From the 1437 symbols used for training, **1435 symbols are available on CoinMarketCap(***ONIT* and *RNDR* are missing**). For all of them, the following data was downloaded via the info and the quotes endpoints:**

- Symbol
- Name
- Total_supply
- Circulating_supply
- Market_cap
- Infinite_supply
- Is_open_source
- Source_code
- Is_active

Although the circulating supply and market cap are more meaningful when tracked over time, we will evaluate their impact on the model's performance when these features are introduced or excluded.

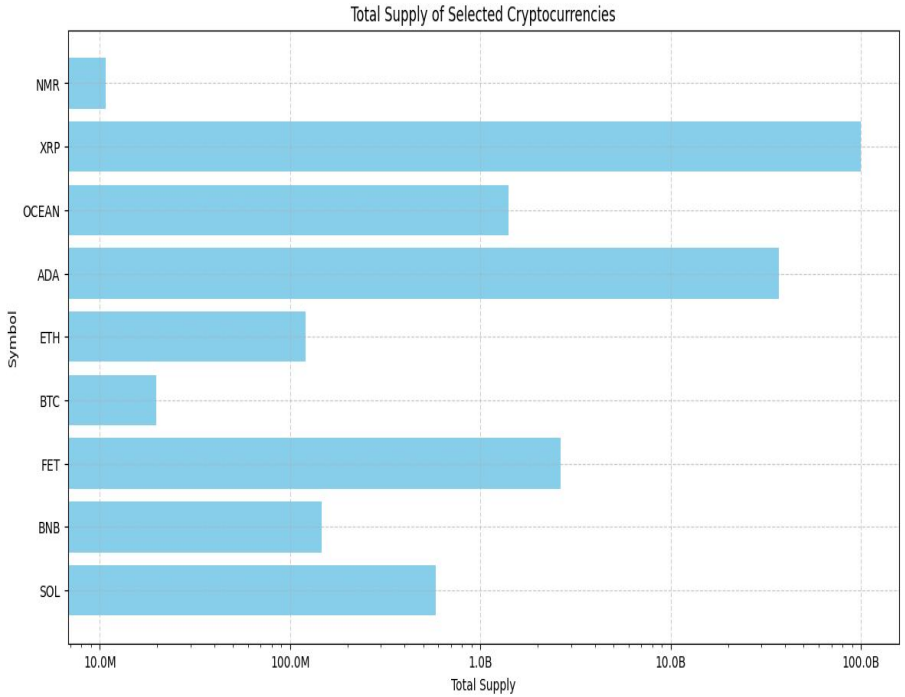
Data Collection

Target: Coin Fundamentals and Market Metrics

Included metrics for understanding the fundamental aspects of a cryptocurrency like: circulating supply, total supply, market capitalization, and infinite supply status.

The data was downloaded from CoinMarketCap.

- **Total Supply:** The total number of coins that will ever be available. This includes coins that are currently circulating, as well as coins that are yet to be released. Understanding total supply is important for grasping the long-term supply constraints and potential future scarcity of the coin. For example, Bitcoin has a fixed total supply of 21 million coins, which contributes to its scarcity and potential value appreciation.
- **Circulating Supply**: The total amount of a cryptocurrency that is currently available in the market and circulating among the public.
- **Market Cap:** The total market value of the circulating supply. It provides a sense of the coin's scale and its relative size in the market.



Visual created for a sample of symbols:
BTC, ETH, SOL, ADA, BNB, XRP, OCEAN, NMR, FET

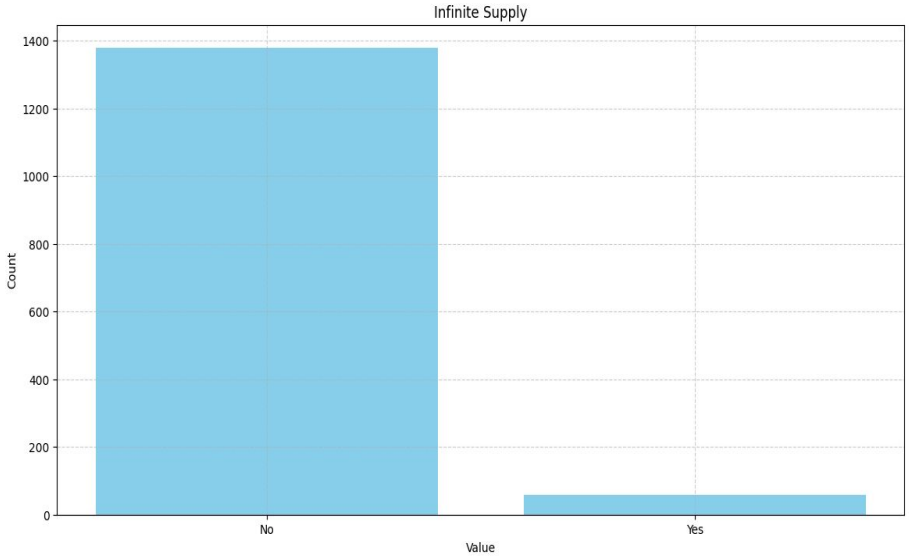
Data Collection

Target: Coin Fundamentals and Market Metrics

Included metrics for understanding the fundamental aspects of a cryptocurrency like: circulating supply, total supply, market capitalization, and infinite supply status.

The data was downloaded from CoinMarketCap.

- **Infinite Supply:** Whether the coin has an infinite supply. Coins with an infinite supply do not have a cap on the number of coins that can be created. This characteristic might lead to inflationary pressures, potentially decreasing the value of the coin over time as more coins are minted. Investors should be cautious with such coins, as their long-term value retention can be uncertain.
- **Is Active:** Indicates if the coin is still active. An active coin is one that continues to be traded and developed. Inactive coins, on the other hand, might have been abandoned by developers or fallen out of favor with the market, making them potentially less useful for prediction and investment. The activity status can impact the coin's liquidity, development updates, and community engagement.



