知能情報実験 III グループ 1
テーマ：機械学習で株価予測をしてみる

225752B PARK CHEOLHWAN
225741G 清水 優馬
225745K 大石根 竜馬
225754J 當山 一朗

2024 年 7 月 18 日

目次

1 はじめに

1.1 実験の目的と達成目標

知能情報実験 Ⅲ は、情報工学分野のより専門的な知識を理解・習得することを目的として、半年間でシステムの開発やデータ解析等に取り組む実施される。その中の一つデータマイニング班においては機械学習外観ならびにその応用を通し、対象問題への理解、特徴量抽出等の前処理、バージョン管理やデバッグ・テスト等を含む仕様が定まっていない状況下における開発方法、コード解説や実験再現のためのドキュメント作成等の習得を目指す。

1.2 テーマ: 株価予測とは

本グループでは、個人投資家の株の投資における将来の株価を予測することを対象問題として設定した。

株価予測とは、市場に出回っている株価の推移を予測することであり、テクニカル分析やファンダメンタルズ分析を使って予測することが一般的である。今回の実験では、数値データとして扱うことができるデータの多いテクニカル分析を用いて株価予測を実施する。テクニカル分析とは、参考文献 [1] によると、 移動平均線、株価チャートなど、株価データの「型」（＝パターン）を基礎に、相場の先行きを予測することである。また、株価の変動は、様々な要因によって引き起こされる。例えば、企業の業績や経済指標、政治情勢、自然災害などが挙げられる。これらの要因を分析し、株価の変動を予測することで、投資家はリスクを最小限に抑えながら、収益を最大化することができる。

2 実験方法

実験方法としては、下記のような手順で進める予定で進むことにした。

- 実験目的：実験を進む前に、実験で進むテーマを明確にし、目的を設定し、テーマとしては、株価予測を選定した。
- データセット構築：株価データを取得し、テクニカル分析を行うためのデータセットを Yfinance API を用いて構築し、実験の対象としては、時事的な要因が少ない株価である 1321.T（日経 225）を選定した。
- モデル選定：株価予測に有効なモデルを探すために、資料などを参考にした結果、LSTM（Long Short-Term Memory）モデルを選定した。
- パラメータ調整：LSTM モデルのパフォーマンスを最適化するために、いくつかのパラメータを調整した。パラメータ調整としては、エポック数、バッチサイズ、LSTM ユニット数、

Dense 層、学習率、データの正規化を調整した。

- 結果：実験結果をまとめ、考察を行い、意図していた実験計画との違いを検討し、まとめを行った。

2.1 実験目的

本グループでは、株価予測モデルの有効性を検証することを目的としている。具体的には、テクニカル分析を用いた移動平均線や株価チャートのパターン認識を通じて、株価の変動をどの程度正確に予測できるかを明らかにする予定である。また、異なる分析手法やパラメータの組み合わせが予測精度に与える影響を確認し、最適な予測モデルを特定することを目指している。

2.2 データセット構築

yfinance API を用いて、株価データを取得する。取得したデータは、Open, High, Low, Close, Volume, Dividends, Stock Splits の 7 つのカラムからなる（Date は除く）。また、取得したデータを元に、直近 3 年間の株価データを取得し、テクニカル分析を行うことができるデータセットを構築する。

yfinance API の URL は、参考文献 [2] である。

2.3 モデル選定

本実験では、株価予測において有効であると判断した LSTM（Long Short-Term Memory）モデルを用いる。LSTM は、リカレントニューラルネットワーク（RNN）の一種であり、特に時系列データの予測に優れた性能を発揮する。LSTM モデルは、過去のデータの長期依存性を捉えることができるため、株価のように時間に依存するデータに対して有効である。また、LSTM は従来の RNN に比べて勾配消失問題を解決する設計がなされており、長期間の依存関係を効果的に学習できるため、株価の変動パターンを正確に予測することが期待できる。このため、本実験では LSTM モデルを選定した。

2.4 パラメータ調整

本実験では、LSTM モデルのパフォーマンスを最適化するために、いくつかのパラメータを調整しました。また、下記は主要なパラメータとその調整過程を示している。また、下記は、パラメータを調整する際のログを示している。

