株価変動要因の分析

― 日次データを用いた投資判断 ―

慶應義塾大学経済学部・池田真志

慶應義塾大学経済学部・畠中律紀

要　旨

一般投資家は IR を主な判断材料として投資を行う。しかし不安定な現代社会が起こす急速な変化に対応するには四半期前の情報では古すぎることがある。よって本稿では金利などの日次データから重要な株価変動要因を見つけることを目的とした。精度の高い分析を行うため、テクニカル分析とファンダメンタル分析を行いその結果をアンサンブルさせた。残念ながらアンサンブルさせた結果予測精度は下がってしまったが、元/円為替レートや原油先物価格は特に株価に強い影響を与えることが分かった。今後は IR 等の長期データを基にした上で、上記 2 つの指標に注目して投資判断を行うことが望ましいだろう。

第 1 章　序論

第 1 節　背景

2020 年 1 月、突如として発生し猛威を振るった新型コロナウイルスは世界中を大混乱に陥れた。その影響は政治面や経済面にまで及び、経済活動の抑制によって経済は停滞しほとんどの企業が業績を大きく落とし、倒産に追い込まれる企業も少なくなかった。しかし、明日世界がどうなっているかもまったくわからない不安定なこの時期に、投資を始めるあるいは投資資金を増やす人が増加した。日経平均株価を始めとして、コロナ禍が始まった 2020 年 2〜3 月は殆どの企業の株価が急降下し、その後は回復傾向にある。その頃に株を買った賢い投資家は今笑いが止まらないだろう。確かな知識とそれに裏打ちされた冷静な投資判断をもってすれば、未曽有の危機をもビジネスチャンスに変えることができるのだと痛感した。

第 2 節　研究目的

デジタル技術の進歩やグローバル化による国際情勢の変化など、目まぐるしく変化する現代社会で私たちが適切な投資判断を行うことが本稿の最終目的である。多くの投資家は、企業が四半期に一度発表する IR(投資家情報)を主な判断材料とする。しかし目まぐるしく変化する現代社会では、四半期前の情報では古すぎる。本稿の目的は私たちが現代社会の日々の変化に沿った適切な投資判断を行うことであるので、国際金利や為替などの日次データを用いて分析を行う。実際に私たちが投資を行う際は、IR 等の長期データをベースにして、日次データから最終判断を行うことが望ましいであろう。

また個別株投資は変動が大きくハイリスクハイリターンである一方、投資信託や ETF に投資するインデックス投資は簡単に分散投資を行うことができ、着実に利益を得ることがでる。本稿では、私たちにとって敷居の低いインデックス投資を扱う。具体的には日経平均株価の変化を予測する。

コロナ禍によって現代社会は大きく変化した。よって現代社会での投資判断精度を上げるため、時系列的にはコロナ禍以降のデータを用いることとする。

第 3 節　本論文の構成

株価分析には 2 種類の方法がある。企業の財務状況や業績をもとにして企業の本質的な価値を分析するファンダメンタル分析と、チャートの変動パターンをもとに将来的な株価の変動を分析するテクニカル分析だ。機械学習を用いた分析では後者のテクニカル分析が主流である。しかし、経済理論の基づいた本質的なファンダメンタル分析も行いたい。

よって第 2 章でテクニカル分析を行い、第 3 章ではファンダメンタル分析を行う。最後に第 4 章でその 2 つをアンサンブルさせる。

第２章　テクニカル分析

第１節　分析手法

1.1　使用モデル

Prophet、ARIMA モデル（自己回帰和分移動平均モデル）、指数平滑法の 3 種類の手法を用いてテクニカル分析を行う。時系列データの 2 割をテストデータ、残りを訓練データとし、MAPE を算出してモデルの精度を比較する。

MAPE とは平均絶対パーセント誤差のことで、予測値と正解値の差を正解値で割ったものの総和をデータ数で割り、絶対値を付け、100%の確率値に直すために 100 をかけたものである。MAPE はデータのスケールに影響されないため誤差の大きさを客観的に判断しやすい。

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

1.2　使用データ

第 1 章 2 節にあるように、コロナウイルスの感染者が国内で初めて報告された 2020 年 1 月 16 日から本稿執筆時の 2021 年 11 月 5 日までの日経平均株価のデータを用いる。株価データを始めとしたファイナンスデータは基本的に非定常である。よって株価データを対数変換し定常性を持たせた。価格変動を図 1 に示す。青いデータが訓練データ、オレンジ色のデータがテストデータである。