Data Collection

Target: Coin Fundamentals and Market Metrics

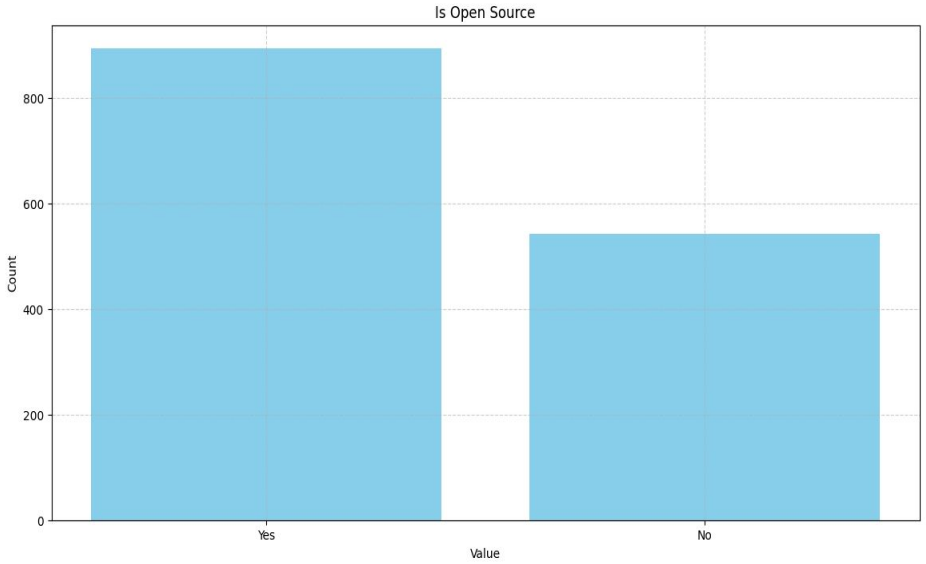
Extracted additional information about each coin: **name**(used for keywords identification of google trends data), **source code** information

The data was downloaded from CoinMarketCap.

Name & Details: While the name itself might not be directly useful for prediction, it can serve as valuable keywords for retrieving data from Google Trends or other search engines. This data can help identify the search frequency and popularity of specific coins, providing insights into **market sentiment** and **interest levels**.

Source Code URL: This can be a proxy for the coin's transparency and developer activity.

- **Open Source**: The cryptocurrency project is open source, meaning its codebase is publicly accessible and can be reviewed, audited, and contributed to by the community.
- **Community Trust**: Projects that provide their source code are generally more transparent and may earn more trust from the community, as anyone can inspect the code for security vulnerabilities or verify the project's claims.



Google Trends

Data Collection

Target: Google Trends Data -Its Importance and Impact on Cryptocurrency Prices

Google Trends data provides valuable insights into the search interest and popularity of specific keywords over time. For cryptocurrencies, this data can be particularly useful for understanding market sentiment and predicting price movements. **Higher search interest** often correlates with **increased attention** and **trading activity**, which can influence a coin's price.

Manual Intervention for Accurate Data

The names of the coins were downloaded from CoinMarketCap. However, for some coins, manual intervention was necessary to ensure accurate data collection from Google Trends. This manual effort was required due to ambiguities in coin names. Some cryptocurrencies have common names or acronyms that could be confused with unrelated search terms. For example, **"Tarot"** could refer to the cryptocurrency or to playing cards, among other things.

By refining the keywords and ensuring they accurately represent the cryptocurrency in question, we can obtain more reliable data that better reflects **market interest.**

Data Collection

Target: Google Trends Data -Its Importance and Impact on Cryptocurrency Prices

Impact on Coin Prices

In a previous challenge, we discovered that Google Trends data often has a significant correlation with cryptocurrency prices. The analysis showed that for several projects, there is usually at least a **0.5 up to 0.87** correlation between search interest and price movements with an **average 1 week lag.**

This mid to strong correlation suggests that increased search interest often precedes or coincides with price changes, making Google Trends a valuable tool for predicting market behavior. **You can find the detailed analysis and findings in the report here: Google Trends - Crypto Prices Report.**

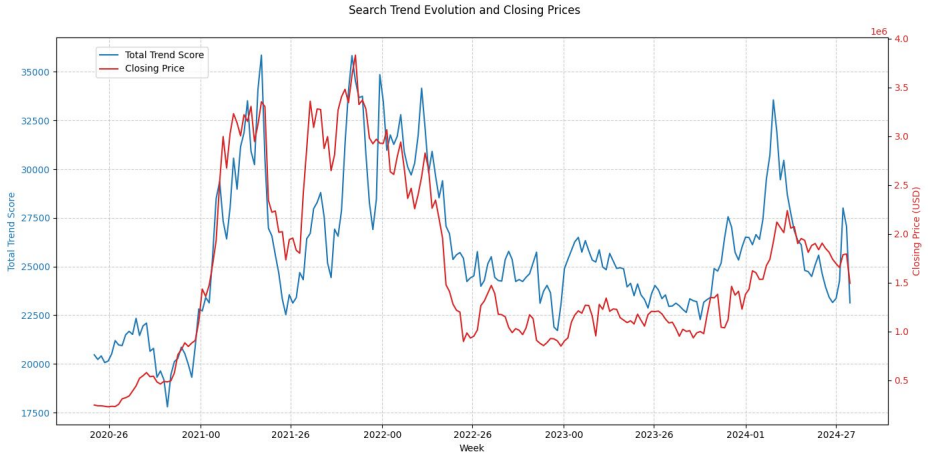
On the overall trend, this is not very visible, but when we look deeper, the data speaks for itself.

Data Collection

Target: Google Trends Data -Its Importance and Impact on Cryptocurrency Prices

Key Insights:

- **Correlation Between Trends and Prices:** It appears to be a noticeable correlation between the Google Trends search scores and the average closing prices of cryptocurrencies.
- **Peaks in search trends often coincide with peaks in cryptocurrency prices, suggesting that increased public interest (as reflected by search trends) may drive up cryptocurrency prices.**
- **Trend and Price Peaks:** Significant peaks in both trend scores and prices are observed around the beginning of 2021 and mid-2021, indicating periods of heightened market activity and interest.
- **Volatility and Market Sentiment:** The volatility in trend scores reflects shifts in market sentiment. Rapid increases and decreases in the trend score might indicate speculative behavior or responses to market news and events.
- **Search Trends as a Leading Indicator:** In several instances, spikes in search trends slightly precede spikes in closing prices, suggesting that search trends could be a leading indicator for price movements.



Details about specific coins and the Google Trends correlation to closing prices are available in the report here.

