LSTMによる株価予測モデルの構築

Trainee業務体験課題

藤居海誠

内容

- 背景
- データの分析結果
- 技術概要
- 評価指標
- 検証内容
- 検証結果1
- 検証結果2
- まとめ

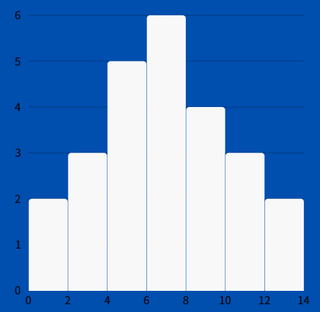
株価予測の重要性や課題とは？

株価予測の重要性

投資戦略の策定	→ トレンドフォロー戦略
リスク管理	→ 分散投資、ロスカットルール
資本の効率的運用	→ 長期的な資産保全

株価予測の課題

01.データの量



精度の高いモデルの構築には十分なデータが必要です

02.予測不能な要素

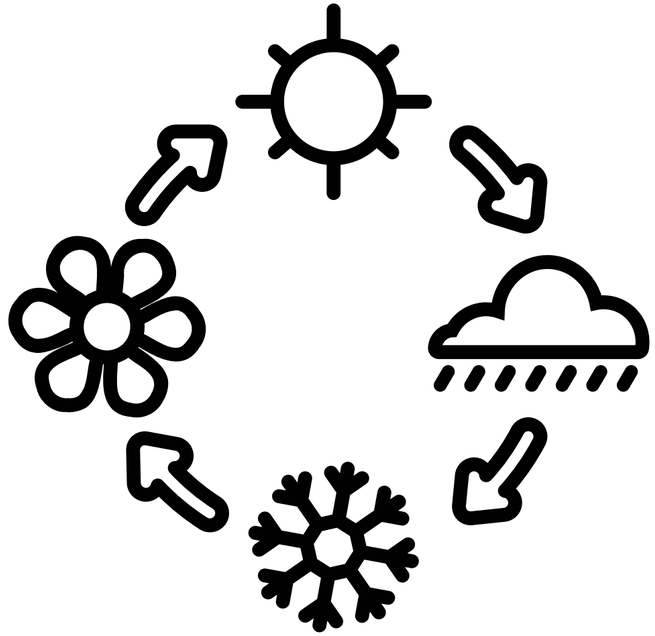
自然災害や政治的な変動等

03.環境の変化

市場変化によって従来の予測モデルが通用しなくなります

EDA

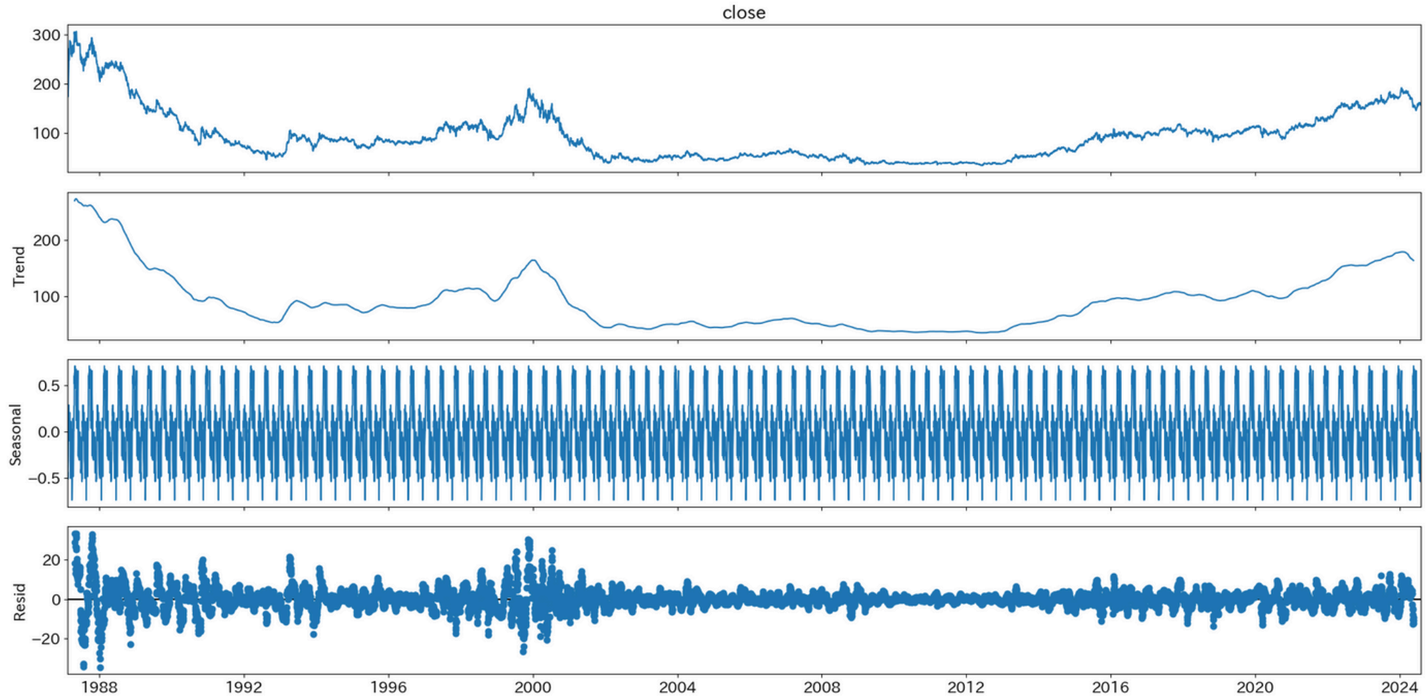
- データの型、欠損値の確認、異常値なし
- 列名の英語化
- トレンド、季節性、ヒストグラム分析
- ACF/PACFで周期性を確認



