LSTMを用いた株価予測

藤居　海誠

01 株価予測の重要性

01

投資・資産運用

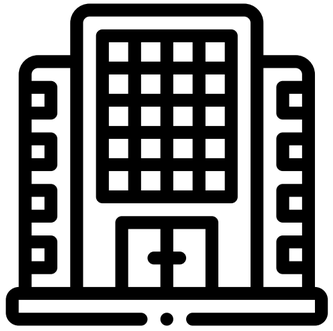
- 売買のタイミング判断
- リスクヘッジ・分散投資

02

戦略・意思決定

- 企業の資金調達
- 戦略判断

03



企業価値マネジメント

- 投資家への適切な情報
- 発信株主価値の向上

02 株価予測の課題

データの不確実性

政治経済ニュースや投資家心理、突発的出来事等
短期的な値動きはランダムに近い(ランダムウォーク)

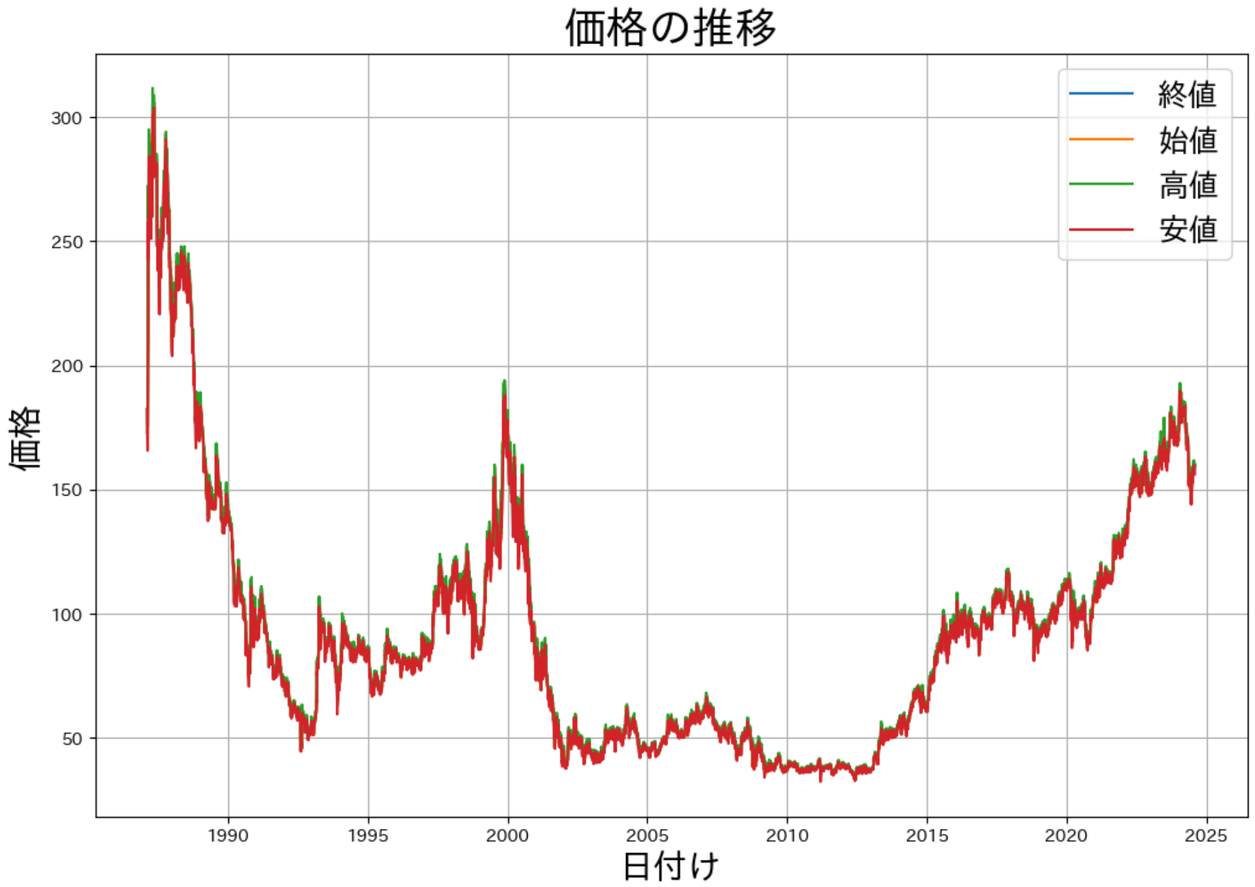
モデルの汎化の難しさ

訓練データで高性能な予測モデルを作っても、未来のデータにはうまく機能しないことが多い(過学習)