Data Collection

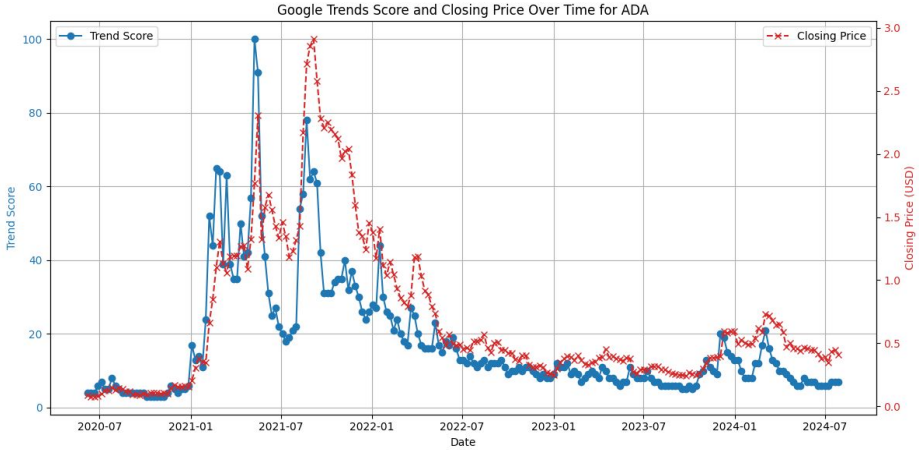
Target: Google Trends Data- Its Importance and Impact on Cryptocurrency Prices

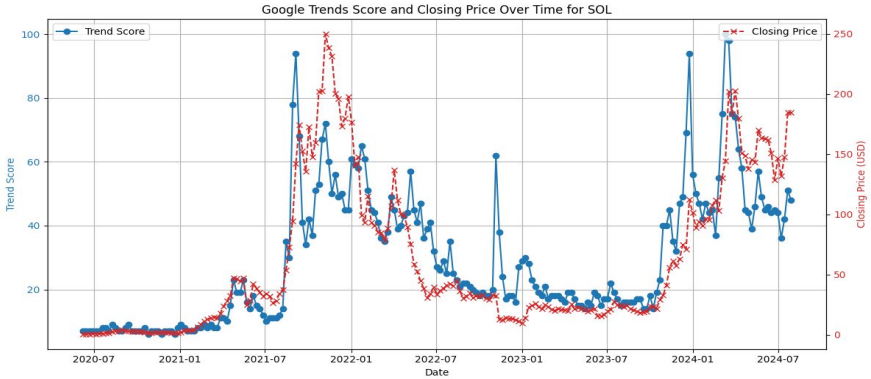
Individual Coins Examples:

- **Public Interest as a Predictor:** For both SOL and ADA, spikes in Google Trends search scores often precede or coincide with increases in closing prices. This reinforces the idea that search trends can serve as a leading indicator of price movements.

- **Volatility Patterns:** Both cryptocurrencies exhibit high volatility in both search trends and closing prices. This reflects the overall nature of the cryptocurrency market, where rapid changes in sentiment and trading activity are common.

- **Significant Peaks and Market Events:** Major peaks in search trends and prices typically align with significant market events or periods of heightened interest. Monitoring these peaks can provide valuable insights into market behavior.





Data Collection

Target: Google Trends Data -Its Importance and Impact on Cryptocurrency Prices

Individual Coins Analysis:

Insights:

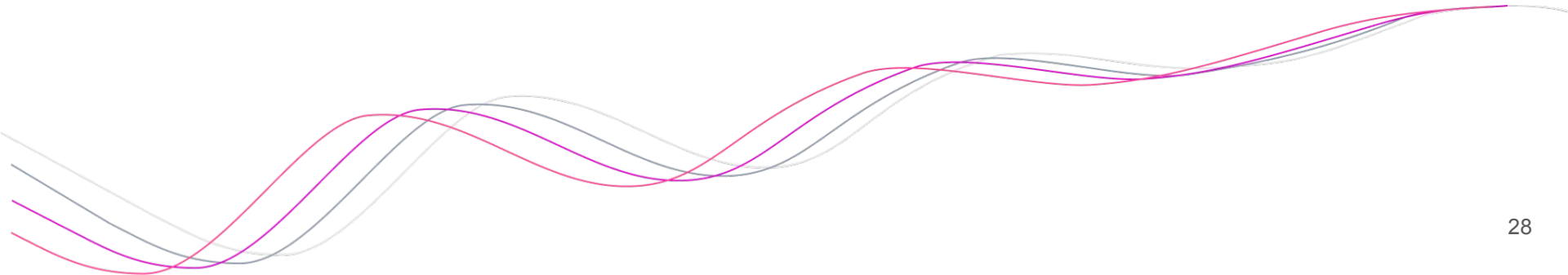
- Most cryptocurrencies exhibit the strongest correlation between Google Trends data and their prices with just a **1-week lag**. This suggests that the impact of changes in search volume on prices is relatively immediate, likely within a week.

- Certain cryptocurrencies like AGIX, DOGE, FET, and KCS show significant correlations at longer lags (2 to 4 weeks), indicating that the effects of search trends on these cryptocurrencies' prices may take longer to materialize.

Optimal Lags and Correlation Insights:

- **Cardano (ADA)**: Strongest correlation at a 1-week lag (Correlation: 0.868)
- **SingularityNET (AGIX)**: Strongest correlation at a 2-week lag (Correlation: 0.752)
- **Binance Coin (BNB)**: Strongest correlation at a 1-week lag (Correlation: 0.499)
- **Bitcoin (BTC)**: Strongest correlation at a 1-week lag (Correlation: 0.636)
- **PancakeSwap (CAKE)**: Strongest correlation at a 1-week lag (Correlation: 0.801)
- **Dogecoin (DOGE)**: Strongest correlation at a 4-week lag (Correlation: 0.598)
- **Polkadot (DOT)**: Strongest correlation at a 1-week lag (Correlation: 0.841)
- **Ethereum (ETH)**: Strongest correlation at a 1-week lag (Correlation: 0.672)
- **Fetch.ai (FET)**: Strongest correlation at a 2-week lag (Correlation: 0.848)
- **Filecoin (FIL)**: Strongest correlation at a 1-week lag (Correlation: 0.713)
- **KuCoin (KCS)**: Strongest correlation at a 2-week lag (Correlation: 0.885)
- **ChainLink (LINK)**: Strongest correlation at a 1-week lag (Correlation: 0.718)
- **Litecoin (LTC)**: Strongest correlation at a 1-week lag (Correlation: 0.799)
- **Ocean Protocol (OCEAN)**: Strongest correlation at a 1-week lag (Correlation: 0.811)
- **Oasis Network (ROSE)**: Strongest correlation at a 1-week lag (Correlation: 0.822)
- **Solana (SOL)**: Strongest correlation at a 1-week lag (Correlation: 0.844)
- **Uniswap (UNI)**: Strongest correlation at a 1-week lag (Correlation: 0.777)
- **Monero (XMR)**: Strongest correlation at a 1-week lag (Correlation: 0.718)
- **XRP**: Strongest correlation at a 1-week lag (Correlation: 0.645)
- **Tezos (XTZ)**: Strongest correlation at a 1-week lag (Correlation: 0.819)

