株価予測モデルの構築

LSTMによるNTT株の予測

市川　蓮刀

概要

- 株価予測の背景
- データの分析
- 技術概要
- 検証内容と評価指標
- 結果と改善
- 結果のまとめと今後の展望

株価予測の背景

重要性

・投資判断を最適化できる

利益を最大化し、損失を最小化できる。

・リスク管理

将来の株価の変動を予測しリスクの高い資産の比率を調整できる

課題

・不確定性

ニュース、企業の業績、人間の心理など様々なものが絡みあって非線形的な変化を見せるため予測が困難

・外部要因の予測不可能性

非線形的な変化の要因となる外部要因は予測困難

データの分析　－モデルの決定に向けて

・基本統計量の確認

　各データの平均、標準偏差、最小値、最大値、四分位数

・時系列データのトレンドの把握

　週平均、月平均などのグラフへの描画

・季節性の確認

　STL分解、時系列分解

・異常値の検出

　閾値以上の残差を検出

基本統計量の確認

基本統計量

	終値	始値	高値	安値	出来高	変化率 %
count	9202.000000	9202.000000	9202.000000	9202.000000	9.202000e+03	9202.000000
mean	92.180961	92.256183	93.176451	91.330146	1.726677e+08	0.017502
std	50.452228	50.598215	51.049837	50.087405	1.251280e+08	1.876667
min	33.000000	33.000000	33.200000	32.200000	9.340000e+06	-14.740000
25%	52.000000	52.100000	52.800000	51.500000	8.073000e+07	-0.940000
50%	85.100000	85.100000	86.050000	84.200000	1.540150e+08	0.000000
75%	110.800000	110.800000	111.900000	109.275000	2.305225e+08	0.900000
max	305.900000	309.800000	311.800000	303.900000	1.280000e+09	16.250000

- 終値

平均：92.190961　標準偏差:50.452228　最小値:33.0　最大値305.9

データセット数：　9202　日分

長期的なトレンドの把握　－トレンド、非線形性

NTT 株価の終値のプロット

週平均：黄色
月平均：青

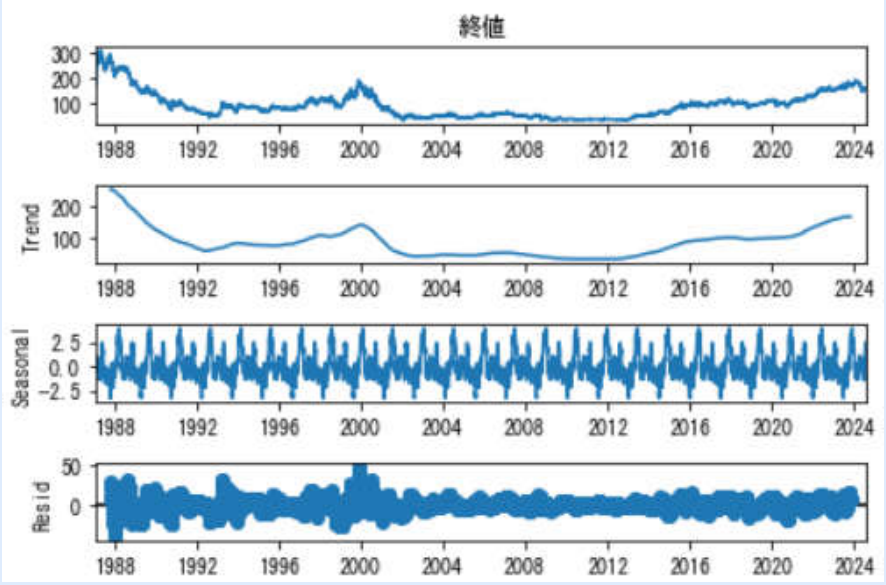
長期的なトレンドが見受けられる。
・バブル崩壊（1990年前後）
・ITバブル期と崩壊（続く2000年前後での山）
・リーマンショック（2008年）
・回復（2010年前後～2020年前後）
・コロナの社会での需要拡大（2021年～）

