Bitcoin Comprehensive Analysis Report

With Data Leakage and Look-Ahead Bias Checks

Generated: 2025-11-15 22:22:59

Data Leakage Analysis

COMPREHENSIVE DATA LEAKAGE AND LOOK-AHEAD BIAS ANALYSIS Overall Status: PASS - No Data Leakage Detected This section checks for all forms of data leakage that could cause overfitting: 1. OHLC LEAKAGE DETECTION Problem: Using same-day open/high/low to predict same-day close. Why it's leakage: Close price is always between high and low. This makes prediction trivial and causes overfitting. □ PASS - No OHLC leakage detected All OHLC features are properly lagged. 2. LOOK-AHEAD BIAS DETECTION Problem: Using future information to predict past. Examples: Forward-looking indicators, negative shifts. □ PASS - No look-ahead bias detected All features use only past data. 3. TARGET LEAKAGE DETECTION Problem: Target column accidentally included in features. □ PASS - No target leakage detected 4. PERFECT CORRELATION DETECTION Problem: Features with correlation > 0.99 are redundant. ⚠ Found 43 perfect correlations: - prev_close vs prev_open: 0.9993 - prev_close vs prev_high: 0.9997 - prev_close vs prev_low: 0.9996 - prev_close vs sma_20: 0.9963 - prev_close vs close_lag_2: 0.9993 RECOMMENDATIONS TO FIX DATA LEAKAGE: - Remove 43 redundant features with perfect correlations

Feature Selection

FEATURE SELECTION WITH LEAKAGE REMOVAL Leaky Features Removed: 0 Initial Features: 31 After Leakage
Removal: 31 Final Features: 7 Total Reduction: 24 features Selection Steps: VIF-based removal:
Removed 24 features Correlation-based removal: Removed 0 features Linear combination removal:

Removed 0 features

FRED vs SMA Analysis

FRED vs SMA REDUNDANCY ANALYSIS Analysis of whether FRED economic indicators and SMA/Technical indicators are redundant. Result: No significant FRED/Technical conflicts found. Conclusion: - FRED and SMA/Technical features provide DIFFERENT information - They are complementary, not redundant -

Both should be included in the model

Model Performance

MODEL PERFORMANCE Best Model: Lasso Performance Metrics: Train+Val R^2: 0.0000 Test R^2: -0.0049 Gap:

100.00% Test RMSE: 0.1145 Test MAE: 0.0907

Overfitting Analysis

OVERFITTING ANALYSIS Status: FAIL - Severe Overfitting Overfitting Detected: severe - Train-Val gap
(100.00%) exceeds threshold (15.00%) - Train-Test gap (100.00%) exceeds threshold (15.00%) Train-Val

Gap: 100.00% Train-Test Gap: 100.00%

Conclusions

CONCLUSIONS AND RECOMMENDATIONS DATA LEAKAGE PREVENTION: 1. Always use LAGGED features (prev_open, prev_high, prev_low, prev_close) 2. Never use same-day OHLC to predict same-day close 3. Ensure all features are from past time periods only 4. Check for look-ahead bias (forward-looking indicators) 5. Verify target column is not in features FEATURE SELECTION: 1. Remove redundant features aggressively 2. Use VIF to detect multicollinearity 3. Remove highly correlated features (>0.85) 4. Check FRED vs Technical redundancy 5. Keep only most predictive features (5-10) OVERFITTING REDUCTION: 1. Lower R^2 scores are better than overfitted high R^2 2. Focus on test performance, not train performance 3. Monitor train-test gap continuously 4. Use strong regularization 5. Accept that perfect predictions are unrealistic MODEL VALIDATION: 1. Only use models that pass all leakage checks 2. Verify no data leakage before deployment 3. Test on out-of-sample data 4. Monitor performance over time 5. Retrain periodically with new data