Macro Economic Factors



Data Collection

Target: Macro-Economic Factors - Importance in the Crypto Market

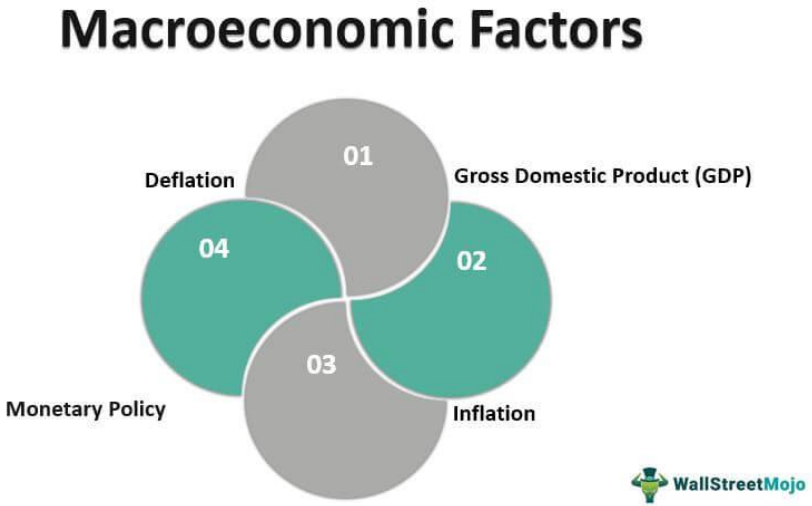
Key Factors:

- Inflation
- GDP (Gross Domestic Product)
- Interest Rates
- CPI (Consumer Price Index)

The key factors data was retrieved from World Bank, FRED (Federal Reserve Economic Data) and Trading Economics.

Countries:: United States, United Kingdom, China, Germany, Japan, India, Brazil, Russia, Canada, Australia, France, Italy, South Korea, Mexico, Saudi Arabia.

These countries were selected due to their significant impact on the global economy and the crypto adoption. Economic changes in these countries can have a ripple effect on the global financial markets, including the cryptocurrency market. Understanding these macro-economic factors helps in predicting the movement of cryptocurrency prices and making informed investment decisions.

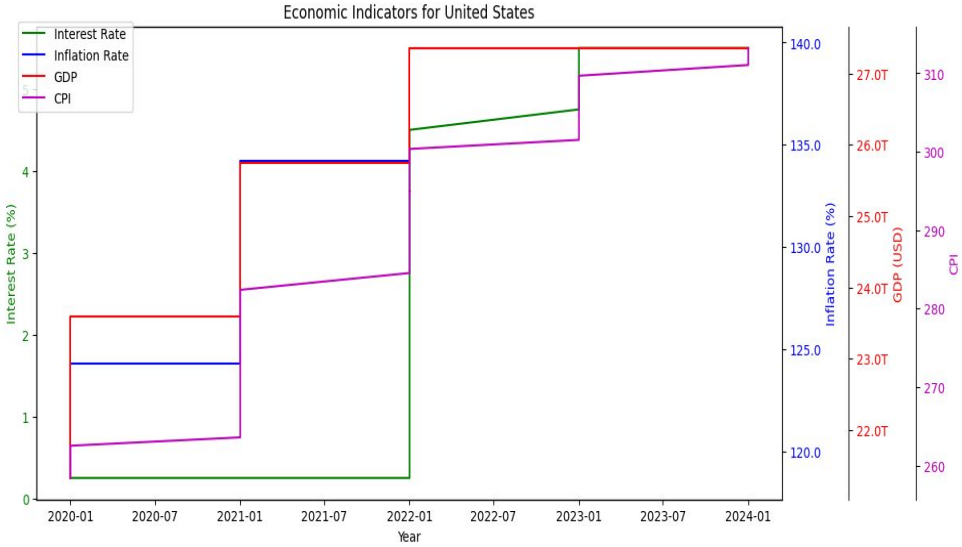


Data Collection

Target: Macro-Economic Factors - Importance in the Crypto Market

Inflation and GDP Data from [World Bank API]

- **Inflation**
  - Measures the rate at which the general level of prices for goods and services is rising, and subsequently, how purchasing power is falling.
  - **Impact on Crypto Market:** High inflation can lead to increased interest in cryptocurrencies as investors seek assets that might protect against currency devaluation. Cryptocurrencies, often perceived as 'digital gold', can be attractive during inflationary periods.
- **GDP (Gross Domestic Product):**
  - Represents the total market value of all finished goods and services produced within a country in a specific period.
  - **Impact on Crypto Market:** Higher GDP indicates a stronger economy, which can increase investment in riskier assets like cryptocurrencies. Conversely, lower GDP may signal economic uncertainty, affecting investor confidence.

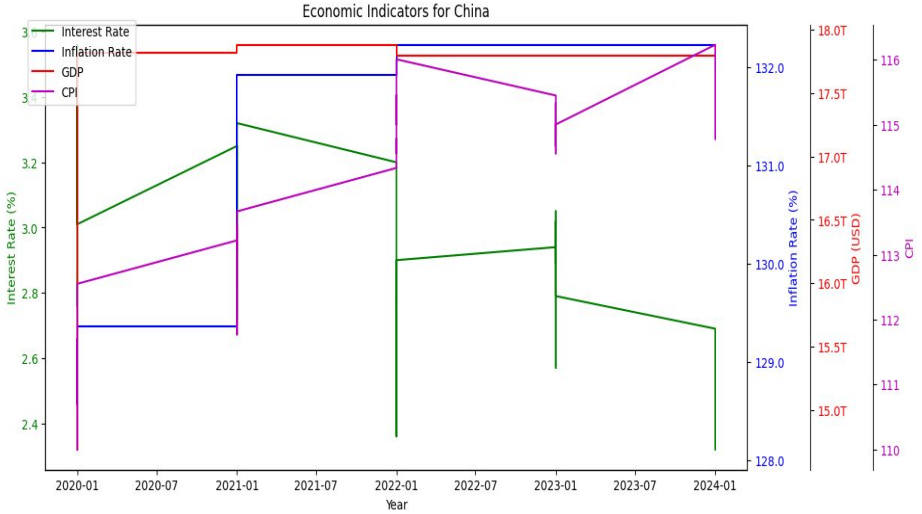


Data Collection

Target: Macro-Economic Factors - Importance in the Crypto Market

Interest Rates & CPI from FRED (Federal Reserve Economic Data)

