Machine Learning Models for Bitcoin Quant Strategy

Arvind Kandala, Claire Oh, Yuqian Wang, Tully Cannon

1 Introduction

Bitcoin exhibits exceptional price volatility, characterized by rapid fluctuations that present a significant challenge for analysts and traders. Unlike traditional assets, Bitcoin does not generate cash flows or possess inherent utility like commodities, making its valuation driven largely by perception rather than fundamentals. Research on Bitcoin market volatility from 2022 identifies key drivers: shifts in user demand (often proxied by online search interest), changes in circulating supply, and broader macroeconomic factors such as U.S. consumer confidence. This suggests that while Bitcoin is highly sensitive to regulatory announcements and social media sentiment, its volatility is not entirely random but linked to specific social interest, supply constraints, and broader financial market sentiment. In the past, immature markets and lower market depth also contributed significantly to this instability.

To address this challenge, we aim to leverage historical data to understand market dynamics and predict future price movements. Utilizing Python, we sourced on-chain and price data from Yahoo Finance spanning November 30, 2014, to the present. We applied three distinct modeling approaches, ranging from simple to complex, to capture different aspects of the data. First, Logistic Regression was employed as a baseline to capture linear relationships. Second, we utilized a Random Forest model, leveraging ensembles of decision trees to handle non-linear interactions and improve performance by reducing the influence of noise. Finally, we implemented a Long Short-Term Memory (LSTM) neural network to learn long-term temporal patterns and dependencies.

The three models achieved directional accuracies ranging from 43% to 61%, which are close to random guessing in terms of predictive capability. However, when the same features were integrated into an investment strategy, the resulting portfolio achieved Sharpe ratios between 1 and 6, significantly outperforming the baseline. This research aims to explore why the models themselves exhibited near-random predictive performance, yet produced meaningful results when applied within a trading strategy framework, and to identify potential avenues for improving the strategy moving forward.

2 Data Sources and Acquisition

2.1 Price and Volume Data

Daily OHLCV (Open, High, Low, Close, Volume) data for BTC-USD was obtained from Yahoo Finance via the yfinance Python library, covering 2015 to 2025 (3,929 days). This represents spot exchange-traded Bitcoin, and the volume metric is total volume summed across major exchanges.

2.2 On-Chain Data

Blockchain data including daily transaction count, active addresses (unique addresses participating in transactions), exchange netflow (aggregate inflow minus outflow across tracked exchanges like Binance and Bitfinex), and exchange reserves (total BTC balance held on exchange wallets) from BigQuery, spanning from 2014 to 2025. The data for reserves and flows was a bit unclean, as the on-chain exchange data used in this study represent only a subset of all exchanges and may not capture the full universe of exchange-related activity. To use the data for training, there needed to be one row per day, so the data multiple entries per day were aggregated together into one row.

3 Target Variable and Evaluation

All models predict the sign of next-day log returns:

$$y_t = \text{sign}\left(\ln\left(\frac{P_{t+1}}{P_t}\right)\right), \quad y_t \in \{0, 1\} \tag{1}$$

where $y_t = 1$ indicates an up day. Log returns were used for their additive property over time. The data was divided into train-test splits by chronological order with the model being trained on the first 80% to 90% of the time window and tested on the last 10% to 20% .

4 Feature Engineering: Common Indicators

The following features were used across multiple models. Formulas are presented once to avoid repetition.

4.1 Returns and Volatility

$$\text{log return}_t = \ln(P_t/P_{t-1}) \tag{2}$$

$$\text{Volatility}_n = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(\text{logreturn}_{t-i} - \bar{r})^2} \times 100 \tag{3}$$

4.2 Moving Averages and Crossovers

Moving Average (MA$_n$): This calculates the average price over the last n days. Intuitively, it smooths out short-term price fluctuations to reveal the underlying trend direction. If the current price is above the moving average, it generally signals an upward trend; if below, a downward trend

$$\text{MA}_n = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i} \tag{4}$$

$$\tag{5}$$

4.3 Momentum Indicators

Rate of Change (ROC_n): This measures the percentage change in price over n days. It quantifies the speed or velocity of price movement. A high positive ROC means the price is surging quickly (strong bullish momentum), while a strongly negative ROC indicates a rapid crash (strong bearish momentum). It helps identify when a trend is accelerating or losing steam.

MACD Signal: The Moving Average Convergence Divergence (MACD) signal line is a derived metric that tracks the relationship between two exponential moving averages. It acts as a sensitive ”trigger” for buy and sell signals. When the MACD line crosses above the signal line, it suggests bullish momentum is building; when it crosses below, it suggests bearish momentum. It is widely used to confirm trend strength and direction.

$$\text{ROC}_n = \frac{P_t - P_{t-n}}{P_{t-n}} \tag{6}$$

$$\text{MACD}_{\text{signal}} = \text{EMA}_{12} - \text{EMA}_{26} - \text{EMA}_9(\text{MACD}_{\text{line}}) \tag{7}$$

4.4 Oscillators

Relative Strength Index (RSI_n): This measures the magnitude of recent price changes to evaluate overbought or oversold conditions. Values range from 0 to 100. A high RSI could mean the asset is overbought and may be due for a correction or pullback. A low RSI could mean the asset is oversold and may be due for a bounce.

$$\text{RSI}_n = 100 - \frac{100}{1 + \frac{\overline{\text{Gains}_n}}{\overline{\text{Losses}_n}}} \tag{8}$$

$$\tag{9}$$

4.5 Volume-Based Indicators

On-Balance Volume (OBV_t): This is a cumulative total of volume, adding volume on ”up” days and subtracting it on ”down” days. The intuition is that volume precedes price. If OBV is rising while price is flat, it suggests ”smart money” is accumulating (buying) positions, predicting a future price breakout. If OBV is falling, it suggests distribution (selling).

$$\text{OBV}_t = \sum_{i=1}^{t} \text{sign}(P_i - P_{i-1}) \times V_i \tag{10}$$

$$\tag{11}$$

4.6 Lagged Returns

Lagged Returns (Lag_Return$_i$): These are simply the returns from i days ago. Including these allows the model to learn auto regressive patterns—how past price moves influence future ones.

