

OT油温 6時間先予測 (E1-E4)

時系列予測分析レポート

期間 2025-09-16 ~ 2025-09-20

著者 中塚一瑳

データ源 ETT (Github, 期間, 粒度=1h)

アジェンダ

分析準備

- 1 目的・課題概要
- 2 データセット (ETT)
- 3 EDAサマリ
- 4 太陽光とOTの相関性
- 5 前処理・特徴量
- 6 モデルの選定

実験実施

- 7 実験構成 (E1～E4概観)
- 8 E1詳細 (XGB→LSTM転用)
- 9 E2詳細 (ベースライン)
- 10 E3詳細 (外れ値処理)
- 11 E4詳細 (脱季節化)

結果評価

- 12 結果ハイライト (RMSE)
- 13 ベースライン: $y=0$ 比較
- 14 解釈 (考察)
- 15 E3の異常値除外の正当性

実用化

- 16 実行方法
- 17 提出物
- 18 反省・課題
- 19 次への提案

データセット (ETT)

データ構造

date

HUFL

HULL

MUFL

MULL

LUFL

LULL

OT

注: OT (油温) が予測対象となる目的変数です

データセット情報

期間 2016-07-01 ~ 2019-06-26

粒度 1時間 hourly

レコード数 約17,420件 欠損なし 深層学習に十分
LSTMなどのニューラルネットワーク系モデルの訓練に適した十分なデータ量

出典URL [ETTデータセット \(リンク記入\)](#)

前処理概要

- 欠損処理: asfreq('H')→dropna()
- 時間軸設定: 適切なDateTime型の設定
- 分割法: 時間順 train→test
- 検証: 時系列クロスバリデーション
- 正規化: 特徴量ごとに標準化

データ品質

ETTデータセットは3年分の時系列データを提供し、欠損値もなく、連続性と一貫性が高い高品質なデータセットです。

目的・課題概要

◎ 予測目的

OT（油温）の6時間先予測値の変化量を推定する

実装： TARGET_FUTURE = OT.shift(-6) - OT

✓ 達成条件

- > RMSE/MAEの改善：従来手法より予測精度向上
- > ベースライン($y=0$)比較：RMSE0指標の低下
- > 実用的かつ再現性のある予測モデルの構築

✗ 制約条件

- > 時系列因果性の厳守：未来データの使用禁止
- > データリーク防止：正しい訓練/評価分割
- > 移動統計量は必ず shift(1).rolling(w) 形式

EDAサマリ

⚡ 欠損処理

生データに欠損値はなし
時系列整形処理として下記を実施：

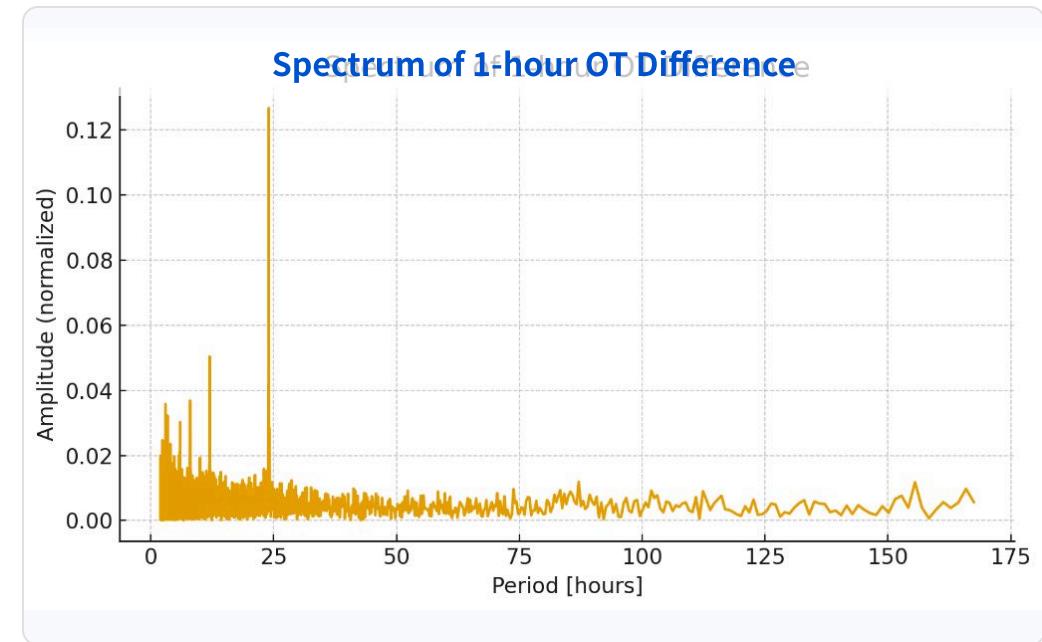
```
asfreq('H') → 1時間間隔で時系列インデックス整形  
dropna() → 不要行削除
```