- **Interest Rates:**
  - The cost of borrowing money, typically expressed as an annual percentage of the loan amount.
  - **Impact on Crypto Market:** Low interest rates can lead to higher investment in cryptocurrencies as borrowing becomes cheaper, and investors search for higher returns. High interest rates may discourage investment in riskier assets, including cryptocurrencies.
- **CPI**
  - CPI is a primary indicator of inflation, showing how much prices have increased or decreased over a specific period.
  - **Impact on Crypto Market**: Cryptocurrencies, particularly Bitcoin, are often viewed as a hedge against inflation. When CPI reports indicate rising inflation, investors may flock to crypto assets as a way to preserve purchasing power. This is especially true if traditional fiat currencies are losing value due to inflation.

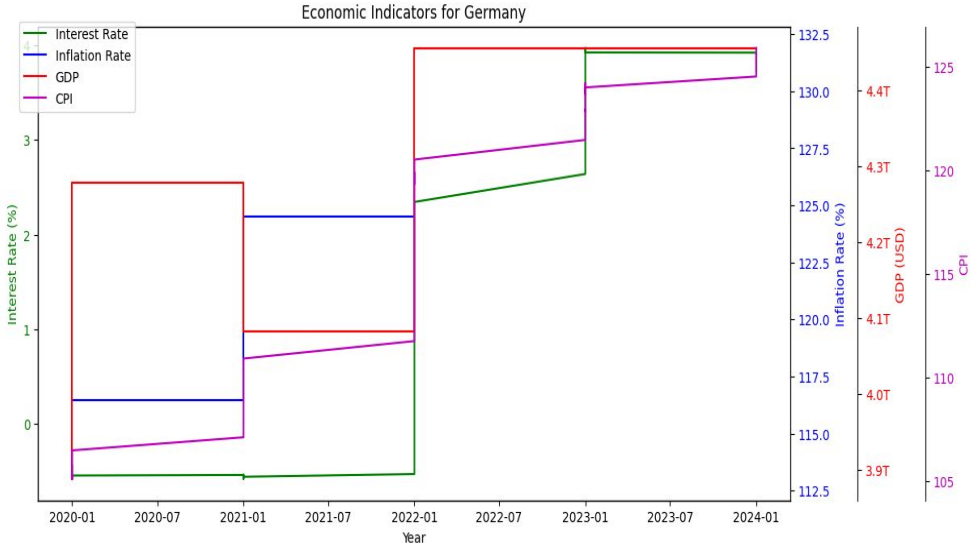


Data Collection

Target: Macro-Economic Factors - Importance in the Crypto Market

Macro Economic Factors Influence

- **Inflation and GDP**: Data was gathered annually to observe long-term trends and their correlation with cryptocurrency market behavior.
- **Inflation and Crypto Prices**: Analysis shows that periods of high inflation often coincide with increased interest and investment in cryptocurrencies.
- **GDP Growth and Market Confidence**: Higher GDP growth correlates with higher market confidence and greater investment in digital assets.**Interest Rates**: Monthly data was collected to capture the more frequent changes in monetary policy and their immediate effects on the market.
- **Interest Rate Changes**: Low interest rate periods have seen significant inflows into the crypto market as investors seek better returns compared to traditional savings accounts.
- **CPI**: When inflation is high, central banks may reduce or stop quantitative easing (the injection of money into the economy), which can reduce the amount of capital flowing into cryptocurrencies.



Data Collection

Target: Macro-Economic Factors - Importance in the Crypto Market

Challenges with Real-Time Data:

- One of the key challenges with these factors is the absence of an API that provides real-time data. Real-time updates, particularly for interest rates, could have a significant impact on market reactions, as we recently observed with changes in **Japan's interest rate**. Being able to monitor these shifts in real-time would be highly advantageous, but the APIs I've explored so far do not offer this capability. Currently, the latest available data is from June 1, 2024.
- **Importance of GDP and Inflation Data:**
  - GDP and inflation are typically available as yearly data.
  - Despite the less frequent updates, this information is still valuable for understanding long-term economic trends and their impact on the crypto market.
  - High GDP growth indicates a strong economy, which can boost investor confidence in cryptocurrencies.
  - High inflation can lead to increased interest in cryptocurrencies as a hedge against currency devaluation.

- **Value of Interest Rate Data:**
  - Interest rate data is available on a monthly basis and is highly valuable for market analysis.
  - Changes in interest rates have a direct and visible correlation with cryptocurrency prices.
  - Lower interest rates tend to drive investment into higher-risk assets like cryptocurrencies.
  - Higher interest rates can lead to reduced investment in cryptocurrencies as investors seek safer returns.

- **Manual Data Input for Real-Time Monitoring:**
  - To achieve real-time monitoring, manual data input is necessary.
  - This involves continuously updating the datasets with the latest available information to maintain accuracy and relevance.
  - While automated APIs can provide historical data, staying updated with real-time changes requires a more hands-on approach.

Data Cleaning

Target: Clean and merge collected data into a single dataset

Data Integration and Imputation:

- An extensive process was undertaken to combine all collected datasets and fill in missing information.
- For instance, the USA interest rate data from the World Bank was incomplete, so additional data was sourced from Trading Economics to fill the gaps, ensuring that this crucial factor was included in the analysis.

Handling Missing Data:

- Columns with over 50% missing data were removed to maintain data integrity. These included:
  - interest_rate_India, interest_rate_Russia, cpi_Russia, cpi_Japan
- This step was essential to reduce noise and focus on the most reliable data.