データの分析結果

時系列のトレンド分析

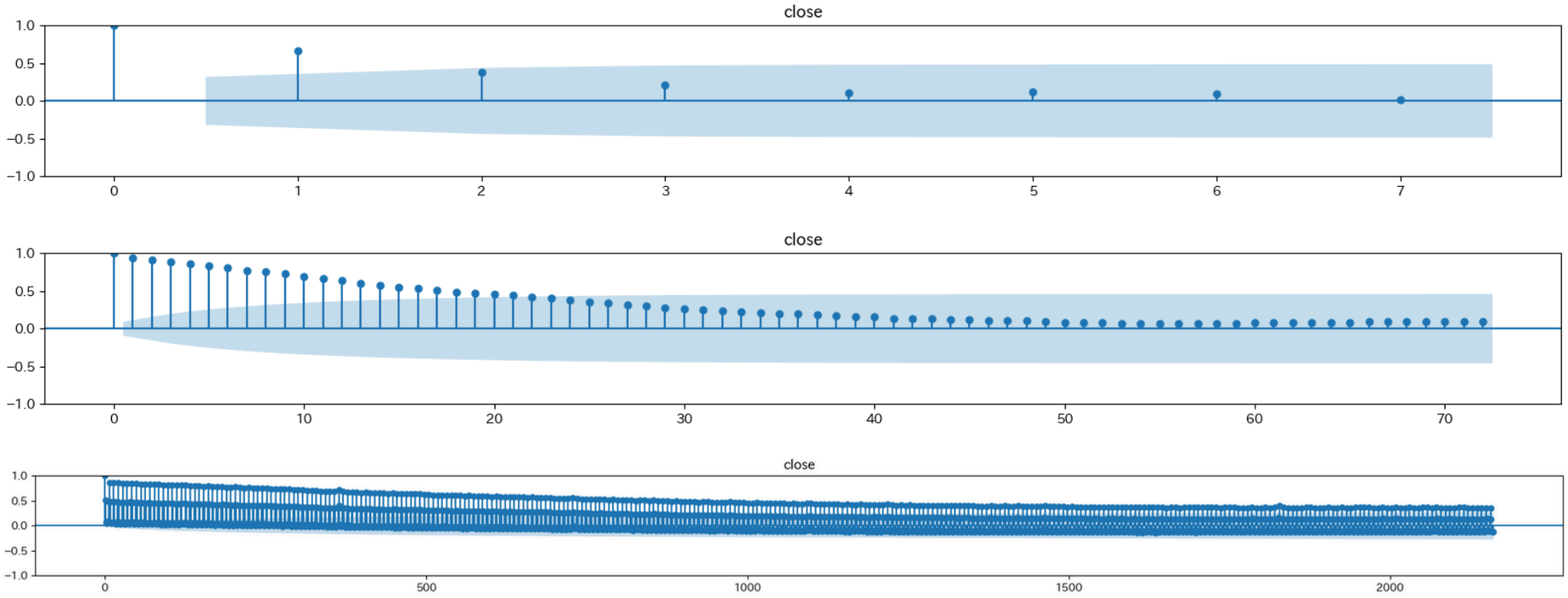
終値から傾向変動、季節変動、循環変動、不規則変動を抽出しました。



データの分析結果