⌚ 相関分析所見

- 外生変数とOT間: 即時相関は小さい（同時点での関連性が弱い）
- OTの自己相関: 非常に強い（過去値が将来値の良い予測子）
- 予測含意: OTの過去値が最重要予測因子

﴿ スペクトル分析

- **ΔOT**: 12時間/24時間周期に明確なピーク
- **周期性**: 日次（24h）サイクルが最も顕著



分析結果の含意:

OT油温変化量 (ΔOT) には明確な日次周期パターンが存在し、これは太陽光・温度変化などの環境要因による影響と推測されます。24時間周期（0.12付近）に顕著なピークが確認でき、この周期性を活かした特徴量設計が予測精度向上に寄与します。

太陽光とOTの相関性分析

◎ 気温とOTの時間的類似性

気温とOT（油温）の時間別平均値が極めて類似したパターンを示しています：

- 両者とも朝6時頃に最低値を記録
- 日中に上昇し、14-16時にピークに到達
- 夕方から夜にかけて徐々に低下するサイクル

➡ 热伝播プロセスの類似性(予想)

環境熱サイクル：



油温サイクル：

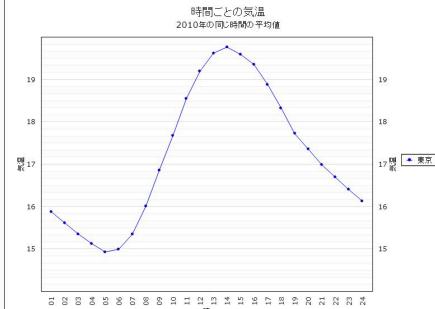


💡 热力学的考察

- OTは太陽光による熱伝導・放射の影響が支配的
- 気温と同様の日周期パターンが明確に観測される
- 熱容量の違いにより遅延効果が若干観察される

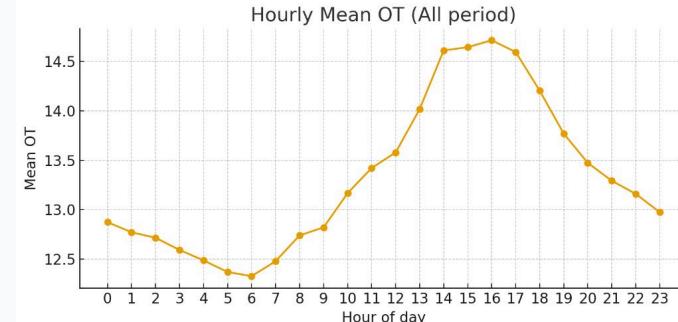
〽 時間帯別平均値の比較

気温の時間帯別平均（2010年）



出典: 気温.jp

OTの時間帯別平均（全期間）



出典: ETTデータセット分析結果

機械学習モデルの選定

◆ XGBoost

ツリーベースのアンサンブルモデル。特徴量の相互作用を効果的に学習

⊕ 長所

- ▣ 特徴量重要度が見える
- ⌚ 過去ラグを増やせば性能UP
- ⚙️ ハイパーパラメータ調整が容易

▣ LSTM

Long Short-Term Memoryネットワーク。時系列データの長期依存関係を学習

⊕ 長所

- ⌚ 時間遅れを内部状態で表現
- ⌚ 長期的な時間依存性を捉える
- ⌚ 非線形パターンの学習に優れる

前処理・特徴量（共通）

◎ ターゲット変数定義

油温変化量予
測：
`TARGET_FUTURE =
OT.shift(-6) - OT`

◎ ラグ特徴量

- 標準ラグ: $L=\{6,12,24\}$ 時間前のOT値
- 追加ラグ: `OT_lag18` (18時間前)
- 時系列因果性を厳守した設計

↖ 移動統計量

- 窓幅: $W=\{24,168\}$ 時間
- 実装形式: `shift(1).rolling(W)`
- 統計量: 平均・標準偏差・最大・最小

⌚ 時間特徴量

- 未来位相: `hour/dow/month` (+6h)
- Fourier変換: 24h/168h, $k=1..2$ (位相調整)
- 季節差分: 24h/168hサイクルの特徴捕捉