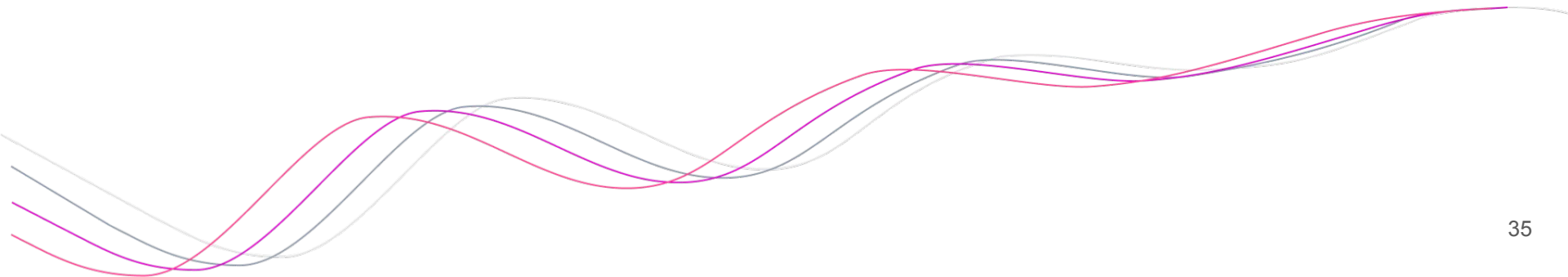
Economic Data Resampling:

- Economic indicators, typically available on a monthly or yearly basis, were resampled to a daily frequency to align with the daily financial data. This was done by forward-filling values, allowing for a merge with the rest of the dataset.

Google Trends Data Resampling:

- Google Trends data, originally provided on a weekly basis, was also resampled to a daily frequency. This ensured consistency across all datasets, making the analysis more robust.

Features



Market data features

Various features were extracted and tested from the complete dataset, including the following key categories:

1. **Market Data Features:** Derived from close prices, trading volumes, and market capitalization.
2. **Macroeconomic Features:** Weighted global GDP, interest rates, and inflation trends.
3. **Sentiment Features:** Derived from the Fear-Greed Index & Google Trends data

Market data features

- **Moving Averages (MA):**
  - Purpose: Captures short-term and long-term trends using averages over 1, 3, 7, 14, and 30 days.
- **Rate of Change (RoC):**
  - Purpose: Measures momentum by calculating the percentage change in price over different periods.
- **Exponential Moving Averages (EMA):**
  - Purpose: Provides a weighted average of prices, emphasizing recent data to quickly react to price changes.
- **Liquidity Factor:**
  - Purpose: Reflects market activity, derived from log-transformed trading volume.
- **Size Factor:**
  - Purpose: Indicates the relative size of an asset, with smaller assets generally perceived as higher risk.
- **Momentum Factor:**
  - Purpose: Assesses price changes over the previous month, based on the idea that trends tend to persist.

Sentiment data features

Sentiment Feature: Extracted from the Fear-Greed Index

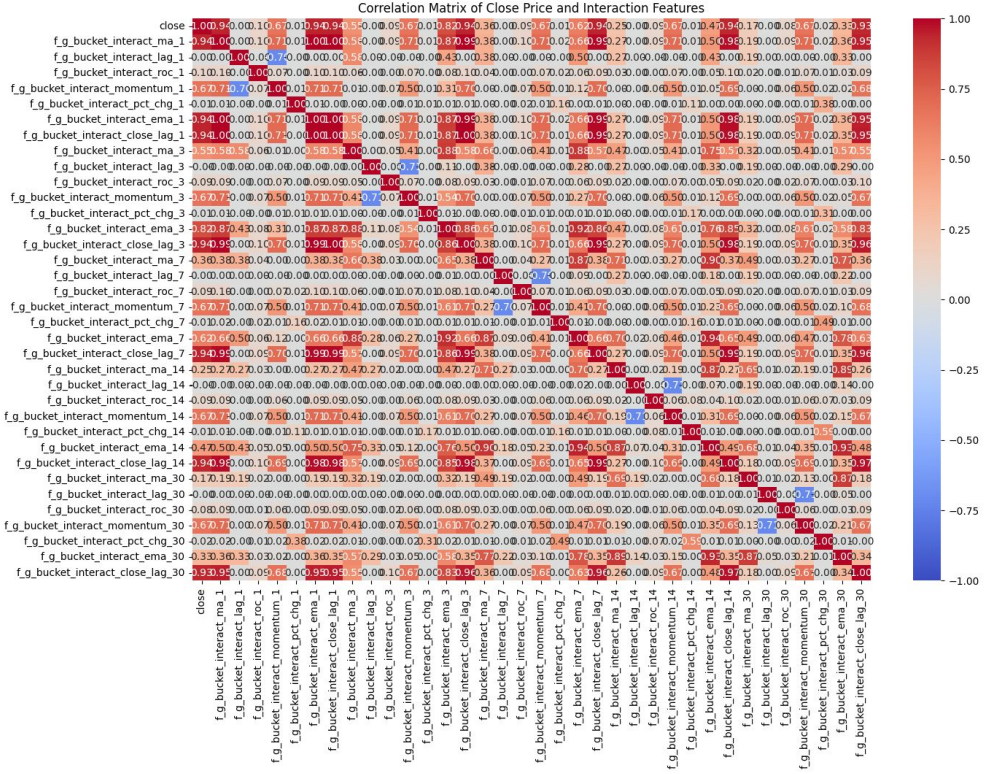
The Fear-Greed Index values was divided into 5 buckets: [0, 0.25, 0.50, 0.75, 1], and its correlation with close price features was analyzed.

The results were notable:

- f_g_bucket_interact_ma_1: 0.940710
- f_g_bucket_interact_ema_1: 0.940710
- f_g_bucket_interact_close_lag_7: 0.938542
- f_g_bucket_interact_close_lag_1: 0.938438
- f_g_bucket_interact_close_lag_3: 0.938429
- f_g_bucket_interact_close_lag_14: 0.935958
- f_g_bucket_interact_close_lag_30: 0.931691

Fear-Greed Index Interaction: The interactions of the Fear-Greed Index with close price features, such as moving averages (ma_1) and exponential moving averages (ema_1), show correlations above 0.94. **This indicates that sentiment, as measured by the Fear-Greed Index, is highly aligned with immediate price trends.**

Predictive Power: The strong correlations imply that these sentiment-based features could be powerful predictors in a model, helping to capture shifts in market behavior driven by investor sentiment. For example, a high Fear-Greed Index score (indicating greed) combined with recent price trends could signal continued upward momentum, while a low score (indicating fear) might align with declining prices.



Macroeconomic and Sentiment Features

These features capture the global economic environment's influence on market dynamics, integrating data on GDP, interest rates, and inflation.
Key Features:

1. **Weighted Global GDP:**
   - **a.** An aggregated measure of GDP from major global economies, weighted by each country's economic size.
   - **b.** Purpose: Reflects the overall health and scale of global economic activity, influencing market sentiment and investment decisions.