因果関係の曖昧さ

株価は多変量かつ非線形な関係性を持ち、どの特徴量が効いているのか解釈しにくい

03 EDAの結果

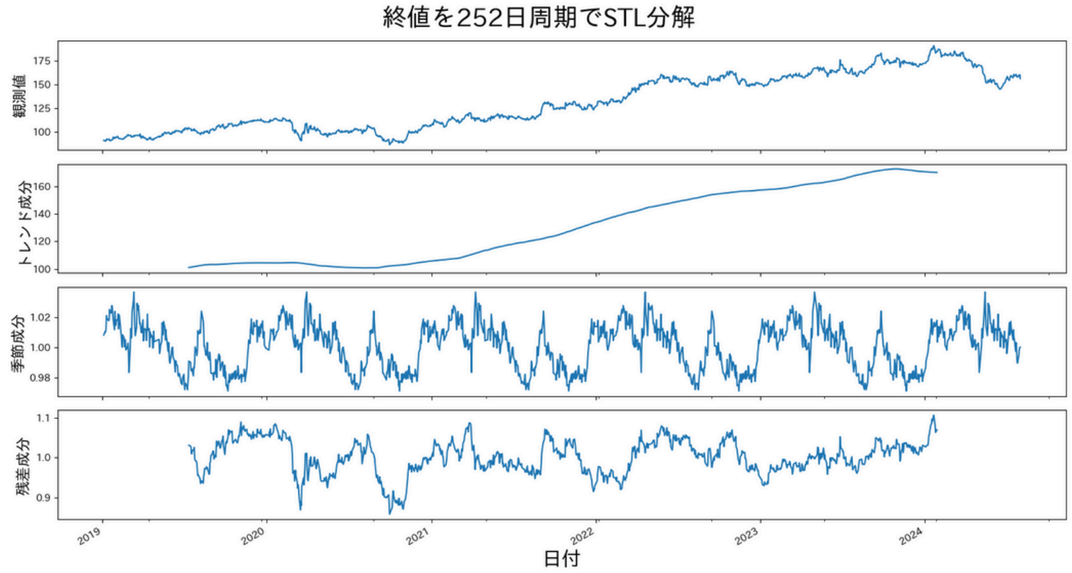


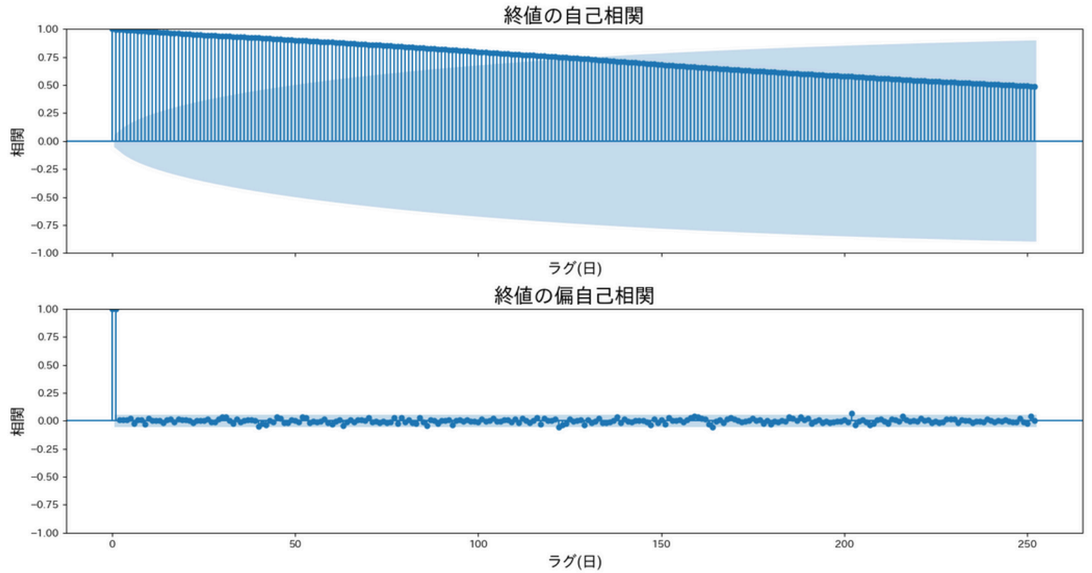
終値などの価格の推移をプロット

年代によってトレンドが異なる
→2019年以降のデータのみ使用

2019年～2022年のデータを学習に使用
2023年1月～7月のデータを検証に使用
2024年1月～7月のデータをテストに使用

03 EDAの結果





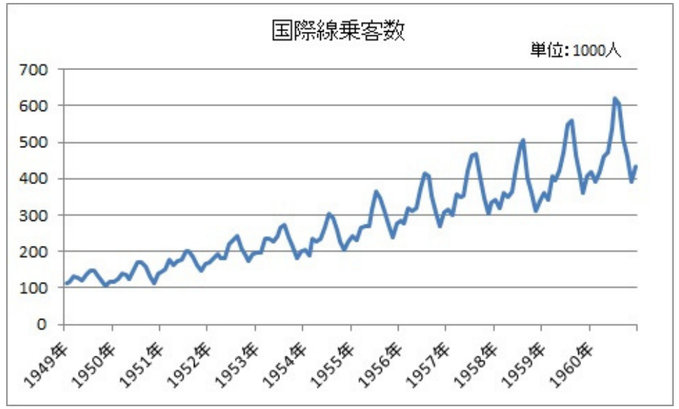
5～252日周期で終値をSTL分解
→いずれも残差成分に対する季節成分が非常に小さい
→目立った周期性は見られず

