ディープラーニングを用いた消費者物価指数の予測

中島　舜介(19X4107)　和田　明夢（19X4141）　　指導教員 五島 洋行

1. はじめに

昨今，内外の感染症の動向やその影響，今後のウクライナ情勢の展開，資源価格や海外の経済・物価動向など，不確実性が極めて高い世の中になっている．そんな中で，全国の世帯が購入する財やサービスなどの物価の動きを把握するための指標であり，国がまとめた消費者物価指数[1]は，経済についての分析や，各種経済施策の指標とされ，金融市場などにも大きな影響を与えることがある．そのため，これを予測し，見通しを立てることが重要である．類似研究では，Reuters 社が発行しているビジネスニュースの日本証券市場に関する文章をベクトル表現し，実際の株価を基にネガポジ判定をしたものを訓練データとして機械学習を行い，投資戦略を求める研究[2]や，アナリストレポートを用いて分散表現を行い，CNN モデルで景況感判定の研究[3]が発表されているが，それを消費者物価指数に応用した研究は見受けられない．

そこで本研究の目的は，日経平均株価，金利，為替（円ドル相場），金価格，原油価格の値動きから予測するモデルと過去のニュースデータを用いた消費者物価指数の予測するモデルを作成し，それを組み合わせることでより正確な予測をし，経済の見通しを立てられるようにすることである．

2. 提案手法

本研究では，予備実験として日経平均株価，金利，為替（円ドル相場），金価格，原油価格の値動きを RandomForest モデルで予測を行い，ニュースデータの本文を用いてディープラーニングで消費者物価指数の予測を行う．最後にそれぞれ予測した値を特徴量として学習し，2 つのモデルを組み合わせて，アンサンブル学習のスタッキングの手法で予測を行う．予測する際は，6 ヶ月分と 12 ヶ月分をそれぞれ予測する．

2.1 使用データ

①～⑤の対象期間及びデータ数は 1992 年 1 月 1 日から 2022 年 2 月 28 日の 7179 個を用いる．

①株価データ：日経平均株価終値．
②金利データ：10 年物国債における利回り．
③為替データ：円ドル相場における終値．
④金価格：東京商品取引所における 1g あたりの金価格．
⑤原油価格：1 バレルあたりの原油価格．
⑥ニュースデータ

NHK 国際, NHK ビジネス, Reuters ビジネス，BBCHome, Reuters トップ，Returns ワールド以上の 6 社のニュースデータを使用する．2008 年 3 月 2 日から 2022 年 8 月 31 日までの 412，671 個のデータを用いる．前処理として，同じ本文は 1 件だけ残し，それ以外は学習データに含まない．

2.2 金利・為替等を用いた予測モデル概要

RandomForest モデルを用いて消費者物価指数の予測を行う．日経平均株価のデータを時系列順に月ごとにまとめ，80%を学習データ，20%をテストデータとして使用する．

消費者物価指数を予測する際，予測する初めの月の前月を入力値として与える．

データを時系列順に月ごとにまとめ，その月の日数分データがあるため 2 次元のデータとなっている．ただし，RandomForest モデルで学習および予測を行うために 1 次元のデータに変換している．また，5 つのデータの打ちどれか 1 つでも欠損値がある場合，そのデータは削除している．よって，月によってデータ数が異なるため，すべての月次データのうち最も日数（データ数）が少ない日数分だけデータを取得する．

表 1:　独自モデルのハイパーパラメーター

Learning rate	0.001
Batch size	128
Epochs	100
Optimizer	Adam
Loss function	MSLE
Loss weights	0.01
Hidden size	400

表 2:　LightGBM のハイパーパラメーター

Learning rate	0.1
Metric	MAE
Num leaves	100
Num iterations	300

表 3:　6 ヶ月分の予測における回帰統計

	重相関 R	重決定 R2	補正 R2	標準誤差	観測数
金利・為替等	0.96513	0.93148	0.92635	0.18108	87
ニュースデータ	0.86683	0.75140	0.75139	0.00746	91,630
複合モデル	0.99652	0.99304	0.99304	0.01322	22,930

表 4:　12 ヶ月分の予測における回帰統計

	重相関 R	重決定 R2	補正 R2	標準誤差	観測数
金利・為替等	0.96449	0.93025	0.91894	0.18535	87
ニュースデータ	0.97325	0.94721	0.94720	0.00662	91,630
複合モデル	0.99628	0.99257	0.99256	0.01290	22,930

2.3 ニュースデータを用いた予測モデル概要

ディープラーニングを用いて消費者物価指数の予測を行う．予測する初めの月の前月のニュースデータを入力値として与える．本章で扱う独自のモデルのハイパーパラメーターは表 1 に記す．

2.4 2 つのモデルを合わせた予測モデル概要

金利・為替等のデータからランダムフォレストで予測した値と，ニュースデータからディープラーニングで予測した値のデータを特徴量とし，アンサンブル学習のスタッキングの手法で予測を行う．メタモデルには LightGBM を採用し，LightGBM におけるハイパーパラメーターは表 2 に記す．

3. 予測結果

表 3，表 4 では各予測値における回帰統計をまとめたものである．

4. 考察

決定係数はサンプル数（観測数）が多くなればなるほど値が大きくなる特徴があるため，サンプル数が異なる今回のモデルでは補正された決定係数（補正 R2）の値を比較する．複合モデルでは他 2 つのモデルよりかなり精度の高いモデルとなっていることが分かる．ニュースデータのモデルでは金利･為替等のモデルと比べ 6 ヶ月分の予測の決定係数が小さくなっている．これよりニュースデータで学習した 6 ヶ月分の予測では，安定的な予測ができないことが考えられる．

5. おわりに

本研究では，二つの手法をもとにアンサンブル学習することで，様々な影響を考慮したより良い消費者物価指数の予測モデルを提案した．一般的に株価の予測は先行研究として散見されたが，消費者物価指数を予測する研究はあまりなかった．

ニュースデータの本文から予測するだけでなく，経済の代表的な指標となる金利や為替等のデータを用いた予測との比較やアンサンブル学習も行った．結果として，アンサンブル学習での予測結果がかなりの高精度で予測することができた．また，本研究ではデータ数がそこまで多くなかったものの，精度の高いモデルを構築することができたことから，データ数や特徴量を増やすことで，これよりもさらに精度の高い予測ができる．

本研究で提案した手法は消費者物価指数の予測だけでなく，様々な経済の指標の予測も可能だと考える．ニュースデータから学習することで精度が高くなったことから，今後 NLP 分野の発展が経済の分析に役立つであろう．

6. 参考文献

[1] 総務省統計局ホームページ / 消費者物価指数 https://www.stat.go.jp/data/cpi/，アクセス日：2022 年 12 月 6 日

[2] 五島圭一，高橋大志，寺野隆雄：“ニュースのテキスト情報から株価を予測する”，全 国知能学会全国大会，2G4-OS-25a-4，vol.29，pp.1-3，June，　2015.