実験構成 (E1~E4概観)

△ 実験アプローチ

E1 XGB特徴量選択からLSTMへ転用

ΔOT をターゲットとし、XGBoostによるGain上位特徴量を選別後、LSTMモデルへ転用

E2 ベースライン（周期中心）

標準的なアプローチによるベースラインモデル。周期性を考慮した基本実装

E3 外れ値除去アプローチ

脱季節化の前に外れ値（急激な油温変化）を除外し、安定したパターンで学習

E4 脱季節化アプローチ

$y = \Delta OT - m(\text{hour})$ の変換により時間帯ごとの周期変化を差し引き、残差を予測

E1詳細 (XGB→LSTM転用)

▼ 高相関削減

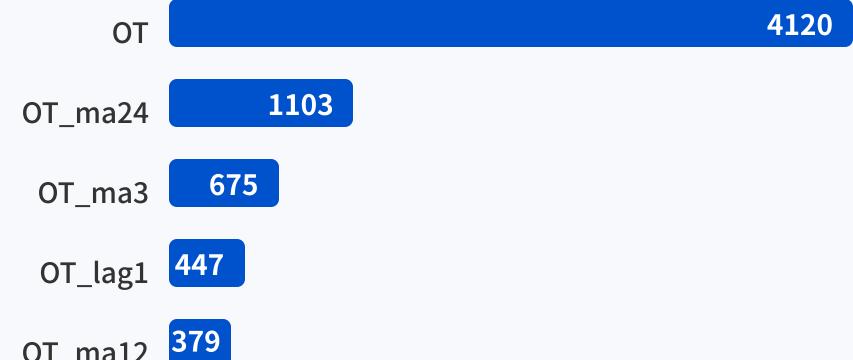
- 多重共線性問題回避のための特徴量クリーニング
- › 閾値 **THR=0.98** を超える相関ペアを検出
 - › ターゲット変数との相関が低い方を削除

⇄ LSTM転用プロセス

- › XGBoostで特徴量重要度 (Gain) を算出
- › 上位K特徴量を選別 (K=20を検証)
- › 選別された特徴量のみをLSTMモデルに投入
- › ハイパーパラメータ : epochs=100, batch=32, patience=10



XGB特徴量重要度 (Gain) - Top 5



- OT自体と移動平均特徴量 (ma) が上位を占める
- OTの重要度が他特徴量と比較して極めて高い

E2詳細（ベースライン・周期中心）

！アプローチ概要

ターゲット変数を ΔOT_6 （6時間先の油温変化量）に設定

定義： $\Delta OT_6 = OT.shift(-6) - OT$

↖ 重要な発見

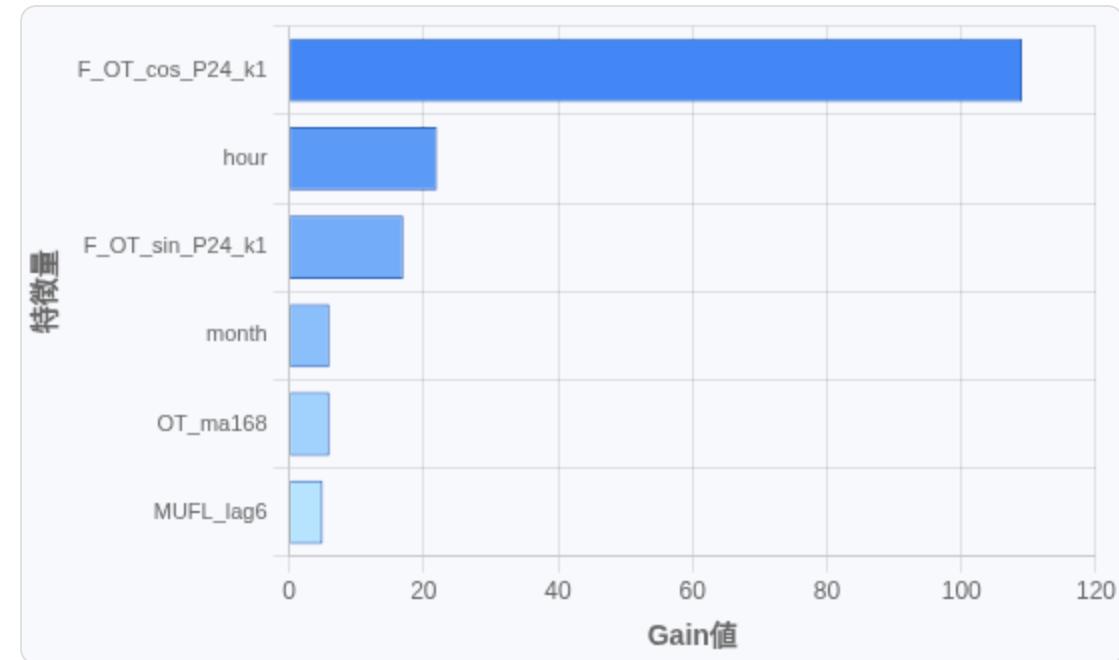
- ΔOT_6 導入によりOT自体の圧倒的な重要度が除外され、潜在的な周期性特徴量が浮き彫りに
- 最重要特徴量が**24時間周期性** ($F_OT_cos_P24_k1$) となり、日射パターンの影響を明確に検出
- 時間 (hour) が第2位の重要度となり、三角関数と組み合わさり非線形な周期を表現