Listing 1　バッチサイズが 1 の場合

```
epochs30
RMAE:473.37
max:1341.05
```

```
min:−990.22

epochs20
RMAE:500.35
max:1442.59
min:−899.36

epochs10
RMAE:509.82
max:1487.57
min:−1021.09

epochs5
RMAE:592.01
max:556.99
min:−1596.20

50
RMAE:1452.70
max:3433.27
min:−737.40
1
RMAE:1932.42
max:3833.04
min:−85.31
```

Listing 2　バッチサイズが 16 の場合

```
epochs30
RMAE:615.52
max:668.24
min:−1735.56

epochs20
RMAE:439.46
max:1143.30
min:−985.56
```

```
epochs10
RMAE:710.27
max:2074.90
min:−1184.34

epochs5
RMAE:961
max:2423
min:−1386
```

Listing 3　バッチサイズが 32 の場合

```
epochs30
RMAE:735
max:
min:−

epochs20
RMAE:
max:
min:−

epochs10
RMAE:
max:
min:−

epochs5
RMAE:
max:
min:−
```

Listing 4　バッチサイズが 2048 の場合

```
epochs30
RMAE:
max:
min:−
```

epochs20
RMAE:2376
max:4549
min:−1084

epochs10
RMAE:
max:
min:−

epochs5
RMAE:
max:
min:−

上記の Listing 1、Listing 2、Listing 3、Listing 4 を通じて、エポック数は 20 30 が最適であることがわかった。また、バッチサイズは 16 が最適であることがわかった。よって、下記のようにパラメータを設定した。

■エポック数　Listing **??**　モデルの訓練において、エポック数は重要なパラメータである。本実験では、エポック数を 20 に設定した。将来的には、エポック数を増やすことでモデルの収束状況や予測精度が向上する可能性があるため、さらに検討する予定である。

■バッチサイズ　バッチサイズもモデルの性能に影響を与える重要なパラメータである。本実験では、バッチサイズを 16 に設定した。この設定は計算効率と予測精度のバランスを考慮したものである。将来的には異なるバッチサイズを試し、最適なバッチサイズを特定することを計画する予定である。

■LSTM ユニット数　LSTM 層のユニット数は、モデルのキャパシティに直接影響する。本実験では、2 つの LSTM 層を使用し、各層のユニット数を 128 および 64 に設定した。将来的には、ユニット数を増やしてモデルの予測性能がどのように変化するかを評価し、最適なユニット数を決定する予定である。

■Dense 層　Dense 層は、LSTM 層の出力を線形変換し、最終的な予測値を生成する。本実験では、LSTM 層の後に 2 つの Dense 層を追加し、最終的な出力層のユニット数を 1 に設定した。これにより、時系列データの特徴を捉えた後、適切な予測値を生成することが可能となった。

■学習率　最適化アルゴリズムの学習率は、モデルの収束速度と安定性に影響を与える。本実験では、デフォルトの学習率（0.001）を使用した。将来的には、異なる学習率（例えば 0.01 や 0.0001）を試し、モデルの収束速度と安定性を最適化することを検討する予定である。

■データの正規化　データの正規化は、モデルの収束速度と予測精度に大きな影響を与える。本実験では、MinMaxScaler を使用してデータを 0 から 1 の範囲にスケーリングした。将来的には、標準スケーリングやロバストスケーリングなどの他のスケーリング手法も試して、モデルの性能向上を図る予定である。

3　実験結果

本実験では、LSTM モデルを用いて 1321.T（日経 225）の終値を予測した。実験の結果としては図 1 にで示している。　図 1 には、訓練データ、実際のデータ、予測データの 3 つの線が示され