For example, the model might learn that a large return yesterday often leads to a small reversal (mean reversion) today, or that a streak of positive returns tends to continue (momentum).

$$\text{Lag\_Return}_i = \ln\left(\frac{P_{t-i}}{P_{t-i-1}}\right), \quad i \in \{1, 2, \ldots, 7\} \tag{12}$$

5 Model 1: Logistic Regression

5.1 Methodology

LR models the probability of an up day via:

$$P(y_t = 1|\mathbf{x}_t) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_t)}} \tag{13}$$

The logistic regression model's features were log return, standard deviation (SD) for 7 days, SD for 30 days, log volume, volume mean ratios for 7 and 30 days, momentum for 7 and 30 days, moving average (MA) for 10 and 50 days, rate of change ROC for 7 days, relative stength index (RSI) for 14 days, and on-balance (OBV), with the training period being from 2015 to 2023 (80%) and test period being from 2023 to 2025 (20%, 794 samples).

5.2 Results

Test accuracy: 51.1%

Table 1: Logistic Regression Performance

Class	Precision	Recall	F1	Support
Down (0)	0.498	0.894	0.639	385
Up (1)	0.602	0.152	0.242	409
Accuracy		0.511		794

5.3 Failure Analysis

With the data used, if one were to guess that the model goes up every single day, they would be correct 51.5% of the time, so the model is not very good. The model exhibited three critical failures. First, it had severe prediction bias: 89.4% of true up days were misclassified as down, indicating learned weights consistently produce negative scores, likely due to negative-skewed feature distributions where large drops exceed large gains. Second, it had linear inadequacy: Bitcoin trends are nonlinear, as high volatility can predicts reversals in bull markets but continuation in crashes, and LR cannot capture these relationships. Third, it had lack of regularization, as the absence of L1 or L2 penalties allowed overfitting.

6 Model 2: Random Forest

6.1 Methodology

RF ensembles 100 decision trees via bootstrap aggregating with random feature selection. Hyperparameters: n_estimators equals 100, max_features equals log2, min_samples_leaf equals 1 (default). Two variants tested: Base model with 30 features (MA crossovers, MACD, ROC, RSI, WLPR, OBV changes, global volume, Fear and Greed Index, Volatility for 30 days, lagged returns), and On-chain model with 36 features (base plus daily transactions, active addresses, exchange netflow, exchange reserves, netflow-to-reserve ratio, 7-day transaction change, 7-day address change).

The on-chain model applied exponential sample weighting:

$$w_i = \lambda^{N-i}, \quad \lambda = 0.995 \tag{14}$$

to emphasize recent data, with oldest sample weight approximately 0.37 times newest.

6.2 Results Across Training Periods

Table 2: Random Forest Performance

Training Start	Base Accuracy	On-Chain Accuracy	RMSE (Base)
2018-06-01	46.25%	—	0.0227
2023-01-01	45.35%	44.19%	0.0274
2023-06-01	43.66%	43.66%	0.0292
2024-01-01	51.02%	44.90%	0.0345
2024-06-01	47.06%	47.06%	0.0341
2025-01-01	61.54%	53.85%	0.0257

6.3 Feature Importance

The best features varied a lot by period but consistently included lagged returns (Lag_Return_1 to 7 with importance 0.037 to 0.051), ROC for 200 days (0.041 to 0.048), and moving average convergence and divergence (MACD) (0.037 to 0.043). On-chain metrics in the expanded model included daily transactions, exchange reserves, active address count, and exchange inflow and outflow, but they did not improve accuracy. how to put an image in overleaf

6.4 Failure Analysis

Four critical issues emerged from the random forest model. First, the model had extreme regime sensitivity. The wide 43 to 61% accuracy range across different training windows suggests that the model learned period-specific patterns that failed to generalize. The 2023 post-FTX collapse time frame was dominated by fear, deleveraging, and exchange risk, whereas the 2024–2025 period was shaped by spot ETF approvals, renewed institutional interest, and different liquidity conditions. In effect, the underlying data-generating process for Bitcoin returns changed between these regimes, so a model trained on one environment no longer described the other. This instability is amplified by the fact that Bitcoin trades globally: new regulations, capital controls, taxation rules, or enforcement actions in any major jurisdiction can quickly alter trading behavior, liquidity, and correlations. As a result, the random forest repeatedly fit high-variance patterns that were specific to one macro or regulatory regime instead of learning relationships that remained stable over time.

Second, the model showed clear signs of overfitting. The best headline result, 61.54% accuracy for the 2025-01-01 training start, was obtained on a test set containing only 13 days. With such a small sample, a single additional correct or incorrect prediction shifts the reported accuracy by more than 7 percentage points, and a simple binomial calculation shows that the 95% confidence interval is roughly ±10%. In statistical terms, this means the apparent "outperformance" is indistinguishable from noise, and the model's true accuracy could easily be very close to the 50% random baseline. More generally, the forest had access to dozens of flexible splits and deep trees, but only a few hundred training observations per window, which makes it easy to memorize idiosyncratic patterns (for example, specific local runs of up or down days) without capturing any robust signal.

Third, despite high feature importance scores, on-chain metrics provided essentially no improvement in predictive accuracy, so they were redundant in practice. Exchange netflow, exchange reserves, and activity measures are conceptually attractive because they summarize supply and demand pressure on centralized venues, but in the data they are strongly correlated with simpler variables like trading volume and realized volatility. Random forests can assign high importance to features that merely duplicate information already present elsewhere, so importance scores do not guarantee incremental signal. In addition, the on-chain data suffered from quality and alignment issues: missing values were forward-filled, which effectively "freezes" some series for long stretches; exchange wallet labels are imperfect, so large cold-storage movements can be misclassified as exchange flows; and the timing of transactions on-chain (confirmed over a 10–60 minute window) does not perfectly line up with the daily close used for returns. All of these factors inject noise and temporal misalignment, making it harder for the model to learn a stable mapping from on-chain flows to next-day price direction.

Fourth, the model faced a mild curse of dimensionality. The on-chain specification used 36 features, but at each split the random forest considered only $\lfloor \log_2(36) \rfloor \approx 5$ or 6 features. This design choice reduces correlation between trees but also increases the chance that genuinely informative variables are simply not available at crucial decision points, while noisy or redundant variables drive the splits instead. Given the relatively small number of training samples in each rolling window, every additional feature effectively dilutes the available signal per dimension, encouraging the trees to fit random fluctuations rather than robust patterns.

7 Model 3: LSTM Neural Network

7.1 Architecture

LSTM processes 20-day sequences of 13 features (same as LR) through the following layers: Input (20 timesteps times 13 features), LSTM with 32 units, Dropout at 0.3, Dense with 16 units and ReLU activation, Dropout at 0.3, and Dense with 1 unit and sigmoid activation.

