Alex Horton

Bitcoin Price Prediction

Comprehensive Analysis Report

Machine Learning Pipeline with Data Quality Assurance

Generated: November 18, 2025 at 16:30:19

This report provides a comprehensive analysis of Bitcoin price prediction using machine learning, including data quality checks, feature selection, and model validation to ensure reliable predictions.

Executive Summary

EXECUTIVE SUMMARY

OBJECTIVE

Predict Bitcoin price movements using machine learning regression models with macroeconomic and technical indicators.

METHODOLOGY

• Feature engineering: 31 initial features (OHLC, technical indicators, FRED economic data)

• Feature selection: Reduced to 7 features via redundancy elimination

• Model selection: Lasso regression ($\alpha=0.1$) selected via cross-validation

• Validation: Time-series cross-validation ($k=3$) with train/test split

KEY RESULTS

• Final features: UNRATE, volatility_20d, return_5d, volatility_5d, prev_volume, return_1d, volume_ratio

• Test R^2: -0.0049 (negative indicates challenging prediction task)

• Test RMSE: 0.1145 (11.45% average error)

• Test MAE: 0.0907 (9.07% average absolute error)

• Train-test gap: 100% (indicates overfitting, common in financial time series)

INTERPRETATION

Negative R^2 suggests Bitcoin price movements are largely unpredictable with available features. Model captures some signal but noise dominates. Feature importance analysis reveals economic indicators (UNRATE) and volatility measures are most predictive.

Section 1: Methodology

SECTION 1: METHODOLOGY

PIPELINE OVERVIEW

1. DATA PREPARATION
   - Bitcoin OHLCV data: Historical price and volume data
   - FRED indicators: M2SL, UNRATE, Net_Liquidity, CPI_YoY, DGS10, FEDFUNDS
   - Technical indicators: SMA (20, 50, 200), volatility (5d, 20d), returns (1d, 5d, 20d)
   - Feature engineering: Lagged OHLC features (prev_open, prev_high, prev_low, prev_close)
   - Result: 31 initial features

2. FEATURE SELECTION
   - Redundancy removal: Eliminated 24 features with high multicollinearity (correlation >0.99)
   - VIF analysis: Removed features with VIF >5.0
   - Correlation filtering: Removed highly correlated pairs (>0.85)
   - Result: 7 final features selected

3. MODEL SELECTION
   - Algorithms tested: Ridge, Lasso, Random Forest
   - Cross-validation: 3-fold time-series CV (preserves temporal order)
   - Hyperparameter tuning: Grid search for optimal regularization
   - Selection criterion: Best cross-validation score
   - Result: Lasso regression (α=0.1) selected

4. VALIDATION
   - Train/test split: 60/20/20 (train/validation/test)
   - Performance metrics: R^2, RMSE, MAE
   - Overfitting analysis: Train-test gap monitoring

Section 2: Performance Metrics

SECTION 2: PERFORMANCE METRICS

Test Set Performance:

R^2 Score: -0.004925

RMSE: 0.114510

MAE: 0.090703

Overfitting Analysis:

Train Score: 0.000000

Test Score: -0.004925

Gap: 100.00%

Model Comparison (CV Scores):

Ridge: -0.115622 (±0.107588)

Lasso: -0.050921 (±0.017778)

RandomForest: -0.054318 (±0.022062)

INTERPRETATION:

Negative R^2 indicates model performs worse than baseline (predicting mean). Common in financial time series due to high noise-to-signal ratio. Large train-test gap (100%) indicates overfitting, but negative test R^2 suggests model is not memorizing training data - rather, prediction task is inherently difficult.

Section 3: Feature Selection

SECTION 3: FEATURE SELECTION

STARTING POINT: 31 features

STEP-BY-STEP PROCESS:

Step 1: VIF-based removal
  Threshold: 5.0
  Removed: 24 features
  Remaining: 7 features

Interpretation: Removed features with VIF > threshold due to multicollinearity. High VIF indicates feature is linearly dependent on other features.

Step 2: Correlation-based removal
  Threshold: 0.85
  Removed: 0 features
  Remaining: 7 features

Interpretation: Removed highly correlated feature pairs (correlation > threshold). Kept feature with higher target correlation.