[3] 高山将丈，小澤誠一，廣瀬勇秀，飯塚正昭：畳み込みニューラルネットワークを用いたアナリスト往訪記録における景況感判定，The 33th Annual Conference of the Japanese Society for Artificial Intelligence，2019

令和 4 年度卒業研究論文

ディープラーニングを用いた消費者物価指数の予測

法政大学　理工学部　経営システム工学科

経営数理工学研究室

19X4107 中島　舜介

19X4141 和田　明夢

指導教員　五島　洋行　教授

学科名	経営システム工学科	学籍番号	19X4107 19X4141
申請者氏名		中島　舜介 和田　明夢	
指導教員氏名		五島洋行	

論文要旨

論文題目	ディープラーニングを用いた消費者物価指数の予測

昨今のウクライナ情勢や新型コロナウイルスによるパンデミックによる経済への影響は計り知れない．円安，エネルギー・資源価格の高騰における家計への影響はここ数年でかなり問題視されてきた．そんな中全国の消費者物価指数は軒並み上昇している．給料が上がらないことが問題とされているこの日本で，給料が上がらずに物価が上昇することは国民にとってはかなりの痛手である．

時を同じくして，NLP 分野における研究では，機械学習の手法として，ディープラーニングを用いた学習を行うことが一般的である．Google が 2018 年に発表した自然言語モデル BERT は，今では世界各国で使用され，BERT に続き GPT3 などのより高精度なモデルのほとんどがディープラーニングを用いて学習を行なっている．類似研究の株価維持予測ではニュースデータを基に予測するモデルの研究が発表されているが，それを消費者物価指数に応用した研究は見受けられない．

そこで本研究では，金利や為替等のデータとニュースデータの両方を用いた消費者物価指数の予測を行う．

比較のための予備実験として，日経平均株価，金利，為替（円ドル相場），金価格，原油価格これら 5 つのデータを用いて RandomForest モデルで消費者物価指数の予測を行う．次にニュースデータの本文を用いてディープラーニングで消費者物価指数の予測を行う．ニュースデータはテキストデータのため，形態素解析と分散表現で前処理をして学習モデルに適合させる．最後にそれぞれ予測した値を特徴量として学習し，2 つのモデルを組み合わせて，アンサンブル学習のスタッキングの手法で予測を行う．

実験の結果，２つのモデルを組み合わせた提案手法の精度が最も高い結果となった．

目次

第1章 はじめに

1.1 研究背景と目的

昨今，内外の感染症の動向やその影響，今後のウクライナ情勢の展開，資源価格や海外の経済・物価動向など，不確実性が極めて高い世の中になっている．そんな中で消費者物価指数とは，全国の世帯が購入する財やサービスなどの物価の動きを把握するための指標であり，国がまとめた消費者物価指数は，日本経済についての分析や，各種経済施策の指標とされ，金融市場などにも大きな影響を与えることがある.そのため，消費者物価指数を予測し，見通しを立てることが重要である．

時を同じくして，NLP 分野における研究では，機械学習の手法として，ディープラーニングを用いた学習を行うことが一般的である．Google が 2018 年に発表した自然言語モデル BERT は，今では世界各国で使用され，BERT に続き GPT3 などのより高精度なモデルのほとんどがディープラーニングを用いて学習を行なっている．類似研究の株価維持予測ではニュースデータを基に予測するモデルの研究が発表されているが，それを消費者物価指数に応用した研究は見受けられない．

そこで本研究の目的は，過去のニュースデータを用いた消費者物価指数の予測を行うことで，より正確な見通しを立てられるようにすることである．

1.2 本論文の構成

本論文は 9 章で構成されている．
第 2 章では先行研究を述べる．
第 3 章では本研究で用いる関連知識を述べる．
第 4 章では本研究の実験の概要及び，使用するデータと実行環境について述べる．
第 5 章では金利や為替等のデータを利用した予測モデルの概要と結果を述べる
第 6 章ではニュースデータを利用した予測モデルの概要と結果を述べる．
第 7 章では実際のデータと予測値の結果を述べる．
第 8 章では結果を比較し，考察を述べる．
第 9 章では本研究のまとめを述べる．

第2章 先行研究

一般的に経済の指標となるもので，日経平均株価の予測が行われることが多い．株価の予測では主にテクニカル分析とファンダメンタルズ分析を用いた先行研究がある．

テクニカル分析とは，過去の値動きをチャートで表し，そこからトレンドやパターンを把握することで今後の値動きや動向を予測するものである．過去に似たパターンがあれば，今後も同じようなパターンになる可能性が高いと予測する．テクニカル分析は数値や株価であればチャートなどのグラフの概形から予測できるため，経済に関する知識がなくても問題がないのが大きなメリットの 1 つとなる．その一方で，パンデミックなどの過去に例のないトレンドやパターンが起きたときに予測精度がかなり下がってしまうデメリットもある．Deep Learning を株価予測に応用した研究[3]では，中間層が 1 層の簡単なフィードフォワードニューラルネットワークで 1 分足ごとの変化率を用いて学習している．予測する値は 1 分前より上がるかまたは下がるかを予測する二値分類で，67%の正解率となっている．

ファンダメンタルズ分析とは，経済成長率，物価上昇率，失業率，財政収支などの経済活動等の状況をもとに分析することである．株式では，PER（株価収益率），PBR（株価純資産倍率），ROE（株主資本利益率）などのファンダメンタル指標をもとに，企業の財務状況や業績状況のデータを分析する．ただしファンダメンタルズ分析では，さまざまな情報や専門的知識を要するため，多大な労力がかかってしまうデメリットがある．

過去には，Reuters 社が発行しているビジネスニュースの日本証券市場に関する文章をベクトル表現し，実際の株価を基にネガポジ判定をしたものを訓練データとして機械学習を行い，投資戦略を求める研究が発表されている[4]．結果は超過収益を獲得しており，機械学習によるニュース記事の評価を用いて将来の株価を予測する可能性を見出した．

他にもアナリストレポートを用いて分散表現を行い，CNN モデルで景況感判定の研究が発表されている[5]．目的は違うものの，テキストデータから分散表現を行い，ディープラーニングモデルでの予測を行うという点で，本研究ではこの研究をもとに提案している手法もある．

第3章 関連知識

本章では，本研究で用いる学習モデルの関連知識について述べる．

3.1　ニューラルネットワーク

ディープラーニングの基となる理論ニューラルネットワークは 1943 年に提唱され，パーセプトロンは 1957 年に開発された．ただし，計算量が膨大な結果，当時のマシンパワーでは実現不可能だった．

ニューラルネットワークは脳の神経回路の一部を模した数理モデルである．また，パーセプトロンを複数組み合わせたものの総称である.入力層, 中間層, 出力層これら 3 つの層から構成され，機械学習を機能させるための一手法であり，現在では，主にディープラーニングで使用されている．

そして，ニューラルネットワークを用いて学習を行う際に，より効率よく精度の高い学習を行うために，活性化関数や損失関数などの手法が取り入れられた．

3.2　活性化関数

ニューラルネットワークにおける活性化関数とは，あるニューロンへと出力する際に，あらゆる入力値を別の値に変換して出力する関数である．

活性化関数は多数存在するが, ここではニューラルネットワークでよく用いられるシグモイド関数, ReLU 関数, GeLU 関数, Softmax 関数の 4 つの活性化関数について説明する．