＊急激な値の変化

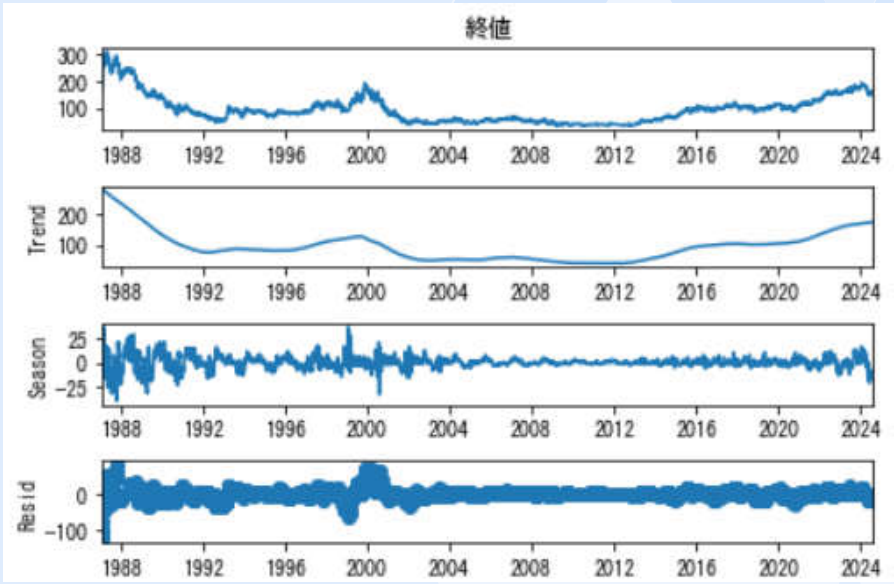
＊非線形性
株価は外的要因が複雑に絡み合って決まる。

季節性の把握　ー明確な季節性を確認

時系列分解

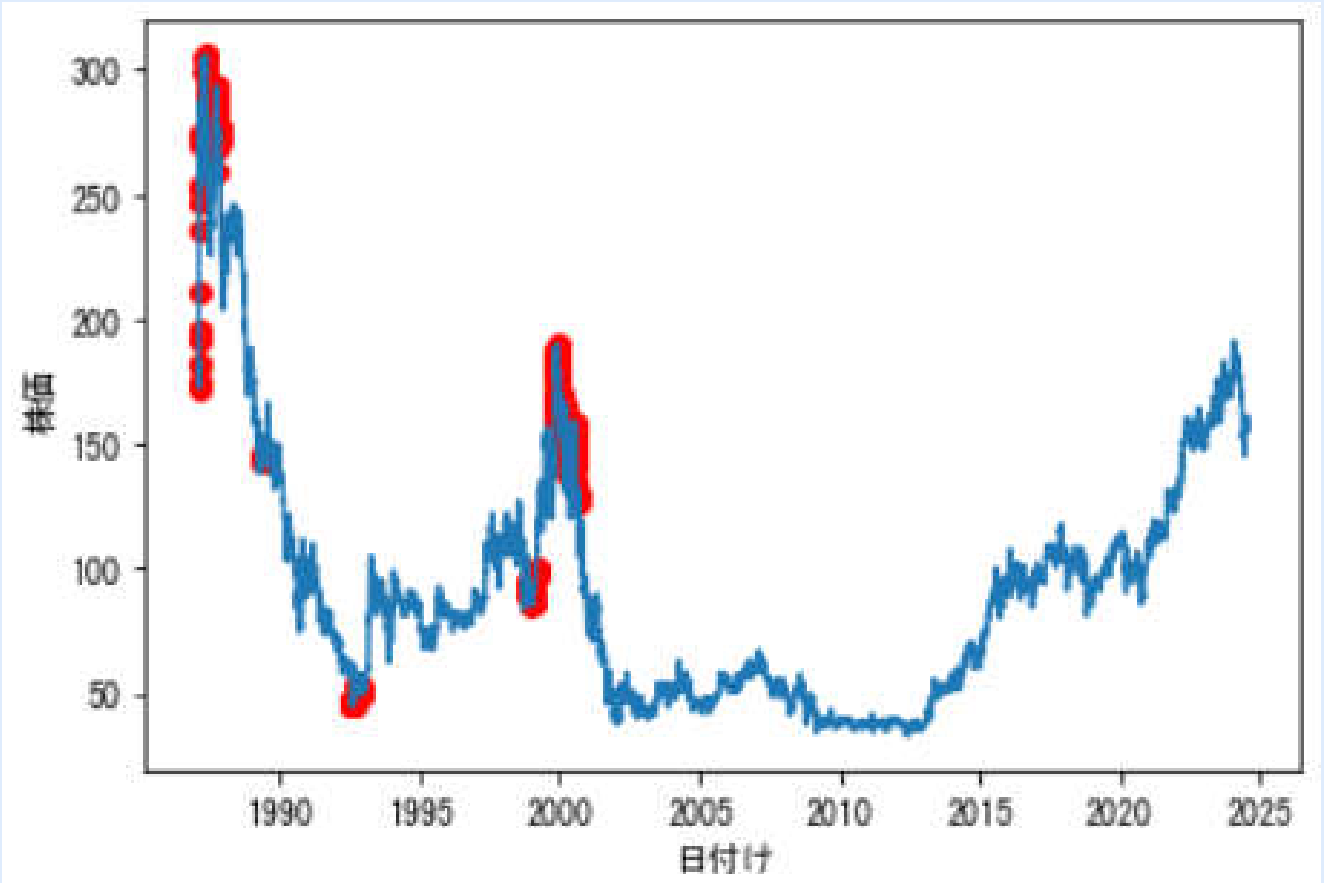


STL分解



seasonal成分に周期的な変化が見受けられることから、季節性が確認できる。

異常値の検出



平均から標準偏差の3倍よりも大きい値、または小さい値を閾値としその範囲外のものを異常値として検出した。

バブル期、バブル崩壊、ITバブルのあたりに異常値が検出されている。

モデルの決定　－LSTMを選択

・明確なトレンド、季節性
・急激な値の変動
・非線形性
・長期的なトレンドに依存

・**LSTM (Long Short-Term Memory)**

高度な予測を実現するリカレントニューラルネットワークモデル。
トレンド、季節性のある時系列データの予測モデルに適し、非線形な変化への対応力や長期的な依存関係に対する学習力という点で優れる。

技術概要
－欠損値の補完と正規化

LSTM

長期依存性の学習を可能にした、拡張リカレントニューラルネットワーク

データの前処理

－LSTMに入力できる形にデータを整える

・欠損値の補完

　－欠損データがある場合その前日の値で欠損データを補完

・正規化

　－データの（最大値、最小値）が（1，0）になるようにスケーリング。

　－入力値を一定の範囲に収めることで学習の効率と速度を上げる。

検証内容と評価指標
－決定係数と二乗平均平方根誤差

・与えられたNTTの株価データの前半80%を学習データとして用いモデルをトレーニング

・残り20%をテスト用データとして用い、予測値とのずれを評価して予測精度の検証とする。

評価指標

以下の二つの評価指標によりモデルの予測精度を評価

・二乗平均平方根誤差

RMSE :0に近いほど予測精度がいい

(Σ（y__pred－y__true)^2)^1/2

・決定係数

R＾2 :1に近いほど予測精度がいい

1-Σ(y_true-y_pred)^2/Σ(y_true-y_mean)^2

結果と改善
－3つの実装とその評価

次の3つのモデルを実装し、lstm_model3が最終成果物。

lstm_model ： 初期実装。1層の単一LSTM層からなる予測モデル。

lstm_model2 ：4層のLSTMモデルと各層の後にdropout率0.2のdropout層を設置