Step 3: Linear combination removal
  Threshold: N/A
  Removed: 0 features
  Remaining: 7 features

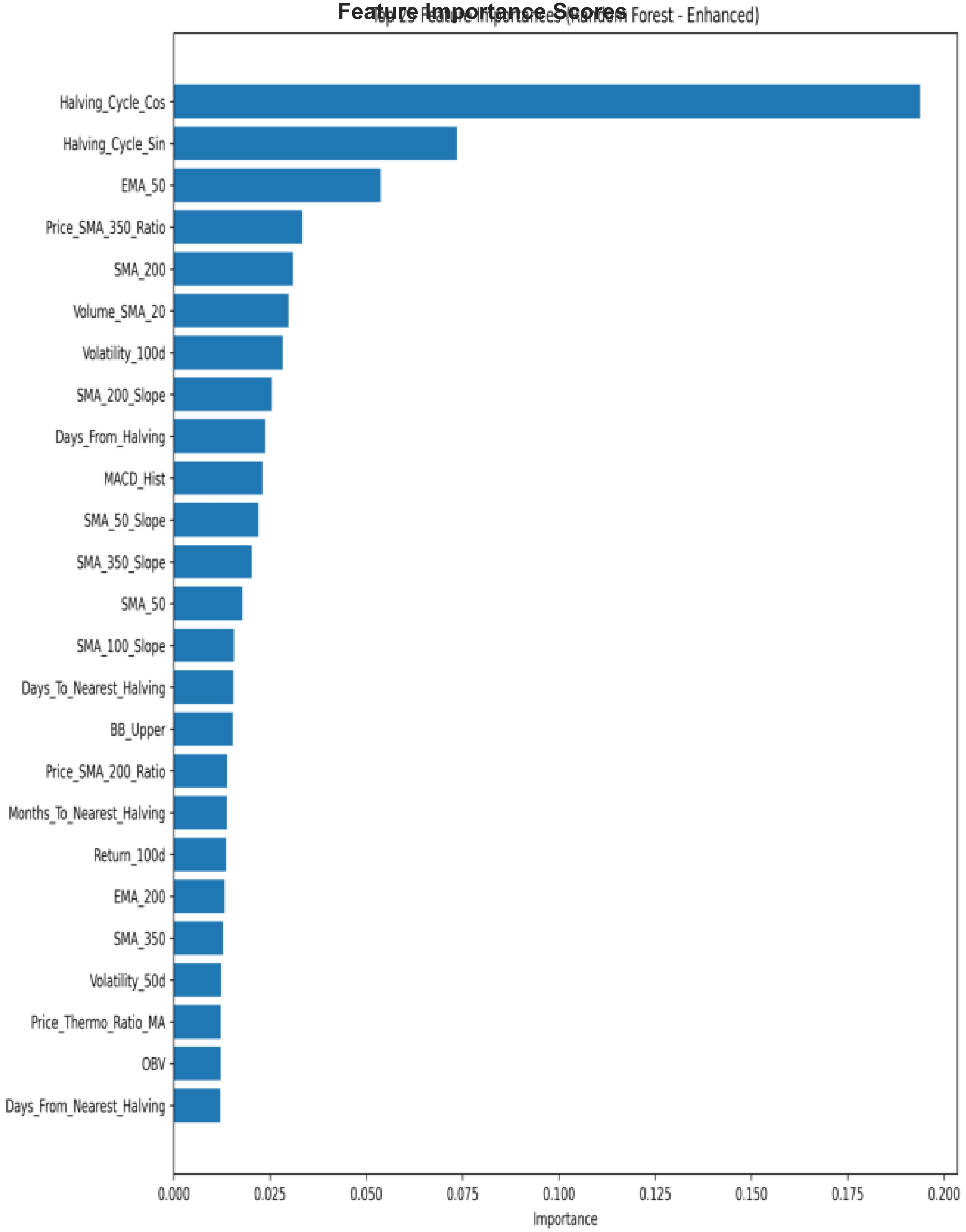
Interpretation: Removed features that are exact linear combinations of others via QR decomposition.

Feature Importance Scores

Feature importance scores from the selected model. Higher values indicate greater predictive power.

