令和２年度 修士論文

Explainable AIを用いた有価証券報告書の分析

An Analysis of Financial Reports with Explainable AI

学籍番号 251902108

久保 穂高

名古屋大学大学院 情報学研究科
複雑系科学専攻
多自由度システム情報論講座
時田研究室

2021年1月29日

概要

機械学習の大きな課題の一つが、AI のブラックボックス問題である。学習済みモデルが正しく予測しても、予測の根拠が説明できない問題のことだ。金融庁が実施した有価証券報告書の分析に関する研究においても、決算数値や記述内容から業績の良い企業の特徴の抽出に成功した一方、数値などの予測の根拠が説明できないことが課題となっていた。これらの問題の解決策として近年注目されている技術が Explainable AI(説明可能な AI) である。これを用いることで学習済みモデルを分析し、予測根拠を人が理解できる形で可視化することが可能となる。本研究では Explainable AI を用いて、金融庁の先行研究で挙げられていた AI のブラックボックス問題の解決を試みた。分析対象は 2020 年 5 月現在の日経平均採用銘柄 225 社で、有価証券報告書に記載されている損益計算書や、経営に関する記述内容、その他株価などのデータから、計 694 個の特徴量を作成した。そして売上高及び売上増加率を予測する 2 つの機械学習モデルを作成し、4 つの Explainbale AI フレームワークを適用することで予測根拠を比較した。結論として、売上高予測には過去の売上高や総資産額、そして営業活動によるキャッシュフローが最も寄与しており、株価の伸び率や所属の業種、また経営に関する記述の印象による特徴量は、予測に影響しないことがわかった。そして売上増加率予測には、株価の変化率や期末残高の増加率、また経営に関する記述が最も予測に寄与しており、過去の売上高の伸び率や売上に対する利益の割合などは、予測に影響しない結果となった。そして Explainable AI の一つである PDP の結果から、売上高の予測値の期待値が最も高い経営に関する記述量などを可視化することができた。さらに別の Explainable AI である SHAP を利用することで一社一社の予測値に対し、どの要因が最も予測に影響を与えたのかを可視化することができた。なお、日経 225 社の決算内容は一つのデータにまとめ、CSV ファイルとして公開している。

Abstract

One of the major problems of machine learning is the black box problem. It is a problem where the learned model predicts correctly, but the reason for the prediction cannot be explained. A research study conducted by the Financial Services Agency on the analysis of securities reports succeeded in extracting the characteristics of companies with good performance, but the problem was that the evidence for the predictions could not be explained. Explainable AI is a technology that has been attracting attention in recent years as a solution to these problems. Using this technology, it is possible to visualize the evidence behind predictions in a form that people can understand. The analyzed data includes companies that are members of the Nikkei 225 as of May 2020. A total of 694 features were created from the income statements, management descriptions, and other data such as stock prices. Using this data, I created two machine learning models which predict sales and sales growth, respectively. Permutation Importance, Partial Dependence Plot, and SHapley Additive exPlanations were applied to those learning models, and the prediction evidence was compared by using multiple Explainable AIs. In conclusion, we found that past sales, total assets, and cash flow from operating activities contributed the most to the sales forecast, while the growth rate of stock prices, the industry to which the company belonged, and the amount of characteristics based on impressions of management descriptions had little effect on the prediction. In addition, the rate of change in stock price, the rate of increase in the balance at the end of the period, and the description of management contributed the most to the prediction of the rate of increase in sales, while the rate of growth of sales in the past and the ratio of profit to sales had little effect on the prediction.

目 次

第1章　本研究について

1.1　本研究の背景

「日本人は投資が嫌いだ」と指摘する識者は多い。金融庁のデータによると、米国の家計資産の中での株式・投資信託の割合は 2015 年末時点で 45.4 %、英国は 35.7 %、日本は 18.8 %である [5]。日本では、現預金として家計金融資産を保有している割合が多く、資産の伸びは低い水準が続いている。実際、米国では 1995 年から家計金融資産は 3 倍に伸びた一方、日本は 1.47 倍しか伸びていない。この現状を踏まえて金融庁は、つみたて NISA、実践的な投資教育、金融機関の顧客本位の業務運営の確立・定着等を総合的に推進し、日本の資産を向上させることを国民に促している。

国の投資に対する施策が推進される中で、株・投資信託の予測に有効な判断材料をつくる取り組みが各所でなされている。例えば、専門家の分析した Web ニュースの着眼点を抽出することで日経平均株価予測の精度が向上するかの実験を行った研究や [6]、大手企業のアニュアルレポートの経営陣の顔写真や文章から感情スコアを計算し企業活動とどう関係しているかを調べた論文も存在する [2]。また、経済物理学という分野では予測の難しい経済現象を物理学の法則と紐付けてモデル化する取り組みなどが行われ [7]、さまざまなバックグラウンドを持った研究者が独自の手法を用いて経済活動の評価を試みている。

本研究では数々の取り組みがなされている中でも、日経 225 社の有価証券報告書を学習データとした売上高に関する予測モデルを作成した。損益計算書 (P/L)、事業内容や研究内容などの自然言語を特徴量に組み込み、それらの重要度を定量的に評価することで、予測に寄与する特徴量を探った。新たな投資判断の知見提供をするとともに、広くは日本全体の投資意欲を掻き立て、国の資産増加の一助となれば嬉しい。

1.2　先行研究

前節で紹介したように、決算発表前に業績を予想する取り組みが各所でなされている。この節ではいくつかの関連研究を紹介する。

1.2.1　Google Trends as Complementary Tool for New Car Sales Forecasting: A Cross-Country Comparison along the Customer Journey.[1]

この研究は、2016 年にオランダの TWENTE 大学で実施されたもので、米国・独国の自動車の売上を Google の検索ワードを使って予測している。いくつかの線形回帰モデルが用いられている。