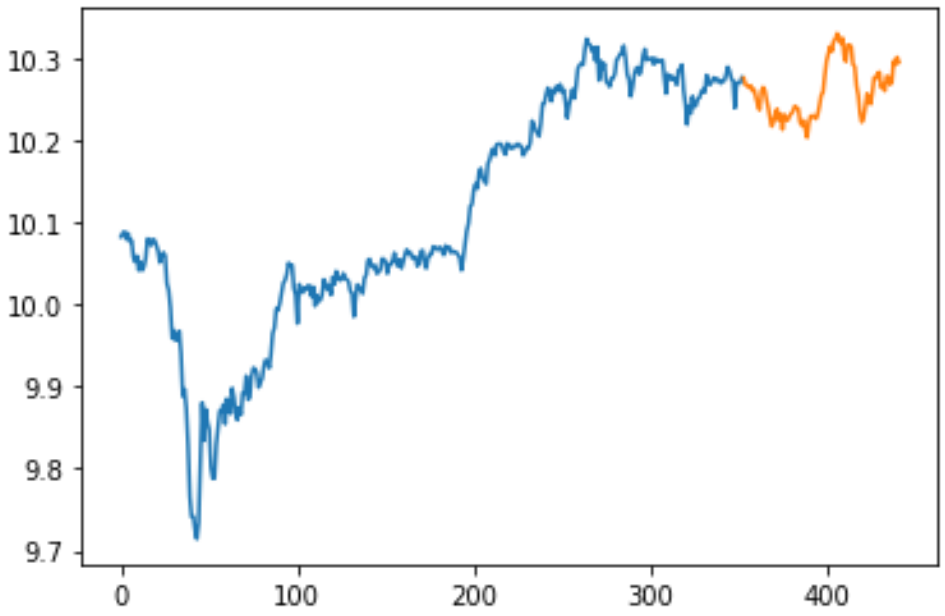

図 1：2020/1/16～2021/11/5 日経平均株価(対数変換後)

第 2 節　各機械学習手法の概念

2.1　Prophet

Prophet とは Facebook 社製の時系列予測ライブラリである。時系列データをトレンド、季節変化、休日効果の 3 つの要素に分けそれらをそれぞれ非線形モデルで表し、その和をとったモデルである。モデル式は以下の通りだ。

$$y(t) = g(t) + s(t) + h(t) + \varepsilon t.$$

- $g(t)$：トレンド関数
- $s(t)$：季節変化
- $h(t)$：休日効果
- εt：誤差項

（1）トレンド関数

以下のロジスティック曲線の式をもとにしてトレンドを表す。

$$y(t) = \frac{1}{1 + \exp(-t)}$$

この式に収容上限 C、成長率 k 、オフセット項 m を追加してグラフの概形を調整する。

$$g(t) = \frac{C}{1 + \exp(-k(t - m))}$$

C：収容上限　k：成長率　m：オフセット項

さらに C、k も時間によって変化するものと捉えて

$$C \to C(t)$$
$$k \to k + \mathbf{a}(t)^{\mathrm{T}}\boldsymbol{\delta}$$

と設定する。

δ は時点 t までに成長率が転換期を迎えた時の変更率を表しており、s 回の転換期の変更率の総和を成長率に加えることで成長率の変化を表す。

$$a_j(t) = \begin{cases} 1, & \text{if} \quad t \geqq s_j, \\ 0, & \text{otherwise.} \end{cases}$$

また、上記のように定義された s 個の要素を持つ 0 または 1 のベクトルを δ に乗ずることで時間によって変化する成長率を表現する。

しかしこのままでは転換期において非連続なグラフになるので、同様に転換期ごとにオフセット項を変化させる Γ を以下のように定義しオフセット項を調節する。

$$\gamma_j = \left(s_j - m - \sum_{l<j} \gamma_l \right) \left(1 - \frac{k + \sum_{l<j} \delta_l}{k + \sum_{l \leqq j} \delta_l} \right)$$

最終的にはトレンド関数は以下のようにあらわされる。

$$g(t) = \frac{C(t)}{1 + \exp(-(k + \mathbf{a}(t)^{\mathrm{T}}\boldsymbol{\delta})(t - (m + \mathbf{a}(t)^{\mathrm{T}}\boldsymbol{\gamma}))}$$

（2）季節変化

季節や曜日による影響など、あらゆる周期性を扱う。周期性を持ったあらゆる波形データは以下のような余弦と正弦の和で表現することができる。

$$f(t) = a_1 \sin(t) + a_2 \sin(2t) + a_3 \sin(3t) + \ldots + b_1 \cos(t) + b_2 \cos(2t) + b_3 \cos(3t) + \ldots$$

これをフーリエ級数展開といい、季節変化の式はこれをもとに以下のように定義されている。

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P} \right) + b_n \sin\left(\frac{2\pi n t}{P} \right) \right)$$

P には周期の値が入り、N は周期の値によって適切な値を設定する。周期に値が大きいほど N も大きな値を設定するのが好ましい。

（例）周期が 1 週間の場合、P＝７、N＝３くらい

周期が 1 年の場合、P＝365.25、N＝10 くらい

（3）休日効果

突発的なイベント等の影響をモデルに組み込むための要素。休日等のイベントは周期性がないが事前に予見できるので分析者自身がイベントカレンダーのリストを作り、モデルに組み込むことができる。