① シグモイド関数

式(1.1)で表されるシグモイド関数は，ニューラルネットワークにおいて古くから利用される活性化関数である.シグモイドは，0 から 1 の間の値を取るため，ディープラーニングにおける複雑な計算コストを削減できる．

$$h(x) = \frac{1}{1 + \exp(-x)} \tag{1.1}$$

② ReLU 関数

式(1.2)で表される ReLU 関数は，シグモイド関数が古くから利用されているのに対し，近年利用されている関数である.ReLU 関数は入力が 0 を超えていれば，その入力をそのまま出力し，0 以下ならば 0 を出力する．

$$h(x) = \begin{cases} x\ (x > 0) \\ 0\ (x \leq 0) \end{cases} \tag{1.2}$$

③ GeLU 関数

式(1.3)で表される GeLU 関数は，BERT や GPT-3 などの近年の自然言語モデルで使われている．ReLU と形が似ていることがわかる．$\Phi(x)$は標準正規分布の分布関数となる．

$$h(x) = x\Phi(x) \tag{1.3}$$

④ Softmax 関数

式(1.4)で表される Softmax 関数は，出力層で使用され，出力層のニューロンの数をnとし，k番目の出力y_kを求める計算式を表している.このとき，分子は入力信号の指数関数，分母は全ての入力信号の指数関数の和となる.したがって，全ての入力信号の指数関数の和で割るので，変換後の出力層の和が 1 になるのが特徴である．

ただし，指数関数の値を計算することは，値が大きくなる可能性が十分にあり，Softmax 関数の恩恵を受けづらい.そこで，全ての入力信号を入力信号の最大値で引くことで，指数関数の値をなるべく小さくするオーバーフロー対策が必須となる．

$$y_k = \frac{\exp(a_k)}{\sum_{i=1}^{n} \exp(a_i)} \tag{1.4}$$

3.3　損失関数

損失関数はニューラルネットワークの性能の指標になり，あるニューラルネットワークが教師データに対してどれだけ適合，一致していないかを表す．つまり，損失関数の値をどれだけ小さくすることができるかが，ニューラルネットワークおよびディープラーニングにおける性能の評価方法となる．また，損失関数にマイナスを掛けた値は，どれだけ教師データに対して適合，一致しているかを表す．

ここで，なぜニューラルネットワークでは認識精度ではなく，損失関数の値を精度の評価にするのかという疑問が生まれる．その理由としては，ニューラルネットワークの学習では，重みとバイアスをパラメーターとし，最適なパラメーターを模索する際に，損失関数の値ができるだけ小さくなるようなパラメータを探す．このとき，できるだけ小さな損失関数を探すために，「勾配」と呼ばれるパラメーターの微分を計算することで，その値を手がかりにパラメーターの値を徐々に更新する．パラメーターの値を少しだけ変化させたときに，損失関数がどれだけ変化するかがパラメーターの更新の指標となる．したがって，損失関数を精度の評価とすることで，連続的な値の変化をもとにパラメータを更新できる．

①交差エントロピー誤差

ここでは，よく用いられる損失関数の 1 つとして，式（2.1）で表される交差エントロピー誤差について説明する．

$$E = -\sum_{k} t_k \log y_k \qquad (2.1)$$

このとき，log は底がeの自然対数とし，ニューラルネットワークの出力をy_k，正解ラベルをt_kとする．また，t_kは正解ラベルとなるインデックスだけが 1 で，その他は 0 である one-hot 表現であるとする．たとえば，「3」が正解ラベルであるとして，それに対するニューラルネットワークの出力が0.6の場合，交差エントロピー誤差は$-\log 0.6 = 0.51$となる．また，「3」のニューラルネットワークの出力が0.1の場合，$-\log 0.1 = 2.30$となる．したがって，ある正解ラベルに対する出力が大きければ損失関数の値は小さくなり，仮に出力が 1 となれば損失関数の値は 0 となる．

②平均二乗対数誤差(MSLE:Mean Squared Log Error)

平均二乗対数誤差とは，各データに対して「予測値の対数と正解値の対数との差」の二乗値を計算し，その総和をデータ数で割った値を出力する以下の関数である.対数の値を取ることで，より小さく連続的な値を取るため回帰問題においてより精度の高いモデルをつくることができる．第 6 章のディープラーニングにおける損失関数ではこの MSLE を使用する．

$$E = \frac{1}{n}\sum_{i=1}^{n}\left(\log(1+\hat{y}_i) - \log(1+y_i)\right)^2 \qquad (2)$$

3.4　誤差伝播法

ニューラルネットワークの学習では，パラメーターの値が決まると，順伝播により入力層，中間層，出力層の順番に計算し，予測した値が実際の値とどれだけ一致しているかを損失関数によって求める．このとき，パラメータを決める際に，闇雲にパラメーターの値を設定し計算することは，計算に時間がかかってしまう．そこで，効率よく学習するために誤差逆伝播法という手法がよく用いられる．誤差逆伝播法とは，損失関数の値から勾配を求め，その勾配を使い逆伝播により出力層，中間層，入力層へとパラメーターを更新する手法である．これにより，どの程度パラメーター更新すればよいのかが分かるため，効率の良い学習が行える．

3.5　正則化

機械学習において過学習が問題になることが多くある．過学習とは，訓練データに適応し

すぎてしまい，訓練データに含まれないデータに対応できない状態のことである．機械学習では，訓練データに含まれないデータを正しく識別する汎用的なモデルが望まれる．そこで重要になるのが過学習を防ぐ正則化である．

ここではディープラーニングに限らず，機械学習でよく用いられる正則化の手法の一つとして，Batch Normalization（以下 Batch Norm）の説明を行う．Batch Norm の利点としては以下が挙げられる．

- 学習を速く進行させることができる
- 重みなどのパラメーターの初期値に依存しない

ディープラーニングでは，学習の際に多くのレイヤーを用意することになり，精度が上がる反面，計算量が膨大になり学習に時間がかかってしまう．また，重みやバイアスなどのパラメータの初期値は，基本的にランダムに決めるため，初期値に依存した学習を行ってしまう．これらの問題を解決できる Batch Norm は，2015 年に提案されて以降，機械学習のコンペティションで優れた結果を達成している．

式 (2.2) で表される Batch Norm は，名前が示すとおり，ミニバッチ単位で正則化を行い，データの分布が平均が 0 で分散が 1 になるように正則化を行う．この処理を活性化関数の前もしくは後に挿入することで，データの分布の偏りを減らすことができる．

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \qquad (2.2)$$
$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$$

ここでは，ミニバッチとして$B = \{x_1, x_2, \cdots, x_m\}$という$m$個の入力データの集合における，平均$\mu_B$，分散$\sigma_B^2$を求め，これらの平均と分散がそれぞれ 0 と 1 になるように正規化する．このとき式（2.2）のεは，0 で除算されることを防止するために小さな値を設定する．