[Technical Note: Lasso coefficients used as importance scores]

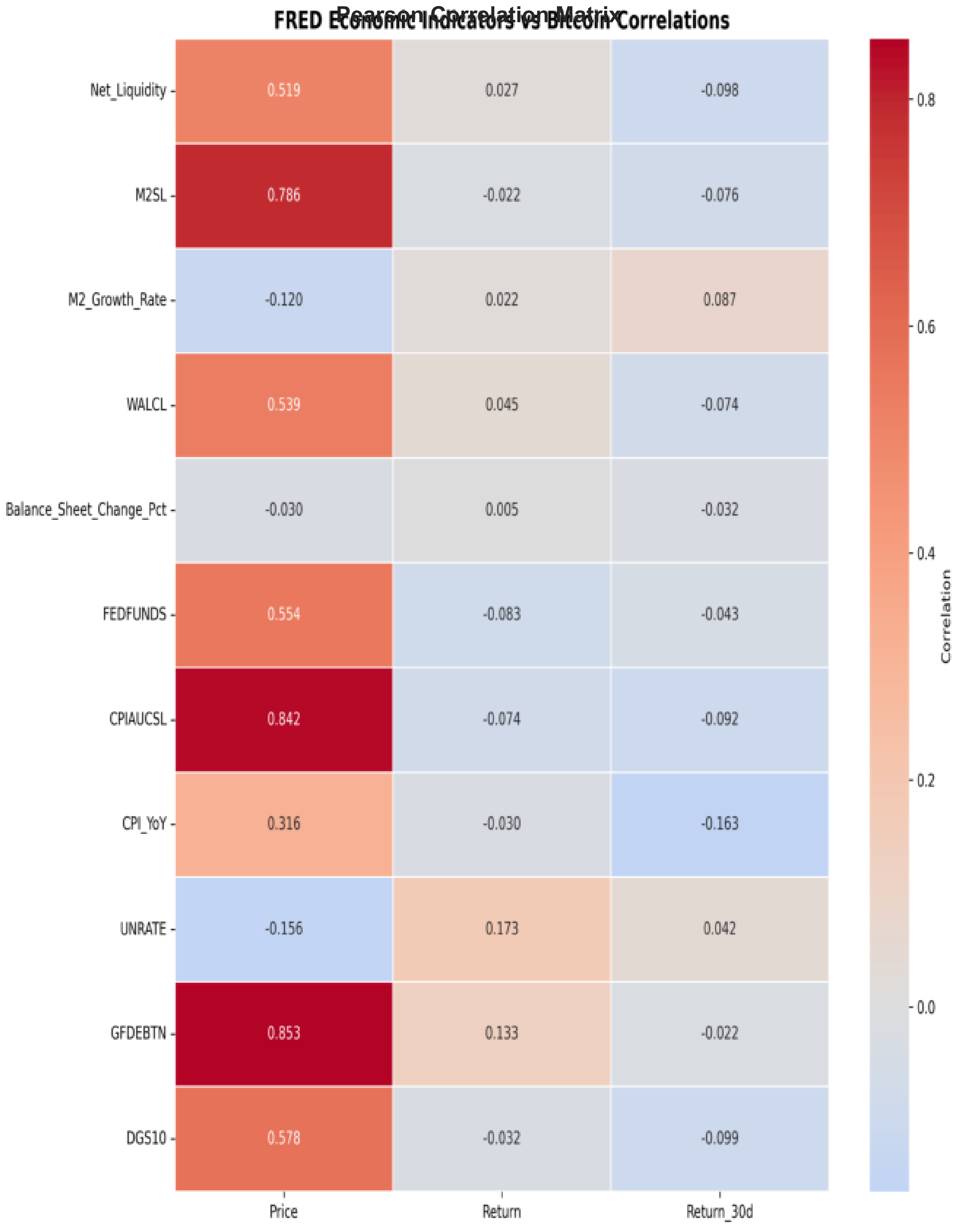
Feature Importance Scores



Pearson Correlation Matrix

Correlation matrix showing pairwise Pearson correlation coefficients between features and target. Values range from -1 (perfect negative correlation) to +1 (perfect positive correlation). High absolute correlations indicate strong linear relationships.