周期性と時間的な相関(ACF)

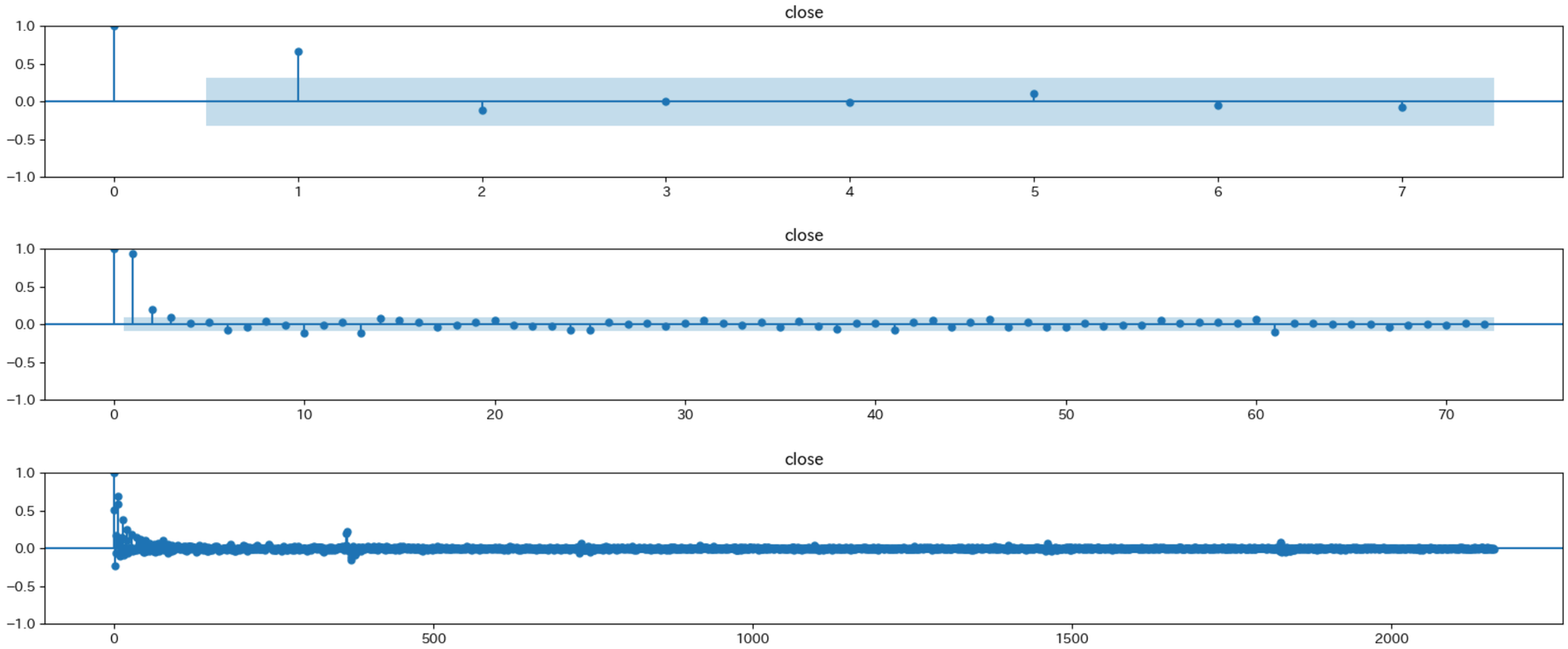
終値から自己相関関数のグラフを作成しました。
グラフは上から順に年、月、週ベースです。



