

会社四季報を用いた理論株価の推定

~Kaggle を参考に~

Calculating Company's Stock Prices from The Japan Company Handbook

吉川 満¹

Mitsuru Kikkawa¹

¹ 大和大学政治経済学部

¹ The Faculty of Political Science and Economics, Yamato University

Abstract: This paper calculates the theoretical stock prices in the fundamentals theory with “The Japan Company Handbook”, which covers all listed companies. Three machine learning methods are used with referring the developed the analytical methods in Kaggle.

1. はじめに

株式の取引を行う際、全上場企業の企業概要や業績データ等が、コンパクトにまとめられている「会社四季報」¹など情報ハンドブックは、投資家の間で、広く利用されている。会社四季報が発売された日は、「四季報相場」²と呼ばれるいつもとは異なる相場が形成されることがあるなど、影響力があることが知られている。ただし、これには、膨大な情報があり、個人投資家はどの項目が重要で、どのように株価に関係しているのか等まで理解するのは困難である。

近年、様々なデータが容易に利用できるように

なってきたこともあり、機械学習³の様々な手法が開発されている。開発された手法を用いて、様々な Web サイトにて、株価や為替等の予測を試みているものも多く公開されている。また、Kaggle (カグル) と呼ばれるデータサイエンスに関するコミュニティが注目されている。特に Kaggle 中で、定期的に企業等がコンペ形式⁴で課題を提示し、参加者が予測精度を競い合っている。

本稿では、会社四季報にある構造化された情報を Kaggle 参加者の手法を参考にし、機械学習を活用することで、ファンダメンタルズ理論における理論株価⁵を予測する。

¹ 東洋経済新報社が年 4 回発行 (新春号 12 月中旬, 春号 3 月中旬, 夏号 6 月中旬, 秋号 9 月中旬) している。また、この会社四季報は、購入せずとも、証券会社が提供している取引ツールからも見ることができるものもある。

² 最近では既に会社四季報 ONLINE [1]において 2 週間程前から先出しの情報があり、事前に株価が値上がりしている銘柄も多いと指摘されている。

³ 機械学習とは、人間が持つ学習にあたる仕組みを機械 (特にコンピュータ) で実現する技術・手法の総称と言われている。一般的に、構造化されたデータを分析する際には、機械学習、非構造化されたデータを分析する際には、深層学習 (ディープラーニング) が適していると言われている。

⁴ コンペには、例えばメルカリによる販売者が投稿した情報を基に「適正な販売価格」を予測するもの [2] やクルート社が運営する「ホットペッパー」と「Air レジ (飲食店向け POS レジアプリ)」のデータを利用して、レストランのお客さんの数を予測するものなどがある [3]。コンペ終了後、成績上位者が提案した分析モデルを見ることが出来る。株価の予想に関するものでは、“Two Sigma: Using News to Predict Stock Movement” [4] というコンペがあり、Two Sigma Investments (米国ヘッジファンド) が主催のスポンサーをしている。このコンペでは、米国株式市場の日足の取引データと、Reuters が提供する経済ニュースを利用して、将来の株価 (10 日先の株価) のリターンを予測するというコンペがある。

⁵ 株価形成理論に関して、大きく分けて次の 3 つに分類することができる。① ファンダメンタルズ理論、② 効率的市場仮説、③ 需給理論

2. データと分析手順

2.1 会社四季報

会社四季報には、図1にあるように、会社基本情報、業績推移、財務詳細、株価詳細、株主情報等のデータと会社四季報編集部のコメントが全上場企業（3723社、第2集）の（構造化された）情報が掲載されている。ただし、算出できない、非回答等に関しては、[-]となっている。

これらの情報から導出される理論株価とは、そ

の企業に投資した場合、将来にわたり投資家へもたらす利潤を予想し、算出した計算上の株価水準のことを指している。現在、投資家の間で広く知られている財務・業績から算出できる理論株価の2つを付録にまとめている。これらの理論株価は、ある独自の計算式にこれらのデータを当てはめることで、導出している。また、このような理論株価は著名な投資家も投資家の多くが参考にして売買していると説明している[5]。

図1 各企業の内容（CD-ROM版、イメージ）

The screenshot displays a software interface for viewing financial data. The main window is titled '7203 トヨタ自動車【東証】【過去の四季報】'. It is divided into three main sections corresponding to the quarters: 2014年4集, 2014年3集, and 2014年2集. The 2014年4集 section shows a table of stock transactions (年月, 資本変動, 万株) and a summary of key financial metrics (売上高, 営業利益, etc.). The 2014年3集 and 2014年2集 sections show detailed financial statements with columns for various metrics like 売上高, 営業利益, 税金利益, etc., and rows for different periods (e.g., 12.3, 13.3, etc.). The interface also includes a 'トヨタ自動車' header with a brief description and a '7203 輸送用機器' sub-header.