3.6 Adam

Adam[6]は，いまやどのモデルにも広く使われているデファクトスタンダードな最適アルゴリズムだ．最適化の手法は世の中にたくさん提案がされており、ニューラルネットワークのパラメータの最適化に関しては入力変数と出力変数の間に起こりうるパターンを見つけ出し、予測精度が最大になるような重みパラメータをどう見つけるかという問題になる.Adam は小さな極小値にトラップされることなく最小値を目指すためのアプローチであり、更にパラメータ毎に最近の勾配の値に比べて、大きかったのか、小さかったのかを考慮したパラメータ更新を行うことで、パラメータ毎に適切な更新ができる最適化手法になっており、AdaGrad や RMSProp の長所を活かし、更にそれらの欠点を発見して克服した手法と言える.これらの理由から、入力される説明変数の分布の偏りや、様々な複雑な形を取

り得るニューラルネットワークの構造に対して頑健に学習ができるのではないかと考えられている.更新式(2.3)は以下になる.

$$
\begin{gathered}
g_t = \nabla_\theta f_t(\theta_{t-1}) \\
m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\
v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\
\widehat{m_t} = \frac{m_t}{1-\beta_1^t},\ \widehat{v_t} = \frac{v_t}{1-\beta_2^t},\ \theta_t = \theta_{t-1} - \alpha \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}+\epsilon}} \\
\left(\theta\text{: 重み},\ f_t\text{: 誤差関数},\ \alpha = 0.001,\ \beta_1 = 0.9,\ \beta_2 = 0.999,\ \varepsilon = 10^{-8},\ m_0,\ v_0 = 0\right)
\end{gathered}
\tag{2.3}
$$

3.7 Doc2Vec

Doc2Vec は，任意の長さの文書をベクトル化する技術で，分野テキストに対して分散表現を獲得することができる．(1)PV-DBOW，(2)PV-DM の 2 種類のモデルから構成されている．

① PV-DBOW

PV-DBOW は Word2Vec の skip-gram に対応するようなアルゴリズムである.PV-DBOW は単語のベクトルを学習の際に用いる必要がないため，PV-DM よりも早く学習が可能．ただ PV-DBOW は学習の際に語順を無視するかたちになっているため，PV-DM の方が精度が良いとされている.以下の手順で学習を行なっている．

1. 同一文章から任意の個数の単語をサンプルしてくる．
2. サンプルした単語を予測するように文章ベクトルと中間層→出力層の重みを最適化.

② PV-DM

PV-DM は Word2Vec の CBOW に対応するようなアルゴリズムである.文章の id と単語を複数個渡し，次に出てくる単語を予測するというタスクを解きながら文章の分散表現を獲得する．学習は以下の手順で行なっている．

1. 文章ベクトルと文書の中から一部をサンプリングした単語のベクトルを用意
2. 1 で用意したベクトルを中間層で結合(平均又は連結，gensim では選択可能)
3. サンプリングした単語に続く次の単語を予測
4. 文章ベクトル及び中間層→出力層の重みを更新

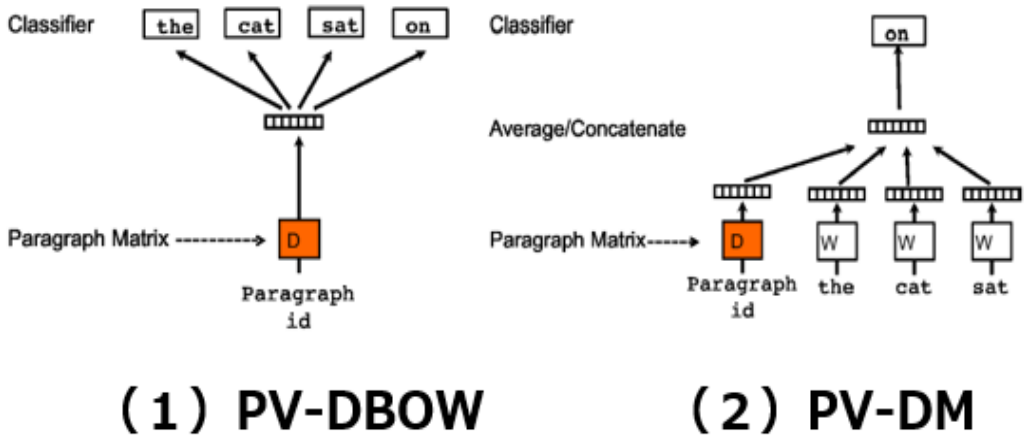

図 1　Doc2Vec のモデル概要[7]

3.8　RandomForest

RandomForest とは機械学習のアルゴリズムの一つで，決定木による複数の弱学習器を統合させて汎化性能を向上させるアンサンブル学習アルゴリズムである．つまり，ランダムサンプリングされたトレーニングデータによって学習した多数の決定木を使用することにより，学習方法は単純だが，一般的な決定木より性能のよい識別・予測ができる．

3.9　勾配ブースティング

ブースティングとは，与えられたデータから決定木分析を行なった後に，予測が正しくできなかったデータに重みをつけて，再度，決定木学習を行い，これを繰り返すことで精度を高める方法である.されに，データに重み付けするのではなく，予測値と実績値の誤差を計算し，誤差を決定木学習する方法が勾配ブースティングである.ブースティングと同様に，誤差に対する学習を繰り返すことで精度を高める．

3.10　LightGBM(Light Gradient Boosting Machine)

LightGBM とは機械学習におけるアルゴリズムの一つで，勾配ブースティングを用いた決定木による手法で，高速なことが特徴である．勾配ブースティングの場合は，誤差を最小化するように分割の要素，基準を見つけるため，データ量に応じて計算量が増える．１つ１つの決定木の精度をなるべく落とさずに，高速に構築できるようにしたことが最大の特徴である．

3.11　アンサンブル学習

アンサンブル学習とは，単独では精度の高くない弱学習機を多数用いることで精度をあげる手法である.アンサンブル学習にはバイアスとバリアンスという概念が関係している．

バイアスは，推定値と実測値との誤差の平均，バリアンスは予測値の分散である．バイアスが低いほどうまく予測できており，複雑なモデルを組んで様々なデータへフィティングさせようとすると予測値がばらつきバリアンスは高くなる.バイアスとバリアンスはトレードオフの関係にある.つまりバイアスが低くてバリアンスが高い状態は過学習に陥っている可能性が高く，アンサンブル学習は，そんなバイアスとバリアンスをうまく調整する手法群であり，うまくバイアスとバリアンスのバランスを取ることが必要である．

アンサンブル学習は，バギング，ブースティング，スタッキングの 3 つの手法がある．バギングは複数の分類器の結果の平均や多数決で予測を行う手法で，それぞれの分類器を並列に学習することが可能である．ブースティングは弱学習器を直列に構築し，直前の弱学習器の結果に重みを加味し新たな弱学習器で学習させることで予測を行う手法．スタッキ ングは複数の異なる学習モデルの結果を特徴量の入力としてメタモデルで学習し，予測を行う手法である．