※STL分解とは時系列データを傾向・季節・残差に分解する手法です

終値の自己相関：なだらかに減少
→非定常なトレンド
終値の偏自己相関：ラグ 2 以降ほとんど0
→高次の自己回帰モデルは不要

04 モデルの選択

EDAの結果から従来の統計手法である ARIMAモデルが使えそう



ARIMAモデルとは

- 機械学習ではなく、統計的な時系列予測手法
- 図のように傾向と周期性・季節性が明確な場合に有効

- しかし、データ量が大きいためARIMAモデルでは学習に時間がかかる
- 季節成分がほとんどない
- トレンド成分だけでなく急激な変動も捉えたい
- 外部の要因が強く影響する

→LSTMを使用

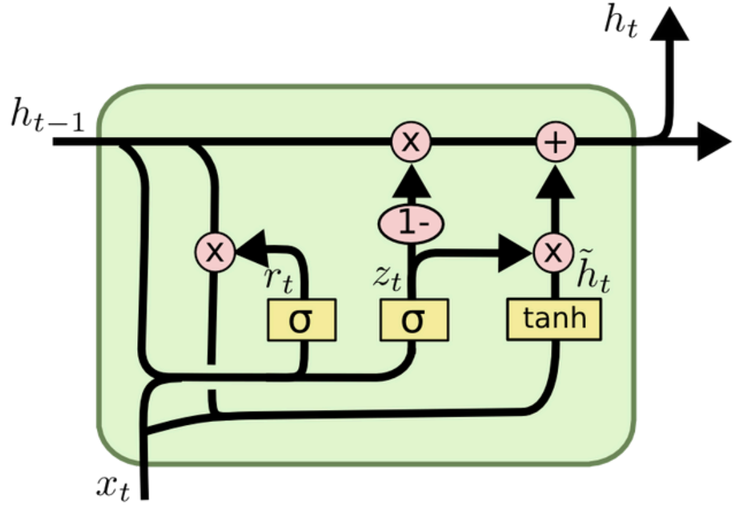
出典：https://www.nag-j.co.jp/naglib/tr/arima/arimaModel.htm