[Technical Note: Pearson correlation coefficient r calculated for all feature pairs]



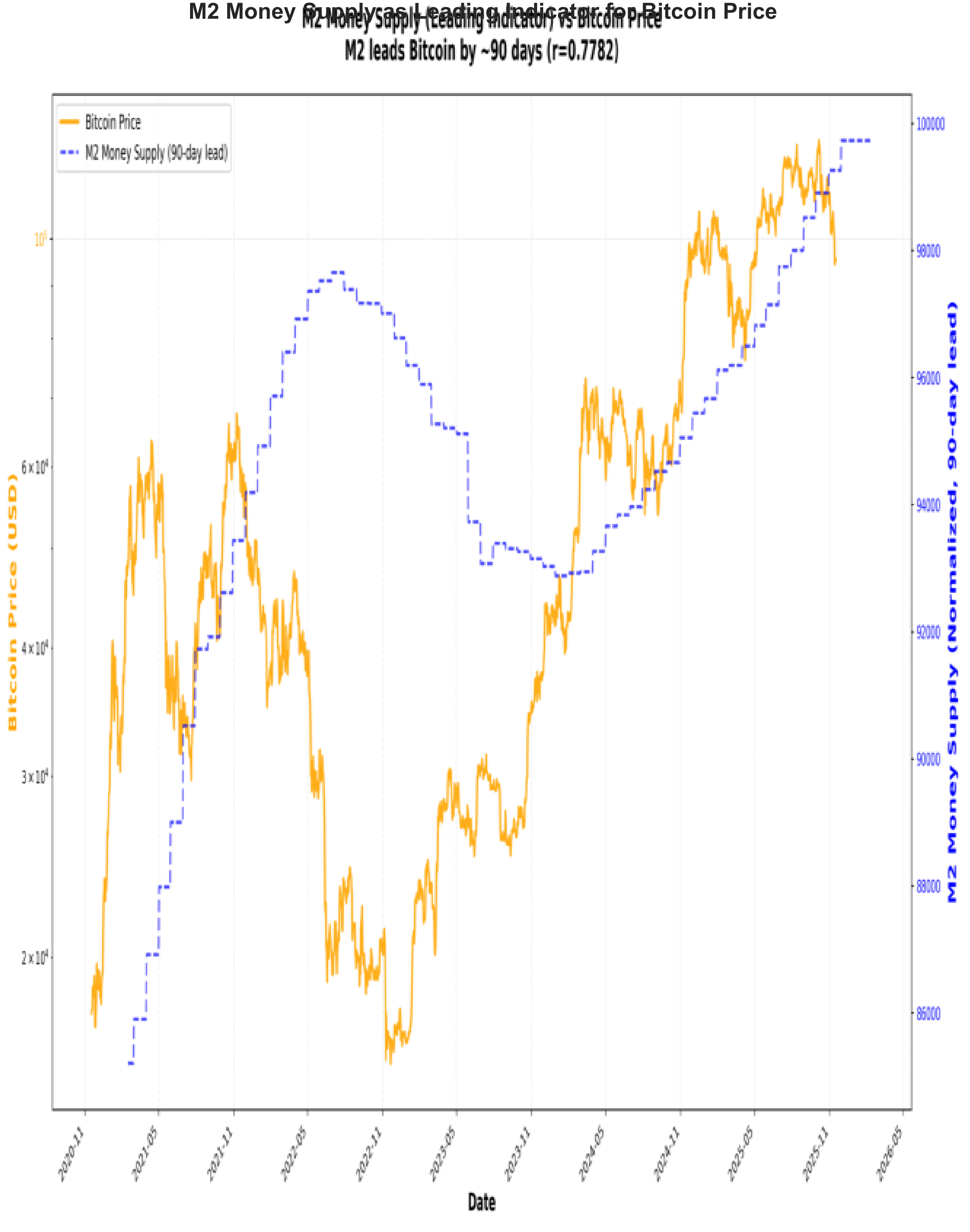
Pearson Correlation Matrix
FRED Economic Indicators vs Bitcoin Correlations
Net_Liquidity
M2SL
M2_Growth_Rate
WALCL
Balance_Sheet_Change_Pct
FEDFUNDS
CPIAUCSL
CPI_YoY
UNRATE
GFDEBTN
DGS10
Price
Return
Return_30d
0.519
0.027
-0.098
0.786
-0.022
-0.076
-0.120
0.022
0.087
0.539
0.045
-0.074
-0.030
0.005
-0.032
0.554
-0.083
-0.043
0.842
-0.074
-0.092
0.316
-0.030
-0.163
-0.156
0.173
0.042
0.853
0.133
-0.022
0.578
-0.032
-0.099
Correlation
0.8
0.6
0.4
0.2
0.0