(出所 株式会社東洋経済新報社 ホームページ)

2.2 データ分析手順

Kaggle 等で活躍されている方の分析手順の概略⁶を次のようにまとめることができる。本稿でもこの手順を参考にして行こう。

(1) 探索的データ分析 (EDA, Exploratory Data Analysis)

データの可視化や相関分析等、簡単な分析を行

①ファンダメンタルズ理論は、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方法のことをいう。

⁶ 例えば、[6], [7]を参照されたい。

い、データの特徴を明らかにする。

(2) ベースラインモデルの構築

最初に与えられた生データを基に有名なモデル、例えばランダムフォレストをそのまま適用するなど、容易に結果が得られるモデルの構築を行う。

ただし、Kaggle では、モデルを提出し、公開されていない未知のデータでのテストの結果の精度を競い合うものであり、本稿の状況とは異なる。

(3) 交差検証 (Cross Validation)

まず訓練データを分割し、その一部をまずベースラインモデルを用いて、予測精度の高い最適なモデルを得る。次に、訓練データのうち、利用していないデータで、その最適なモデルを用いて、テストを行い、モデルの妥当性の検証を行う。

(4) 特徴量エンジニアリング[8]

会社四季報のデータにおいても、産業・上場区分等の数値以外のデータを取り扱うため、数値に置き換える必要がある。

また、検証を行いながら、数値の対数化、変化率等に行ったり2つ以上の種類の特徴量を組み合わせ、例えば、1株あたりに関するものを入れることで予測精度が上がるのであれば、新しい特徴量を作るなどを行う。ここでは、ドメイン知識と呼ばれる専門分野に関する知識が重要であると指摘されている。

(5) ハイパーパラメータ調整

モデルのパラメータを調整し、予測精度を上げていく。

(6) アンサンブル

テストデータは未知であり、訓練データで最も予測精度が高いものが必ずしも、テストデータでも予測精度が高いとは限らない。そのため複数のモデルの予測を組み合わせ、全体の予測精度を上げていく。

3. 評価指標と特徴量エンジニアリング

本稿では、個人投資家が会社四季報を次のような使い方をしてしていると想定している。会社四季報が発売され、業績、状況等のファンダメンタルを精査し、今後どのような株価となるのかを予想する。そこから株価は、割安・割高なのかを判断する。

そこで、会社四季報のデータを機械学習、教師あり学習の手法⁷を用いて、理論株価を導出する。ここでは、訓練データを「2019年第2集(春号)」(2019年3月15日発売)とデータ作成時点の終値(2019年2月25日)、テストデータを「2019年第3集(夏号)」(2019年6月18日発売)における

⁷ 教師あり学習とは、答えを含めたデータセット(訓練データ)を用意し、それを用いて学習を行い、モデルを作

全上場企業の財務データ等(説明変数・特徴量)を用いて、データ作成時点の終値(2019年5月30日)の株価(終値、被説明変数)を推計する。ただしその日に取引がなく、終値がない場合、最後に取引が行われた日の株価を用いる。

3.1 評価指標

モデルの評価は、機械学習等の分野でよく使われている次の3つの指標を用いる(n:企業数)。

- (1) 平均絶対誤差率(MAPE, Mean Absolute Percentage Error)

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|\text{理論株価}_i - \text{実際の株価}_i|}{\text{実際の株価}_i}$$

- (2) 平均二乗誤差(MSE, Mean Square Error)

$$MSE = \frac{1}{n} \sum_{i=1}^n (\text{理論株価}_i - \text{実際の株価}_i)^2$$

- (3) 決定係数 (R²)

実際の株価が理論株価に近い、予測精度が高い場合、このMAPE, MSEの値は小さく、R²の値は1に近くなる。本稿では、MAPE, MSEの値が小さく、R²の値が1に近いモデルを予測精度の高いモデルとする。