Training used Adam optimizer (learning rate 0.001), binary cross-entropy loss, batch size 32, early stopping (patience 20), and class weights (0: 1.07, 1: 0.94). Training period: 2015 to 2023 (3,066 sequences), validation: 2023 to mid-2024 (346 sequences), test: mid-2024 to 2025 (497 sequences).

7.2 Results

Table 3: LSTM Performance (Epoch 22, Early Stopped)

Dataset	Loss	Accuracy
Training	0.6891	52.32%
Validation	0.6920	50.87%
Test	0.6969	47.48%

Test classification showed Down class with precision 0.485, recall 0.931, F1 0.638, support 247; Up class with precision 0.261, recall 0.024, F1 0.044, support 250; and overall accuracy 0.475 over 497 samples.

7.3 Failure Analysis

LSTM performed worse than random (47.5%) with prediction bias showing 2.4% recall on up days. First, the sigmoid outputs were consistently below 0.5, exploiting training set 53.4% up-day bias, with class weighting insufficient to correct this. Also, this likely is suffering from the same issue as the random forest model– inability to generalize because of such different trends in different time periods. One may have to address this by just training and testing on short time windows but with higher frequency data (hourly).

Second, there was likely too little data for deep learning, as 3,066 training sequences is much less than typical DL requirements (tens of thousands), with dropout preventing memorization but also learning. But this much data cannot be obtained without going to drastically different time periods with drastically different policies unless hourly data is obtained. Third, the model had vanishing gradients despite LSTM design, so 20-day sequences may be too short to capture Bitcoin multi-week cycles like 2-week difficulty adjustments.

Third, having only 13 features provide limited representational capacity, as deep learning excels with high-dimensional raw inputs like tick-by-tick prices, not hand-crafted indicators.

8 Cross-Model Analysis

Table 4: Model Comparison Summary

Model	Best Accuracy	Avg Accuracy	Training Time
Logistic Regression	51.1%	51.1%	less than 1 sec
RF (Base)	61.5%	47.5%	approximately 10 sec
RF (On-Chain)	53.8%	46.7%	approximately 15 sec
LSTM	47.5%	47.5%	approximately 2 min
Random Baseline		50.0%	

9 Bitcoin Quant Strategy

9.1 Methodology

This study develops a rule-based trading strategy for the Bitcoin spot market using daily BTC–USD price data from 30 November 2014 to 30 November 2025. The target variable is the daily log return of the closing price, defined as

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right).$$

As predictors, the model uses up to five lags of past log returns, which are fed into a single-hidden-layer neural network implemented in PyTorch. The network takes the feature vector as input and produces a scalar prediction of the next day's log return, denoted by $\hat{r}_{t+1}$. A fixed random seed (42) is used throughout the training process to ensure consistent data splitting and model estimation.

The trading rule assumes a constant 1x long position in Bitcoin, regardless of the sign or magnitude of the predicted return, effectively holding BTC with a fixed weight at each point in time. All trading frictions—including taker and maker fees—are set to zero, so that the performance evaluation reflects only the theoretical returns implied by the model's forecasts.

9.2 Feature Stages

To assess the incremental value of different information sets, the strategy is evaluated under three feature stages:

- **Stage 1: Price-based Rules**
  In Stage 1, only lagged returns are used to construct basic momentum or mean-reversion rules. The strategy takes a long position when the recent k-period return is positive and a short position when it is negative.

- **Stage 2: Market Condition Features**
  In Stage 2, the feature set is expanded to include additional market variables such as trading volume, volatility, liquidity, and momentum. These features capture changing market conditions and allow the strategy to scale risk up or down as conditions evolve.

- **Stage 3: On-chain Information Integration**
  In Stage 3, on-chain data is incorporated into the signal set. Key features include active addresses, new addresses, and large wallet activity. The underlying hypothesis is that certain on-chain metrics may lead or complement price-based information, thereby improving the timing or sizing of positions.

9.3 Model training and Validation

The training sample covers the early portion of the dataset, while the final segment is reserved strictly for out-of-sample evaluation. A fixed random seed (42) ensures that the data split, parameter initialization, and training dynamics remain fully reproducible.

The neural network consists of a single hidden layer with ReLU activation. The input is the feature vector of lagged returns (additional features in later stages), and the output is a one-step-ahead forecast of the log return. The model is optimized using the Adam optimizer with a mean-squared-error loss function. Training proceeds for a fixed number of epochs with batch gradient descent.

Because the architecture is intentionally simple, no explicit regularization techniques such as dropout or weight decay are applied. The model's limited depth and the use of only lagged features help reduce the risk of overfitting.

To evaluate the model's forecasting accuracy, several performance metrics are computed on the out-of-sample period:

- **Mean Squared Error (MSE)** between predicted and actual returns.
- **Mean Absolute Error (MAE)** to assess typical prediction deviations.
- **Directional Accuracy (Hit Ratio)** measuring whether the model correctly predicts the sign of the next day's return.

9.4 Trading Strategy Evaluation

The trading strategy is evaluated under the three feature stages defined earlier. Performance is benchmarked against a passive buy-and-hold Bitcoin position.

- **Stage 1: Price-based Rules**
  As shown in Table 5, the price-based lagged return strategy using `Close log return lag 1 to 5` has a modest win rate of just above 50% and limited expected value (0.0013 to 0.0014). Although Sharpe ratios are moderate (0.79 to 0.94), maximum drawdowns are high (-38% to -45%), indicating that relying solely on short-term price changes carries significant downside risk.