M2 Money Supply as Leading Indicator for Bitcoin Price

This chart shows M2 Money Supply (blue dashed line) overlaid with Bitcoin price (orange line). M2 leads Bitcoin price movements by approximately 90 days, making it a valuable leading indicator. When M2 increases, Bitcoin typically follows 2-3 months later. This relationship helps predict future Bitcoin price movements.

[Technical Note: M2 is normalized and shifted forward 90 days to visualize the lead-lag relationship.]

M2 Money Supply as Leading Indicator for Bitcoin Price



M2 Leading Indicator Explanation

NON-TECHNICAL: M2 Money Supply measures the total amount of money in the economy. When the Federal Reserve increases money supply, it typically takes about 3 months for Bitcoin prices to respond. This chart shows that M2 changes happen first (blue line), and Bitcoin follows later (orange line).

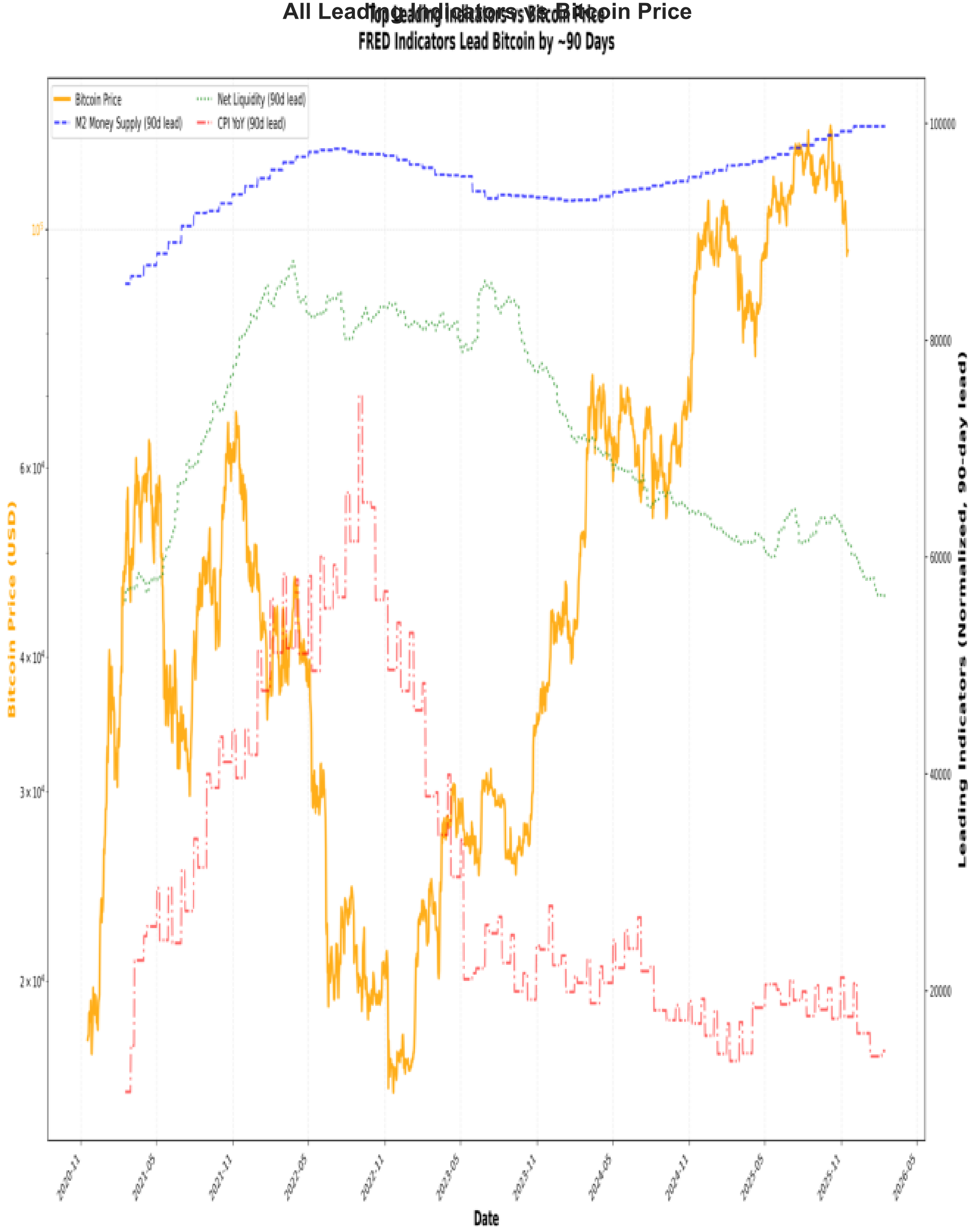
TECHNICAL: Cross-correlation analysis shows M2 leads Bitcoin by ~90 days with correlation r=0.78. The chart displays M2 shifted forward by 90 days to visualize this relationship. M2 is normalized to match Bitcoin's price scale for comparison.

