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 8 月 1 日

目次

1 はじめに

1.1 概要

本レポートは、知能情報実験 III の一環として取り組んだ「機械学習で株価予測をしてみる」というテーマに関する実験の詳細をまとめたものである。本実験の目的は、テクニカル分析を用いて株価の変動を予測し、その有効性を検証することにある。実験では LSTM（Long Short-Term Memory）モデルを選定し、yfinance API を利用して日経 225 の株価データを収集し、データセットの構築から始まり、モデルの選定とパラメータ調整、予測結果の評価と考察に至るまでの詳細な手順を報告している。

実験の結果、LSTM モデルは特に安定している株価データに対して高い予測精度を示しているが、急激な価格変動に対する予測精度の低下が観察された。このため、外部要因を考慮したモデルの改良が必要であると結論付けた。また、実験計画と実行のバランスの重要性を再認識し、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求められることを学んだ。

最終的に、LSTM モデルによる株価予測の有効性を確認し、パラメータ調整とデータ前処理の重要性を再認識した。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改良が今後の課題となる。また、グループワークを通じてコミュニケーションや役割分担の重要性を学び、これらの知見を次回以降のプロジェクトに活かすことを目指している。このレポートは、LSTM モデルを用いた株価予測の実験を通じて得られた知見を基に、さらなる研究と応用に向けた具体的な方向性を示している。

1.2 テーマ: 株価予測とは

本グループでは、中長期で株式投資をする個人投資家の投資に対して、将来の株価を予測することの判断材料の一つとして機械学習を活用でき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。これらの要因を分析し、株価の変動を予測することで、投資家は株価の変動から状況を判断し、収益の最大化を目指すことができる。

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、参考文献 [2] によると、1321.T（NEXT FUNDS 日経 225 連動型上場投信）は、株式会社日本経済新聞社が発表している株価指数で、東京証券取引所プライム市場に上場する銘柄のうち株式市場を代表する 225 銘柄を対象に算出されている連動型上場投信である。企業単独の業績ではあまり影響がでず、日本企業全体の株価状況が一目でわかる指標かつ、多くの投資家が参考しているのもあり今回の実験で選定した。
- 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[*1]を用いた移動平均線[*2]や株価チャート[*3]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、1 日あたりの株価データをそれぞれ取得する。取得したデータは、Open（始値）、High（高値）、Low（低値）、Close（終値）、Volume（取引量）、Dividends（配当金）、Stock Splits（株式分割）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。yfinanceAPI の URL は、参考文献 [3] である。

また、Listing 1 は 1321.T の直近 5 日間のデータを示しているものである。

[*1] テクニカル分析とは、価格や取引量などの過去の市場データを分析することで、将来の価格動向を予測する方法である。

[*2] 移動平均線とは, 一定期間の株価の平均値を連続的に計算し、その推移をグラフで表示する方法である。

[*3] 株価チャートとは、株価の変動をグラフ形式で視覚的に表現したものである。これにより、トレンドやパターンを簡単に把握することができ、テクニカル分析において重要な役割を果たす。

Listing 1　1321.T の直近 5 日間のデータ

```
Date=2024-07-22, Open=41130.0, High=41170.0, Low=40710.0, Close=40770.0, Volume=247432.0
Date=2024-07-23, Open=41110.0, High=41140.0, Low=40710.0, Close=40800.0, Volume=220860.0
Date=2024-07-24, Open=40520.0, High=40840.0, Low=40280.0, Close=40300.0, Volume=381478.0
Date=2024-07-25, Open=39370.0, High=39480.0, Low=39000.0, Close=39130.0, Volume=913969.0
Date=2024-07-26, Open=39030.0, High=39270.0, Low=38760.0, Close=38810.0, Volume=506248.0
```

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。株価のデータが時系列データであるため、まず RNN（Recurrent Neural Network）と LSTM の二つのモデルを考慮した。それぞれのモデルには特性と利点があり、最適なモデルを選定するために比較検討を行った。

RNN は時間的な依存関係を学習するために設計されたモデルであるが、長期的な依存関係を学習する際に「消失勾配問題」という問題が発生する。これは、時間が経つにつれて勾配が消失し、学習が進まなくなる問題である。この問題により、RNN は長期間の依存関係を効果的に学習することが難しくなる。

この消失勾配問題に対処するために開発されたのが LSTM である。LSTM は RNN の構造に改良を加え、情報の流れを効果的に制御するためのゲート機構を導入している。具体的には、入力ゲート、忘却ゲート、出力ゲートの 3 つのゲートを使用し、重要な情報を保持し、不要な情報を忘却することができる。この設計により、LSTM は長期的な依存関係をよりよく学習することができ、株価のような複雑な時系列データに対しても高い予測性能を発揮する。

したがって、LSTM は RNN に比べて長期的な依存関係を保持するのに優れており、株価予測のようなデータにおいて、より正確な予測を実現する可能性が高いと判断した。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。

■エポック数　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。

■RMSE　参考文献 [7] によると、RMSE とは実際の値と何らかのモデルに基づく予測値があるとき、両者の差を二乗して平均し、平方根（ルート）を取った値である。また、実行結果には誤差の最小値（min）と最大値（max）を追加している。