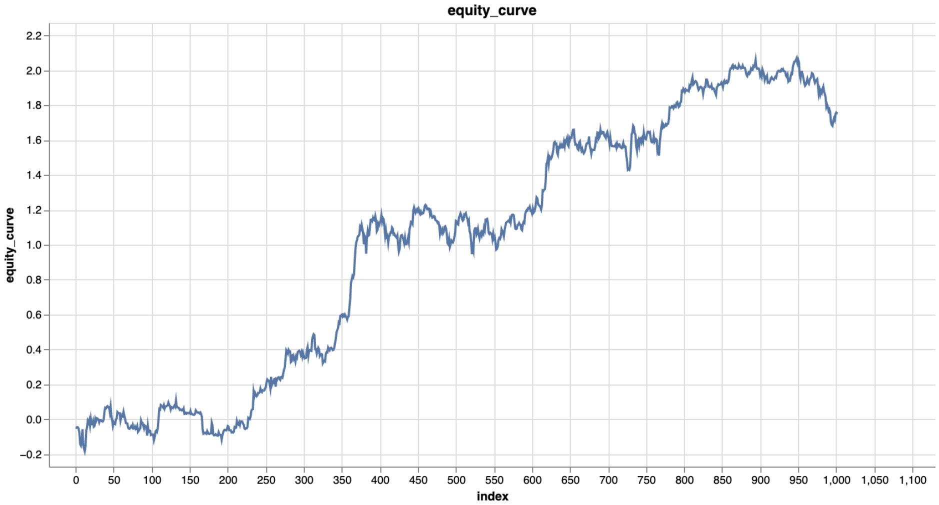

Figure 1: Effect of Equity Carve-Out Signals on Close Log Return (Lag 1)

Table 5: Price-Based Lagged Return Performance

Feature	Win Rate	Expected Value (EV)	Sharpe	Max Drawdown
Close_log_return_lag_1	0.5125	0.0014	0.9378	-0.3856
Close_log_return_lag_2	0.5055	0.0013	0.8773	-0.3878
Close_log_return_lag_3	0.5045	0.0012	0.7913	-0.4522
Close_log_return_lag_4	0.5055	0.0013	0.8773	-0.3878
Close_log_return_lag_5	0.5055	0.0013	0.8773	-0.3878

- **Stage 2: Market Condition Features**
  As shown in Table 6, feature combinations incorporating lagged returns and technical indicators achieve higher win rates (0.6095–0.6337) and expected values (0.0078–0.0085) compared to strategies based solely on price lags. While Sharpe ratios are strong (5.44–5.97) for these top-performing combinations, downside risk remains present, and strategies relying on indicators with longer-term signals show lower performance and larger drawdowns, highlighting the trade-off between risk and potential reward.

Table 6: Top 5 Performance Metrics for Selected Feature Combinations

Features	Win Rate	EV	Sharpe	Max Drawdown
Close_log_return_lag_1, bb_high, bb_low	0.6337	0.008505	5.9698	-0.118534
Close_log_return_lag_1, ROC_7d, bb_low	0.6297	0.008342	5.8396	-0.118536
Close_log_return_lag_1, Mom_7d, bb_high	0.6206	0.008171	5.7041	-0.133623
ROC_7d, bb_high, bb_low	0.6165	0.007857	5.4589	-0.106730
Close_log_return_lag_1, ROC_7d, bb_high	0.6095	0.007840	5.4453	-0.133623

- **Stage 3: On-chain Information Integration**
  As shown in Table 7, incorporating on-chain data alongside lagged returns and technical indicators does not lead to substantial improvements in the top-performing combinations. Although some metrics such as Sharpe ratios and expected values are comparable to Stage 2 (5.42–5.98 and 0.0078–0.0085, respectively), the same features appear with minor variations, and none of the top 5 combinations are dominated by purely on-chain indicators. This suggests that on-chain features, while potentially informative, are not among the most predictive signals for this dataset, and their contribution is mostly context-dependent rather than directly enhancing performance.

Table 7: Top 5 Performance Metrics for Selected Feature Combinations

Features	Win Rate	EV	Sharpe	Max Drawdown
Close_log_return_lag_1, bb_high, bb_low	0.6276	0.008514	5.9773	-0.124473
ROC_7d, bb_high, bb_low	0.6085	0.008043	5.6035	-0.106730
Close_log_return_lag_1, Mom_7d, bb_high	0.6095	0.007925	5.5117	-0.133623
Close_log_return_lag_1, ROC_7d, bb_high	0.6065	0.007905	5.4959	-0.133623
Close_log_return_lag_1, ROC_7d, RSI_14d	0.6236	0.007804	5.4177	-0.168461

10 Limitations

Data limitations include survivorship bias (Bitcoin is the only 2015-era cryptocurrency still dominant), daily aggregation missing intraday volatility (HFT operates on millisecond scales), on-chain data quality issues (forward-filling introduces staleness, exchange coverage inconsistent over time), and outliers (extreme events like negative 50% flash crashes dominate training but are rare).

Model limitations include no transaction costs (55% accuracy may be unprofitable after 0.1 to 0.5% fees per trade), binary simplification (predicting direction ignores magnitude), no uncertainty quantification (confidence intervals not provided except LSTM probabilities which are poorly calibrated), and independent predictions (path of returns ignored).

Experimental design limitations include no cross-validation (single train-test split has high variance; walk-forward analysis needed), no hyperparameter optimization (default parameters likely suboptimal), and no ensemble stacking (combining LR, RF, LSTM predictions might improve robustness).

11 Recommendations for Future Work

11.1 Model Improvements

First, use gradient boosting (XGBoost or LightGBM), which typically outperforms RF on tabular data and handles missing values natively. Second, implement regime detection by training separate models for bull, bear, and ranging markets, using a meta-model to select which applies. Detect regimes via Hidden Markov Models on returns, moving average slopes (200-day MA direction), or volatility quantiles.

Third, add attention mechanisms: Transformer architectures like Temporal Fusion Transformer learn which historical days matter most, avoiding LSTM fixed 20-day window. Fourth, increase LSTM sequence length to test 60 or 90-day windows to capture longer cycles.

11.2 Alternative Problem Formulations

First, multi-class classification: predict magnitude buckets like large up (greater than 2%), small up (0 to 2%), small down (0 to negative 2%), large down (less than negative 2%), focusing model on actionable large moves. Second, volatility prediction: forecast realized volatility (Parkinson high-low estimator) or VIX-style implied volatility proxies, which are often more predictable than direction.

Third, multi-horizon prediction: simultaneously predict 1, 3, and 7-day returns via multi-task learning to capture different timescale patterns. Fourth, quantile regression: predict 10th, 50th, and 90th percentile returns to quantify uncertainty.

11.3 Enhanced Features

First, high-frequency data: use minute-level OHLCV to capture intraday patterns (requires institutional data access). Second, order book data: bid-ask spread, depth at best bid or ask, order imbalance (buy to sell ratio), requiring exchange API access (not available historically).

Third, social media sentiment: Twitter or Reddit NLP via FinBERT or sentiment transformers (requires scraping and GPU compute). Fourth, macro indicators: Fed funds rate, DXY dollar index, S and P 500 returns, gold prices, 10-year Treasury yields, as Bitcoin increasingly correlates with risk-on or risk-off regimes.

