MGMT 467 Final Project - Bitcoin Price Prediction Pipeline

Governance Report

Team 4

Anurag Koripalli, Lily Larson, Sanjana Mohan, Kundada Nittala

This governance report contains the ethical, security, data quality and compliance decisions governing the operation and processes of the Bitcoin Price Prediction Pipeline.

Project Scope

The Bitcoin Price Prediction Pipeline aims to predict short term BTC price direction and volatility to support algorithmic trading and risk management. The historical data used for this purpose is a Kaggle dataset (https://www.kaggle.com/datasets/mczielinski/bitcoin-historical-data) which is the batch data and we use a CoinCap API to get our streaming data. Our model type is a BQML regression/classification model using a blended feature set which include lagged batch features and real-time streaming metrics. The pipeline is designed to run 24/7 with 1 minute refresh cycles.

Data Ethics and Privacy

- **Privacy Assessment**
  - The privacy assessment for this pipeline is low risk.
  - The pipeline handles exclusively publicly available, non-personally identifiable information (PII) related to cryptocurrency prices, volume, and market indicators.
  - No PII or sensitive personal data is collected, stored, or processed.
  - Therefore, standard privacy regulations (like GDPR or CCPA) are not applicable.
- **Bias Mitigation and Fairness**
  - The model relies on historical financial data, which is susceptible to biases arising from market volatility, historical trading patterns, and global economic events.
  - Mitigation: The use of ML.EXPLAIN_PREDICT is crucial to audit the model's predictions and ensure the influence of real-time streaming features is balanced against historical features. This allows the team to detect if the model develops a reliance on an unpredictable or unstable feature.

Security Notes

- **Authentication & Secrets Management:**

    - The chosen secondary source, the CoinCap Live Prices API, does not require an authentication key or secret.
    - Security Principle Applied: We adhere to the principle of "No secrets in source code." Had the API required a secret, it would have been mandated to be stored in Google Cloud Secret Manager, and the Cloud Function's Service Account would have been restricted to only read that secret.
- **IAM Least Privilege:**
    - All service accounts (Cloud Function, Dataflow, Looker Studio) are configured with the minimum necessary permissions (e.g., read-only for BQ analytics, write-only for Dataflow to the streaming table). This limits the blast radius if any component is compromised.

Data Quality and Transformation Logics

Batch Ingest (Kaggle CSV → BQ)

- **Quality Check/Transformation:** All records must be non-null in ‘timestamp’, ‘open’ and ‘close’ columns. The ETL script also includes steps to remove duplicate timestamps and validate monotonic timestamp order.
- **Rationale:** Rows with zero and duplicates corrupt derived features (volatility, lag) and inflate the training data, leading to poor model performance.

Streaming Ingest (CoinCap to Pub/Sub)

- **Quality Check/Transformation:** The raw JSON payload from CoinCap is immediately **normalized** within the Cloud Function. Nested fields (e.g., data.price_usd) are extracted and converted to the correct BigQuery data types (e.g., price_usd as FLOAT64) before publishing to Pub/Sub.
- **Rationale:** Ensures a flat, predictable schema for the Dataflow pipeline, simplifying streaming logic and ensuring type consistency for BQML.

Risks & Challenges

The main challenges faced by the pipeline can be divided across the system architecture. Data Pipeline integrity is threatened by high velocity and low latency, which means that network congestion can risk delaying data and nullifying its real-time value, alongside the need to efficiently manage cloud costs for continuous stream processing. For machine learning model risks, the most critical issue is concept drift, where changes in market behavior cause the model's accuracy to quickly decline, made worse by the fact that complex models are hard to explain which can contribute to reducing confidence and trust amongst executives.. Finally, dashboard limitations need complex pre-filtering in BigQuery to get past Looker Studio's functionality limits, sometimes resulting in slow dashboard load times and "Too Many Rows" errors.

Conclusion

This governance report successfully establishes relevant controls governing the ethical use, security and data quality of the Bitcoin Price Prediction Pipeline. All stakeholders are required to adhere to the mandated IAM policies, bias mitigation strategies, and data validation logics to ensure continuous compliance and the reliability of our predictive analytics. This entire governance document will undergo a mandatory annual review to ensure its continued relevance against evolving market dynamics and future platform changes.