データの分析結果

周期性と時間的な相関(PACF)

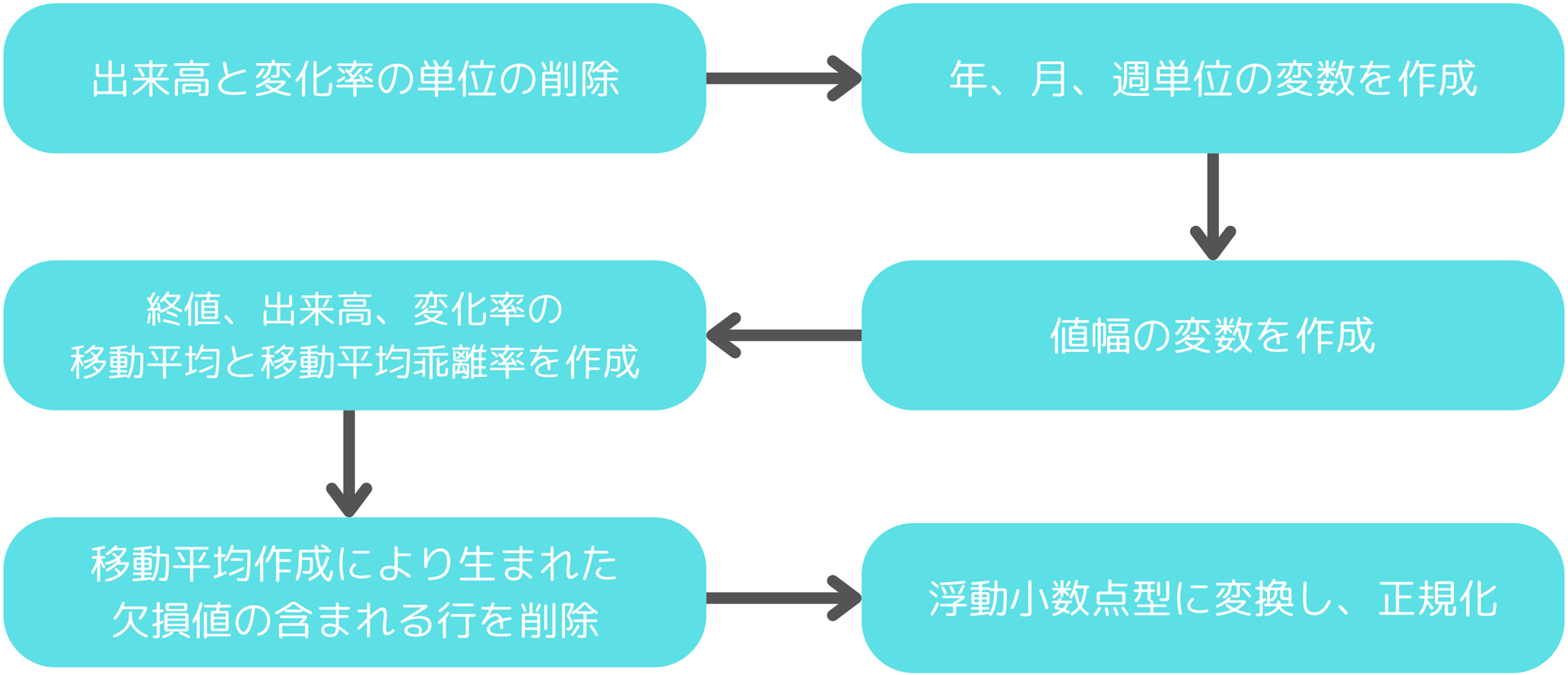
また、偏自己相関関数のグラフも作成しました。
グラフは上から順に年、月、週ベースです。