Fifth, funding rates: perpetual swap funding rates from Binance or Bybit indicate leveraged positioning, with positive rates suggesting overleveraged longs (bearish). Sixth, Google Trends: search volume for terms like Bitcoin, buy Bitcoin, or Bitcoin crash as retail sentiment proxy.

12 Conclusion

Three machine learning paradigms (Logistic Regression as linear, Random Forest as nonlinear ensemble, and LSTM as deep sequential) were applied to Bitcoin directional prediction using 13 to 36 engineered features. Despite comprehensive feature engineering including technical indicators, sentiment scores, and blockchain fundamentals, all models achieved 43 to 61% accuracy with average approximately 48%, indistinguishable from random.

Key findings: lagged returns consistently ranked as most important features but provided insufficient predictive power; on-chain data added no value, likely due to noise, data quality issues, and redundancy with volume; LSTM underperformed simpler models due to data scarcity (3,066 sequences insufficient for deep learning); and regime sensitivity was pervasive, with accuracy ranging from 43 to 61% across training periods with no consistent improvement.

The consistent failure suggests Bitcoin daily direction may be fundamentally unpredictable from historical data alone, as price is driven by exogenous news events. Future work should explore higher-frequency data to capture microstructure effects, regime-aware modeling to handle non-stationarity, alternative targets like volatility prediction, and multi-horizon forecasts to exploit different timescale patterns. Most critically, models must be evaluated on economic metrics (Sharpe ratio, maximum drawdown) rather than directional accuracy, as transaction costs can render even 55% accurate models unprofitable.

13 Non-Technical Report

13.1 Cryptocurrency Introduction and Market Overview

Cryptocurrency is a form of digital token currency that allows people to directly make payments to one another online, using cryptography to ensure secure payments and digital ledgers called ‘blockchains’ to document and verify all transactions.

This blockchain ledger serves as the backbone for all cryptocurrencies by creating a decentralized way of storing a permanent history of every single transaction. So what exactly is this chain of blocks? When cryptocurrency transactions occur, they are recorded and grouped together into blocks, with each block having its own unique cryptographic hash and timestamp. This hash serves as a digital fingerprint to secure the block, and the timestamp allows blocks to be arranged in chronological order into a chain. The blockchain is not stored in one space, but rather has copies on many different computers called ‘nodes’, preventing attackers from compromising the system from any one point. This means that the entire blockchain is visible to all users, allowing individuals to ensure the validity of blocks being added and decreasing the need for centralized intermediaries like banks, all while ensuring user privacy by making transaction-makers anonymous. Additionally, the lack of a central authority controlling transactions has made cryptocurrency popular in areas with government-imposed restrictions for people seeking greater financial freedom.

The cryptocurrency blockchain idea was developed in 2008 by a pseudonymous person or group of people named Satoshi Nakamoto when they published “Bitcoin: A Peer-to-Peer Electronics Cash System,” and then Nakamoto proceeded to ‘mine’ the first Bitcoin in 2009. When a transaction is made by the general public, it is first placed into a pool of unverified transactions. Cryptocurrency mining is the process in which people called ‘miners’ compete to verify these cryptocurrency transactions by trying to find a hash that is deemed to be valid by the Bitcoin network. The first miner that finds a hash that is deemed valid is given the ability to add the block to the chain and is rewarded with a chunk of Bitcoin. Through this process, Bitcoin has grown greatly since 2009, and many other tokens have entered the world by replicating this process.

Stablecoins are a particular type of cryptocurrency designed to solve the problem of extreme price volatility in the crypto market. Instead of allowing for a heavily fluctuating price as Bitcoin does, stablecoins aim at maintaining the same value as a government-issued currency, such as the U.S. dollar, by pegging to it. This means that one unit of a U.S. dollar-pegged stablecoin, like USD Coin, is intended to fluctuate so that it will always be worth approximately one U.S. dollar. To ensure that they are capable of maintaining this one-to-one transaction value, stablecoin issuers hold liquid assets like cash and short-term government securities. This structure allows traders to effectively exchange the USD value around the world through a blockchain without needing to go through a bank. Because of their stability, stablecoins now play a large role in the broader cryptocurrency world. Many trading pairs on cryptocurrency exchanges are quoted against stablecoins rather than against traditional fiat currencies, meaning that investors move in and out of risky positions using Tether or USD Coin as their ‘cash’ on the exchange. Stablecoins are also used in decentralized financial services that run on public blockchains, where users can use them as collateral, earn interest on them through lending, or use them as a medium of exchange. They effectively function as a bridge between fiat money and volatile cryptocurrency by allowing users to hold something close to dollars while still residing within the blockchain-based financial system.

The current global cryptocurrency market capitalization is estimated to fluctuate between $3-4 trillion USD with a daily trading volume of approximately $100 billion USD. Millions of different cryptocurrency tokens have been created, however there are a few tokens that account for most of this market capitalization and trading activity, including Bitcoin, Ethereum, Tether, USD Coin, and Binance Coin. Bitcoin has been the largest cryptocurrency since its inception with a market capitalization of about $1.9 trillion, and Ethereum follows in second place with about $380 billion. While the market capitalization is relatively small in comparison to the SP 500 and the US Treasury debt market, it is still large enough to make a substantial impact in firms and governments trading and investment decisions.

Circle Internet Financial is a global financial technology firm that sits at the center of this stablecoin revolution as the issuer of USD Coin (USDC). Unlike decentralized stablecoins governed by algorithms, Circle operates on a fully reserved model, meaning that every USDC token in circulation is backed 1:1 by cash and short-term U.S. Treasury bonds held in regulated financial institutions. This structure allows Circle to bridge the traditional financial system with the emerging blockchain economy, providing a trusted digital dollar that can be moved globally 24/7 with the speed of the internet. By prioritizing transparency and regulatory compliance (publishing monthly attestation reports on its reserves) Circle has positioned USDC as a key infrastructure layer for digital commerce, enabling businesses to settle payments, manage treasury operations, and access decentralized finance (DeFi) markets with reduced counterparty risk.

Beyond issuance, Circle provides a suite of Web3 services that help developers and enterprises integrate blockchain technology into their operations. These include programmable wallets, cross-chain transfer protocols that allow USDC to move seamlessly between different blockchains like Ethereum, Solana, and Avalanche, and smart contract management tools. Essentially, Circle functions not just as a digital central bank for the crypto economy but as a platform company, building the base that allows traditional dollars to operate natively on the internet. As stablecoins increasingly facilitate real-world economic activity, from cross-border remittances to humanitarian aid, Circle's role has expanded from serving crypto traders to becoming a critical node in the future of global money movement.

