GDELT-BTC: Business understanding

Osvald Nigola & Leo-Martin Pala

Identifying business goals

Background

The cryptocurrency market, particularly Bitcoin (BTC), is known for its high volatility and sensitivity to global events. While traditional financial markets have established relationships with news and world events, the cryptocurrency market's reactions are less understood and often more dramatic. The Global Database of Events, Language, and Tone (GDELT) provides a comprehensive, near real-time catalog of world events, offering an opportunity to analyze their impact on Bitcoin price movements. This project aims to bridge the gap between world events and cryptocurrency market behavior, potentially uncovering actionable patterns for trading decisions.

Business goals

1. Develop a data-driven system that predicts Bitcoin price direction (increase, decrease, or stability) based on significant world events captured in the GDELT database
2. Generate actionable insights for cryptocurrency traders by providing clear relationships between event types and price reactions within each 15-minute window
3. Build an automated trading strategy that outperforms buy-and-hold by at least 5% over 6 months

Business success criteria

The success of this project will be measured through the following criteria:

- Demonstrate a 5% superior profit performance compared to a passive bitcoin holding strategy within a six-month period
- Identification of at least three distinct categories of world events that show strong correlation with Bitcoin price movements
- Development of a reliable method to filter and prioritize significant events from the GDELT database that impact cryptocurrency markets
- Creation of clear, interpretable relationships between event types and price movements that can be used for trading decisions

Situation assessment

Inventory of resources

Hardware and Infrastructure

- M1 MacBook Air with 16GB RAM
- HP EliteBook 840 G8 Notebook PC with 11th Gen Intel Core i5-1135G7
- Free online notebook hosting sites: Google Colab, Kaggle Notebooks
- Cumulative 60$ modal credits
- 93€ of Azure Cloud credit

Data Resources

- GDELT Event Database through Google BigQuery
- Historical Bitcoin price data with 1-minute intervals from Kaggle

Software Tools

- PyCharm Professional (student license) as primary IDE
- Anaconda Python distribution with included libraries
- Jupyter Notebooks for analysis
- Version control through Git hosted on GitHub

Requirements, assumptions, constraints

Requirements

- Data must be processed in 15-minute intervals to match GDELT event timeframes
- System must handle real-time GDELT data filtering efficiently
- Models must provide interpretable outputs for trading decisions

Assumptions

- GDELT data accurately captures significant world events
- Historical patterns between events and price movements remain relevant
- 15-minute intervals provide sufficient granularity for analysis
- World events have measurable impacts on cryptocurrency prices

Constraints

- Limited computational resources
- Memory constraints when processing large amounts of GDELT data
- Project completion deadline within a few weeks

Risks and contingencies

Technical Risks

- Risk: Computational limitations with full GDELT dataset
- Contingency: Implement efficient filtering and data streaming

Data Risks

- Risk: GDELT data quality or availability issues
- Contingency: Maintain data backups and alternative data sources
- Risk: Inconsistent or missing price data
- Contingency: Multiple data sources

Implementation Risks

- Risk: Complex model deployment and maintenance
- Contingency: Focus on simple, robust models

Terminology

- GDELT: Global Database of Events, Language and Tone

- BTC: Bitcoin cryptocurrency
- Price Direction: Categorical outcome (Up/Down/Stable)
- Event Impact: Calculated significance of world events
- Trading Signals: Generated buy/sell/hold recommendations
- CAMEO: Conflict and Mediation Event Observations
- FIPS 10-4: US Government standard for 2-letter country codes

Costs and benefits

Costs

- Development time: Approximately 20 hours per week per person
- Computational resources: Local processing and cloud services
- Data storage: Historical and processed data
- Documentation effort: Maintaining clear project documentation

Benefits

- Creation of a novel approach to cryptocurrency trading
- Potential for profitable trading strategies
- Academic contribution to financial data science
- Enhanced understanding of global event impacts on crypto markets
- Skills development in machine learning and data processing