使用したモデル：LSTM

今回学習に用いた株価データは長期的かつ非定常的なパターンを持ち
ボラティリティが高いため、
そのようなデータの予測に適したLSTMを選択しました。

特徴量エンジニアリングの手法



特徴量エンジニアリングの手法

追加した特徴量

- 1日前の始値と終値の差
- 1日前の高値と安値の差
- 1日前の始値と2日前の終値の差

} データリークを防ぐため

- 年月日を三角関数に変換したもの
- 1年間、1か月、1週間の中で何日目にあたるかを表す変数
- 終値、出来高、変化率の移動平均と移動平均乖離率

特徴量エンジニアリングの手法

各特徴量を作成した理由

- 大きな値幅が発生した場合、市場はその銘柄に対して注目が集まる可能性が高まります。
- また、ACF/PACFではっきりしなかった周期性を確認するため、終値に年、月、週の周期性があるという仮説のもとに三角関数化したものを作成しました。
- 終値・出来高・変化率に関してもデータリークが起こらないように移動平均とその乖離率を作成しました。

特徴量エンジニアリングの手法

データリークを防ぐため、以下の説明変数を削除しました。

- 始値
- 高値
- 安値
- 出来高
- 変化率

これらの変数は過去のデータとして
新しく追加した特徴量に含まれています。

評価指標

以下の評価指標を用いました。

学習時の損失関数

- MSE

MSEはAdamなどの最適化アルゴリズムに適しており、外れ値に敏感ですが今回のデータには外れ値が無いため。

テスト時の損失関数

- MSE
- RMSE
- MAE
- R2

RMSEは終値と単位が同じであるため、誤差を解釈しやすいです。
MAEは外れ値の影響を比較的受けません。
決定係数R2はモデルの説明力を測定したもので、比較が容易なので用いました。