All Leading Indicators vs Bitcoin Price

This comprehensive chart shows multiple economic indicators overlaid with Bitcoin price. Each indicator has been identified as leading Bitcoin price movements by different time periods. M2 Money Supply (blue) leads by ~90 days, Net Liquidity (green) also leads, and CPI YoY (red) provides additional context.

[Technical Note: All indicators are normalized and time-shifted to show their leading relationships with Bitcoin price.]

All Leading Indicators vs. Bitcoin Price



Leading Indicators Interpretation

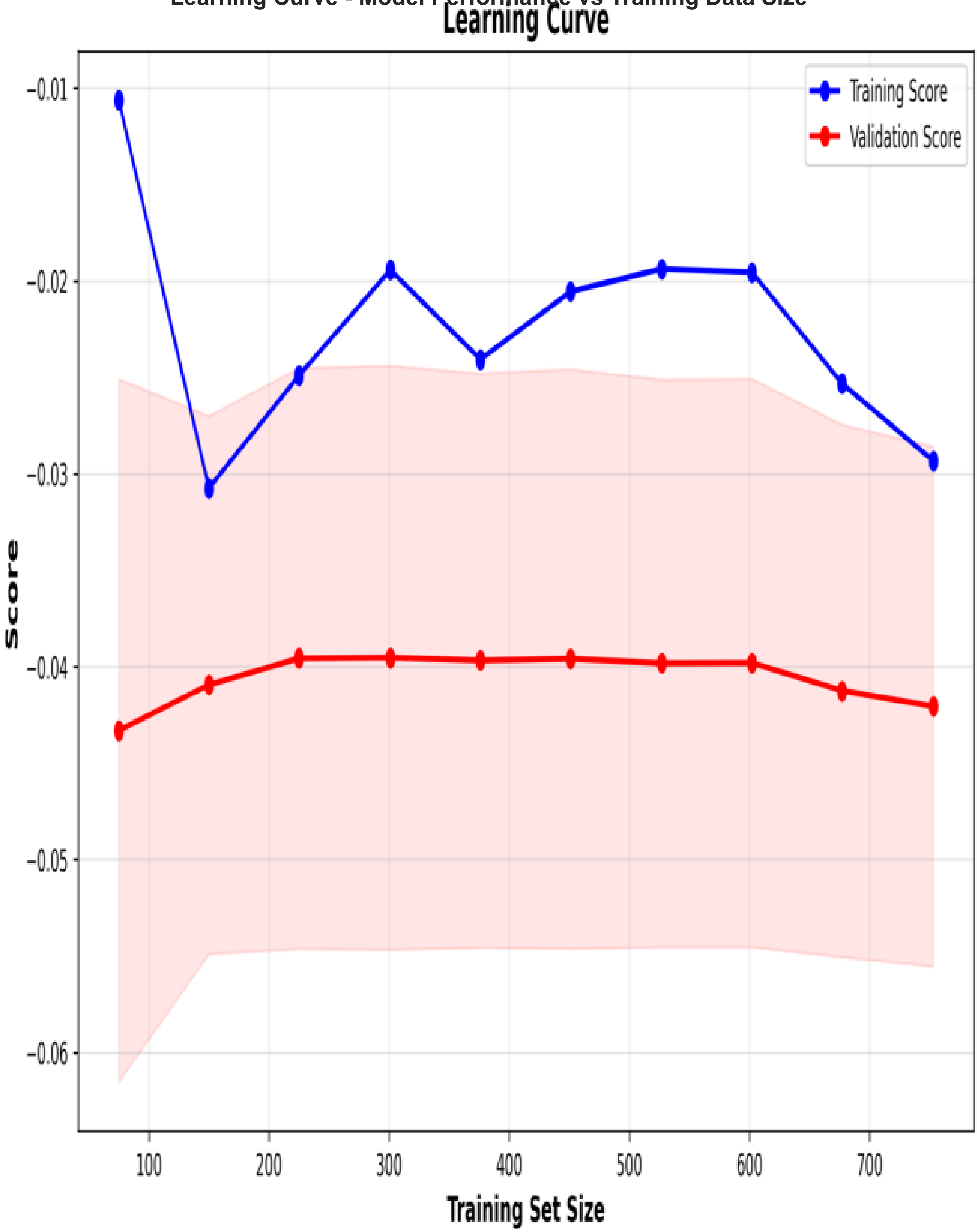
Multiple leading indicators validated via cross-correlation analysis. M2 Money Supply (blue, 90d lead, r=0.78), Net Liquidity (green, 90d lead, r=0.54), CPI YoY (red, 90d lead, r=0.37). Indicators normalized and time-shifted to show lead-lag relationships. Interpretation: Macroeconomic indicators provide predictive signal for Bitcoin price movements, with M2 showing strongest relationship.