あるイベントを i とし、それに該当する日付をすべて含んだベクトルを作る。

（以下の例はクリスマス）

$$D_i = [\ldots, 1975/12/25, 1976/12/25, \ldots, 2020/12/25, \ldots]$$

次に各時点 t が Di に該当するか否かを表すインディケーターとして

$$Z(t) = [1(t \in D_1), \ldots, 1(t \in D_L)]$$

というベクトルを定義する。時点 t がイベント i に該当するかどうかを 1 または 0 で表現する。最後

に各イベントｉに対する係数パラメータをκiとし、そのベクトルをκで表す。

最終的に休日効果は以下のように表現される。

$$h(t) = Z(t)\boldsymbol{\kappa}$$

2.2　ARIMA モデル（自己回帰和分移動平均モデル）

一般的な時系列モデルである AR モデル（自己回帰モデル）、MA モデル（移動平均モデル）を組み合わせたモデルである ARMA モデルを改良したモデル。AR モデルは過去のデータに重みを付けて注目しているデータを表現する方法で、MA モデルは過去のデータと注目するデータに共通項を持たせて関係性を表現する方法である。これらはそれぞれ、定常性のある時系列データしか扱うことはできないが、データの差分をとってから ARMA モデルを推定する ARIMA モデルは定常性のない時系列データも扱うことができる。各データの係数のパラメータを求める。

$$y_t - y_{t-d} = c + \varepsilon_t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

また、係数を調整することで AR モデルや MA モデルを表現することも可能。

※定常性とは

(1)　任意の時点ｔにおける期待値がｔに依存しない。

(2)　ｋだけ離れた任意の２つの時点ｔ、ｔ－ｋにおけるデータの相関係数が時点ｔに依存しない。

という条件を満たしていること。

2.3　**指数平滑法**（Exponential Smoothing Forecast）

指数平滑法（以下 ES）とは過去のデータを用いる加重平均法の１つである。より新しいデータに重みを大きくし、より古いデータは重みを小さくする。似た手法に過重移動平均法があるが、加重移動平均法が過去に遡るにつれて重みを等間隔で減少させるのに対して、ES では、過去に遡るにしたがって重みを指数関数的に減少させる。それにより、加重移動平均法ではある時点で重みが 0 となり、それより過去のデータを考慮することができないが、ES は重みが 0 になることはないので、過去のすべてのデータを予測に反映させることができるという特徴がある。

以下の通り、時点ｔの予測値と観測値の差に重みパラメータｐをかけて時点ｔの予測値に足したものを時点 t+1 での予測値と定義する。

$$\hat{y}_{t+1} = \hat{y}_t + \rho(y_t - \hat{y}_t)$$

同様にして時点ｔ、ｔ－1、ｔ－2…についての式を立てて変形し、繰り返し代入していくと以下の ES のモデル式を定義できる。

$$\hat{y}_{t+1} = \rho y_t + \rho(1-\rho) y_{t-1} + \rho(1-\rho)^2 y_{t-2} + \cdots + \rho(1-\rho)^{k-1} y_{t-k+1} + (1-\rho)^k \hat{y}_{t-k}$$

第 3 節　実証分析

3.1　分析結果

上記の分析手法で分析した結果は以下の表 1：（各モデルの MAPE）の通りである。

Prophet	0.2470843535
ARIMA	0.319046643
ES	0.3066216738

表 1：各モデルの MAPE

1.2 で説明した通り、日経平均株価のデータを対数変換したものを分析した。Prophet を用いた分析における MAPE は 0.24、ARIMA を用いた分析における MAPE が 0.319、ES を用いた分析における MAPE は 0.30 と Prophet を用いた分析の精度が最も良いという結果となった。

3.2　考察

今回の分析では Prophet での分析の精度が最も高いという結果となった。今回扱った時系列データはコロナ禍のものであり、日経平均株価は異常な動きをしたと考えられる。Prophet にはトレンドを抽出するだけでなく、トレンドの変化点を自動で検出する機能も備えている。よって新型コロナウイルス感染拡大によるトレンドの変化点を検出し、予測に活かすことができたためにこのような結果になったのではないかと考えられる。

第 3 章　ファンダメンタル分析

第 1 節　分析手法

1.1　使用モデル

第 1 章 3 節で述べたように機械学習を用いたファンダメンタル分析は先行研究が少なく、明確に有効といえる手法が確立されていない。しかし数少ない先行研究を参考にして本論文では決定木、ランダムフォレスト、Extra Trees、 Gradient Boosting、 LightGBM の 5 種類の分析手法を利用する。

まず株価に影響を与える可能性のある日次データを様々用意し、ノイズに強いとされる LightGBM を用いて特徴量選択を行う。その選択した特徴量を用いて 7 種類の学習器を学習させ、再び特徴量の重要度を求め、株価を予測するために重要な要素を分析する。学習器同士の比較には、第 2 章と同じ様に MAPE を用いる。

重要度の計算には、それぞれの実装ライブラリに用意されているものを利用する。LightGBM 以外は全て Scikit-learn を使用した。Scikit-learn の決定木系モデルではジニ不純度を基に重要度が計算されている。また LightGBM では、特徴量の使用頻度から重要度が測定されている。

1.2　使用データ

被説明変数は第 2 章同様に日経平均株価を対数変換したものとする。用意した説明変数は以下の（表 2: 特徴量一覧）の通りである。これらのデータを標準化して分析を行った。