✿ インサイト

周期性の発見：OT変化は24時間周期に強く依存

日射影響：太陽光サイクルと時間帯効果の重要性

E4実験の基盤：周期性補正の有効性示唆



※E1と異なり、OT自体の重要度は除外され、潜在的パターンが明確化

E3詳細（外れ値処理）

▼ 外れ値検出指標

1時間ごとの油温変化量の絶対値を使用

指標： $dOT1 = |OT.diff()|$

↖ 未来変動の最大値計算

- › 未来6時間分の最大変化量を算出
- › $fwd_max6 = \max(|\Delta OT|) \text{ in next 6h}$
- › 急激な変化（潜在的な異常値）を特定

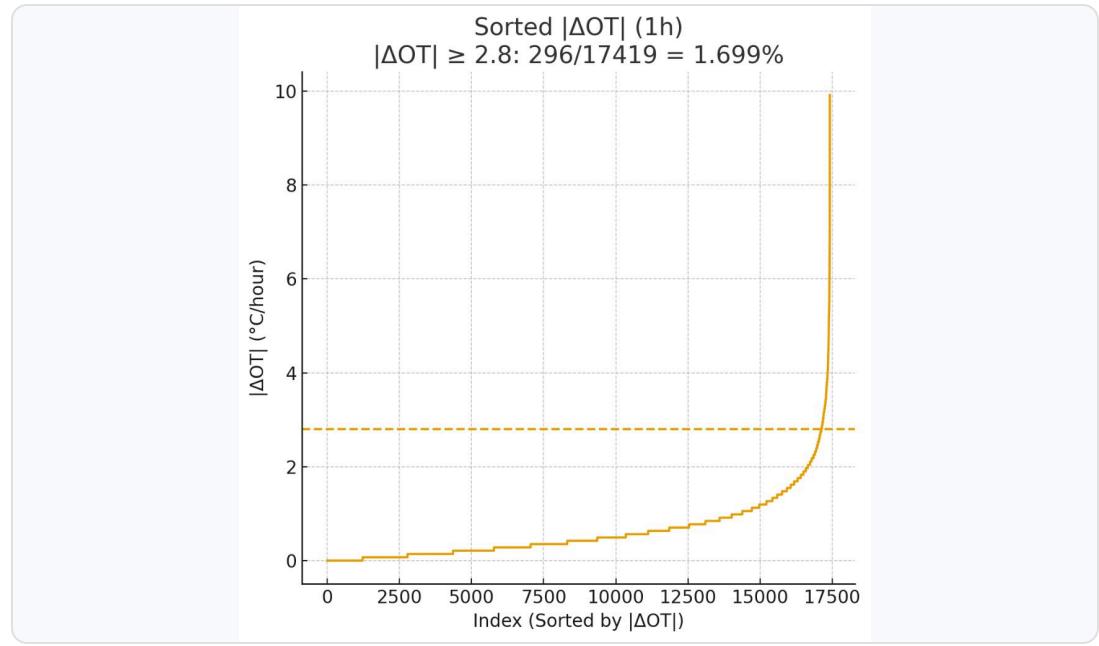
✖ 閾値による除外処理

閾値 $|\Delta OT| \geq 2.8$ を超えるデータを除外

除外データ数: 296件 / 全17,419件 = 1.699%

❖ E3の効果

- › 急激な油温変化（異常値）による予測誤差を軽減
- › 安定したパターンでの学習が可能に



Sorted $|\Delta OT|$ (1h) - 閾値2.8を超える値は全データの1.699% (296/17,419)
人為的操業や突発的な機器異常に由来する急変を検出