04 モデルの選択

LSTMとは

- 過去の重要な情報を長期的に覚えておくのが得意なAIモデル
- ノイズ（いらない情報）を無視できる
- 時間の流れを考慮した判断が可能

過去22日分のデータを入力として
未来1日分の終値を予測するモデルを構築



出典：https://colah.github.io/posts/2015-08-Understanding-LSTMs/

05 特徴量エンジニアリング

まず多数の特徴量を作成し、評価指標をもとに不要なものを削除していく

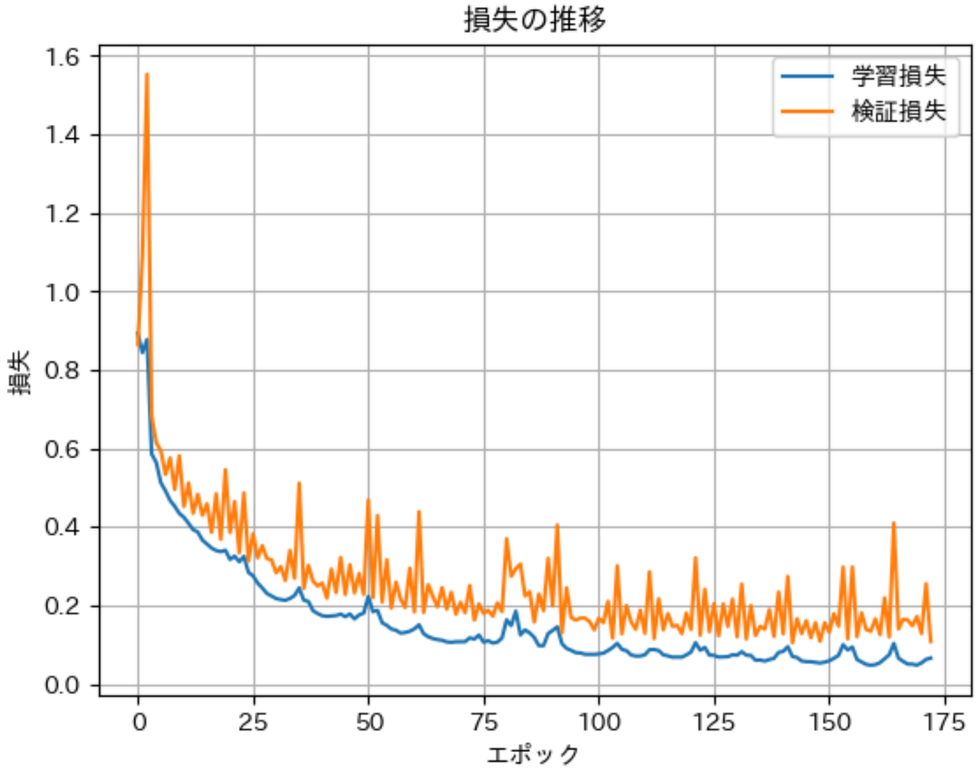
- 終値の単純・指数移動平均→市場の瞬間的な方向感、全体的なトレンド
- 終値の相対力指数→売られすぎか買われすぎか
- 終値の標準偏差→価格変動の大きさ(ボラティリティ)
- ゴールデンクロス、デッドクロス→トレンド転換のシグナル
- MACD→上昇・下降傾向
- ADX→トレンドの強さ
- ストキャスティクス→過去一定期間の高値・安値との位置関係
- 時間的特徴量→日・週・年の周期的なリズム
- 日本の休日→祝日前後の特殊な取引傾向