これらの要素が日経平均株価の変動にどれだけの影響を与えるか分析する。

コロナ新規感染者数	日本国債1年利回り	WTI原油先物（終値）
コールレート	日本国債2年利回り	WTI原油先物（始値）
ユーロ/米ドル	日本国債3年利回り	WTI原油先物（高値）
中国人民元/円	日本国債4年利回り	WTI原油先物（安値）
円/メキシコペソ	日本国債6年利回り	WTI原油先物（騰落率）
円/ユーロ	日本国債7年利回り	
円/台湾ドル	日本国債8年利回り	
韓国ウォン/円	日本国債9年利回り	
英ポンド/円	日本国債10年利回り	
スイスフラン/円	日本国債15年利回り	
	日本国債20年利回り	
	日本国債25年利回り	
	日本国債30年利回り	
	日本国債40年利回り	

表 2：特徴量一覧

第 2 節　各機械学習手法の概念

2.1 決定木分析

決定木分析とは決定木と呼ばれる樹形図を用いてデータを分析する手法である。決定木は分類木と回帰木に分けられ、A か B かというように区分の分類を扱うものを分類木、連続して変化しうる値の分析を扱うのが回帰木と呼ばれ、その 2 つを総称して決定木という。結果の解釈が容易で、数値データやカテゴリデータが混ざっていても問題なく利用できるという特徴がある。

2.2　ランダムフォレスト

1.1 で記述した決定木を複数組み合わせ、それぞれで分析を行い分類なら多数決、回帰なら平均をとることで、より精度の高い分析を行うアンサンブルモデル。過学習を起こしやすいという決定木分析の欠点を克服することができる。アンサンブルにはバギングとブースティングという 2 種類の方法があり、ランダムフォレストはそれぞれの決定木を並列に扱うバギングという方法でアンサンブルモデルを構築する。

2.3 Extra trees

ランダムフォレストに似た手法で、決定木を複数組み合わせて構築するアンサンブルモデルであるが、ランダムフォレストが特徴軸を分割する際にジニ係数やエントロピーなどの特徴量を用いて利得が最大となる法を選択するのに対し、Extra trees はランダムに用いる特徴量を選択する。

2.4 Gardient Boodting（勾配ブースティング）

勾配ブースティングとは

１，予測値と実測値の誤差を求める

２，求めた誤差を用いて決定木を構築

３，構築した決定木をそれ以前の予測結果とアンサンブルしてより精度の高い予測結果を出す

４，1～3 を繰り返す。

というようにして誤差を段階的に小さくしていく手法。

2.5 LightGBM

LightGBM はランダムフォレストと同様に決定木を複数組み合わせて構築するアンサンブルモデルであり、XGBoost を改良したものである。XGBoost は決定木を用いて得られた予測結果をその精度に応じて重みを付け、次の学習機に学習させるという方法でアンサンブルモデルを構築する。ランダムフォレストと異なり、複数の決定木を直列に扱うブースティングという方法でアンサンブルをするため、ランダムフォレストと比べ学習に時間がかかるが、その分精度が高くなりやすいという特徴がある。さらに、LightGBM はヒストグラムを用いて分岐すべきノードにあたりをつけ葉の数を減らすことで、精度を落とさずに学習に時間がかかるという XGBoost の欠点を解消した。