lstm_model3 ： dropout率を調整

初期実装lstm_modelから始め、次にlstm_model2,lstm_model3を順に実装した。

LSTM_MODEL　－高水準な予測精度

初期実装

単一のLSTM層からなる

Epoch＝100　:100回繰り返しで学習

Batch_size=32　:一回の重み更新で使用するサンプル数

LSTMの内部ユニット数は50。

活性化関数：ReLU関数

Window=60:過去60日分のデータをもとに61日目を予測

結果

- rmse=1.918667448741605
- r2=0.9954236745834351

高い精度で予測モデルを実現できた。改善案としては、より高い表現力を求めるために
LSTM層を多段化すること。

ＬＳＴＭ__ＭＯＤＥＬ2　－予測精度の悪化

より高い表現力を実現するためにLSTM層を4層に拡張。

多層構造にしたことによる過学習を抑制するために、各層の後に重み0.2のdropout層を導入

Epoch=100　:100回繰り返しで学習

Batch_size=32　:一回の重み更新で使用するサンプル数

結果

- rmse=2.0589180715249484
- r2=0.9947301745414734
- 依然として高い水準を保っているものの、初期実装に比べて予測精度の悪化がみられた。

ＬＳＴＭ＿MODEL3　ー予測精度の改善

過剰なドロップアウトが原因で学習が阻害された可能性があるので、ドロップアウト層を3層に減らし、ドロップアウト率を（0.2，0.1，0.05）に設定。

結果

- rmse=1.5762338540803946
- r2=0.9969114065170288

・初期実装と比較しても精度の改善が明確にみられた。
・やはりlstm_model2においてドロップアウト層による学習の阻害が起きていたと考えられる。
・本実装を今回の提出の最終成果物とする。

課題と今後の展望　– 長期的な予測について

- **課題**

  今回は過去60日分のデータから６１日目の株価を予測した。

  このように過去数十日分のデータから未来数時間～数日の予測は高精度でできる。

  数週間から数年の長期的な予測については未確認であるが、かなり困難であると予想している。

- **展望**

  長期的な予測が可能であるか確認するとともに、可能でないならばその手法について探求していきたい。