13.2 Regulatory Landscape: The Genius Act and Compliance

Before the GENIUS Act, stablecoins were regulated as money transfer institutions, and mainly at the state level. There was no single federal framework for stablecoins. Stablecoin issuers like Circle had to operate inside a patchwork of old rules that were not made for crypto assets. They were mostly treated like money transfer businesses. They had to get different licences from different states and register for different networks and systems at the federal level.

On July 18, 2025, the President of the United States signed the GENIUS Act into law, which prioritizes consumer protection, strengthens the US dollar's reserve currency status, and bolsters US national security. The GENIUS Act is the first federal regulatory system for stablecoins. The integration of the stablecoin frameworks at both the state and federal levels ensures the fairness and consistency of regulation across the entire country.

The GENIUS Act ensures the stability and trust of stablecoins through strong reserve requirements, which require a 100% reserve fund must be set aside as a guarantee, either in US dollars or in short-term Treasuries. Moreover, the issuer is required to publicly disclose the composition of the reserve fund on a monthly basis. Furthermore, the issuers of stablecoins must adhere to strict marketing regulations to protect consumers from fraudulent activities. It is particularly important that they refrain from making any misleading statements, claiming that their stablecoins are supported by the US government, insured by the federal government, or have the status of legal tender.

By regulating and registering stablecoin issuers and collaborating with the Ministry of Finance to implement sanctions measures, the crackdown on illegal activities in digital assets is strengthened, enhancing national security. In the GENIUS Act, it is clearly stipulated that the issuers of stablecoins must comply with the relevant provisions of the Bank Secrecy Act, explicitly requiring the establishment of effective anti-money laundering and sanctions compliance procedures, including risk assessment, verification of sanctions lists, and customer identity verification etc. This enhances the ability of the Treasury Department to combat illegal stablecoin activities and improves its law enforcement capabilities in terms of sanctions evasion and money laundering suppression. All stablecoin issuers, like Circle, must possess the technical capability to seize, freeze, or burn the payment stablecoins when required by law, and must comply with relevant legal instructions to carry out such operations.

At the same time, the GENIUS Act not only has an impact in the US, but it has also made the US a leader globally, which enhances global regulatory coordination. It ensures that stablecoins denominated in US dollars are of the highest standard. This measure strengthens the position of the US dollar. Some people consider this measure a strategic move for the economy and peripheral politics.

The impact of the GENIUS Act on stablecoin issuers, especially Circle, is a positive influence. It's good for Circle because it plans long-term and works closely with regulators under a more regulated and trusted environment. The federal law raises the standard for everyone, and due to Circle already making everything transparent even before the law, this puts them in a more competitive situation among all the issuers.

13.3 Outlook for Circle and the Stablecoin Market

With regulatory clarity increasingly delivered through the 2025 legislation, stablecoins are evolving from speculative crypto-assets into institutional-grade infrastructure. Rather than serving primarily as high-volatility trading instruments or speculative assets, stablecoins are now gaining adoption as a form of programmable, digitally native dollar liquidity. As stablecoin usage spreads into corporate treasury operations, cross-border payments, and decentralized finance, the broader financial system is beginning to incorporate stablecoins as part of its structural plumbing.

Despite the excitement, the growth trajectory is not without constraints. J.P. Morgan, in a recent analysis, expressed caution about overly optimistic forecasts. While some market participants project stablecoin supply reaching multitrillion-dollar levels by 2028, J.P. Morgan argues that such predictions may overestimate mainstream adoption. Their assessment underscores that payments adoption remains limited among conservative liquidity investors and stablecoin activity is still mainly in novel environments rather than broad-based retail or institutional payments, something that will take time to fully develop. Accordingly, J.P. Morgan suggests a more modest market cap ceiling of approximately $500 billion to $700 billion.

Nonetheless, even a conservative scenario implies significant structural impact. Stablecoins may reinforce dollar demand, cement the dollar's global role, and provide a digital payment and settlement infrastructure that augments current fiat-rail systems. In that context, private firms that have built stablecoin issuance on transparent, fully-reserved models and have robust infrastructure are well positioned to benefit.

Circle is a prominent example of such a firm. Its Q3 2025 financial results offer a concrete demonstration of how reserve-backed stablecoin issuance can be both scalable and profitable under favorable macro conditions. As of the end of Q3, Circle reported USDC circulation of $73.7 billion, a 108% year-over-year increase. In the same quarter, the company generated $740 million in total revenue and reserve income, marking a 66% increase from the prior year period. Its net income reached $214 million, a 202% year-over-year rise, while adjusted EBITDA rose to $166 million, up 78% YoY, yielding an adjusted EBITDA margin of roughly 57%.

These results reflect the inherent leverage in the "narrow-bank" stablecoin model: Circle holds fully reserved assets, like Treasurys and cash, backing USDC liabilities. It earns yield on those assets and does not pay interest on the circulating stablecoins. The business benefits significantly when interest rates remain elevated and demand for digital liquidity increases.

Moreover, Circle's revenue diversification is visible. Besides reserve income of $711 million, it earned "other revenue", about $29 million, from subscriptions, services, and transaction-related fees.

The significance of these results is magnified when viewed against Circle's broader strategic expansion. The firm is not simply issuing a stablecoin, but building a full-stack infrastructure for digital-dollar mobility. Its efforts include cross-chain transfer protocols (CCTP), programmable wallets, and the development of a blockchain platform (Arc) intended for tokenized financial workflows. As stablecoins become more deeply embedded in global finance, the demand for data-driven insights into on-chain behavior, liquidity flows, and risk exposure will only intensify.

13.4 Data Science Career at Circle

The role of Staff Data Scientist at Circle is part of the foundation for governance, compliance, product design, and institutional credibility. Professionals in this role must analyze high-dimensional blockchain data such as transaction graphs, cross-chain transfers, wallet behaviors, liquidity distribution, and flow dynamics. They must model token distribution, identify anomalous activity, assess protocol-level risk, and anticipate how shifts in macroeconomic or market conditions affect stablecoin circulation and reserve utilization.