一つ目の仮説検定では、Google における検索数 (相対値) と売上では有意な相関があることを明らかにしており、二つ目の仮説検定ではその傾向は独国よりも米国の方が強いことを示している。また、14 メーカー 24 種の国別の検定も実施し、加えて平均し 6 か月のタイムラグを加えることでより高い相関が得られることも明らかにしている。さらに自動車の価格帯に応じて最適なタイムラグが違うことも明らかにしている。

このように一般消費者向けに財を売る企業は、検索エンジンの検索量を活用し、検索時期と売上の価格帯ごとの最適な時間差を見つけることで、売上の予測の精度を上がることが示されている。

1.2.2 感情分析を通じて見える経営者の特性と企業の関係 [2]

この論文は、2018 年に慶應義塾大学の研究チームが行ったもので、2013 年から 2017 年までの日経 225 社のアニュアルレポートにおける、写真や文章の感情スコアなどを指標とした分析を実施した。それらの指標を元に「設備投資額売上高比率」「研究開発費売上高比率」「販管費率」「負債比率」「ROA」との有意な関係を調査している。アニュアルレポートの感情スコアには「経営者の登場回数」「1 ページにおける経営者の写真やサインの占める割合」に加え、Microsoft 社の EmotionAPI を用いて文章を測定したものが加えられている。

検定結果として、ページに占める写真やサインの割合が大きいほど「設備投資額売上高比率」が、そしてレポートにおける経営者の登場回数が大きいほど「負債比率」が大きいことが統計的に示されている。また、経営者の登場回数が高いほど「ROA(総資産利益率)」が小さくなることも統計的に明らかになっている。

1.2.3 深層学習を用いたアンサンブルモデルによる企業価値推定モデルの提案 [3]

この研究は、2017 年に東京大学の松尾豊教授の研究室で行われたもので、「有報キャッチャー」と呼ばれる財務情報を取得する API を用いてデータベースを作成し、XGBoost などの学習モデルを用いたアンサンブル学習を行った。予測対象は各企業の投資判断 (買いか売りか) 及び目標株価とし，各企業の決算短信発表日から営業 60&240 日で検証している。既存の経済学的ファンダメンタル分析と機械学習を用いた筆者のモデルとの予測精度を比較しており、既存の経済的手法では平均して約 50 %だった正解率が、この研究で構築した予測モデルでは平均約 64 %の正解率を出し、筆者の提案したモデルの有効性が実証された。

また、決定木ベースのモデル XGBoost を用いていることから、予測に寄与した特徴量も紹介されており、以下の 表 1.1 がその特徴量の上位 10 位である。

順位	項目
1	EV
2	連結総従業員数
3	株価
4	株式発行数
5	PBR
6	PER
7	従業員数
8	配当利回り
9	法人税等の支払額
10	配当金の支払額

表 1.1: 2014 年第 4 四半期・営業 240 日平均の回帰予測におけるファンダメンタル指標の重要度ランキング [3]

1.2.4 金融庁政策オープンラボ 「有価証券報告書等の審査業務等における AI 等利用の検討」実証実験の結果概要について [4]

この研究は、2019 年 9 月に金融庁が主導で行ったもので、有価証券報告書の効果的・効率的な審査や、投資家等にとって有用な記述情報の充実に向けて、AI 等のテクノロジーの利用が考えられないか検討したものである。1 か月という短い期間の中で情報通信業社から監査法人まで計 20 社で実証実験を行った。機械学習を用いた数値分析や自然言語処理を実施し、ニューラルネットワークや決定木アルゴリズムなどさまざまな学習モデルを用いて知見の獲得を試みている。

機械学習や深層学習では、教師データのとおり分類できたとしても、何故それが適切なのか、どこが適切でないのかを明示することは困難であることが挙げられており、さらに実証期間が 1 か月だったことやデータが不十分だったことから十分なフィードバックが得られなかったことを課題として挙げている。業績の良い企業の報告書の分類手段として、経営者がどのような考えで経営指標を設定しているかでクラスタリングを行い、文章の特徴や内容などを解析できたこと、さらには固有名詞を用いて自社のサービスをアピールしていることがわかった。

今後の展望として、有価証券報告書は形容詞が少ない一方で固有名詞が多く、動詞も「名詞＋する」というように使われていることが多かったため、名詞とその属性について着目することでより有用な知見が得られるのではないかという指摘や、業績の良い企業の報告書には「増減」「変動」「損失」など具体的な分析に使用される単語が見られ、業績の悪い会社では「努める」「図る」など曖昧な表現がされていたことなども結果としてわかっている。また論文の中で、業績予想や報告書の分類などのタスクは AI にまかせ、人が AI の出力した知見を用いることで人が AI を活用する社会を目指すべきではと提言している。

1.3　本論文の概要

本研究では、初めに日経 225 社の会計に関するデータをまとめる。そして、2015 年度から 2018 年度までの損益計算書 (P/L) に書かれている数値や経営に関して書かれている記述に加えて、先行研究で明らかとなっている「文章の感情分析」や「株価推定に有効な特徴量（表 1.1 より）」を特徴量として加えたデータから、2019 年度の売上高及び売上率の予測モデルを作成する。さらに、[4] で課題となっていた予測の根拠を複数の「説明可能な AI」を用いることで可視化し、予測に寄与した特徴量を定量的に評価することを試みる。有価証券報告書の分析の一例として、そして Explainable AI の利用例として有用な資料になれば嬉しい。

第2章　データの取得

本章では、日経225社の有価証券報告書の取得先と、機械学習可能なPandas形式のデータへの加工方法について述べる。

2.1　取得先

データのダウンロードはEDINET[8]の書類検索ページから行った。EDINETとは金融庁が運営する「金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システム」である。分析対象は日経225社としたが、これは日本経済新聞社によって定期的に入れ替えが実施されており、著作権も日本経済新聞社が保有している。補足として、日本の重要な経済指数として広く知られている日経平均株価はこれら225社の株価から算出されたもので、東証一部上場企業の株価指数はTOPIXとして知られている。