第 3 節　実証分析

3.1　特徴量抽出

LightGBM で特徴量の重要度を算出した結果が以下の図 2：（LightGBM による特徴量の重要度比較）である。

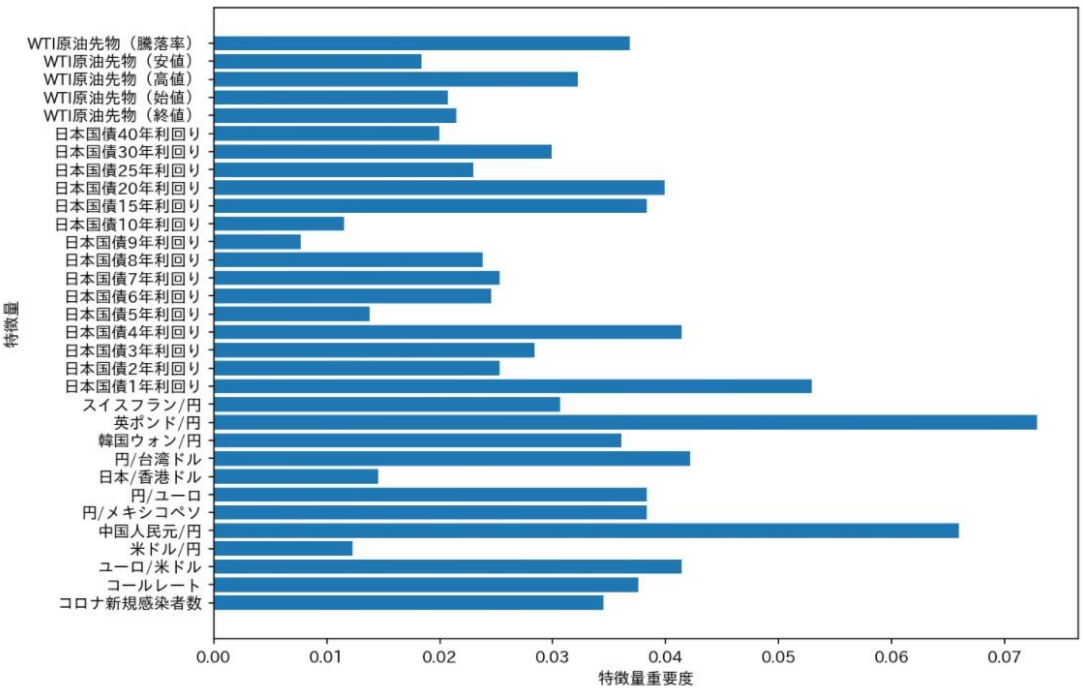

図 2：LightGBM による特徴量の重要度比較

金利は重要度がかなり高く、中期の国債は重要度が低いという結果になった。本稿で記載はしないが、他の学習器で試してみた際も同じような傾向になった。

この結果をもとに重要度の低い説明変数を取り除く。今回は日本国債 9 年利回り、日本国債 5 年利回り、日本/香港ドル、米ドル/円の 5 つを取り除くこととした。

3.2　分析結果

残りの特徴量を他の学習器にも学習させ、重要度を求めた。以下がその結果である。

3.2.1 決定木

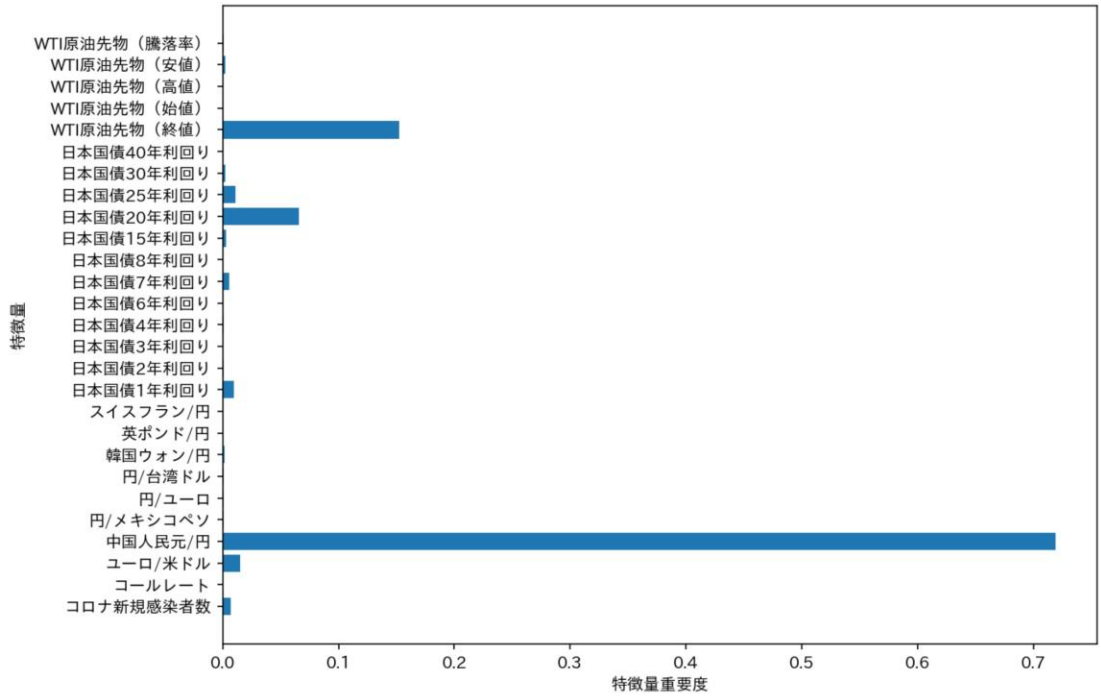

図 3：決定木分析