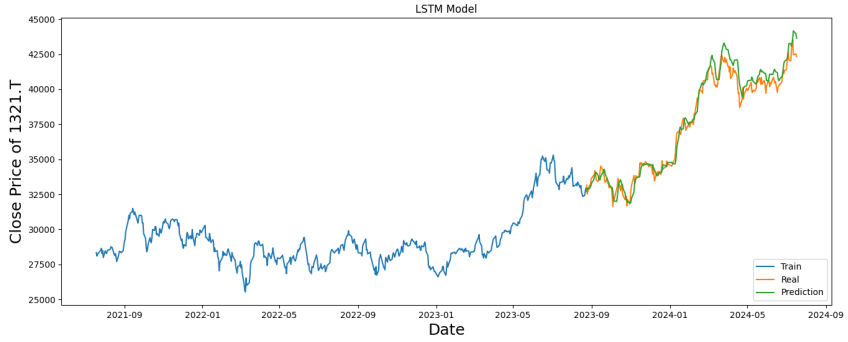

図 1　exericise の test_two.py の実行結果

ている。青色の線は訓練データを表し、オレンジ色の線は実際のデータ、緑色の線は LSTM モデルによる予測データを示している。軸ラベルはそれぞれ横軸が日付（Date）、縦軸が 1321.T の終値（Close Price of 1321.T）である。

4　考察

今回の実験を通じて、LSTM モデルが 1321.T の終値予測に対して高い精度を持つことがわかった。特に 2023 年以降のデータにおいて、予測データと実際のデータが非常に近い値を示していることが確認できる。近い値を得ることができた理由としては、LSTM モデルの適用により、時系列データのパターンを捉えやすくなったからである。一方で、予測における誤差もいくつか見られた。特に、急激な価格変動が発生した部分では、モデルの予測精度が低下する傾向にあった。この点については、外部要因（経済ニュースや市場のイベントなど）を考慮したモデルの改良する必要がある。

今後の展望としては、予測精度の向上を目指し、他の時系列予測モデルとの比較検討を行うことが挙げられる。また、予測結果を活用した投資戦略の構築にも取り組みたいと考えている。また、実験を通して得られた知見を基に、次のステップとしてさらなるモデル改良と応用研究を進める予定である。

5　意図していた実験計画との違い

当初の実験計画では、LSTM モデルのパラメータ調整や外部要因を考慮した予測モデルの改良に十分な時間を割くことを予定していたが、LSTM モデルのトレーニング中に予期せぬエラーが発生し、その対応に多くの時間を取られた。これにより、パラメータ調整の時間が不足した。また、株価の変動には経済ニュースや市場のイベントなどの外部要因が大きく影響するため、これらをモデルに組み込むことを計画していたが、モデルの基本的なトレーニングに多くの時間を割いたため、外部要因を考慮した実験を十分に行うことができなかった。

6　まとめ

データマイニング班として設定したテーマ「株価予測」を通じて、多くの知見を得ることができた。特に、LSTM モデルの適用により時系列データの予測精度を高める方法について深く学ぶことができた。今回の実験では、LSTM モデルのパラメータ調整やデータ前処理の重要性を再確認し、適切なモデル構築のプロセスを実践することができた。

実験の結果、LSTM モデルは 2023 年以降の株価データに対して高い予測精度を示し、株価の変動パターンを捉える能力があることがわかった。一方で、急激な価格変動に対する予測精度が低下する課題が浮き彫りになった。この課題を解決するためには、経済ニュースや市場イベントなどの外部要因を考慮したモデルの改良が必要であると考えている。

グループワークを通じて、メンバー間のコミュニケーションや役割分担の重要性を学んだ。計画と実行のバランスを取ることの難しさを実感し、次回以降のプロジェクトにおいては、リスク管理と柔軟な対応が重要であることを認識した。

参考文献

[1] 野村證券株式会社、証券用語解説集、テクニカル分析、https://www.nomura.co.jp/terms/japan/te/tec_analysis.html。

[2] yfinance API、https://pypi.org/project/yfinance/。

[3] Keras ライブラリ、Sequential model、https://keras.io/guides/sequential_model/。

[4] Keras ライブラリ、Dense layer、https://keras.io/api/layers/core_layers/dense/。

[5] Keras ライブラリ、LSTM layer、https://keras.io/api/layers/recurrent_layers/lstm/。