ダウンロードした日経225社の決算データは、日本取引所グループの東証上場銘柄一覧[9]の2020年5月末のデータに記載してあるものを採用した。日経225社リストはA.1に記載している。

検索結果に含まれるダウンロード可能なファイルはPDFとXBRLの二種類である。中でもスクレイピングをするのに適したXBRlファイルを2019年度・2018年度の2年分×225社 = 450ファイルを全てダウンロードした。

検索結果

4件中(1～4件表示)

XBRL一括ダウンロード

提出日時	提出書類	コード	提出者／ファンド	発行／対象／子会社／臨報提出事由	PDF	XBRL	比較	備考
R2.06.25 13:18	有価証券報告書−第40期(平成31年4月1日−令和2年3月31日)	E02778	ソフトバンクグループ株式会社		PDF	XBRL	COMPARE	
R2.02.13 14:40	四半期報告書−第40期第3四半期(令和1年10月1日−令和1年12月31日)	E02778	ソフトバンクグループ株式会社		PDF	XBRL	COMPARE	
R1.11.12 12:41	四半期報告書−第40期第2四半期(令和1年7月1日−令和1年9月30日)	E02778	ソフトバンクグループ株式会社		PDF	XBRL	COMPARE	
R1.08.09 13:25	四半期報告書−第40期第1四半期(平成31年4月1日−令和1年6月30日)	E02778	ソフトバンクグループ株式会社		PDF	XBRL	COMPARE	

4件中(1～4件表示)

図 2.1: EDINETにおける有価証券報告書の検索結果

2.2 学習データの取得方法

各企業の決算フォルダは図 2.2 のような構造になっており、XBRL/PublicDoc 以下に決算のデータが保存されている。ファイルは xml という拡張子で保存されている。2019 年度の損益計算書の数値と 2018 年度の経営に関する項目の自然言語データをスクレイピングし分析に用いるデータを抽出した。

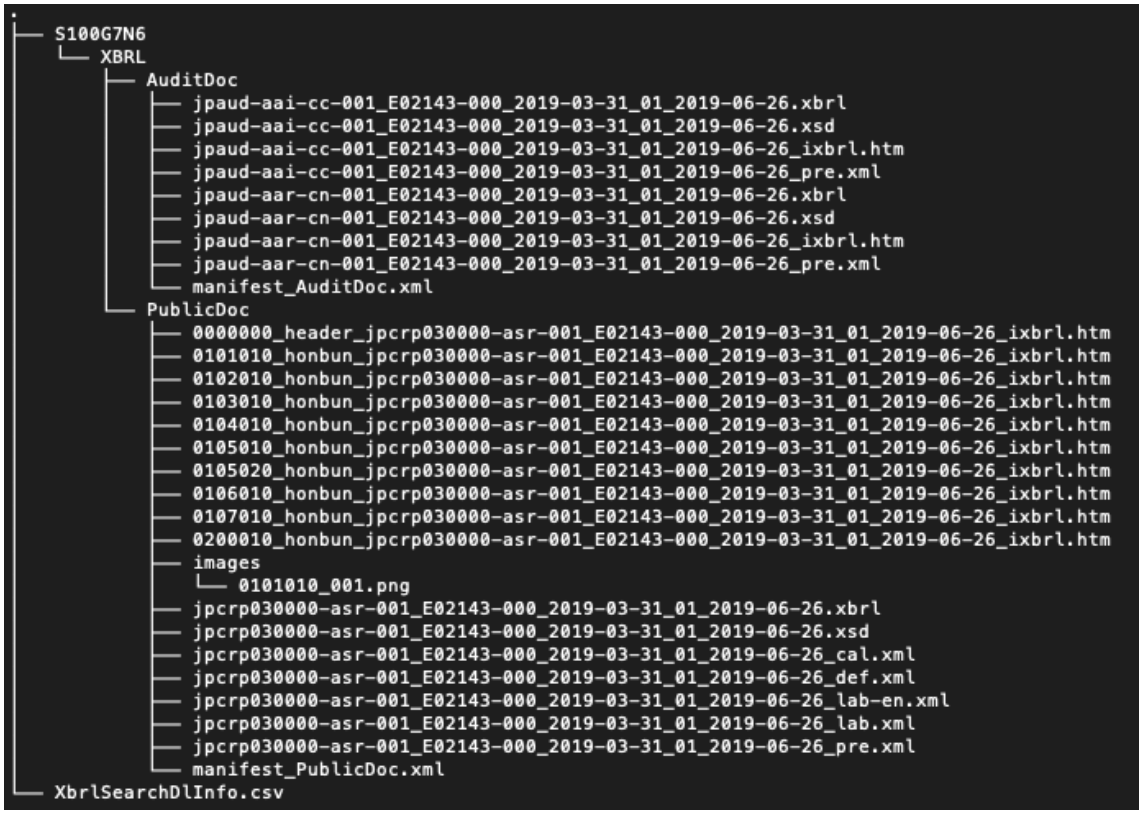

図 2.2: 各企業の決算ファイルの構造

2.2.1 損益計算書

図 2.2 の中でも 0101010 で始まるファイルに損益計算書が記録されている。損益計算書のフォーマットは各社で異なる。単語の違いや全角半角などの違いに注意しながら項目をまとめ、Python 言語のフレームワークである Pandas[10] の DataFrame と呼ばれる形式にまとめた。