According to Circle's actual job postings, Staff Data Scientists require deep expertise in blockchain-specific analysis, typically 6+ years of data science experience with at least 3 years focused explicitly on blockchain data. This specialization distinguishes the role from traditional financial data science positions. While conventional quantitative analysts at banks or asset managers work with price series, trading volumes, and balance sheet data, Circle's data scientists must maintain labels and tags for on-chain addresses, understand stablecoin mechanics at the protocol level, work with blockchain analytics platforms like Chainalysis, TRM Labs, Arkham, and Elliptic, and conduct transaction tracing across multiple blockchain networks simultaneously. The work involves not merely analyzing historical patterns but building real-time monitoring frameworks that track USDC dispersion across Ethereum, Solana, Avalanche, Polygon, and emerging chains, while simultaneously assessing competitive dynamics as rival stablecoins like Tether's USDT and emerging algorithmic alternatives compete for market share.

The technical demands extend beyond standard statistical modeling. Circle's data scientists partner with blockchain researchers to conduct network analysis, anomaly detection, and blockchain forensics. They address questions spanning Circle's Cross-Chain Transfer Protocol (CCTP) (which enables efficient USDC movement between blockchains), programmable wallet infrastructure, and chain expansion strategies. Each of these products serves distinct stakeholder groups—developers building decentralized applications, exchanges facilitating liquidity, protocols integrating stablecoin functionality, and individual users seeking dollar-denominated stability in volatile crypto markets. Understanding usage patterns, conversion funnels, and behavioral economics across these diverse constituencies requires a hybrid skill set that fuses computer science, statistics, domain expertise in distributed systems, and financial expertise.

Compensation reflects this specialized expertise, as data scientists in blockchain and cryptocurrency startups earn an average of $119,583, approximately 31.5% higher than the $90,917 average for data scientists across all industries within the crypto sector. Senior and staff-level positions at Circle likely command salaries ranging from $130,000 to $205,000, with total compensation packages including equity that can appreciate substantially if the stablecoin market continues its institutional adoption trajectory. This premium compensates not only for technical complexity but also for the rapid pace of innovation, regulatory uncertainty, and the reputational risks inherent in an industry still establishing legitimacy within traditional financial circles.

Comparatively, data scientists at traditional financial institutions likeGoldman Sachs, JPMorgan Chase, BlackRock work within established regulatory frameworks and leverage decades of historical data to optimize portfolio allocation, credit risk assessment, and fraud detection. Their models are typically supervised learning systems trained on structured datasets like credit bureau records, transaction histories, macroeconomic indicators, and market microstructure data. While sophisticated, these environments offer greater predictability. Regulatory reporting standards are codified, data lineage is traceable through audited systems, and validation frameworks follow industry-standard practices developed over decades. In contrast, Circle's data scientists must navigate evolving blockchain protocols where smart contract bugs can lock millions of dollars, where regulatory guidance remains ambiguous across jurisdictions, and where data integrity depends on correctly interpreting on-chain events that lack standardized schemas.

The closest analogs to Circle's data science function may be found in other fintech companies like Stripe, Plaid, Tether, or Coinbase, where real-time transaction monitoring, fraud prevention, and compliance automation drive product integrity. However, even within fintech, there are gradations. Stripe's data scientists optimize payment routing, detect fraudulent merchants, and model credit risk for its lending products, but they operate within traditional payment rails where settlement is governed by card networks and banking regulations. Plaid's data scientists analyze consumer financial data aggregated from bank accounts, building models to predict cash flow, assess affordability for lending decisions, and flag anomalous spending patterns. Coinbase's data scientists focus on cryptocurrency exchange dynamics: order book liquidity, price discovery mechanisms, user acquisition funnels, and regulatory compliance for a trading platform. Circle occupies a unique niche, operating the infrastructure layer itself—USDC is not merely traded on Circle's platform but issued, redeemed, and tracked across dozens of independent blockchains where Circle has no direct control over user behavior, only visibility through public ledgers.

These tasks draw directly on the mathematical and statistical foundations typical of a graduate program in mathematics, statistics, or data science. Techniques such as stochastic modeling inform how token flows might behave under stress scenarios. For example, one might be simulating USDC redemption cascades during market panics or modeling liquidity shocks when major exchanges experience outages. Graph theory provides the framework for analyzing transaction networks: identifying centrality measures to detect influential wallets, detecting community structures that reveal ecosystem clustering (DeFi protocols, centralized exchanges, retail wallets), and tracing fund flows to satisfy anti-money laundering (AML) requirements. Network analysis extends to cross-chain dynamics, where data scientists must understand how liquidity fragments across blockchains and how bridging mechanisms introduce systemic risk. Machine learning offers the conceptual understanding and technical skills necessary to build predictive models for anomaly detection like flagging wash trading, Sybil attacks, or sanctions evasion. ML also helps for classification systems for address labeling (exchange, DeFi protocol, mixer, personal wallet), and time-series forecasting for reserve utilization and circulation patterns.

Moreover, because stablecoin infrastructure must now meet comprehensive regulatory standards under frameworks like the Markets in Crypto-Assets Regulation (MiCA) in Europe and forthcoming U.S. legislation, data scientists at Circle also contribute to compliance, transparency, and risk-management frameworks. They build automated monitoring systems that flag transactions potentially linked to sanctioned entities, leveraging tools from specialized blockchain forensics firms. They design dashboards that provide regulators with real-time visibility into reserve composition, collateralization ratios, and redemption rates. They conduct stress testing by simulating extreme market scenarios: a sudden 20% drop in Treasury bond prices affecting Circle's reserve holdings, or a surge in redemption requests exceeding daily operational capacity. Their work thus supports not only product innovation but also operational integrity and regulatory trust. This dual mandate of advancing blockchain-native financial infrastructure while satisfying traditional financial oversight defines the unique professional identity of data scientists in the stablecoin sector.

The intellectual challenges are notable too. Unlike traditional finance where historical correlations often persist, blockchain ecosystems exhibit regime changes that can invalidate models rapidly. A new blockchain gaining adoption can shift USDC liquidity distribution within weeks. A protocol exploit on one chain can trigger cross-chain contagion as users flee to perceived safe havens. Regulatory announcements such as a jurisdiction banning stablecoin usage or mandating reserve audits can reshape user behavior overnight. Data scientists at Circle must therefore build models with regime-aware architectures, incorporating structural breaks, deploying ensemble methods that weight recent data more heavily, and maintaining continuous monitoring systems that trigger alerts when model assumptions are violated. The pace of iteration far exceeds that of traditional finance, where quarterly model recalibrations are standard. Circle's environment may demand weekly or even daily model updates during periods of market turbulence.