決定木分析の結果、中国人民元と円の為替レート重要度が非常に高く、次点で WTI 原油先物価格の終値の重要度が高いという結果となった。

3.2.2 ランダムフォレスト

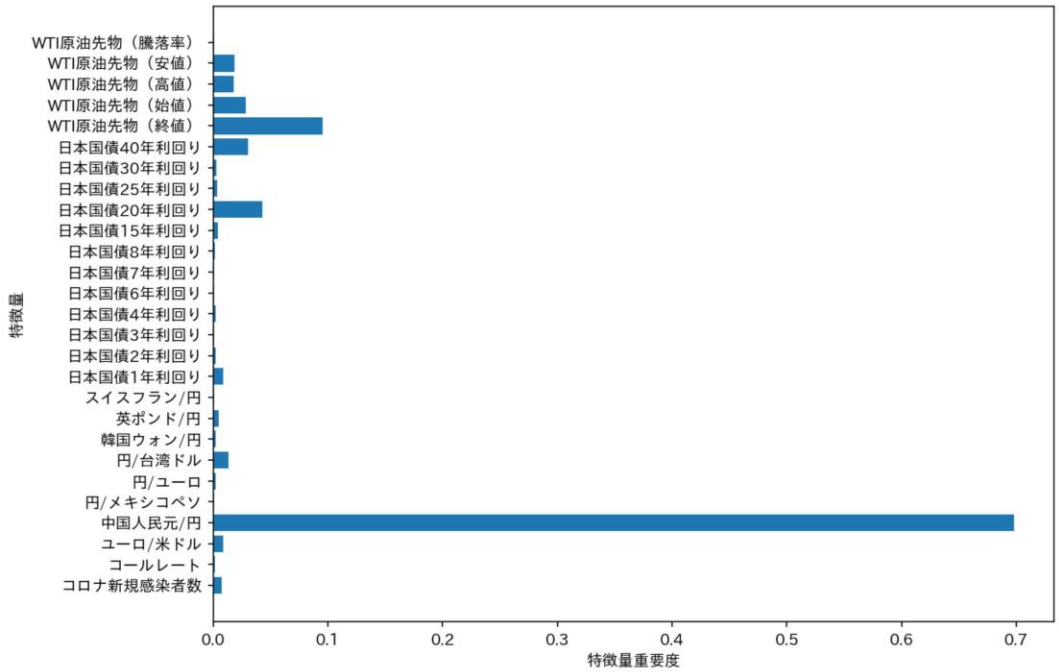

図 4：ランダムフォレスト

ランダムフォレストの結果、中国人民元と円の為替レートが最も重要度が高く、次点で WTI 原油先物価格の終値の重要度が高いという結果となった。

3.2.3 Extra Trees

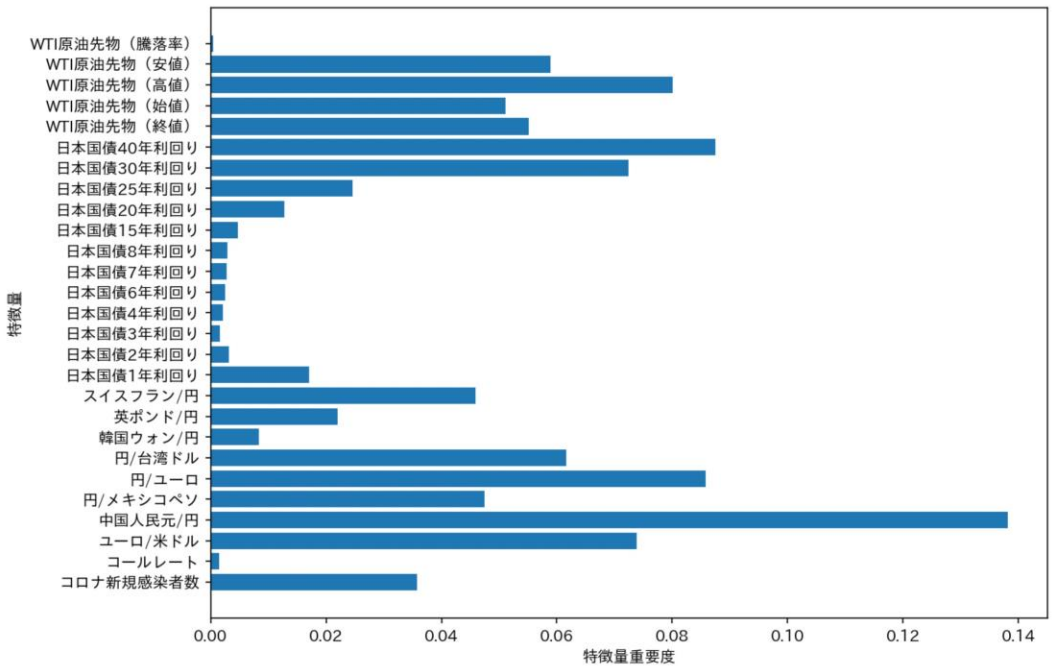

図 5：Extra trees

Extra trees の結果、中国人民元と円の為替レートの重要度が最も高く、次点で WTI 原油先物価格の高値、日本国債 40 年利回り、日本国債 30 年利回り、円とユーロの為替レート、ユーロと米ドルの為替レートの重要度が高いという結果となった。