評価指標

テストの結果は以下のようになりました。

MSE:0.000377

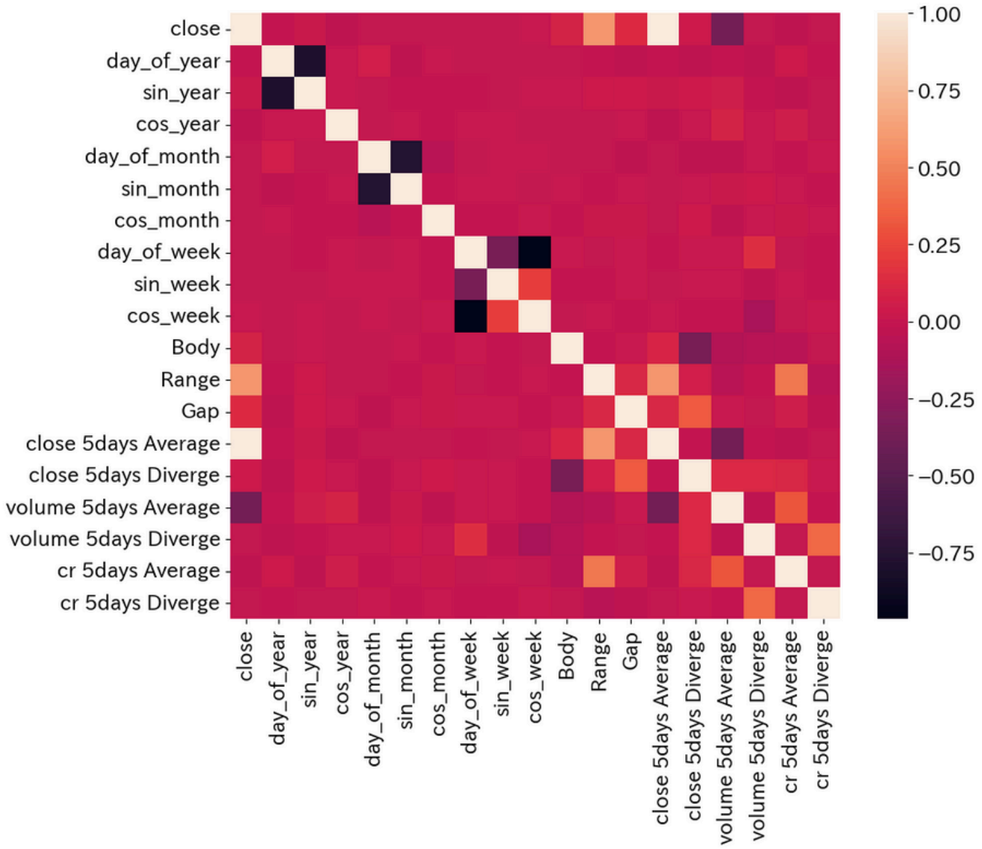
RMSE: 0.0194

MAE:0.0159

R2: 0.965

評価指標

相関係数をヒートマップにプロットした結果



- 高値と安値の差
- 過去5日間の終値の移動平均
- 過去5日間の出来高の移動平均

以上の3つの変数が終値との相関が最も大きいことがわかりました。

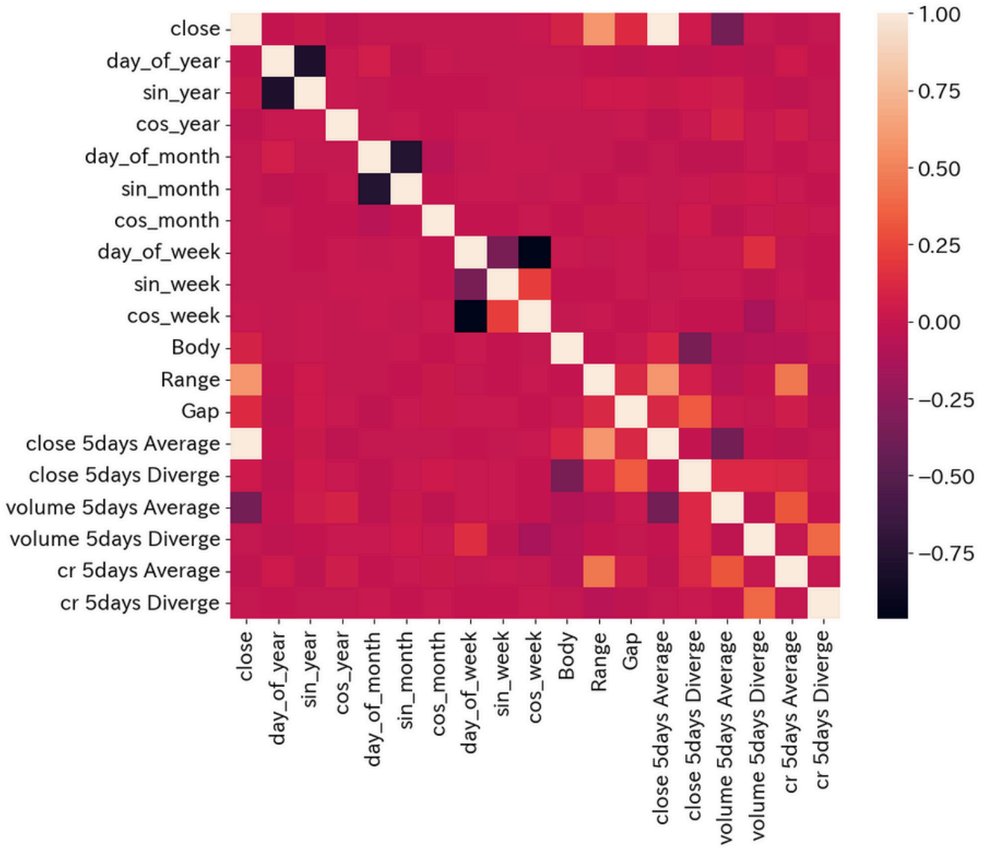
検証内容

評価の結果から得た仮説

- 仮説１：多重共線性
- 仮説２：不適切な学習率

検証内容

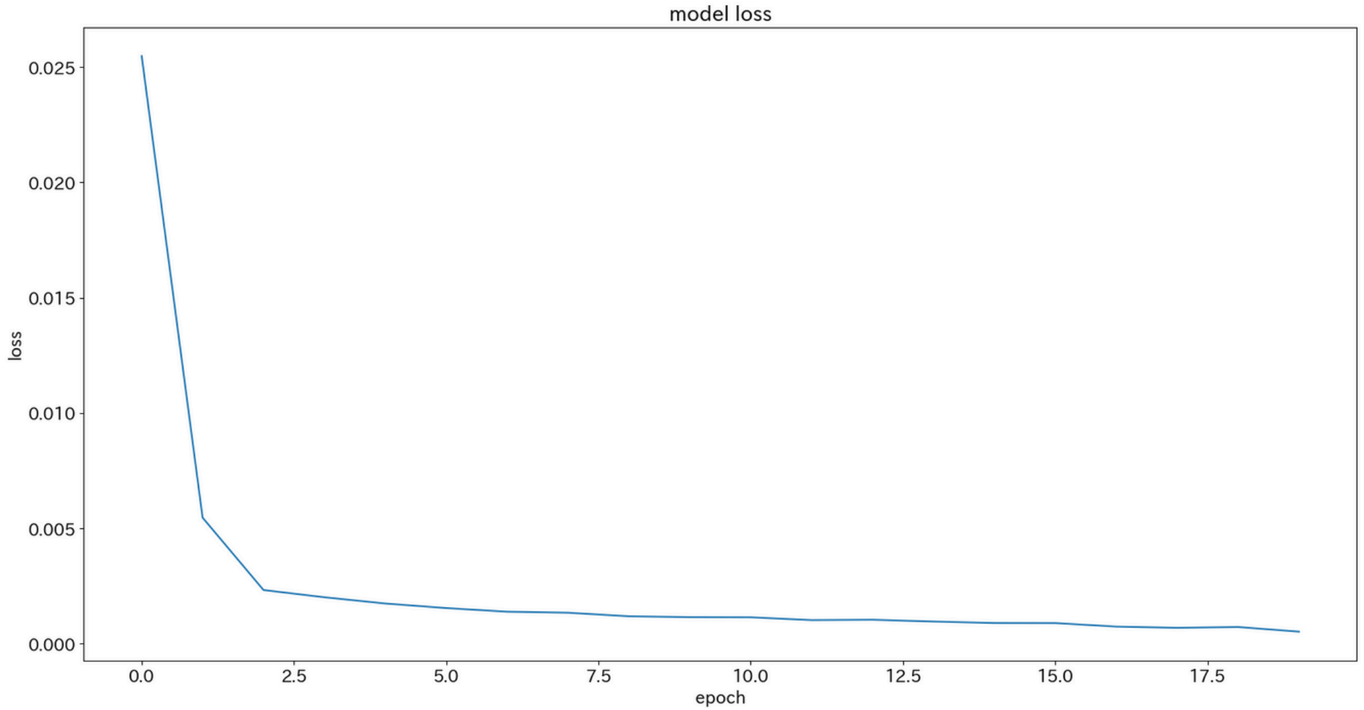
ヒートマップから得た仮説



- 説明変数の中で相関の大きい組がいくつか見られます。
  →多重共線性が存在している？

検証内容

損失関数のグラフから得た仮説



損失が減少しない区間が存在
→より最適な学習率が存在するのではないか

検証結果1

仮説１：多重共線性

各説明変数のVIFを計算し、それらの中で最もVIFの高い変数を削除することを繰り返しました。

VIFが10以上である説明変数は、
1日前の始値と2日前の終値の差
高値と安値の差
過去5日間の終値の移動平均乖離率
1週間の何日目かを表す変数
でした。

→

それらを削除した結果
MSE: 8.05e-05
RMSE: 0.00897
MAE: 0.00719
R2: 0.993
となりました。

検証結果2

仮説２：不適切な学習率

学習率スケジューリングを用いて学習率を学習の途中で変化させることを試みました。
→学習率は変化しましたが、最終的な損失の値はほとんど変化しませんでした。

最終的に
MSE: 8.05e-05
RMSE: 0.00897
MAE: 0.00719
R2: 0.993
となった

まとめ

成果

- 過学習を起こすことなく、テストデータに対して汎化性能の高いモデルを構築することができました。
- 株価の終値を少ない誤差で予測することができました。

今後の展望

- 他の株価データを予測することによって、投資戦略の立案、リスク管理、自動売買システムの構築に役立てることができるはずです。