06 結果

- 大きな誤差にペナルティを与えたい
→評価指標にMSEを使用（小さいほど性能が良い）
※スケーリング後のデータに対して使用

  - 訓練データのMSE：約0.0701
  - 検証データのMSE：約0.1050

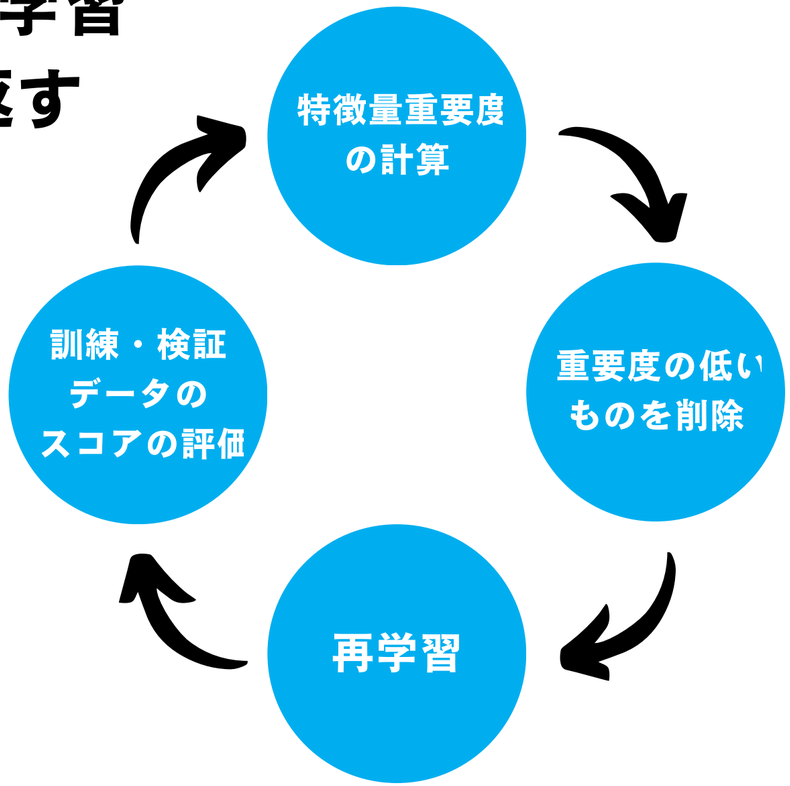
学習曲線を見ると
損失が不安定な部分が見られる



07 仮説の検証

・仮説１：特徴量が多すぎる
→（順列重要度の計算→値の低いものを削除→再学習
→訓練・検証データのスコアの評価)を繰り返す

ゴールデンクロスやMACDなどの
冗長な特徴量を削除
→最終的に検証データのMSEが0.0838まで改善



07 仮説の検証

- 仮説２：検証データに偏りがある

→バッチサイズを64→128に変更

→訓練MSE：0.0478、検証MSE：0.0612

- 仮説３：検証データに外れ値が含まれている

→2023年5月下旬から終値の変動が大きく不安定

→検証期間を2023年1月～5月15日に限定

→訓練MSE：0.0478、検証MSE：0.0460

07 仮説の検証

- 仮説 4 ：予測が過去の情報の影響を受けすぎている

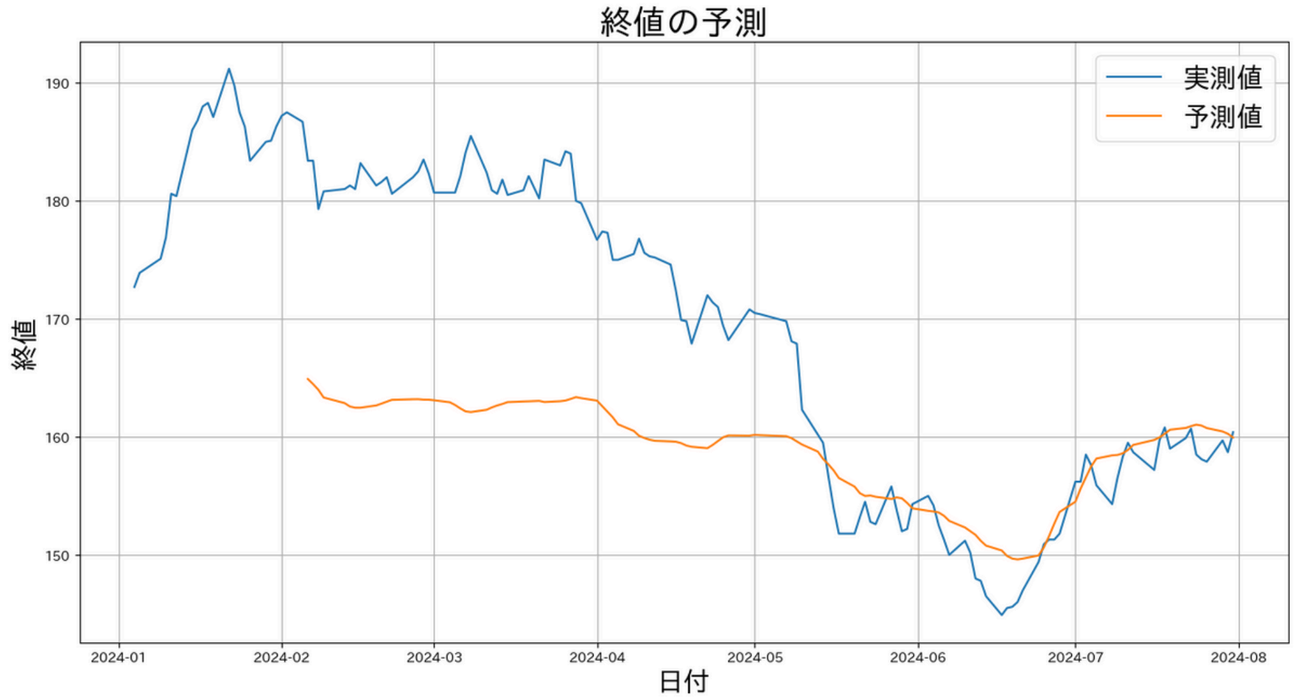
→20日移動平均を削除

→訓練MSE：0.0367、検証MSE：0.0395

特徴量選択や検証データの修正により、
ベースラインに比べて訓練・検証データの
スコアが改善

08 検証結果と考察

入力期間を22日としたモデルに学習させ、テストデータを用いて汎化性能を測った



- テストデータのMSE：0.3322
- 逆スケーリング後のRMSE：12.0041