3.2.4 Gradient Boosting

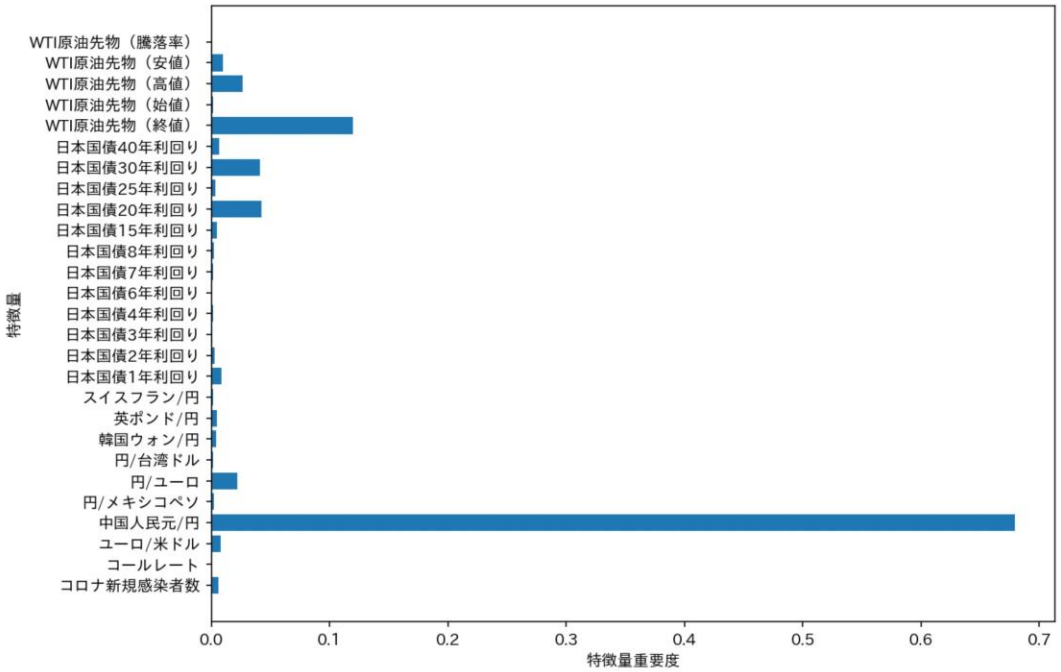

図 6：Gradient boosting

Gradient boosting の結果、中国人民元と円の為替レートの重要度が高いという結果となった。

3.2.5 LightGBM

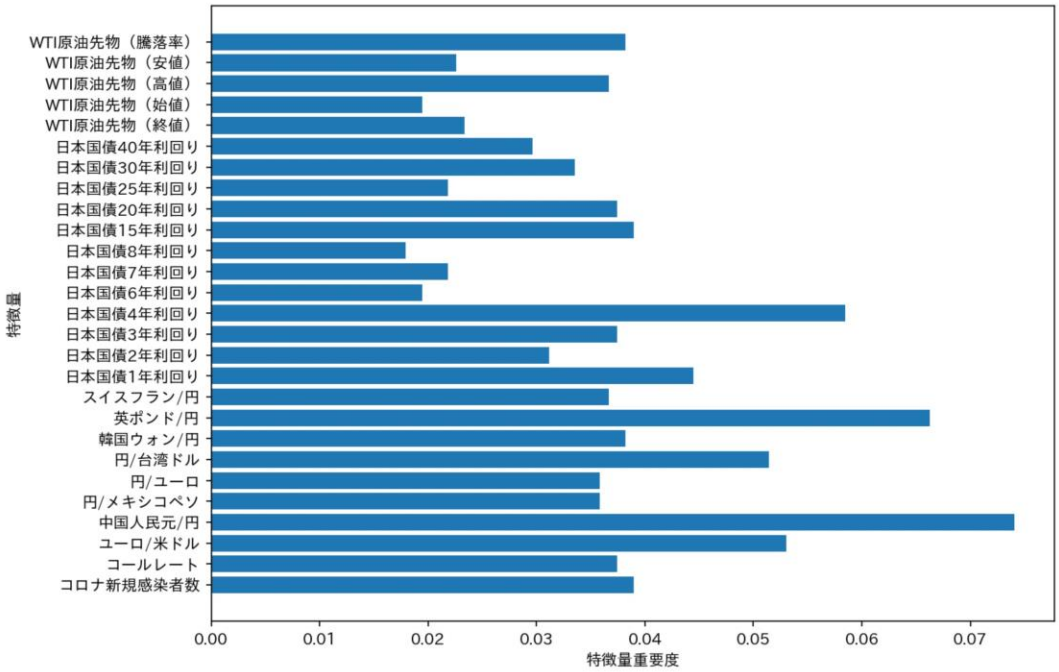

図 7：LightGBM

LightGBM の結果、中国人民元と円の為替レート、英ポンドと円の為替レート、次点で日本国債４年利回り、ユーロと米ドルの為替レート、円と台湾ドルの為替レートの重要度が高いという結果になった。

3.2.6 予測精度比較

上記の分析手法の MAPE は以下の表 3:(各モデルの MAPE)の通りである。Extra Trees、Gradient Boosting、LightGBM、ランダムフォレスト、決定木の順で精度が良かった。

決定木	0.7199509756
ランダムフォレスト	0.5231402108
Extra Trees	0.26435172
Gradient Boosting	0.3183257154
LightGBM	0.3968711174

表 3：各モデルの MAPE

3.3 考察

本稿のファンダメンタル分析ではすべてのモデルで中国人民元と円の重要度が高いという結果となった。近年の著しい成長で世界経済の中心となった中国のマーケットがやはり日本の経済へも大きな影響力を持っているということがわかる。日経平均株価の構成比率上位 5 つの企業の内 3 つは中国での売上比率が 13～19%を占めているため、中国との為替レートが日経平均株価に強い影響を与えるのは明らかであるだろう。

また WTI 原油先物価格の重要度も比較的高いことが分かる。コロナウイルスの影響で原油価格は下落したが、その後の世界経済の持ち直しによって現在は高騰している。原油の大半を輸入に頼る日本の経済にとっては、原油価格の変動も大きな影響力を持つことが確認できた。意外にも新型コロナウイルスの感染者数は日経平均株価への影響力があまり大きくないという結果となった。