	包括利益	収益	営業利益	営業活動によるキャッシュ・フロー	基本的1株当たり当期利益	基本的1株当たり親会社の所有者に帰属する当期利益	売上高	希薄化後1株当たり当期利益	希薄化後1株当たり当社株主に帰属する当期純利益	当期純利益	...	1株当たり当期純利益	1株当たり当社株主に帰属する当期純利益	1株当たり株主資本	1株当たり潜在株式調整後当期純利益金額	1株当たり純資産額	1株当たり親会社所有者帰属持分	1株当たり配当額
0_(株)ＩＨＩ	26829.0	NaN	NaN	63589.0	NaN	NaN	1455844.0	NaN	NaN	NaN	...	58.84	NaN	NaN	58.77	NaN	NaN	NaN
1_(株)ＩＨＩ	-15228.0	NaN	NaN	95338.0	NaN	NaN	1539388.0	NaN	NaN	NaN	...	9.90	NaN	NaN	9.90	NaN	NaN	NaN
2_(株)ＩＨＩ	4628.0	NaN	NaN	65373.0	NaN	NaN	1486332.0	NaN	NaN	NaN	...	33.98	NaN	NaN	33.96	NaN	NaN	NaN
3_(株)ＩＨＩ	16774.0	NaN	NaN	99018.0	NaN	NaN	1590333.0	NaN	NaN	NaN	...	53.71	NaN	NaN	53.67	NaN	NaN	NaN
4_(株)ＩＨＩ	39597.0	NaN	NaN	46402.0	NaN	NaN	1483442.0	NaN	NaN	NaN	...	258.53	NaN	NaN	258.37	NaN	NaN	NaN
0_いすゞ自動車(株)	219711.0	NaN	NaN	151558.0	NaN	NaN	1879442.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1_いすゞ自動車(株)	77561.0	NaN	NaN	132972.0	NaN	NaN	1926967.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2_いすゞ自動車(株)	106315.0	NaN	NaN	151352.0	NaN	NaN	1953186.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

図 2.3: 一つの Pandas オブジェクトに変換したデータ

損益計算書の項目は以下である。フォーマットが定まっていないが故に、データの項目ごとに欠損値の頻度が異なる。欠損値の割合をまとめた表が図 2.4 である。

損益計算書の項目

決算年月 / 売上高 / 収益 / 営業利益 / 経常利益 / 税引前利益 / 当期純利益 / 包括利益 / 営業活動によるキャッシュ・フロー / 投資活動によるキャッシュ・フロー / 財務活動によるキャッシュ・フロー / 純資産額 / 総資産額 / 従業員数 / 親会社株主に帰属する当期純利益 / 親会社の所有者に帰属する当期包括利益 / 親会社の所有者に帰属する持分 / 親会社所有者帰属持分比率 / 親会社所有者帰属持分当期利益率 / １株当たり当期純利益 / 基本的１株当たり当期利益 / 基本的１株当たり親会社の所有者に帰属する当期利益 / 希薄化後１株当たり当社株主に帰属する当期純利益 / 親会社の所有者に帰属する当期利益 / １株当たり当社株主に帰属する当期純利益 / １株当たり潜在株式調整後当期純利益金額 / １株当たり純資産額 / １株当たり株主資本 / １株当たり配当額 / １株当たり親会社所有者帰属持分 / 希薄化後１株当たり当期利益 / 株価収益率 / 現金及び現金同等物の期末残高 / 自己資本比率 / 株主資本比率 / 自己資本利益率 / 最高株価 / 最低株価 / 株主総利回り / 資本金 / 発行済株式総数 / 配当性向

	Total	Percent
月	0.0	0.000000
社名	0.0	0.000000
総資産額	0.0	0.000000
年	0.0	0.000000
決算年月	0.0	0.000000
従業員数	0.0	0.000000
営業活動によるキャッシュ・フロー	2.0	0.188147
現金及び現金同等物の期末残高	5.0	0.470367
財務活動によるキャッシュ・フロー	7.0	0.658514
投資活動によるキャッシュ・フロー	7.0	0.658514
包括利益	11.0	1.034807
親会社の所有者に帰属する当期利益	79.0	7.431797
株主総利回り	122.0	11.476952
売上高	234.0	22.013170
純資産額	293.0	27.563500
自己資本比率	309.0	29.068674
自己資本利益率	310.0	29.162747
経常利益	314.0	29.539040
1株当たり当期純利益	484.0	45.531515
税引前利益	798.0	75.070555
1株当たり親会社所有者帰属持分	800.0	75.258702
親会社所有者帰属持分比率	800.0	75.258702
親会社所有者帰属持分当期利益率	807.0	75.917215
親会社の所有者に帰属する当期包括利益	820.0	77.140169
1株当たり潜在株式調整後当期純利益金額	824.0	77.516463
親会社の所有者に帰属する持分	824.0	77.516463
収益	826.0	77.704610
基本的1株当たり当期利益	886.0	83.349012
希薄化後1株当たり当期利益	907.0	85.324553
基本的1株当たり親会社の所有者に帰属する当期利益	960.0	90.310442
希薄化後1株当たり当社株主に帰属する当期純利益	964.0	90.686736
1株当たり純資産額	968.0	91.063029
営業利益	974.0	91.627469
株価収益率	994.0	93.508937
親会社株主に帰属する当期純利益	995.0	93.603010
当期純利益	1003.0	94.355597
資本金	1005.0	94.543744
株主資本比率	1021.0	96.048918
1株当たり株主資本	1026.0	96.519285
1株当たり当社株主に帰属する当期純利益	1048.0	98.588899
最高株価	1058.0	99.529633
最低株価	1058.0	99.529633
1株当たり配当額	1058.0	99.529633
発行済株式総数	1058.0	99.529633
配当性向	1058.0	99.529633

図 2.4: 各項目の Null 値の割合

2.2.2 事業内容

また、以下の 5 点についての記述は 0102010 から始まる 2018 年度のファイルから取得した。経営方針から研究活動までの 225 社の記述も同じ Pandas の DataFrame ファイルにまとめた。

1. 経営方針，経営環境及び対処すべき課題等
2. 事業等のリスク
3. 経営者による財政状態，経営成績及びキャッシュ・フローの状況の分析
4. 経営上の重要な契約等
5. 研究開発活動

	name	id	keiei	risk	money	contract	reserch
0	(株)ＩＨＩ	50	3013	7130	8254	1107	2225
1	いすゞ自動車(株)	57	1329	3767	7086	434	1703
2	出光興産(株)	124	5597	7569	8504	2178	3753
3	花王(株)	205	2952	9462	8905	258	4541
4	東宝(株)	186	1388	1327	10014	197	11
...	...	...	...	...	...	...	...
220	昭和電工(株)	176	2515	4231	6185	1421	4865
221	(株)ＳＣＲＥＥＮ ＨＤ	53	1998	3393	6961	11	1396
222	(株)ブリヂストン	24	2600	5340	5854	11	2072
223	(株)日本製鋼所	40	4875	2065	7615	651	1493
224	(株)三井住友ＦＧ	27	3486	12090	15618	15640	15646