第4章 実験概要

本章では本研究の実験概要及び，使用するデータ，評価関数，モデル概要，実行環境について述べる．

4.1 提案手法

予備実験として，第 5 章では日経平均株価，金利，為替（円ドル相場），金価格，原油価格これら 5 つのデータを用いて RandomForest モデルで消費者物価指数の予測を行う．第 6 章ではニュースデータの本文を用いてディープラーニングで消費者物価指数の予測を行う．ニュースデータはテキストデータのため，形態素解析と分散表現で前処理をして学習モデルに適合させる．

また，第 7 章では第 5 章，第 6 章でそれぞれ予測した値を特徴量として学習し，2 つのモデルを組み合わせて，アンサンブル学習のスタッキングの手法で予測を行う．複合モデルの予測の流れは図 2 のとおりである．

消費者物価指数を予測する際，6 ヶ月分と 12 ヶ月分をそれぞれ予測する．

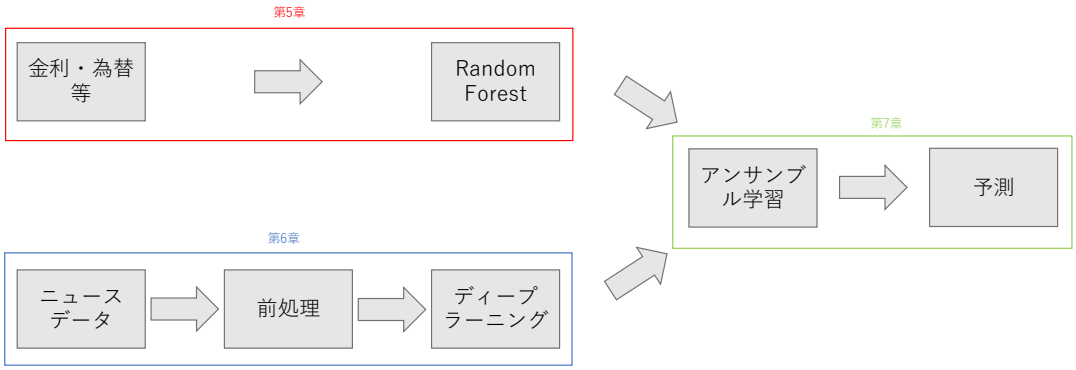

図 2　複合モデルの予測の流れ

4.2 使用データ

① 株価データ

1992 年 1 月 1 日から 2022 年 2 月 28 日までの日経平均株価終値 5,520 個のデータを用いる．日経平均株価は日本経済新聞社が東京証券取引所一部の上場している企業の中から選んだ 225 銘柄で構成されており，終値は取引時間終了時の取引価格である．

② 金利データ

10 年物国債における利回りのデータを用いる．対象期間およびデータ数は，株価データと同様である．

③ 為替データ

円ドル相場における終値のデータを用いる．対象期間及びデータ数は，株価データと同様である．

④ 金価格

東京商品取引所における 1g あたりの金価格のデータを用いる．対象期間及びデータ数は，株価データと同様である．

⑤ 原油価格

1 バレルあたりの原油価格のデータを用いる．対象期間及びデータ数は，株価データと同様である．

⑥ ニュースデータ

NHK 国際,NHK ビジネス，Reuters ビジネス，BBCHome，Reuters トップ，Returns ワールド以上の 6 社のニュースデータを使用する．2008 年 3 月 2 日から 2022 年 8 月 31 日までの 412,671 個のデータを用いる．また前処理として，同じ本文があった場合は 1 件だけ残しそれ以外は学習データに含まない．

第 5 章，第 6 章，第 7 章いずれにおいても 80%を学習データ，20%をテストデータとして使用する．

4.3　評価関数

本研究では予測精度を比較する際に以下の評価関数を用いる．数値が小さいほど誤差が小さく，精度が良いといえる．一般的に機械学習で回帰モデルを作成する場合，以下で紹介される評価関数が用いられることが多く，機械学習に限らず回帰統計においても用いられることが多い．

N はデータ数，y_iは実測値，$\hat{y}_i$は予測値を表す．

① 平均絶対誤差(MAE:Mean Absolute Error)

MAE とは,「予測値と実測値の差の絶対値」を計算し，その総和を平均したものである．

MAE は以下の式で求められる．

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

② 平均二乗誤差 (MSE:Mean Squared Error)

MSE とは，「予測値と実測値の差の 2 乗」を計算し，その総和を平均したものである．2 乗の計算を行うため，値が大きくなりやすいデメリットがある一方，小数点以下であれば値が小さくなるというメリットがある．

MSE は以下の式で求められる．

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

③ 二乗平均平方根誤差 (RMSE:Mean Squared Error)

RMSE とは，MSE の計算結果の平方根を計算したものである．MSE では値が大きくなりやすいデメリットがあると前述したが，平方根をとれば値の大小に限らず，値を小さくすることができるため，評価関数においては MSE よりも RMSE が利用されることが多い．

RMSE は以下の式で求められる．

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{3}$$

複数の評価関数を用いることでモデルの精度を様々な観点から評価することができる．したがって，本研究ではこれら 3 つの評価関数を用いて精度の評価を行う．

4.4　消費者物価指数

消費者物価指数[1]は，全国の世帯が購入する家計に係る財及びサービスの価格等を総合した物価の変動を時系列的に測定するものである．品目は様々なものがあり代表的なものでいうと食料，飲料，家賃，ガス代などの光熱費が挙げられる．世帯の消費支出上一定の割合を占める重要なものから構成されており，すべてで 582 品目ある．消費者物価指数は基準年の物価を 100 として比較計算とした値となっており，基本的に99.7，100.3のように小数点第 1 位までで計算される．

本研究を行うにあたり，総務省統計局の消費者物価指数のデータを用いる．今回扱う消費者物価指数の基準年は 2020 年とする．

4.5　実験環境

本研究における実験環境は表 1 のとおりである．

表 1　実験環境

CPU	Apple M1

OS	macOS Ventura 13.0.1
Memory	16GB
使用言語	Python 3.8.13

第5章 金利・為替等のデータを用いた予測

本章では，日経平均株価，金利，為替（円ドル相場），金価格，原油価格 5 つのデータを用いた予測モデルの概要と予測結果について述べる．

5.1 予測モデルの概要

RandomForest モデルを用いて消費者物価指数の予測を行う．金利・為替等のデータを時系列順に月ごとにまとめた．

消費者物価指数を予測する際，予測する初めの月の前月を入力値として与える．例えば，2020 年 6 月から 2020 年 11 月までの 6 ヶ月分の消費者物価指数を予測するときは，2020 年 5 月のデータを用いる．また，2020 年 6 月から 2021 年 5 月までの 12 ヶ月分の予測を行うときも同様である．

データを時系列順に月ごとにまとめ，その月の日数分データがあるため 2 次元のデータとなっている．ただし，RandomForest モデルで学習および予測を行うために 1 次元のデータに変換している（図 3）．また，5 つのデータの打ちどれか 1 つでも欠損値がある場合，そのデータは削除している．よって，月によってデータ数が異なるため，すべての月次データのうち最も日数（データ数）が少ない日数分だけデータを取得する．