最初にバッチサイズ 1 に固定し、エポック数を 1、5、10、20、30、50 の中で適切な値を探すことになった。その結果、エポック数が 1、50 の時は RMSE が 1000 を超えており、良好な結果とは言えないことが判明した。次に、バッチサイズを 1 、16、32、64、128、256、512、1024、2048 に調整したが、64 以上のバッチサイズでは精度が低下することが明らかだったため、これらのログは省略している。

Listing 2、Listing 3、Listing 4、Listing 5 は、パラメータを調整する際のログを示している。

Listing 2　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 3　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 4　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 5　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 2、Listing 3、Listing 4、Listing 5 を通じて、エポック数は 20 から 30 が最適であり、バッチサイズは 16 が最適であることがわかった。

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。実験結果は図 1 に示しているものである。

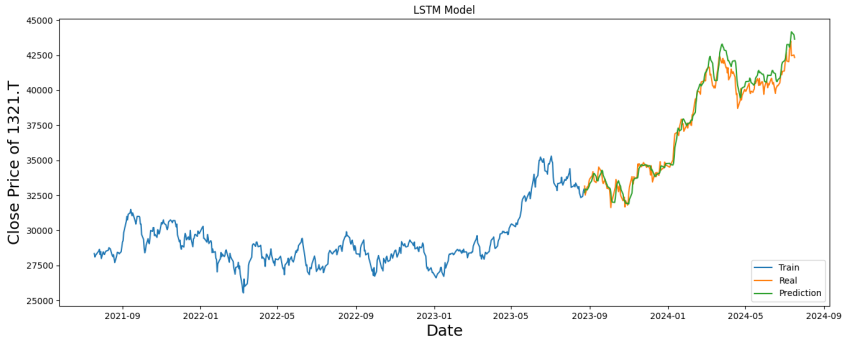

図 1　exericise の test_two.py の実行結果

また、実際にこのモデルと Streamlit を用いて、次のようなページを作成したものである。

■Web ページの URL：　https://finance-svkvjevmpsyc9f9ru8yefz.streamlit.app/

上記のページでは、図 2 のように 3 年間の株価データを表示し、それを用いて株価を予測するグラフを表示できるものである。

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ものである。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）であり、縦軸の単位は円である。

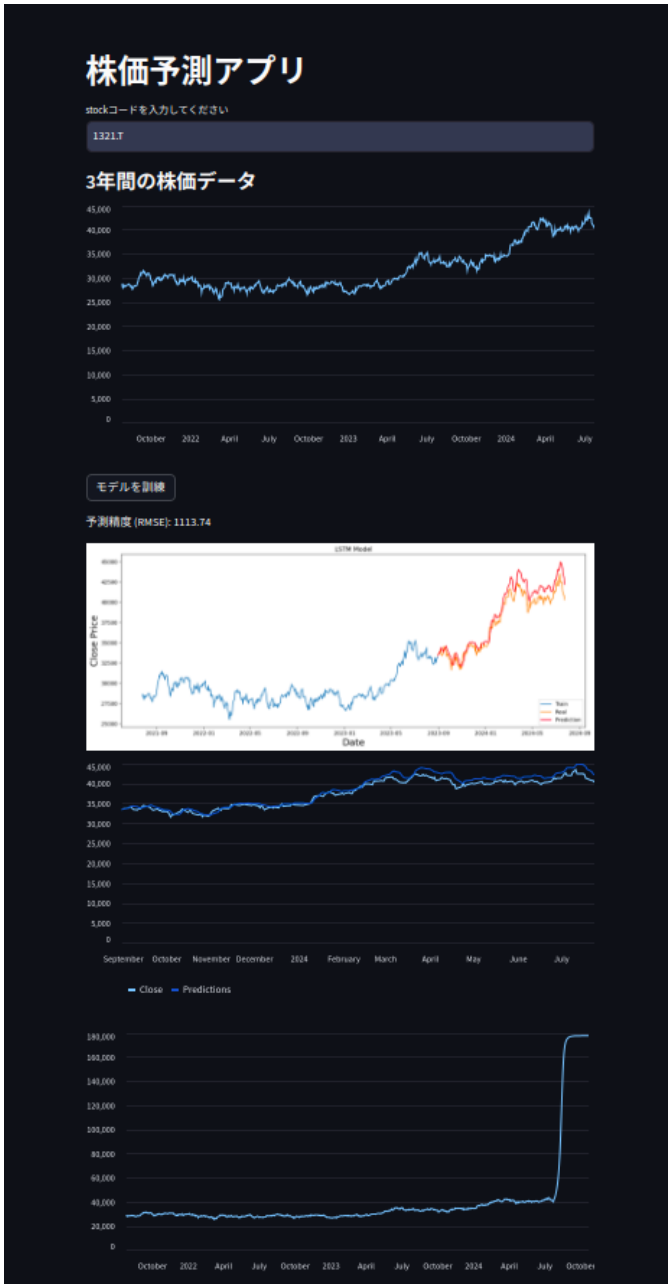

図 2　Streamlit を用いた Web ページ

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] 日本取引所グループ、日経 225 上場投資信託 (1321) 、https://www.jpx.co.jp/equities/products/etfs/issues/files/1321-j.pdf。

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

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

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

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

[7] IT 用語辞典、二乗平均平方根誤差 【RMSE】、https://e-words.jp/w/%E4%BA%8C%E4%B9%97%E5%B9%B3%E5%9D%87%E5%B9%B3%E6%96%B9%E6%A0%B9%E8%AA%A4%E5%B7%AE.html