図 2.5: 各社の事業内容の項目別の文字数

取得した 225 社分の総文字数を以下の表にまとめた。

内容	225 社の総文字数
経営方針，経営環境及び対処すべき課題等	868775
事業等のリスク	1240678
経営者による財政状態	2340886
経営上の重要な契約等	189169
研究開発活動	548329

以上の過程を経て、有価証券報告書からデータを抽出した。次章ではそれを用いて学習を行うための前処理をする。

第3章　前処理

前章で作成したデータから機械学習可能な形式にデータを処理する。本章の「3.1 特徴量生成」では特徴量の生成方法を、続いて「3.2 特徴量追加」では有価証券報告書に追加したデータについて紹介する。

3.1　特徴量生成

特徴量生成は予測精度を高める上で重要なプロセスである。目的変数と相関の強い特徴量が見つかれば、予測精度の向上が期待できるからだ。この章では、第 2 章で生成したデータから予測精度をあげる特徴量を様々な手法を用いて生成する。

3.1.1　簡単な演算による特徴量生成

目的変数となる売上高について、欠損値の場合は収益、尚欠損値の場合は営業利益を代入している。その他にも営業/投資/財務活動によるキャッシュフローを総資産額や利益で割った値、さらに従業員数を利益で割った特徴量など、簡単な四則演算による特徴量を追加している。

3.1.2　Aggregation

第 2 章で生成したデータは、図 3.1 のようにそれぞれの行がある会社のある年度の業績を表している。各会社の 2018 年度までのデータで、2019 年度の売上高と売上率を予測するモデルを作成するので、図 3.2 のように複数行あるデータを会社ごとにまとめる必要がある。そこで Aggregation と呼ばれる方法を用いて、2018 年度までの各項目の平均値、分散、最大値、最小値を求め、特徴量に加えた。これによって 1083 行のデータを 225 行に変換した。

	CompanyID	Year	NumberOfEnployee	NumberOfEnployee/Profit_	NetCashByOperating	NetCashByInvestment	NetCashByFinancing
0	50	2015	28533.0	1.063513	63589.0	-74611.0	33443.0
1	50	2016	29494.0	-1.936827	95338.0	-35513.0	-47530.0
2	50	2017	29659.0	6.408600	65373.0	-28961.0	-21941.0
3	50	2018	29706.0	1.770955	99018.0	-47977.0	-57326.0
4	50	2019	29286.0	0.739601	46402.0	-79280.0	16463.0
...	...	...	...	...	...	...	...
1058	206	2015	47565.0	0.640770	223613.0	-212912.0	1689.0
1059	206	2016	47456.0	0.452725	259880.0	-233219.0	-31315.0
1060	206	2017	47382.0	0.514479	234144.0	-295808.0	44304.0
1061	206	2018	47869.0	0.419275	275101.0	-166352.0	-71422.0
1062	206	2019	47842.0	0.456434	289728.0	-247420.0	-7174.0

1063 rows × 7 columns

図 3.1: Aggregation する前のデータ

	CompanyID	bool_ratio	hist_NetCashByOperating/Profit_max	hist_NetCashByOperating/Profit_min	hist_NetCashByOperating/Profit_mean	hist_NetCashByOper
0	50	False	14.125540	-6.260704	4.034515	
1	57	True	1.714418	0.689806	1.251446	
2	124	True	0.759630	-2.499601	-0.711006	
3	205	False	1.975762	1.043956	1.643016	
4	186	True	2.106560	1.046881	1.425449	
...	...	...	...	...	...	
220	208	False	37.753865	-3.527843	10.062503	
221	207	True	1.403237	-1.764116	0.219498	
222	80	True	32.307423	4.692648	15.293616	
223	33	True	12.036806	-1.472685	4.987493	
224	206	True	3.012394	2.409552	2.610884	

225 rows × 98 columns

図 3.2: Aggregation した後のデータ

3.1.3 TF-IDF

TF-IDF とは文章に含まれる各単語の出現頻度とその希少性から、文章をベクトル化するアルゴリズムである。一つひとつの文章を特徴量で表現することを考えたときに、単純に単語の出現回数を特徴量とすると、英語でいえば the や a のような文章において重要でない単語に重みが付けられてしまう。そこで前文章におけるその単語の希少性と出現回数を掛け合わせることで、文章のオリジナリティをうまく特徴量として表現するよう考案されたアルゴリズムが TF-IDF である。ちなみに、TF-IDF とは Term Frequency/Inverse Document Frequency の略で、それ「単語の出現頻度」と「単語の逆文書頻度」と直訳される。

TF 値

TF 値は、文章に占めるその単語の割合のことで以下の式で定義される。

$$tf(t_i, d_i) = \frac{\text{文章 } d_i \text{内における単語 } t_i \text{の割合}}{\text{文章 } d_i \text{全体の単語数}} \tag{3.1}$$

したがって、文章中に含まれる単語の割合が多ければ多いほど、TF 値は高くなる。

IDF 値

一方、IDF 値はその単語を含む文章の割合の逆数を表す。式にすると下のように定義できる。

$$idf(t_i) = \log(\frac{\text{全ての文章数}}{\text{単語 } t_i \text{が出現する文章の割合} + 1}) \tag{3.2}$$

対数がついているのは、分母の値が小さい場合に値が発散することを防ぐためであり、また分母に 1 が加算されているのは、分母が 0 となるのを避けるためである。これらの演算をによる $idf(t_i)$ 値の大小関係は変化しない。