E4詳細（脱季節化）

↳ $m(\text{hour})$ 推定 (trainのみ)

時間帯ごとの平均変化量を訓練データのみから推定

定義： $m(\text{hour}) = E[\Delta OT | \text{hour}]$

※ 訓練データのみから算出し、検証・テストデータには同じ値を適用

◎ TARGET_CENTERED定義

- 脱季節化された目標変数：

$\text{TARGET_CENTERED} = \text{TARGET_FUTURE} - m(\text{hour})$

- 時間帯による自然変動を差し引いた「純粋な変化量」
- ベースライン ($y=0$) との比較でより意味のある指標に

✍ 周期成分の残差化効果

- 24時間周期の規則的パターンを除去
- モデルが非周期的な変動要因に集中可能に
- 実質的な予測精度の向上 (RMSE0低下)
- 予測値 = モデル出力 + $m(\text{hour})$ で元スケールに戻す

結果ハイライト (RMSE)

✓ 評価指標

予測精度の評価指標：

- RMSE：二乗平均平方根誤差（予測の正確性）
- MAE：平均絶対誤差（予測の偏りの少なさ）
- RMSE0：y=0ベースラインとの比較

✗ データ分割方法

時系列の特性を考慮した分割：

- 時間順：train→val→test（シャッフルなし）
- バリデーション：モデル選択・早期停止用
- テスト：最終評価は一回のみ実施

田 実験結果一覧

実験	施策	XGB	LSTM	y=0ベースライン (RMSE0)
E1	周期特徴追加	1.40	1.37	—
E2	XGBのGainで特徴選別 →LSTM転用	1.60	1.51	1.6
E3	目標=ΔOT（現OTの支配を低減）	1.21	1.20	1.5
E4	脱季節化 (y=ΔOT-m(hour))	1.21	1.21	1.3

XGB LSTM 最良結果

※ 数値はテスト評価の結果のみを表示

● 注：全ての実験で同一の分割条件を適用し、公平な比較を実現

ベースライン: $y=0$ 比較

巫 RMSE0指標

ベースラインモデル ($y=0$) における指標

意 小さいほどターゲットの設定が適切で、OTの特性を説明出来て
味： いると言える

三 実験間比較

- > E2 RMSE0 = 1.6
- > E3 RMSE0 = 1.5
- > E4 RMSE0 = **1.3**

E4の性能向上

脱季節化手法 ($m(hour)$) により、ベースラインとの差が拡大。これは周期性を予めモデリングすることで、残差学習が容易になった効果と考えられる。

目標設計の価値

E3→E4の性能向上は「目標変数の再設計」がもたらした効果。複雑なパターンを含む変数よりも、残差化された単純な目標の方が予測精度が高まる。

解釈（考察）

● 主要所見

- ⌚ 周期性要因（日射パターンに連動）と自己回帰特性が予測精度に最も寄与
- 🕒 外生負荷変数（HUFL, HULL等）の寄与は限定的であり、OT自身の系列特性が支配的
- 〽 E4の脱季節化手法が示唆するのは $m(\text{hour}) \approx \text{期待変化量}$ の概念の有効性