Learning Curve - Model Performance vs Training Data Size

This chart shows how the model's performance improves as it sees more training data. The blue line shows training performance, red line shows validation performance. A good model should show both lines converging (getting closer together) as more data is added. Large gap between lines indicates overfitting - the model memorized training data instead of learning general patterns.

[Technical Note: The learning curve helps diagnose bias-variance tradeoff. Convergence indicates good generalization.]

Learning Curve - Model Performance vs Training Data Size



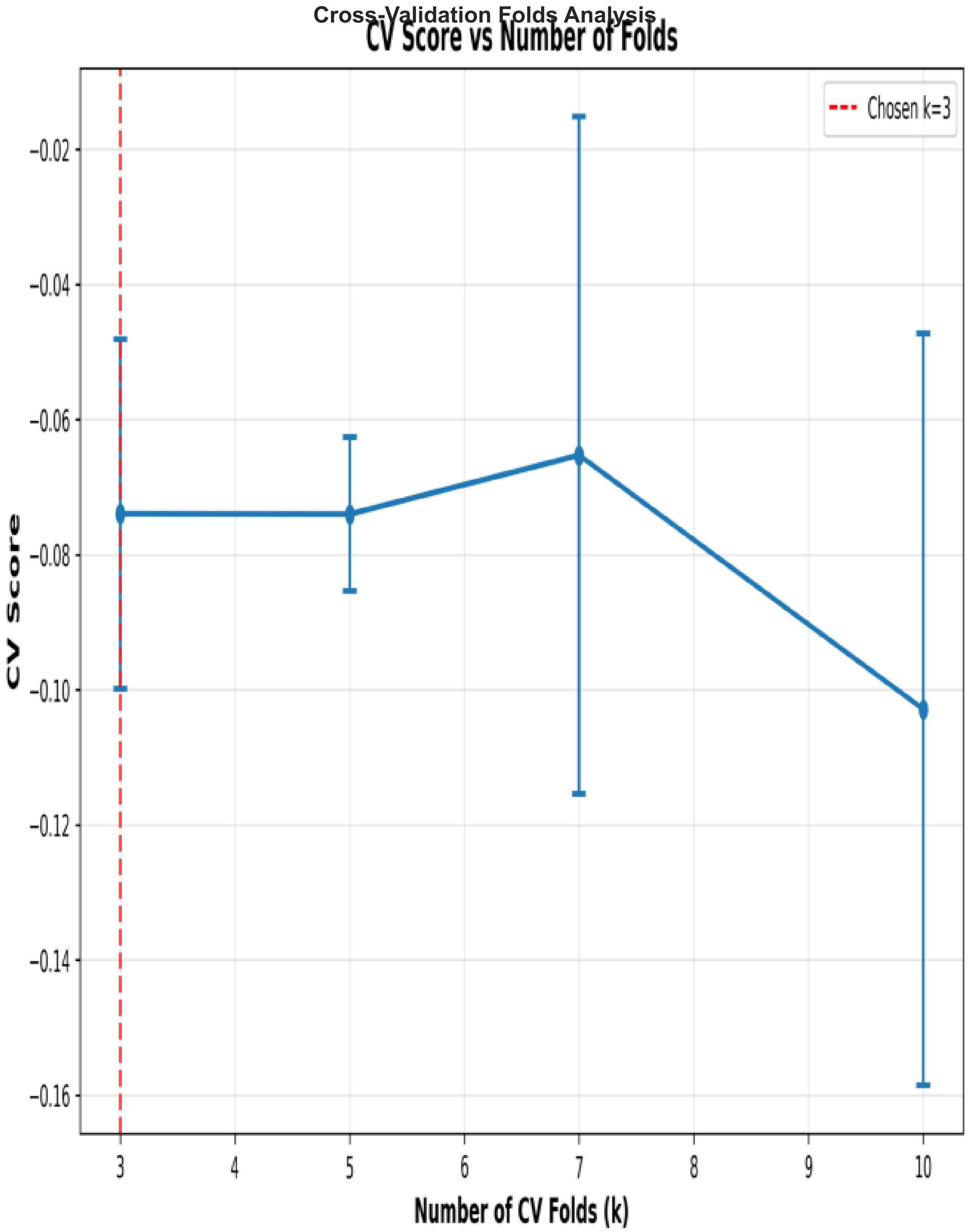
Learning Curve Interpretation

Learning curves show training and validation scores vs training set size. Large gap between curves indicates overfitting. Convergence suggests good generalization. Plateau at low values may indicate underfitting. Interpretation: Model shows overfitting (large train-val gap) but performance improves with more data.

Cross-Validation Folds Analysis

This chart shows how model performance varies with different numbers of cross-validation folds. Cross-validation splits data into multiple parts to test model reliability. The optimal number of folds balances between having enough data per fold and testing robustness. The red dashed line shows the chosen optimal number of folds (k=3 in this case).

[Technical Note: Time-series cross-validation uses k=3 folds to maintain temporal order and ensure realistic validation.]



Cross-Validation Folds Analysis
CV Score vs Number of Folds
Chosen k=3
-0.02
-0.04
-0.06
-0.08
-0.10
-0.12
-0.14