3.2 探索的データ分析

会社四季報のデータの主な特徴や相関分析の結果を次にまとめた。

- (1) 会社四季報から得られる説明変数・特徴量は137あり、数が多いものの、会社数は3721(第2集(春号))と少ない。
- (2) 最近、上場した会社など多くの会社において、欠損値が多い。
- (3) 株価(終値、被説明変数)と説明変数・特徴量と相関分析の結果、相関係数(0.5~0.7)が高いものは、「1株益(連結)」、「1株当たり純資産(連結)」、「1株配当(円)(連結)」であり、これらが高いほど株価も高い関係にあることが分かった。また、これらのデータは、既存の理論株価の導出に用いられている。

成する。作成したモデルを元に主に分類や回帰を行う。

3.3 ベースラインモデル

ベースラインモデルとして、容易に推計できる重回帰分析を用いて行った。しかしデータに欠損値が多く、予測をすることができなかった。そのため欠損値の多い銘柄や説明変数・特徴量を削除した。その結果、MAPE は約 206.8 となり、予測精度は 2 倍、半分以上異なった。

3.4 特徴量エンジニアリング

説明変数・特徴量に関して、主に次のことを行った。

- (1) 上場区分、業種、日経 225 採用銘柄などをカテゴリー化した。
- (2) 株価と発行済株数に関しては、数値のばらつきが大きいため、対数化した。
- (3) データには、マイナスの値があり、欠損値が多い。そこで、特徴量・データをカテゴリー化し、欠損値自体も 1 つのカテゴリーとして取り扱った。具体的には、データに順位をつけ、10% 毎にカテゴリー化し、欠損値を 11 番目のカテゴリーとした。

4. 推定結果と考察

4.1 推定結果

ベースラインモデルとして、ランダムフォレスト、勾配ブースティング(2 種類)の計 3 つの手法にて、理論株価の推計を行なった。

特にこの勾配ブースティング (Gradient Boosting) は、予測手法として、近年注目されている決定木アルゴリズムに基づいた勾配ブースティングの機械学習フレームワークである (1) XGBoost と (2) LightGBM を用いる。この勾配ブースティングは Kaggle コンテストで最も成功した最近の機械学習アルゴリズムの 1 つであると言われている [9]。また、実際 Kaggle 入賞者たちのアルゴリズムによく使われている。

3 つのモデルにおける結果は次のようになった。

(1) ランダムフォレスト

	train	test
MAPE	2.312	4.492
MSE	0.050	0.163
R ²	0.947	0.825

(2) XGBoost

	train	test
MAPE	5.014	5.615
MSE	0.216	0.249
R ²	0.767	0.731

(3) LightGBM

	train	test
MAPE	3.294	4.771
MSE	0.111	0.185
R ²	0.881	0.801

第 2 集(春号)と第 3 集(夏号)の終値に関する MAPE は約 13.3 であり、終値は、平均約 13.3%変動しており、上記 3 つのモデルにおいて、誤差が約 5%程度であり、高い予測精度であることが分かる。また、決定係数も高い。

4.2 重要な説明変数・特徴量

上記 3 つのモデルにおいて、機械学習ライブラリ scikit-learn を用いて、全 137 個の説明変数・特徴量の重要な上位 10 個を求め、その中で共通する説明変数・特徴量を表 1 にまとめた。相関分析と同様、「1 株益」や「1 株当たり純資産」が重要な説明変数・特徴量であることが分かった。これらは、代表的な理論株価でも利用されており、株価を説明する際、重要な特徴量であることが分かった。

表 1 各モデルに共通する重要な説明変数・特徴量

3 つ	連結・1 株益(修正, 最新) 連結・1 株当たり純資産 (修正, 最新)
2 つ	上場区分 業種 発行済株式数(対数) 連結・1 株配当(修正, 最新) 連結・1 株配当(修正, 1 期前) 連結・1 株配当(修正, 4 期前) 連結・剰余金の配当 連結・1 株当たり純資産 (修正, 5 期前) 連結・売上高原価率(%)

5. まとめと今後の課題

以上のように、会社四季報のデータを、Kaggleで利用されている手法を活用し、理論株価を導出した。指摘されているように、本稿においても分析アルゴリズムよりも説明変数・特徴量を工夫することにより、大きく精度が改善することが確かめられた。