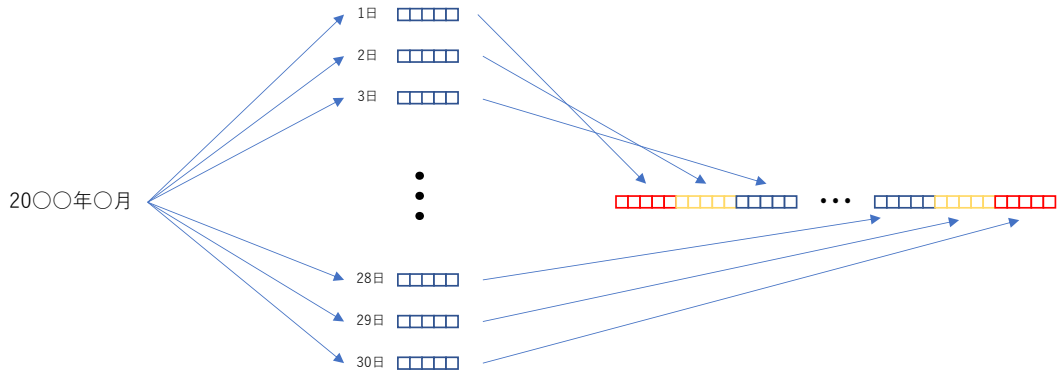

図 3　2 次元データ構造および 1 次元データへの変換の流れ

5.2 予測結果

テストデータに対する予測値の評価関数の値を表 2 に記す．

MAE の値の平均値から 1 つあたりの予測値が 6 ヶ月分の予測では 0.05，12 ヶ月分の予測では 0.02 ずれていることが分かる．消費者物価指数は小数点第 1 位までしか計算されないことから，この手法では正確に値を予測できているとは言えない．特に 6 ヶ月分の予測では 1 つあたりの予測値が 0.05 ずれているので，小数点第 1 位の値が大いに変わる可能性

がある.

表 2　金利・為替等のデータを用いた予測の評価値

評価関数	6 ヶ月	12 ヶ月
MAE	0.34707	0.35467
MAE(Average)	0.05785	0.02956
MSE	0.21710	0.22576
RMSE	0.46527	0.47391

※小数点第 6 位を四捨五入

第6章 ニュースデータを用いた予測

本章では，ニュースデータを用いた予測モデルの概要と予測結果について述べる．

6.1 予測モデルの概要

ディープラーニングを用いて消費者物価指数の予測を行う．5 章と同様，消費者物価指数を予測する際，予測する初めの月の前月のニュースデータを入力値として与える．

ニュースデータは NHK 国際,NHK ビジネス，Reuters ビジネス，BBCHome，Reuters トップ，Returns ワールドこれら 6 社のニュースデータを使用し日本語のみを扱う．

本章で扱うモデルは独自のモデルを採用しており，このモデルのハイパーパラメーターは表 3 に記し，モデルの概要は図 4 で示す．

表 3　ディープラーニングのハイパーパラメーター

Learning rate	0.001
Batch size	128
Epochs	100
Optimizer	Adam
Loss function	MSLE
Loss weights	0.01
Hidden size	400

図 4　ディープラーニングのモデル概要図

6.2　前処理

ニュースデータは同じ記事があるため，同じ記事があった場合は 1 件だけ残しそれ以外は学習データに含まない．以下の 3 つに従って，ニュースデータの本文の前処理を行う．

1. 半角カタカナを全角カタカナに，全角英数を半角英数に，機種依存文字を複数文字に，全角の記号を半角に変換
2. スペース，記号，html タグ，ハイパーリンクは除く
3. ニュースデータの本文に名詞，動詞，形容詞のみを抽出する形態素解析を行い，Doc2Vec でベクトルへと変換する

Doc2Vec で分散表現を行う際のパラメーターを表 4 で記す．

表 4　Doc2Vec のパラメーター

Vector size	400
Min count	1
Window	5
Epochs	10

6.3　予測結果

テストデータに対する予測値の評価関数の値を表 5 に記す．

MAE の値の平均値から 1 つあたりの予測値が．金利・為替等のデータを用いた予測に比べ，それぞれ 6 ヶ月分の予測では半分，12 ヶ月分の予測では 3 分の 1 に減っている．ニュースデータを用いることで精度は高くなったが，それでも小数点第 1 位に影響を及ぼす可能性があるので正確に値を予測できるとは言えない．

MSE と RMSE に関しては，共に，6 ヶ月分の予測では値が大きく増加している．一方で１２ヶ月分の予測では値が減少している．この結果から MSE と RMSE の値をもとに今回のモデルを評価するのは適切とは言えない．

表 5　ニュースデータを用いた予測の評価値

評価関数	6 ヶ月	12 ヶ月
MAE	0.15002	0.12224
MAE(Average)	0.02500	0.01019
MSE	0.99369	0.20298

RMSE	0.99671	0.43104

※小数点第 6 位を四捨五入

図 5，図 6 のとおり学習が進むにつれ損失関数の値が減少している．テストデータに対する損失関数の値は 6 ヶ月分の予測より 12 ヶ月分の予測のほうが良い結果となっている．12 ヶ月分の予測では 6 ヶ月分の予測と比べ予測する値が増えるため，その分誤差が生じやすい．それでも 12 ヶ月分の予測のほうが損失関数の値が小さいだけでなく，評価値の値も優位となっている．金利・為替等のデータを用いた予測でも MAE の値は 6 ヶ月分の予測のほうが小さかったものの，MAE の平均値では 12 ヶ月分の方が良い結果となっていた．