Professional development in this role also differs markedly from traditional data science careers. At a bank, a data scientist might specialize deeply in credit risk or market risk, building expertise over years within a stable regulatory and institutional context. At Circle, the rapid expansion of blockchain infrastructure means data scientists must become generalists across a widening array of chains, protocols, and use cases. They must learn Solidity to audit smart contracts, understand Ethereum's EIP-1559 fee mechanism to model transaction costs, grasp Solana's Proof-of-History consensus to analyze transaction throughput, and monitor emerging layer-2 solutions like Arbitrum or Optimism that introduce new liquidity fragmentation dynamics. Continuous learning is not optional but essential, as the competitive landscape evolves with each new blockchain launch and each protocol upgrade.

Mentorship and collaboration structures at Circle reflect this interdisciplinary demand. Staff Data Scientists partner with and mentor teams of analysts and junior data scientists, fostering a culture where technical rigor meets pragmatic business impact. They work closely with blockchain researchers- specialists who focus on protocol-level analysis, cryptographic security, and network performance- to translate academic findings into operational intelligence. They collaborate with product managers to define metrics that measure success for new features like programmable wallets or CCTP adoption. They interface with compliance teams to ensure that anomaly detection models meet AML/KYC standards. And they engage with executive leadership to communicate complex technical risks in accessible terms, supporting strategic decisions about chain expansion, partnership priorities, and reserve management. This cross-functional integration distinguishes Circle's data science function from siloed analytics teams in traditional finance, where regulatory, risk, and product functions often operate in parallel with limited interaction.

In sum, Circle's Q3 2025 results and strategic posture illustrate a plausible, near-term scenario in which stablecoins come to represent a meaningful layer of global financial infrastructure that parallels traditional banking, yet operates on programmable blockchain rails. The data science challenges at Circle are emblematic of a broader transformation in financial services where decentralized systems, real-time settlement, and programmable money introduce complexities that exceed those of legacy banking. Whether the stablecoin ecosystem expands to the lofty projections advanced by some remains uncertain; mainstream adoption and utility beyond crypto-native contexts remain constraints. However, even under modest assumptions, the institutionalization of stablecoins appears likely to reshape dollar liquidity, cross-border payments, and the architecture of digital finance. For firms like Circle and well-trained quantitative analysts seeking to shape the next generation of financial infrastructure, the coming years promise both challenge and opportunity. Those entering this field will find themselves at the intersection of cryptography, economics, data science, and regulatory innovation—a rare confluence of disciplines that defines the frontier of 21st-century finance. The professional rewards extend beyond compensation to include the intellectual satisfaction of solving unsolved problems, the reputational capital of establishing best practices in an emerging industry, and the potential societal impact of building infrastructure that could increase financial inclusion, reduce remittance costs, and enable programmable economic systems at global scale.

13.5 Conclusion

The cryptocurrency ecosystem has begun to mature from decentralization, speculation, and technical experimentation into a regulated and increasingly institutionalized segment of global finance. The foundational mechanics of blockchain technology introduced a novel architecture, but the emergence of stablecoins has translated this into economically meaningful infrastructure. By enabling the circulation of a digitally native dollar, stablecoins occupy a unique link between financial markets and distributed technology, offering liquidity, programmability, and international accessibility.

The passage of the GENIUS Act marked a decisive regulatory shift from ambiguity to formal oversight. The imposition of reserve, disclosure, and compliance requirements established stablecoins and legally recognized financial instruments and set a national standard for operational integrity. This framework strengthens consumer protections and national security in addition to enhancing international confidence in U.S.-denominated digital assets. Thus, legislation supports broader trends in dollarization and cements the role of regulated issuers like Circle.

Circle's trajectory exemplifies how private-sector actors are responding to and helping to shape this new landscape. The firm's 2025 financial performance demonstrates the scalability of a fully reserved stablecoin model and the economic durability of narrow-bank-like structures in a high-rate environment. Its continued expansion the the Circle Payments Network, CCTP, programmable wallets, and the Arc chain signals that the future of stablecoins lies in both the issuance and creation of programmable financial rails capable of supporting institutional payments, tokenized assets, and multi-chain liquidity systems.

This evolution elevates the importance of rigorous data science within the digital-asset ecosystem. As stablecoins become more deeply integrated, the analytical demands associated with monitoring system behavior, identifying risks, and informing strategic decisions expands correspondingly. The Staff Data Scientist role at Circle illustrates how advanced quantitative strategy can be applied to the pressing problems shaping the next generation of monetary systems. Such work is now fundamental to ensuring the stability, transparency, and utility of digital-dollar infrastructure operating under federal oversight.

Ultimately, the maturation of stablecoins reflects a broader transformation in how value moves globally. Whether stablecoins eventually reach multitrillion-dollar scale or grow more gradually, the infrastructure being built today will meaningfully influence payment systems, liquidity networks, and cross-border financial integration. Circle's emergence as a key architect of this environment illustrates both the opportunities and responsibilities facing regulated stablecoin issuers in the years ahead. The capacity to bridge technical, regulatory, and analytical domains will be essential and the institutions and professionals who steward stablecoins will play a defining role in the outlook of digital finance.

14 References

References

[1] Baur, D. G., & Dimpfl, T. (2021). The volatility of Bitcoin and its role as a medium of exchange and a store of value. *Empirical Economics, 61*(5), 2663–2683. doi:10.1007/s00181-020-01990-5.
PMID: 33424101; PMCID: PMC7783506.

[2] Bakas, D., Magkonis, G., & Oh, E. Y. (2022). What drives volatility in Bitcoin market? *Finance Research Letters, 50*, 103237. doi:10.1016/j.frl.2022.103237

[3] Crypto.com. (n.d.). Bitcoin volatility: What causes BTC price swings? Crypto.com. Retrieved December 9, 2025, from `https://crypto.com/us/bitcoin/why-is-bitcoins-price-volatile?utm_source`

[4] Nguyen, K. Q. (2024, June 30). Why are cryptocurrencies so volatile? *ACFR Working Paper.* Retrieved from `https://acfr.aut.ac.nz/data/assets/pdf_file/0004/926140/Why-are-cryptocurrencies-so-volatile-03-Jul2024.pdf`

[5] Gradojevic, N., Kukolj, D., Adcock, R., & Djakovic, V. (2023). Forecasting Bitcoin with technical analysis: A not-so-random forest? *International Journal of Forecasting, 39*(1), 1–17. doi:10.1016/j.ijforecast.2021.08.001

[6] Sebastião, H., & Godinho, P. (2021). Forecasting and trading cryptocurrencies with machine learning under changing market conditions. PMC Article. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7785332/`