本分析は、あるモデルの結果を示したものであり、より精度の高いモデルもあると考えられ、試行錯誤、様々なデータ分析を行うことが重要である。

実際、Kaggleは、データ分析を行うものにとって、失敗も含め、様々な情報が蓄積されてきており、データ分析を行う上で、大変有益な情報が多い。

本稿での分析における欠点として、例えばデータに関する問題がある。

- (1) 第2集(春号)から第3集(夏号)へのデータの更新・変更があまりない。そのため、数字の修正があったとしても、カテゴリーが変わるほど影響がない。
- (2) 例えば第2集(春号)では欠損値であったものが、第3集(夏号)となり、データが入力されることがある。この変化はファンダメンタルの変化ではなく、時間が経過し、データが入力されたに過ぎない。

また、残された課題として、

- (1) 会社四季報編集部による業績予想数字と企業評価の手がかりとなるような材料についてのコメントには変更があるため、これに関するテキストマイニングを行うとより、今後の予測精度向上となるかもしれない⁸。
- (2) 実際にこのような分析結果を利用する際には、本稿では、2019年第2集(春号)から第3集(夏号)のみの予想を行なったが、データの期間を遡り検証を重ねていくことが必要である。

注意事項

本稿における理論株価は、あくまで「会社四季報」から得られる財務・業績数値等から独自に算出した「理論的な株価水準」を計算したものであり、その正確性、実現性を保証するものではない。

参考文献

- [1] <https://shikiho.jp/>
- [2] <https://www.kaggle.com/c/mercari-price-suggestion-challenge/>
- [3] <https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting>
- [4] <https://www.kaggle.com/c/two-sigma-financial-news>
- [5] 片山晃, 小松原周「勝つ投資 負けない投資」クロスメディア・パブリッシング, 2015年
- [6] <https://drive.google.com/file/d/1G1uJoC2umBh3zsg0oGH2dKP4nfpmb3L/view>
- [7] 門脇大輔, 阪田隆司, 保坂桂佑, 平松雄司「Kaggleで勝つデータ分析の技術」技術評論社, 2019年
- [8] Alice Zheng, Amanda Casari「機械学習のための特徴量エンジニアリング—その原理とPythonによる実践」オライリージャパン, 2019年
- [9] <https://www.kdnuggets.com/2016/03/xgboost-implementing-winningest-kaggle-algorithm-spark-flink.html>
- [10] https://www.gsam.com/content/gsam/jpn/ja/gsitm/resources/investment-ideas/gsbdsj.html#section-background_aa2c_background_sectionheaderwithab
- [11] https://str.toyokeizai.net/files/user/dl/magazine/risc_riron.pdf

⁸ 会社四季報のコメント欄にある言葉がポジティブかネ

ガティブかを判断し、銘柄評価に活用しているファンドもある[10]。

付録 広く知られている理論株価

	東洋経済新報社「会社四季報プロ 500」 ⁹	ダイヤモンド社「ダイヤモンド ZAi(ザイ)」 ¹⁰
理論株価	理論株価＝直近期 BPS ＋将来の 1 株当たりの超過利益の現在価値	理論株価＝資産価値＋利益価値＋成長価値
詳細	将来の 1 株当たりの超過利益の現在価値＝（今期 予想 EPS－最近 BPS×株主資本コスト） / （1＋株主 資本コスト）＋（来期予想 EPS－今期予想 BPS×株 主資本コスト） / （1＋株主資本コスト） ² ＋ 以下 n 期 まで 株主資本コスト＝長期国債（10 年債）利回り ＋β 値×株式リスクプレミアム EPS＝1 株当たり利益 BPS＝1 株当たり純資産 注 特殊要因による過大な理論株価を避けるため、時価 に対する倍率を最大 3 倍に制限している。	資産価値：1 株あたりの株主純資産(BPS) 利益価値：今期の予想 1 株益(EPS)×年数 成長価値：売上高成長率から算出した、各年 の 1 株益成長分の合計
利用データ	純資産（前期末、最新） 発行済株式数（最新） 予想利益(今期、来期) 長期国債（10 年債）利回り 業種	純資産（前期末） 発行済株式数（最新） 前期売上高（前期末） 今期予想売上高(最新) 予想利益(最新) 業種
MAPE	75.07	83.79

⁹ 「会社四季報プロ 500」は編集部が選んだ「会社四季報」にある 500 銘柄について、図を多く入れることで、よりビジュアル的に分かりやすく解説した投資情報誌で

ある。この理論株価に関しては、[11]で説明されている。

¹⁰ 四半期毎に、別冊付録として理論株価を掲載している。