TF-IDF 値

TF-IDF 値は、以上の TF 値・IDF 値の積で表される。

$$tfidf(t_i, d_i) = tf(t_i, d_i) \cdot idf(t_i) \tag{3.3}$$

本研究では、2018 年度有価証券報告書の「事業の状況」に含まれる「経営方針，経営環境及び対処すべき課題等」「事業等のリスク」「経営者による財政状態」「経営上の重要な契約等」「研究開発活動」の 5 つの項目を対象に TF-IDF 値を計算した。全ての単語の値を計算すると計算コストが莫大になるので、自然言語の中でも名詞のみを抽出し、加えて 1 回しか出現しない単語は除外した。

内容	225 社の項目別単語種数
経営方針，経営環境及び対処すべき課題等	7719
事業等のリスク	7164
経営者による財政状態	13577
経営上の重要な契約等	4098
研究開発活動	9425

SVD

さらにこれら全ての TF-IDF 値を特徴量にすることを考えると

$$7719 + 7164 + 13577 + 4098 + 9425 = 41983 \tag{3.4}$$

と、特徴量がとても膨大でスパースなデータとなってしまう。そこで、225 社× 41983 の TF-IDF 値の行列を SVD アルゴリズムで次元削減をすることで、より密で機械学習の実行が容易な行列に変換した。

SVD とは次元削減をする行列 A を下の三つの行列に近似することで行われる。A は m(文章数) 行 n(単語数) 列の行列である。r は削減した次元数である。[11]

$$A_{m\times n} \approx U_{m\times r}\Sigma_{r\times r}V^T_{n\times r} \tag{3.5}$$

U は $m \times r$ で m 次元空間における A の列の正規直交基底となっている。V は $n \times r$ で n 次元空間における A の業の正規直交基底となっている。Σ は $r \times r$ の対角行列で、対角成分は固有値の平方根 (特異値) であり寄与度 σ が $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ と高い順にソートされている。

$$A_{m\times n} \approx U_{m\times r}\Sigma_{r\times r}V^T_{n\times r} = \Sigma_i \sigma_i u_i \circ v_i^T \tag{3.6}$$

$$= \{u_1, u_2, \ldots, u_i\} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_i \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_i^T \end{bmatrix} \tag{3.7}$$

このアルゴリズムを用いることで 41983 の特徴量を 417 に次元削減した。

3.1.4 印象分析

有価証券報告書の「事業の状況」に含まれる「事業の状況の 5 項目」について日経 225 社全てのデータを印象分析し特徴量として追加する。印象分析には Microsoft Azure の Text Analytics を用いて、文章のポジティブ値、ネガティブ値、ニュートラル値を出力する。

また、Text Analytics では 5120 文字以上を一度に分析することができない。記述内容が 5120 文字を超える場合は、初めの 5120 文字を採用した。

- 経営方針，経営環境及び対処すべき課題等
- 事業等のリスク
- 経営者による財政状態，経営成績及びキャッシュ・フローの状況の分析
- 経営上の重要な契約等
- 研究開発活動

	CompID	keiei-pos	keiei-neu	keiei-neg	risk-pos	risk-neu	risk-neg	money-pos	money-neu	money-neg	contract-pos	contract-neu	contract-neg	reserch-pos	reserch-neu	reserch-neg
0	(株)ＩＨＩ	0.01	0.98	0.01	0.00	0.99	0.01	0.00	0.99	0.01	0.00	1.00	0.00	0.01	0.99	0.00
1	いすゞ自動車(株)	0.50	0.44	0.06	0.03	0.97	0.00	0.02	0.97	0.01	0.09	0.90	0.01	0.07	0.93	0.00
2	出光興産(株)	0.03	0.96	0.01	0.07	0.92	0.01	0.01	0.99	0.00	0.03	0.96	0.01	0.03	0.96	0.01
3	花王(株)	0.04	0.95	0.01	0.05	0.95	0.00	0.02	0.98	0.00	0.00	1.00	0.00	0.04	0.94	0.02
4	東宝(株)	0.06	0.93	0.01	0.06	0.92	0.02	0.02	0.97	0.01	0.00	1.00	0.00	0.21	0.67	0.12
5	凸版印刷(株)	0.05	0.95	0.00	0.04	0.96	0.00	0.03	0.96	0.01	0.07	0.92	0.01	0.04	0.95	0.01
6	(株)三井Ｅ＆Ｓ ＨＤ	0.04	0.96	0.00	0.03	0.96	0.01	0.04	0.95	0.01	0.02	0.98	0.00	0.03	0.97	0.00
7	(株)ファミリーマート	0.04	0.95	0.01	0.03	0.97	0.00	0.03	0.96	0.01	0.00	1.00	0.00	0.07	0.91	0.02
8	(株)デンソー	0.06	0.93	0.01	0.05	0.94	0.01	0.02	0.97	0.01	0.24	0.73	0.03	0.09	0.90	0.01
9	(株)新生銀行	0.02	0.97	0.01	0.06	0.93	0.01	0.03	0.96	0.01	0.21	0.67	0.12	0.21	0.67	0.12
10	(株)電通グループ	0.03	0.97	0.00	0.05	0.94	0.01	0.03	0.97	0.00	0.02	0.97	0.01	0.03	0.96	0.01
11	三菱倉庫(株)	0.02	0.98	0.00	0.03	0.97	0.00	0.03	0.96	0.01	0.20	0.71	0.09	0.21	0.67	0.12
12	アサヒグループＨＤ(株)	0.01	0.97	0.02	0.03	0.96	0.01	0.01	0.91	0.08	0.01	0.98	0.01	0.63	0.31	0.06
13	第一三共(株)	0.02	0.98	0.00	0.02	0.98	0.00	0.01	0.98	0.01	0.03	0.97	0.00	0.00	1.00	0.00

図 3.3: 「事業の状況の 5 項目」の各社の印象数値

3.2 特徴量追加

3.2.1 業種区分

日本取引所グループの公開している東証一部上場企業のリスト (図 3.4) には、17 業種区分や 33 業種区分といった業種区分が含まれている。それらの情報も特徴量として追加する。