Data-mining goals

Goals

1. Event Impact Analysis
    - Create a scoring system for event significance based on GDELT metrics
    - Identify and categorize events that consistently affect BTC price
2. Feature Engineering Pipeline
    - Transform GDELT event data into meaningful numerical features
    - Generate derived features capturing event intensity and market sentiment
3. Predictive Modeling
    - Build a regression model predicting price change percentage
    - Implement ensemble methods combining different prediction approaches
4. Real-Time Processing Framework
    - Design an efficient system for processing incoming GDELT events
    - Create methods for rapid feature calculation and prediction

Success criteria

1. Trading Performance
    - Strategy must outperform "buy and hold" by at least 5% over 6-month test period
    - Strategy must remain profitable after accounting for:
        i. Trading fees (assumed to be 0.1% per trade)
        ii. Minimum transaction amount (assumed to be 10$)

2. Risk Management
    - Position sizing must be dynamic based on prediction confidence
3. Technical Performance
    - Predictions must be generated within each 15-minute window

GDELT-BTC: Data understanding

Osvald Nigola & Leo-Martin Pala

Gathering Data

Data Requirements Outline

To predict Bitcoin price movements based on world events, we require:
asdfasdf

1. GDELT Event Data (2019-2023)
    - Event details with timestamps (15-minute granularity)
    - Actor information (countries, organizations)
    - Event characteristics (type, impact, tone)
    - Event metadata (mentions, sources, articles)
2. Bitcoin Price Data (2019-2023)
    - Price data with high temporal resolution (15-minute granularity)
    - Price change metrics

Data Availability Verification

- GDELT 2.0 Data
    - Available through Google BigQuery
    - Free and open usage
    - 15-minute update frequency
    - Complete coverage for required period
- Bitcoin Price Data
    - Available through Kaggle
    - 1-minute granularity
    - Complete coverage for required period
    - Includes all required price metrics

Selection Criteria Definition

1. GDELT Events
    - Time range: 2019-2023
    - Required fields:
        - Date and time
        - Actor country codes and types
        - Event codes and types
        - Impact metrics (GoldsteinScale, NumMentions, etc.)
        - Geographic information about the event itself
2. Bitcoin Data
    - Time range: 2019-2023

    - Required fields:
        - Date and time
        - Open price
        - Close price

Describing Data

GDELT 2.0 Dataset

- Temporal Coverage: 2015 - present
- Update Frequency: 15 minutes
- Key Fields:
    - DateAdded (string): Event date in YYYYMMDDHHMMSS (15-min resolution)
    - Actor1CountryCode (string): CAMEO country code for initiating actor
    - Actor1Geo_CountryCode (string): 2-char FIPS 10-4 country code
    - Actor1TypeCode (string): CAMEO type code for initiating actor
    - Actor2CountryCode (string): CAMEO country code for receiving actor
    - Actor2Geo_CountryCode (string): 2-char FIPS 10-4 country code
    - Actor2TypeCode (string): CAMEO type code for receiving actor
    - EventGeo_CountryCode (string): 2-char FIPS 10-4 country code for event location
    - IsRootEvent (boolean): A measure roughly correlated with significance
    - EventRootCode (string): One of 20 event categorizations
    - QuadClass (string): One of 4 event categorizations
    - GoldsteinScale (float): Event type score -10..+10 describing potential impact on the stability of a country
    - Quality metrics (integer): NumSources, NumArticles
    - AvgTone (float): Average tone -100..+100 (commonly -10..+10) of articles describing the event with magnitude being correlated to significance
    - SourceURL (string): Link to the article about the event

Bitcoin Dataset

- Temporal Coverage: 2012 - present
- Resolution: 1-minute intervals
- Key Fields:
    - Timestamp (float)
    - Price metrics (float): Open, High, Low, Close
    - Volume (float)