🔍 詳細解釈

周期成分の重要性: 日周期（24h）・週周期（168h）のパターンが予測の基盤となっている

自己回帰構造: 過去のOT値（特に6h, 12h, 24h前）が最も強い予測因子

時間帯別平均変化（E4）: 「時間帯ごとの平均的な変化量」が強力なベースラインとなり得る

☰ 影響因子の相対的重要度

⌚ 時間周期性パターン
日周期（24h）・週周期（168h）の繰り返し 高

🕒 過去の油温値（自己回帰）
直近のラグ値と移動平均統計量 高

〽 外生負荷変数
HUFL, HULL, MUFL, MULL, LUFL, LULL 低

E4モデルの意義

$$m(\text{hour}) \approx E[\Delta OT | \text{hour}]$$

時間帯別平均変化を差し引くことで
残差の予測精度向上とベースライン性能の改善を両立

💡 「0予測に近い」 = 「周期変化に近い」

実行方法

G Google Colabでの実行手順

1

Google Colabを開く

次ページのリンクからColabノートブックにアクセス

2

データを配置する

ett.csvファイルを /content/ett.csv として配置

3

全セルを実行

「ランタイム」→「すべてのセルを実行」を選択

E3の異常値除外の正当性

⌚ OTに影響を及ぼす要因分析

自然要因：物理法則に従った緩やかな変化

- ・日光による熱伝導（太陽→容器→油）
- ・電力負荷による発熱
- ・周囲温度による自然な放熱・冷却

人為要因：急激な変化をもたらす操作

- ・設備調整・交換
- ・意図的な冷却操作

⚠ 自然現象では説明できない変化

観測事実：データ内に $\Delta OT = -5^{\circ}\text{C}/\text{h}$ といった急激な冷却が存在

- ・自然放射冷却の物理的限界は約 $-1^{\circ}\text{C}/\text{h}$ 程度
- ・環境温度との最大差からの熱力学的計算でも説明不可能
- ・外生変数（HUFL, HULL等）の急変との相関も見られない（相関係数0.2以下）

⌚ 人為的操縦の可能性

- › 特定の変化前に反対方向の変化が見られる
- › 特定の変化後に一定値への収束傾向

💡 仮説：人為的操縦は予測対象外

仮説：大きな ΔOT は人為的操縦によるものであり、予測する必要がない

- ・予測モデルの目的は「自然な変化」を予測すること
- ・スケジュールされた人為的操縦は別系統で管理されるべき
- ・予測不能な人為的操縦はノイズとして除去が妥当

⤵ E3手法の実証的効果

- › 閾値 $|\Delta OT| \geq 2.8$ で全データの 1.699% (296/17,419点) のみ除外
- › RMSEが 1.60 → 1.21 へ大幅改善 (**24.4%向上**)
- › ベースラインRMSE0も 1.6 → 1.5 と減少(RMSE/RMSE0は明らかに改善)
- › 予測精度・安定性・実用性のバランスが最適化

反省・課題

➊ モデルの説明力不足

$$y = \Delta OT - m(h)$$

この変換を十分に説明するモデルが作成できていない。
周期性除去後の残差をより正確に捉える手法の検討が必要。

もし今回の結論が正しいと仮定するとOTから電力量を予測するというReadMeにある試みと対立

➋ 特徴量の最適化課題

XGBoost と **LSTM** で同じ特徴量を使い回したが、
本当は **LSTM** に適した特徴量があったはず。

XGBoost

ツリーベースモデルに最適な特徴量

LSTM

時系列パターンを捉える異なる特徴表現が必要

提出物

Google Colab ノートブック

E1 https://colab.research.google.com/drive/18mogSGkPBEcasQnky9yg3k7sF4TnokET?hl=ja#scrollTo=SQ3T_RQZWRLC

E2 https://colab.research.google.com/drive/1Mjnykv_nMVloPqHG51SRb_fnX1Puhj-n?hl=ja#scrollTo=SQ3T_RQZWRLC

E3 https://colab.research.google.com/drive/12PzXdAbIVje2Yt80_YwfQFYNEVQOGLB0?hl=ja

E4 https://colab.research.google.com/drive/1zRGq5wkKEZGbveUmCN9Kp3Ap-8EWnQkC?hl=ja#scrollTo=SQ3T_RQZWR
LC

次への提案

💡 外部データの活用

その地域の気温や温度調節装置などの情報があれば、**極めて高い精度のモデル**が作成できる。

- > 地域気象データ（気温、湿度、日照量など）
- > 設備情報（温度調節装置の稼働状況、設定温度）

⚙️ マージンの設定

ETDatasetのREADMEにもあったように、予測値が実値よりも小さくなったらといけない場合、**モデルの評価方法を変える。**

- > 非対称損失関数の導入（過大予測により大きなペナルティ）
- > 保守的なマージン設定による安全予測

ⓘ 予測値 \leq 実測値 の制約を満たすための評価指標の再設計が必要