

機械学習モデル構築 要件定義書

1. 背景と目的

- 背景: BTCUSDの短期的な価格変動は、取引戦略やリスク管理において重要な要素である。
- 目的: BTCUSDの5分足データに基づき、将来の価格変動を予測する機械学習モデルを構築する。具体的には、5分後から15分後の価格変動の大きさと方向性を予測し、短期的な市場動向の把握や意思決定支援に役立てる。

2. スコープ

- 対象資産: **BTCUSD**
- 入力データ: **5分足 OHLCV** データ (及びそこから派生する特徴量)
- 予測対象時間軸:
 - **5分後 (1行先)**
 - **10分後 (2行先)**
 - **15分後 (3行先)**
 - 備考: 各時間軸に対して個別のモデルを構築・評価する。
- 予測内容:
 1. 価格変動率 (%): 各予測時間軸における終値の変動率(回帰予測)。
 2. 価格変動カテゴリ (方向性): 各予測時間軸における価格変動の方向性を分類(分類予測)。
 - 基本カテゴリ (3分類):
 - 上昇 (**Up**): 変動率 > +0.05% (CLASSIFICATION_THRESHOLD)
 - 下落 (**Down**): 変動率 < -0.05% (-CLASSIFICATION_THRESHOLD)
 - 横ばい (**Stable**): 上記以外
 - 検討カテゴリ (5分類): (初期要望に基づく。閾値定義が必要)
 - 大幅上昇 (Significantly Up)
 - 上昇 (Slightly Up)
 - ほぼ変動なし (Stable / Almost Unchanged)
 - 下落 (Slightly Down)
 - 大幅下落 (Significantly Down)
- スコープ外:
 - 正確な価格の絶対値予測。
 - モデルによる自動取引執行。

3. データ要件

- 入力データソース: BTCUSD 5分足 OHLCV データ (CSV形式等を想定)。
- データ期間: モデル学習および検証に十分な期間(例: 過去数年分)。
- データ品質: 欠損値、異常値に対する前処理方針を定義する(例: 検証コードではNaN除去、無限大値の置換を実施)。
- 特徴量 (候補):
 - 過去の価格変動率 (Lag Features: price_change_pct_*)
 - 過去の出来高変動率 (Lag Features: volume_change_pct_*)
 - 移動平均線 (SMA: sma_*) および乖離率 (sma_diff_pct_*)

- RSI (rsi)
- 高値/安値からの距離 (dist_from_high, dist_from_low)
- 備考: 特徴量の選択とエンジニアリングはモデル性能に影響するため、継続的な検討が必要。

4. 機能要件

- モデル入力: 定義された特徴量。
- モデル処理:
 - 各予測時間軸(5分後, 10分後, 15分後)に対して、それぞれ独立した予測モデルを学習する。
 - 検証コードでは LightGBM (Regressor/Classifier) を使用したが、他のアルゴリズムも検討可能。
- モデル出力:
 - 各予測時間軸に対する予測価格変動率。
 - 各予測時間軸に対する予測価格変動カテゴリ(上昇/下落/横ばい)。

5. 非機能要件

- 性能:
 - 評価指標:
 - 変動率予測 (回帰): MAE (平均絶対誤差) - より小さい方が良い。
 - 方向性予測 (分類): 正解率 (Accuracy) - より大きい方が良い。クラス不均衡がある場合は、適合率(Precision)、再現率(Recall)、F1スコア、混同行列 (Confusion Matrix)も考慮する。
 - 目標値: 検証フェーズの結果に基づき、各予測時間軸に対する具体的なMAEおよび Accuracy (または他の分類指標) の目標値を設定する。
 - 検証方法: 時系列データの特性を考慮し、時間ベースの Train/Test Split (検証コードで使用) またはウォークフォワード検証、時系列クロスバリデーションを用いる。
- 予測速度 (Latency): 最新データ取得後、許容される予測計算時間 (要定義)。
- 運用・保守: モデルの再学習頻度、性能監視方法 (モデルドリフト検知) を定義する。

6. 制約条件

- 利用技術スタック: Python, Pandas, Scikit-learn, LightGBM (現状の検証コードベース)。
- 開発期間・予算: (別途定義)

7. 承認基準

- 合意された性能目標値 (MAE, Accuracy等) を、指定された検証方法で達成すること。
- 定義された各予測時間軸 (5, 10, 15分後) に対して、予測結果 (変動率、方向性カテゴリ) が出力されること。

8. 検証からの決定事項・確認事項

- 複数モデル: 予測時間軸 (5, 10, 15分後) ごとに個別のモデルを構築し評価する方針とする。

- 評価指標: 変動率予測は**MAE**、方向性予測(3分類)は**Accuracy**を主要指標として比較検証する。
- 分類閾値: 方向性(3分類)の閾値は現在 +/- 0.05% (CLASSIFICATION_THRESHOLD = 0.0005) としている。この値の妥当性は要継続検討。
- (要検討) 初期要望にあった5分類(大幅上昇/下落など)の導入可否と、その場合の明確な閾値定義。
- (要検討) 最適な特徴量セット、モデルアーキテクチャ、ハイパーパラメータ。
- (要検討) より厳密な時系列検証方法(ウォークフォワード等)の導入。
- (要定義) 具体的な性能目標値(MAE, Accuracy)。
- (要定義) モデルのデプロイ方法と予測結果の利用インターフェース。

この要件定義は、これまでの議論と検証コードの内容を基にしたものです。必要に応じて更新してください。