Exploring Data

GDELT Analysis

- Actor country codes available in 30-50% of events

- Actor types available in 30-50% of events
- Event codes and Goldstein scale present in most entries
- Action geographic codes available in 94% of events
- Impact scores calculated using Goldstein scale, articles, and tone
- Event distribution shows expected patterns (more verbal than material events)

Bitcoin Analysis

- Complete price data without gaps
- Normal trading volume patterns
- Expected volatility patterns
- Consistent data quality across the period

Relationships

- Initial exploration shows potential correlations between significant events and price movements

Verifying Data Quality

GDELT Data Quality

Strengths:

- Consistent 15-minute updates
- Rich event categorization
- Reliable source and mention counts
- Good geographic coverage

Issues:

- Missing actor information in some events
- Varying detail levels in event descriptions
- Some duplicate events across updates

Bitcoin Data Quality

Strengths:

- Complete price history
- Consistent data format
- Reliable volume data
- No missing values

Issues:

- Verification of time zone alignment required

Quality Management Plan

1. Data Cleaning Strategy:
    - Filter significant events based on impact scores

    - Align timestamps between datasets

2. Data Validation:
    - Ensure consistent time zone handling

GDELT-BTC: Project plan

Osvald Nigola & Leo-Martin Pala

Data Collection, Preprocessing, and Exploration

GDELT Data Acquisition and Initial Filtering (Osvald - 18h)

- Setting up BigQuery (4h)
- Script and permissions for getting GDELT data from BigQuery (4h)
- Initial Filtering (2h)
- Data Quality Checks, Cleaning and Validation (8h)

BTC price data collection and initial cleaning (Leo - 10h)

- Download and Clean BTC Price Data (4h)
- Resample to 15-Minute Intervals (3h)
- Timestamp Synchronization and Alignment (3h)

Output: Clean, merged datasets ready for analysis

Exploratory Data Analysis and Feature Engineering

Event impact scoring, Event type categorization (Osvald - 17h)

- In-depth EDA of GDELT Data (10h)
- Event Impact Scoring (4h)
- Event Categorization and Aggregation (3h)

BTC EDA and correlation with GDELT (Leo - 23h)

- BTC Price Time Series Analysis (8h)
- Correlating BTC and GDELT data (10h)
- Data Encoding for Modeling (5h)

Output: Feature-rich dataset ready for modeling

Model Development and Selection

Baseline Model Development (Leo - 10h)

- Implement a simple baseline model

Traditional Machine Learning Model Training (Leo - 15h)

- Model Selection (3h)
- Hyperparameter Tuning and Cross-Validation (8h)
- Model Evaluation and Comparison (4h)

Neural Network Model Training (Osvald - 20h)

- Architecture Design and Implementation (8h)
- Hyperparameter Tuning and Training (8h)
- Model Evaluation and Selection (4h)

Output: Multiple trained models with performance metrics

Trading Strategy Development and Backtesting

Signal Generation and Trading Logic (Leo - 20h)

- Develop Signal Generation Logic (10h)
- Implement Trading Logic (10h)

Risk Management and Performance Evaluation (Osvald - 20h)

- Implement Risk Management Rules (5h)
- Develop Performance Evaluation Metrics (5h)
- Implement Backtesting Framework (10h)

Output: Complete trading strategy with risk management

Documentation and Reporting

Comprehensive Documentation (Osvald - 15h)

- Create detailed documentation.

Results Visualization and Report Compilation (Leo - 12h)

- Create clear and informative visualizations of the project results.

Output: Comprehensive performance analysis with Final report and presentation

Tools

- Python (Anaconda)
- Pandas
- ScikitLearn
- Pytorch
- cuML
- MatPlotLib
- NumPy
- Google BigQuery
- Kaggle
- PyCharm
- GitHub Copilot
- SimTheory
- Claude
- …

Total Hours

Leo: 90 hours
Osvald: 90 hours
Combined: 180 hours