2. **Overall Interest Rate:**
   - a. The average of interest rates across different economies(please see slide 29 for the countries selected).
   - b. Purpose: Indicates the general level of global interest rates, affecting borrowing costs, consumer spending, and investment flows.

3. **Interest Rate Trend:**
   - a. An aggregated trend indicator, showing whether global interest rates are generally rising, falling, or stable
   - b. Purpose: Captures the direction of monetary policy across major economies, which can impact market liquidity and investor risk appetite.

4. **Overall Inflation Rate:**
   - a. Description: The average inflation rate across multiple countries, providing a view of global inflationary pressures.
   - b. Purpose: Reflects the cost of goods and services globally, influencing central bank policies and consumer purchasing power.

5. **Inflation Rate Trend:**
   - a. Description: An aggregated trend indicator for inflation rates, showing whether global inflation is rising, falling, or stable.
   - b. Purpose: Tracks the general trend in global inflation, which is crucial for understanding potential shifts in monetary policy and economic stability.

Identifiers & date

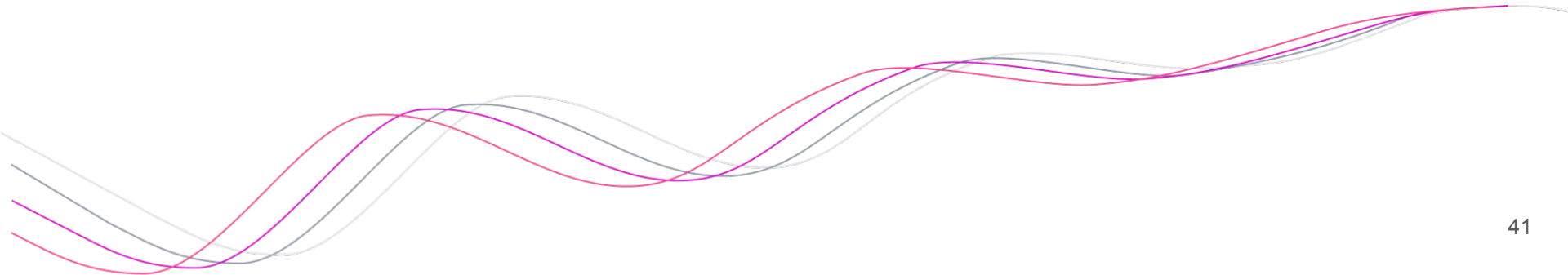
In addition to the other features, it was necessary to train the machine learning model with symbol and date information, given that the dataset includes over 1,000 coins. To accomplish this, a **symbol_encoded** feature was created, and the date was transformed into a **timestamp**.

Additionally, factors such as **is_active** and **is_open_source**, sourced from CoinMarketCap, were included to enhance the validity and credibility of the coins in the dataset.

	symbol_encoded	timestamp	size_factor	liquidity_factor	is_active	is_open_source	close	ma_7	lag_7	momentum_7
0	7919830	1.593562e+09	13199	13720	1	1	0.104406	0.704446	0.001840	0.102566
1	8744466	1.593562e+09	15803	17707	1	0	0.098087	0.553378	1.228615	-1.130528
2	2758212	1.593562e+09	18837	15489	1	1	0.096486	0.038826	0.000084	0.096402
3	888652	1.593562e+09	23208	19835	1	1	0.095954	0.330056	0.030132	0.065822
4	7158091	1.593562e+09	16851	13466	1	1	0.099389	33.639187	0.297936	-0.198547

close_lag_30	fear_greed_bucket	f_g_bucket_interact_ma_1	f_g_bucket_interact_ema_1	f_g_bucket_interact_close_lag_7	weighted_global_gdp	overall_interest_rate	interest_rate_trend	overall_inflation_rate	inflation_rate_trend
0.145287	0.5	0.052203	0.052203	0.055886	1.405724e+13	0.816792	0	131.517715	0
0.112986	0.5	0.049043	0.049043	0.049484	1.405724e+13	0.816792	0	131.517715	0
0.105139	0.5	0.048243	0.048243	0.051744	1.405724e+13	0.816792	0	131.517715	0
0.081168	0.5	0.047977	0.047977	0.041292	1.405724e+13	0.816792	0	131.517715	0
0.093641	0.5	0.049695	0.049695	0.052731	1.405724e+13	0.816792	0	131.517715	0

Model Development



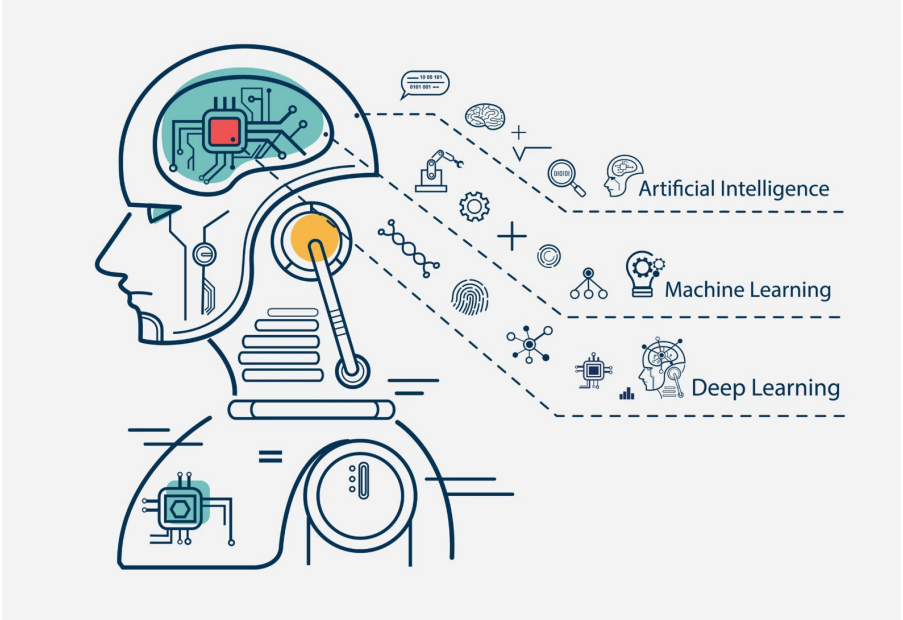
Model Development

Several LGBMRegressors were trained and evaluated using different sets of features to identify the best ones. The LGBMRegressors used followed the configuration mostly used by participants and recommended by Numerai. Details on the model notebook in the github repository.

