Systematic Multi-Timeframe Ablation Study for Bitcoin Price Prediction

Student Name: Jeawn Lee

1. Background & Motivation

Bitcoin experiences repeated significant drawdowns due to extreme volatility and fat-tail distributions, degrading risk-adjusted performance while increasing the practical value of decline avoidance. While recent research has demonstrated the effectiveness of multi-timeframe analysis in Bitcoin prediction, existing studies often focus on short-term timeframes (5min-4h) and assume that adding more timeframes continues to improve performance. However, the fundamental question remains: **does adding more timeframes actually continue to improve prediction performance, or does it introduce noise and overfitting?** Additionally, **does utilizing historical information (24h) also contribute meaningfully to risk prediction?** This study employs a controlled Ablation Study (A0→A1→A2→A3) to rigorously test whether each timeframe addition provides meaningful performance gains or introduces diminishing returns.

Research Questions:

- **RQ1:** What is the optimal MTF combination for long-term risk prediction when expanding H4 → D1 → W1?
- **RQ2:** Does utilizing historical information (24h) contribute meaningfully to risk prediction?

2. Methods & Hypotheses

2.1. Methods

Features & Ablation Study:

- **Core Technical Indicators:** RSI, MACD, MA(7/14/20/60/120), Ichimoku, OHLCV derivatives
- **Multi-Timeframe Hierarchy:** W1 → D1 → H4
- **Ablation Study Design:**
    - **A0 (Baseline)**: H4 timeframe technical indicators only
    - **A1 (MTF-1)**: A0 + D1 timeframe indicators
    - **A2 (MTF-2)**: A1 + W1 timeframe indicators
    - **A3 (MTF-3)**: A2 + 24h historical features

Model & Validation:

- **Model:** LightGBM multiclass (class_weight=1/prevalence)
- **Cross-Validation:** Purged & Embargoed Time-Series CV (H=30, E=30)
- **Evaluation:** Macro-F1 (primary), Balanced Accuracy, OvR ROC-AUC, Brier Score

2.2. Hypotheses

H1 (Incremental Contribution of MTF): Multi-timeframe addition will show either meaningful improvement (A0 < A1 < A2) or diminishing returns (A0 ≥ A1 ≥ A2), indicating whether additional timeframes provide value or introduce noise.

H2 (Historical Features Contribution): Historical features will either enhance performance (A2 < A3) or introduce noise (A2 ≥ A3), testing the value of 24h lookback information.

H3 (Feature Sensitivity for Sell Signal): SHAP analysis will identify the most important features for 'sell' signal prediction, revealing which features are most critical for decline risk prediction.

3. Materials & Data Sources

3.1. Data Source and Period

- **Asset:** BTC/USD spot price data
- **Period:** 2020-05-12 ~ 2024-04-20 (post-3rd halving period)

3.2. Technical Indicators & Feature Engineering

Technical Indicators: RSI (14-period), MACD (12,26,9), Moving Averages (7,14,20,60,120), Ichimoku Cloud, OHLCV

Multi-timeframe Feature Engineering: Follows the Ablation Study design (A0→A1→A2→A3) with hierarchical structure (W1→D1→H4) and leakage prevention methods (shift=0).

Labeling Scheme:

- **Prediction Horizon:** 30-day
- **Thresholds:** -15% (sell), +5% (buy), otherwise (wait)

4. Expected Results & Contribution

Expected Results:

- **Scenario 1**: Meaningful performance improvement with MTF information addition (A0 < A1 < A2 < A3)
- **Scenario 2**: Diminishing returns or noise introduction (A0 ≥ A1 ≥ A2 ≥ A3)
- **Scenario 3**: Mixed results where some timeframes help while others introduce noise
- SHAP analysis will reveal which features are most important for 'sell' signal prediction

5. Validation & Evaluation Protocol

Time-Series Cross-Validation:

- **Method**: Purged & Embargoed CV (H=30, E=30) to prevent data leakage
- **Fold Design**: Monthly test periods with embargo periods for temporal independence
- **Fair Comparison**: All ablation steps (A0→A1→A2→A3) use identical fold indices

Evaluation Metrics:

- **Primary**: Macro-F1 (class imbalance consideration)
- **Secondary**: Balanced Accuracy, OvR ROC-AUC, Brier Score
- **Statistical Testing**: Directional Analysis, Effect Size (Cohen's d), Bootstrap CI (95%), Noise Detection