→検証データのMSEに比べ
はるかに性能が低下
約12円分の誤差

- 振幅が大きく、急激な動きに追従できていない
- 2024年5月ごろまで低く予測されている
- 2024年7月ではやや改善
- 全体的になめらかな予測になっている

09 結果のまとめ

- 入力期間22日のLSTMモデルを用いて株価予測を行った結果、
  テストデータで性能が低下し、約12円の誤差が生じた
- 予測結果が全体的になだらかな推移となっており、実際の株価に見られる
  振幅の大きさや急激な価格変動に十分に追従できていない
- 特に価格の水準が異なるデータに対しては予測が外れやすい傾向
- 2024年5月頃までは実測値より低い予測値となっているが、
  7月にかけては徐々に改善している様子が見られる

10 今後の展望と追加検証

今後の展望

- 予測に寄与していない特徴量をさらに除外し、より有効な特徴量（テクニカル指標や市場センチメント指標など）を追加することで精度向上が期待できる
- 近年時系列予測で高い性能を示しているTransformerベースのモデルの導入
- LSTM、GRU、Transformerなどの複数のモデルのアンサンブルにより、安定した予測が可能になることが期待できる

追加検証としてTransformerモデルを構築して予測
→テストデータに対する逆スケーリング後のRMSEが5.1642に改善
（予測誤差が約5.2円に改善）

新しいデータを学習させることにより、モデルに現在のトレンドを理解させ、モデルを更新していくことができます
Transformerモデルで特徴量エンジニアリングを行うことにより5.2円よりも少ない予測誤差を達成することが期待できます