Each model was developed by combining a varied set of features, each intended to capture different aspects of the data to improve predictive performance. These features included basic information like symbol identifiers, as well as more complex calculations such as moving averages over various periods. The models were designed to evaluate the impact of several key factors, including the Fear and Greed Index for market sentiment and the Liquidity Factor for market activity.

In addition, we incorporated Size Factor metrics, which account for the relative scale of assets, and global economic indicators like the average interest rate worldwide to provide a broader economic context. The aim was to identify the best combination of these features to achieve accurate and reliable predictions.

Each model uses a set of features to determine which combinations are most effective in prediction. The following slides provide details on the various models that were evaluated.



Model Development

Model 1

Features:

- **symbol_encoded, timestamp, size_factor, liquidity_factor, close**

Starting with a smaller feature set, the model was built using core factors like symbol identifiers, size and liquidity metrics, along with basic market data such as the close price, to establish a foundational understanding of the key drivers in the dataset.

Results:

- **Mean Squared Error (MSE): 0.02714**
  - MSE measures the average squared difference between the actual and predicted values. A lower MSE indicates better predictive accuracy. An MSE of 0.027 suggests that the model's predictions are relatively close to the actual values, with small errors on average.

- **R-squared (R2): 0.53**
  - R-squared indicates the proportion of the variance in the target variable that is explained by the model. An R-squared value of 0.53 means that approximately 53% of the variance in the data is explained by the model. This is a quite solid R-squared value, indicating that the model is capturing a good amount of the underlying patterns in the data.

- **Mean Absolute Error (MAE): 0.0822**
  - MAE measures the average magnitude of errors in the predictions, without considering their direction. The MAE of 0.0822 suggests that, on average, the model's predictions are off by about 0.0822 units from the actual values. This is a relatively low error, indicating good accuracy.

- **Normalized MSE (NMSE): 0.36**
  - NMSE compares the MSE to the variance of the actual values. A NMSE of 0.36 indicates that the model's MSE is about 36% of the variance in the data, suggesting that the model is performing significantly better than just predicting the mean value of the target variable.

Model Development

Model 2

Features:

- **symbol_encoded, timestamp, size_factor, liquidity_factor, close, fear_greed_bucket, f_g_bucket_interact_ma_1, f_g_bucket_interact_ema_1, f_g_bucket_interact_close_lag_3**

  Building on the initial feature set, additional sentiment-driven factors were introduced, including the Fear and Greed Index and its interactions with key metrics like moving averages and price lags. This expanded feature set aims to capture the influence of market sentiment on price movements, providing a more nuanced understanding of the dynamics at play.

  The model results are improved by more than **10%. R2 0.53 -> 0.636**

Results:

- Mean Squared Error (MSE): **0.02744266563626544**
- R-squared (R2): **0.6360100260719439**
- Mean Absolute Error (MAE): **0.08387744079949767**
- Normalized MSE (NMSE): **0.3639899893364456**

Model Development

Model 3

Features:

- **symbol_encoded, timestamp, size_factor, liquidity_factor, close, fear_greed_bucket, f_g_bucket_interact_ma_1, f_g_bucket_interact_ema_1, f_g_bucket_interact_close_lag_3, weighted_global_gdp, overall_interest_rate, interest_rate_trend, overall_inflation_rate, inflation_rate_trend**

  Macroeconomic factors, such as the overall interest rate, interest rate trend, and inflation rate and trend, were added to the previous model. The results show a slight improvement, indicating that these additional features do contribute value without introducing significant noise. However, their impact on the overall performance is not substantial.

Results:

- Mean Squared Error (MSE)**: 0.026857850735258603**
- R-squared (R2): **0.6437668075519802**
- Mean Absolute Error (MAE): **0.08256785060353117**
- Normalized MSE (NMSE): **0.35623320752804977**

Model Development

Model 4

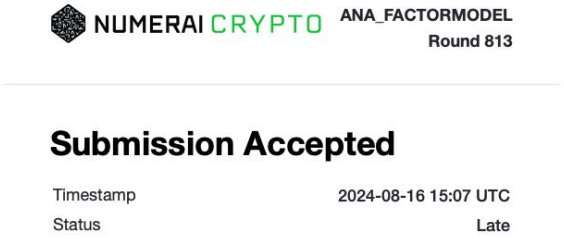
Features:

- **Symbol_encoded, timestamp, size_factor, liquidity_factor, close, pct_chg_30, close_lag_30, close_lag_14, close_lag_7, overall_interest_rate, momentum_30, f_g_bucket_interact_ma_1, f_g_bucket_interact_close_lag_3**

  The feature selection for the model focused on capturing a balanced mix of market behavior, economic indicators, and sentiment factors. By including key elements such as the size and liquidity factors, along with specific price lag indicators like close_lag_7, close_lag_14, and close_lag_30, the model aims to account for both immediate and longer-term trends. Additionally, macroeconomic inputs like the overall interest rate and sentiment-driven features such as the Fear and Greed Index interactions were incorporated to provide a more comprehensive view of the market dynamics.

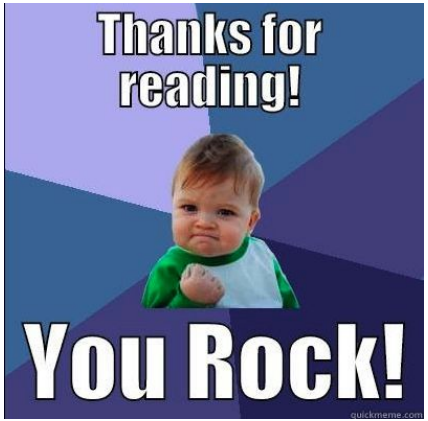
Results:

- Mean Squared Error (MSE)**: 0.02674862674154639**
- R-squared (R2)**: 0.6452155166223954**
- Mean Absolute Error (MAE): **0.08270396894745392**
- Normalized MSE (NMSE): **0.3547844983963081**



The results aren't where I'd like them to be yet, as this report is still a work in progress. I wanted to share what I've accomplished so far, even though it's not fully finished.

See you on the Numerai crypto leaderboard! I've kicked off the submission process and plan to automate it for some smooth, consistent entries 😎



For the code and additional resources, head over to the GitHub repository. If you have any question, please reach out via Discord (white_rider_) or Twitter!