	日付	コード	銘柄名	市場・商品区分	33業種コード	33業種区分	17業種コード	17業種区分	規模コード	規模区分	企業名
20	20200529	1332	日本水産	市場第一部（内国株）	50	水産・農林業	1	食品	4	TOPIX Mid400	日本水産(株)
21	20200529	1333	マルハニチロ	市場第一部（内国株）	50	水産・農林業	1	食品	4	TOPIX Mid400	マルハニチロ(株)
173	20200529	1605	国際石油開発帝石	市場第一部（内国株）	1050	鉱業	2	エネルギー資源	2	TOPIX Large70	国際石油開発帝石(株)
242	20200529	1721	コムシスホールディングス	市場第一部（内国株）	2050	建設業	3	建設・資材	4	TOPIX Mid400	コムシスＨＤ(株)
272	20200529	1801	大成建設	市場第一部（内国株）	2050	建設業	3	建設・資材	4	TOPIX Mid400	大成建設(株)
...	...	...	...	...	...	...	...	...	...	...	...
3838	20200529	9613	エヌ・ティ・ティ・データ	市場第一部（内国株）	5250	情報・通信業	10	情報通信・サービスその他	4	TOPIX Mid400	(株)ＮＴＴデータ
3903	20200529	9735	セコム	市場第一部（内国株）	9050	サービス業	10	情報通信・サービスその他	2	TOPIX Large70	セコム(株)
3920	20200529	9766	コナミホールディングス	市場第一部（内国株）	5250	情報・通信業	10	情報通信・サービスその他	4	TOPIX Mid400	コナミＨＤ(株)
4020	20200529	9983	ファーストリテイリング	市場第一部（内国株）	6100	小売業	14	小売	2	TOPIX Large70	(株)ファーストリテイリング
4021	20200529	9984	ソフトバンクグループ	市場第一部（内国株）	5250	情報・通信業	10	情報通信・サービスその他	1	TOPIX Core30	ソフトバンクグループ(株)

225 rows × 11 columns

図 3.4: 日本取引所グループが公開している東証一部上場企業のデータ

3.2.2 株価

Stooq[12] の API を用いて日経 225 社の過去 5 年間の株価を取得した。2015 年の 10 月からのみ取得可能だったことから 2015 年 10 月、2016 年 10 月、2017 年 10 月、2018 年 10 月の第一金曜日の株価をデータとして追加した。図 3.5 はそのデータを示している。

株価は価格の変動が予測の数値を左右する要因になるため、直近一年、二年、三年の株価の伸び率を計算し、それも特徴量に加えた。また、2015 年から 2018 年度までに株価の伸びた年度数も加えた。

	id	stock_2015	stock_2016	stock_2017	stock_2018
0	1332	337.90	422.20	633.59	699.88
1	1333	1674.71	2665.90	3437.05	4055.33
2	1605	959.55	853.91	1088.10	1323.80
3	1721	1292.30	1662.38	2567.17	3105.74
4	1801	3533.56	3464.56	5513.13	4827.52
5	1802	916.01	897.42	1280.03	1019.65
6	1803	932.45	850.46	1207.31	971.89
7	1808	1172.15	848.40	1394.98	1348.63
8	1812	1137.85	1266.13	2076.96	1516.39

図 3.5: 取得した株価のデータ

以上の過程を踏まえて、生成した特徴量を以下の表にまとめる。以上の 694 の特徴量で

予測を行う。

特徴量の種類	数
損益計算表の項目	19
簡単な四則演算で生成したもの	74
Aggregation	73
事業に関する 5 項目の記述量	5
TF-IDF → SVD	417
印象分析によるもの	93
株価に関するもの	13
合計	694

第4章　学習

「売上高」の予測モデルと「2018 年度比較の 2019 年度の売上高の増加率」の予測モデルの学習過程について述べる。

4.1　学習モデル

予測の根拠の可視化が可能な決定木構造のアルゴリズムを 2 つ採用した。LightGBM と XGBoost である。双方のアルゴリズムは現在多くの予測コンペティションで用いられている手法で、適用できる Explainable AI のフレームワークも多く存在する。

4.1.1　XGBoost

XGBoost[13] とは決定木構造を使った勾配ブースティングのアルゴリズムである。決定木を複数作成し、それらをアンサンブルすることで予測値を計算する。有名な決定木構造のモデルの一つに Random Forest がある [14]。Random Forest では決定木モデルを並列化しアンサンブルを取る (バギングという) 一方で、XGBoost ではそれぞれの決定木を成長させる際に、前に生成した決定木を参考にその損失関数 (4.1) を小さくする葉を追加する (ブースティング) という。

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{4.1}$$

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{4.2}$$

l は「予測値 $\hat{y}_i$ と正解値 y_i の差」を表し、Ω は決定木の成長に制限を設ける関数である。第一項は葉の数を制限し、第二項は葉の重みを制限する。4.1.2 の LightGBM との一番大きな違いは、葉の成長方法である。XGBoost では図 4.1 のように深さ単位で葉を増やす一方で、LightGBM では図 4.1 のように葉の階層を増やすことで葉を増やす。

4.1.2　LightGBM

LightGBM[15] も、決定木構造を使った勾配ブースティングのフレームワークである。このアルゴリズムの特徴は大きく 2 つある。

一つが、決定木の分岐生成の際にヒストグラムでビンニングした集合を分類していることである。先ほど紹介した XGBoost では、決定木の分岐生成は分岐になりうる全ての連

続的な数値を見ていた。しかし、LightGBM ではサンプルを一つ一つで扱うのではなく、値の近いサンプルで括って処理することで、高速化を実現した。

もう一つが、深さ単位でなく葉単位の分岐で精度を向上した点である。勾配ブースティングの学習過程において、決定木の成長は「Level-Wise」と「Leaf-Wise」が存在する。XGBoost(4.1.1) では深さ単位で葉を増やしていたが（図 4.1)、LightBGM は後者のアルゴリズムを採用し目的関数を減少させる分割を優先して増やす（図 4.2)。