また各モデルの精度を比較したところ、Extra Trees, Gradient Boosting, LightGBM の精度が高かった。特に Gradient Boosting と LightGBM は Kaggle でもよく使われる勾配ブースティング回帰モデルであるので、この結果はおおむね想定通りだ。

第 4 章　アンサンブルモデル

今回のアンサンブルモデルは、簡易的なものを実施する。第 2 章、第 3 章の計 8 つの学習器の予測値を説明変数として LightGBM で分析を行うという方法だ。被説明変数には第 2 章、第 3 章と同様に対数変換した日経平均株価を用いる。

結果は MAPE が 0.751429124 となり、これまでのモデルで一番性能が悪くなってしまった。

第 5 章　考察と今後の課題

第 1 節　考察

本稿の目的は、目まぐるしく変化する現代社会で私たちが適切な投資判断を行うことであった。日々の変化に対応するため、一般に公表されていて参照しやすい日次データのみを用いて分析を行った。

第 2 章では複数のモデルを用いてテクニカル分析を行い、株価の変動パターンから日経平均株価の予測を試みた。Facebook 社が公開している Prophet というモデルを用いると比較的優位な結果が得られた。

第 3 章ではファンダメンタル分析を行い、日経平均株価の変動に影響を与える要素を抽出した。今回の分析では中国人民元と円の為替レートが最も影響力が大きく、次点で WTI の原油先物価格の影響も受けているという結果が得られた。また 5～15 年辺りの国債金利は重要度が低いということも分かった。Extra Trees、GradientBoosting, LightGBM の順で精度が高かった。

最後にテクニカル、ファンダメンタル両方のモデルをアンサンブルしたモデルを構築し日経平均株価の予測を行ったが、精度が一番落ちてしまった。

今回の研究によって、私たちは日経平均株価のインデックスファンドに投資を行う際は、中国人民元と円の為替レートや、WTI 原油先物価格に特に着目してタイミングを選択すべきだということが分かった。勿論日次データのみを用いて株式投資することは危険である。消費者物価指数や政府の政策、四季報など広範囲にアンテナを貼って判断すべきである。その上で最後に投資タイミングを判断する際に、今回の結果を使用してほしい。

第 2 節　今後の課題

本稿では、以下のような課題がある。

1. データの量、質
2. ハイパーパラメータの調整
3. アンサンブルモデルの機能不全

今回は原油先物価格、日本国際利回り、主要為替レート、コロナ感染者数、コールレートを説明変数としている。国内で初めてコロナウイルス感染者が出た 2020 年 1 月 16 日以降のデータしか用いていないため、無暗に説明変数を増やすと分析精度が落ちる恐れがあると考えこのような対応を行った。コロナ禍のデータはまだ 2 年弱分ほどしかないため、今後さらにデータを追加することで説明変数も増やし分析精度を上げたい。また、説明変数は基本的に加工をせず元データを使用していた。コロナ感染者数の変化率を説明変数とした方が、より適切なデータになったのではないかと考えている。

また、分析モデルは基本的にハイパーパラメータを調整することが出来なかった。よって分析モデルの性能を活かせていない可能性は否定できない。

最後に、第 4 章においてアンサンブルモデルが逆に精度を落としてしまったという課題がある。先行研究ではアンサンブルモデルを構築する前に、株価そのものの予測だけでなく株価の上下の分類予測も行っていた。また今回のアンサンブルモデル手法は簡易的に 7 つの学習器の予測値を説明変数とするも

のだったが、アンサンブルモデルとは、バギング、ブースティング等を用いたより高度な分析方法である。これらの点を踏まえて、今後はこのアンサンブルモデルの改良にも注力したい。

参考文献

著者名(最終更新日)「ウェブサイト名」URL、最終閲覧日

1、DateStrategy 岡氏（2019.12.20）「Prophet のモデル式を 1 から理解する」、https://devblog.thebase.in/entry/2019/12/20/110000_1、2021.11.5

2、kanbe（2020.4.14）「ARIMA モデル（自己回帰和分移動平均モデル）について分かりやすく解説」、https://ai-trend.jp/basic-study/time-series-analysis/arima_model/、2021.11.8

3、不明（2020.9.15)「需要予測の基礎知識（単純移動平均、加重移動平均、指数平滑法）」、https://mirukognosis.com/?p=1020、2021.11.8

4、Cacoo 編集部（2020.4.14）「決定木分析（ディシジョンツリー）とは？概要や活用方法、ランダムフォレストも解説」、https://cacoo.com/ja/blog/what-is-decision-tree/、2021.11.9

5，土井健（2019.12.2）「GBDT の仕組みと手順を図と具体例で直感的に理解する」、https://www.acceluniverse.com/blog/developers/2019/12/gbdt.html、2021.11.9

6、codexa チーム(2019.2.13)「LightGBM 徹底入門 – LightGBM の使い方や仕組み、XGBoost との違いについて」、https://www.codexa.net/lightgbm-beginner/、2021.11.9