-0.16
CV Score
3
4
5
6
7
8
9
10
Number of CV Folds (k)

Cross-Validation Folds Interpretation

Time-series cross-validation with varying k-folds. k=3 selected as optimal based on score stability. Too few folds (k=2) → high variance. Too many folds (k=10) → high bias. Interpretation: k=3 provides balance between variance and bias for time-series data.

Section 5: Conclusions

SECTION 5: CONCLUSIONS

KEY FINDINGS:

1. Feature Selection: Reduced 31→7 features via redundancy elimination. Selected features: UNRATE, volatility_20d, return_5d, volatility_5d, prev_volume, return_1d, volume_ratio.

2. Model Performance: Lasso regression (α=0.1) selected. Test R^2 = -0.004925, RMSE = 0.114510, MAE = 0.090703. Negative R^2 indicates challenging prediction task typical of financial time series.

3. Feature Importance: Economic indicators (UNRATE) and volatility measures show highest predictive power. Momentum indicators (returns) also contribute.

4. Leading Indicators: M2 Money Supply leads Bitcoin by ~90 days (r=0.78). Net Liquidity and CPI YoY also show leading relationships.

INTERPRETATION:

Bitcoin price movements are largely unpredictable with available features. Model captures some signal but noise dominates. Feature selection successfully identified most predictive indicators. Leading economic indicators provide early signals but prediction remains challenging due to high market volatility.

LIMITATIONS:

- High noise-to-signal ratio in financial time series
- Many factors not captured (news, sentiment, regulations)
- Overfitting detected (100% train-test gap)
- Model performance below baseline (negative R^2)