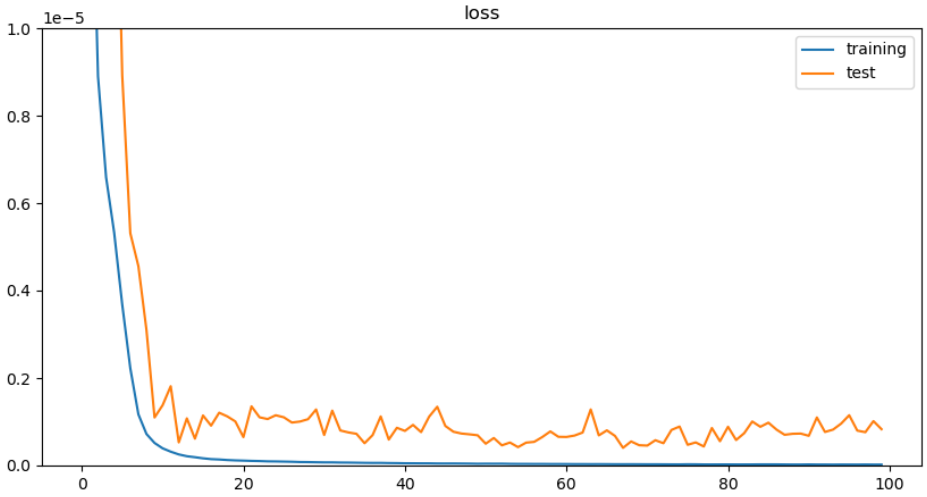

図 5　6 ヶ月分の予測における損失関数の値の推移

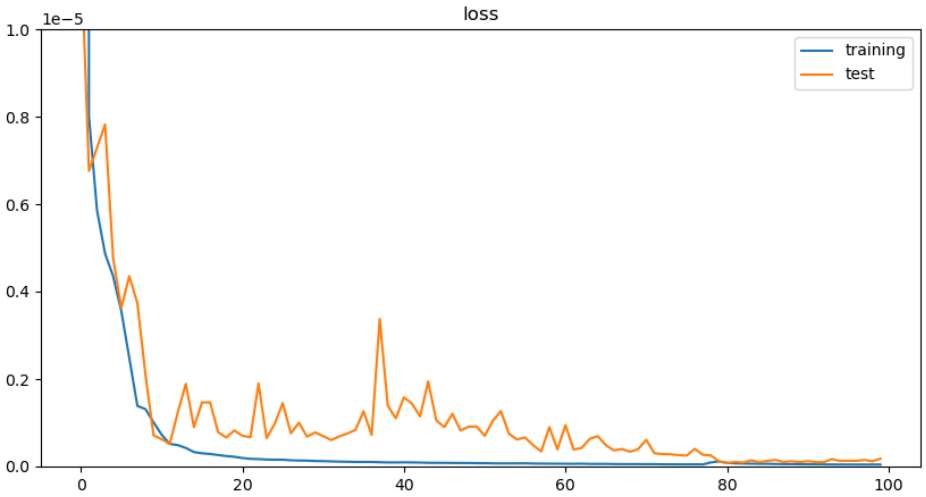

図 6　12 ヶ月分の予測における損失関数の値の推移

ディープラーニングで学習を進めていく過程で，テストデータに対する評価値の最小値とそれに対する epoch 数は表 6 のとおりである．MAE の値は実際に値を予測したときとさ

ほど差はない．12 ヶ月分の予測では MAE の値が同じである．それに対して MSE と RMSE に関しては，実際に値を予測したときと大きく異なる．（MSE，RMSE の値が増加していることを書く）

図 7，図 8 のとおり損失関数の値の減少に比例して MAE の値も減少している．機械学習において評価関数の値と損失関数の値が密接な関係になっていることが読み取れる．損失関数の値は限りなく小さい値を扱うことで，評価関数の値を連続的に変化させることができる．その結果精度の高いモデルをつくることができる．

表 6　テストデータにおける評価値の最小値

	6 ヶ月		12 ヶ月	
	最小値	最小値を記録した epoch 数	最小値	最小値を記録した epoch 数
MAE	0.14272	78	0.12224	100
MSE	0.46157	68	0.07887	81
RMSE	0.67939	68	0.28084	81

※小数点第 6 位を四捨五入

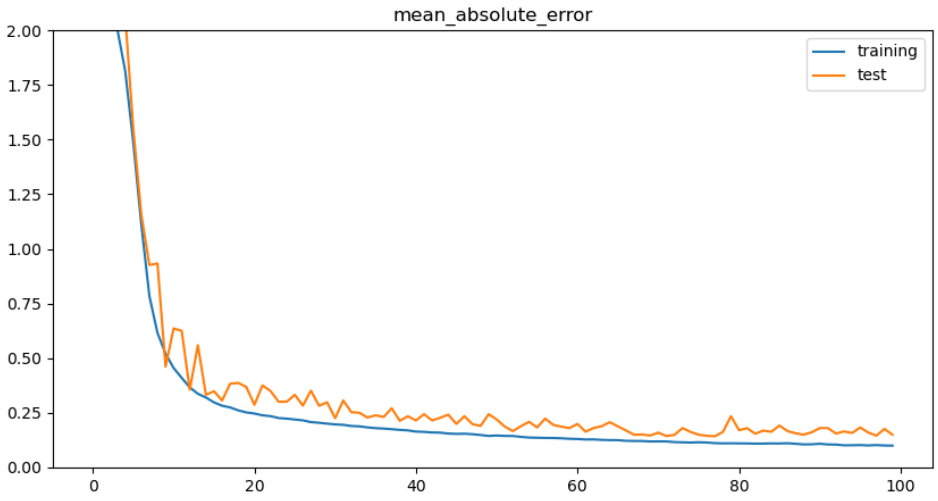

図 7　6 ヶ月分の予測における MAE の値の推移

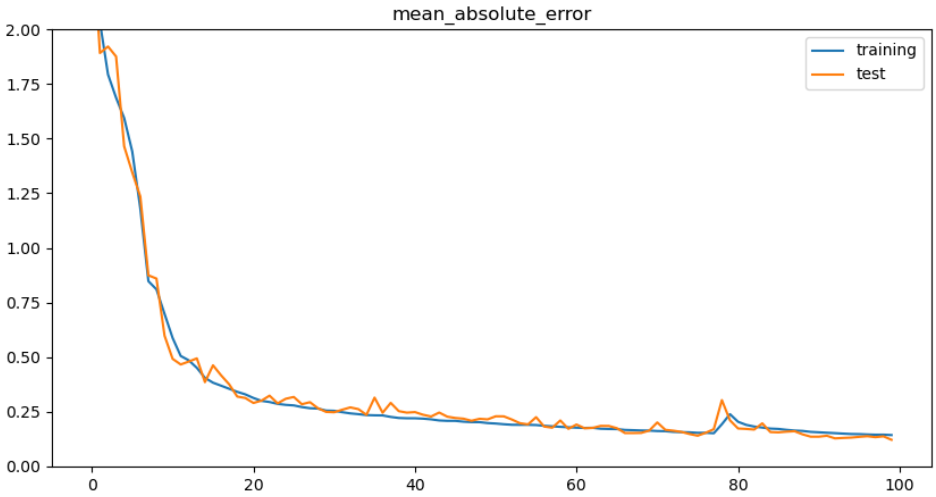

図 8　12 ヶ月分の予測における MAE の値の推移

第7章 2つのモデル組み合わせた予測

本章では第5章，第6章で扱った2つのモデルを組み合わせた予測モデルの概要と予測結果について述べる．

7.1 予測モデルの概要

日経平均株価，金利，為替（円ドル相場），金価格，原油価格5つのデータからランダムフォレストで予測した値と，ニュースデータからディープラーニングで予測した値のデータを特徴量とし，アンサンブル学習のスタッキングの手法で予測を行う．メタモデルにはLightGBMを採用し，LightGBMにおけるハイパーパラメーターは表7に記すとともに，複合モデルの予測の流れを図9で示す．

表7 LightGBMのハイパーパラメーター

Learning rate	0.1
Metric	MAE
Num leaves	100
Num iterations	300

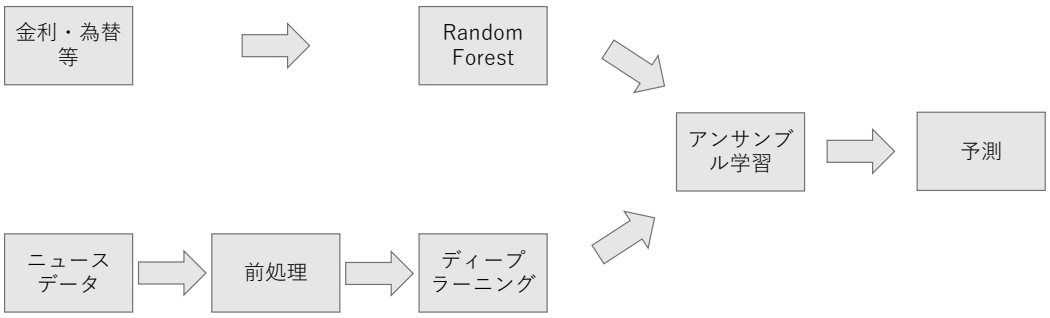

図9 複合モデルの予測の流れ

7.2 予測結果

テストデータに対する予測値の評価関数の値を表8に記す．

MAEの値の平均値から1つあたりの予測値が小数点第1位に影響を及ばさないことが読み取れる．したがって，これまでのモデルと比較すると非常に精度が上がっている．一般的に出力した値を用いて新たに値を出力すると精度が下がると言われているが，今回の場合は出力した値をスタッキングでアンサンブル学習することで精度の高いモデルを実現する

ことができた．また今回のモデルでは，MSE と RMSE でモデルの精度を評価することは適切ではないと述べたが，MSE と RMSE の値もこれまでのモデルと比べ値が大きく減少していることが分かる．

表 8　複合モデルによる予測の評価値

評価関数	6 ヶ月	12 ヶ月
MAE	0.04551	0.06344
MAE(Average)	0.00758	0.00529
MSE	0.02796	0.02823
RMSE	0.16716	0.16794

※小数点第 6 位を四捨五入

第8章 実験結果の考察

本章では第 5 章，第 6 章，第 7 章それぞれの実験結果をもとに考察を述べる．

表 9 から各評価関数の値を比較すると，金利・為替等のデータから予測した値とニュースデータから予測した値を用いた複合モデルでの予測が最も精度が高い結果となった．

表 9　各予測における評価関数の値

	MAE		MSE		RMSE	
	6 ヶ月	12 ヶ月	6 ヶ月	12 ヶ月	6 ヶ月	12 ヶ月
金利・為替等	0.34707	0.34707	0.21710	0.22576	0.46527	0.47391
ニュースデータ	0.15002	0.12224	0.99369	0.20298	0.99671	0.43104
複合モデル	0.04551	0.06344	0.02796	0.02823	0.16716	0.16794

※小数点第 6 位を四捨五入

表 10, 表 11 では各予測値における回帰統計をまとめたものである．これまでは評価関数でモデルの精度を評価していたが，決定係数や標準誤差などの回帰統計の結果を考慮した上でモデルの精度を評価する．

決定係数はサンプル数（観測数）が多くなればなるほど値が大きくなる特徴がある．よってサンプル数が異なる今回のモデルでは補正された決定係数（補正 R2）の値を比較する．今までの考察も含め，複合モデルでは他 2 つのモデルよりかなり精度の高いモデルとなっていることが分かる．ニュースデータのモデルでは MAE の値が金利・為替等のモデルと比べ大きく減少したものの，6 ヶ月分の予測では決定係数の値が小さくなっている．これよりニュースデータで学習した 6 ヶ月分の予測では，安定的な予測ができないことが考えられる．

表 10　6 ヶ月分の予測における回帰統計

	重相関 R	重決定 R2	補正 R2	標準誤差	観測数
金利・為替等	0.96513	0.93148	0.92635	0.18108	87
ニュースデータ	0.86683	0.75140	0.75139	0.00746	91,630
複合モデル	0.99652	0.99304	0.99304	0.01322	22,930

表 11　12 ヶ月分の予測における回帰統計

	重相関 R	重決定 R2	補正 R2	標準誤差	観測数
金利・為替等	0.96449	0.93025	0.91894	0.18535	87
ニュースデータ	0.97325	0.94721	0.94720	0.00662	91,630
複合モデル	0.99628	0.99257	0.99256	0.01290	22,930

※小数点第 6 位を四捨五入

第9章 おわりに

昨今のウクライナ情勢や新型コロナウイルスによるパンデミックによる経済への影響は計り知れない．円安，エネルギー・資源価格の高騰における家計への影響はここ数年でかなり問題視されてきた．

本研究では，金利や為替等のデータを用いてランダムフォレストモデルで学習させる一般的な機械学習，ニュースデータを用いたディープラーニングでの学習，これら 2 つの手法をもとにアンサンブル学習することで，様々な影響を考慮したよりよい消費者物価指数の予測モデルを提案した．一般的に株価の予測は先行研究として散見されたが，消費者物価指数を予測する研究はあまりなかった．

先行研究では，株価の予測にトレンドやパターンを把握することで予測するテクニカル分析，経済成長率，物価上昇率，失業率，財政収支などの経済活動等の状況をもとに予測するファンダメンタルズ分析について言及した．ただ，これらの分析は過去に例のない動きを予測できないことや専門的な知識を要するという問題点がある．これらの問題を解決するために文章を入力データとし機械学習を用いて予測する手法が提案された．フレームワークからなる従来の分析とは大きく変わって，専門的な知識を要することなく，過去に例のないパターンにも対応した予測が可能になった．これらの先行研究に付随して，ニュースデータの本文から予測するだけでなく，経済の代表的な指標となる金利や為替等のデータを用いた予測との比較やアンサンブル学習も行った．結果として，アンサンブル学習での予測結果がかなりの高精度で予測することができた．また，本研究ではデータ数がそこまで多くなかったものの精度の高いモデルを構築することができたことから，データ数や特徴量を増やすことで，これよりもさらに精度の高い予測ができる．

本研究で提案した手法は消費者物価指数の予測だけでなく，様々な経済の指標の予測も可能だと考える．ニュースデータから学習することで精度が高くなったことから，今後 NLP 分野の発展が経済の分析に役立つであろう．

参考文献

[1] 総務省統計局ホームページ / 消費者物価指数
https://www.stat.go.jp/data/cpi/, アクセス日：2022 年 12 月 6 日

[2] 斎藤康毅：「ゼロから作る Deep Learning」, オーム社（2016）

[3] 池戸鉄一, 林田実：“ディープラーニングの株価予測への応用”, 北九州市立大学「商経論集」vol.52, pp.13-26, March, 2017.

[4] 五島圭一, 高橋大志, 寺野隆雄：“ニュースのテキスト情報から株価を予測する”, 全 国知能学会全国大会, 2G4-OS-25a-4, vol.29, pp.1-3, June, 2015.

[5] 高山将丈, 小澤誠一, 廣瀬勇秀, 飯塚正昭, 神戸大学大学院 工学研究科, 神戸大学 数理・データサイエンスセンター, 三井住友 DS アセットマネジメント株式会社：畳み込みニューラルネットワークを用いたアナリスト往訪記録における景況感判定, The 33th Annual Conference of the Japanese Society for Artificial Intelligence, 2019.

[6] D.P. KIngma, and L. J. Ba：”Adam：A method for Stochastic Optimization” ,International Conference on Learning Representations, 13pages, 2015.

[7] Quoc Le, Tomas Mikolov: “Distributed Representations of Sentences and Documents”, Google Inc,1600 Amphitheatre Parkway, Mountain View, CA 94043, pp3-4, 2014.