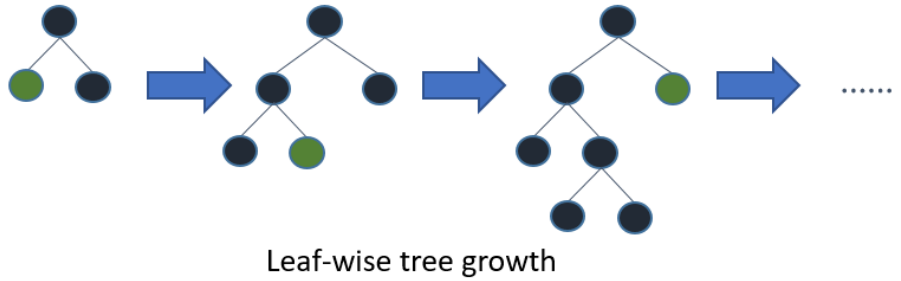

図 4.1: 深さ単位の木の成長

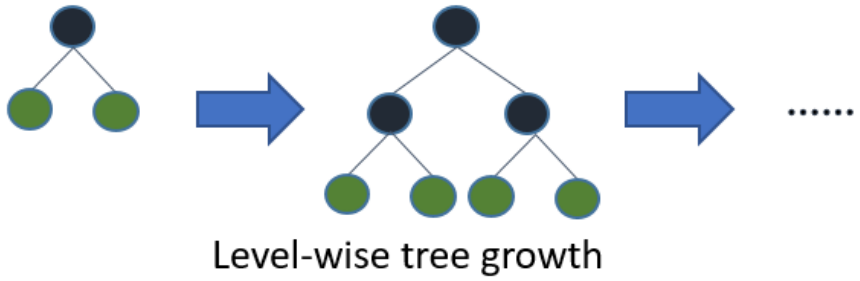

図 4.2: 葉単位の木の成長

4.2 学習方法

4.2.1 評価関数

評価関数には、二つの予測モデルで別の関数を用いた。

初めに、売上率の評価関数には最小二乗誤差 (Root Mean Squared Error 以下、RMSE) を用いた。RMSE を用いることで正解ラベルと予測値が離れるほど 2 乗のオーダーで損失関数は大きくなる。RMSE は (4.3) で与えられる。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{true})^2} \tag{4.3}$$

続いて、売上高の評価関数には対数平均二乗誤差 (Root Mean Squared Log Error 以下、RMSLE) を用いた。RMSLE は式 (4.4) で与えられる。

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(y_{pred}+1) - log(y_{true}+1))^2} \tag{4.4}$$

RMSLE を評価関数とするにあたり、225 社の売上高の分布が図 4.3 の左図のように偏っていたことから、対数変換を行い図 4.3 の右図のような分布となるようにした。

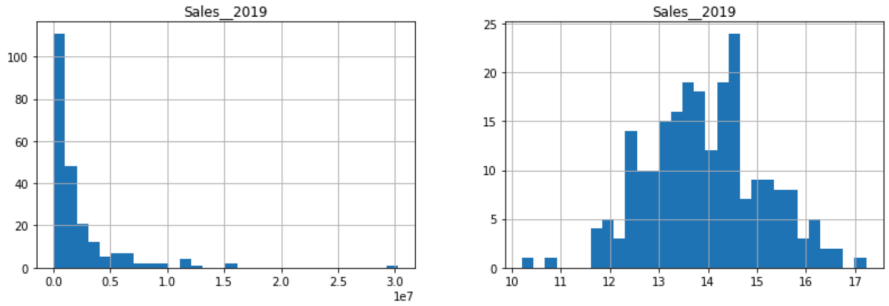

図 4.3: 変換前 (左) と変換後 (右) の売上高の分布

変換後に 4.3 を用いることで結果として RMSLE として計算をすることとした。

4.2.2 交差検証

用いるデータのサンプル数は 225 と少ないため、なるべく多くの分割数で交差検証が行えるよう、学習データとバリデーションデータの比率を 14：1」となるように分割した。また、サンプルが少ないことから学習データとバリデーションデータの分割方法によって結果が大きく変わる可能性がある。それを防ぐために Seed 値を 15 回変更し結果の平均をとる。

4.2.3 ハイパーパラメータ最適化

学習のパラメータを決定する際、パラメータチューニングには Preferred Networks 社が開発した Optuna を使用した [16]。Optuna とはハイパーパラメータの最適化を自動で行うフレームワークである。指定した範囲内でハイパーパラーメータを探すが、それまでに完了している試行の結果を元に Tree-structured Parzen Estimator と呼ばれるベイズ最適化アルゴリズムを用いて行っている。

4.2.4 学習結果

以上の過程を踏まえて学習を行い、以下の表にあるような精度の学習済みモデルを作成した。

予測対象	学習モデル	精度
売上高	LightGBM	0.0534
	XGBoost	0.0537
売上率	LightGBM	0.0144
	XGBoost	0.0143

第5章　Explainable AI

本章では、本研究で用いる Explainable AI の解析手法と結果について論じる。本研究では、特徴量の重要度を評価する Feature Importance、Permutation Importance に加え、特徴量を一律で変化させた場合の予測値の変化を計測する PDP、そして各サンプルに対し予測に寄与した特徴量を可視化する SHAP というフレームワークを用いた。各節でその詳細を述べる。

5.1　Explainable AI を使う利点

結果に入る前に、これらのフレームワークを使う利点について簡単にまとめる。大きく 3 つに分けられる。

5.1.1　特徴量選択の基準として

Explainable AI を用いる最大の利点が特徴量選択の基準にできる点である。いわゆるビッグデータを用いた学習では、より高精度でかつ計算時間が短い学習済みモデルを生成するために、データ容量の削減が多々行われる。また、計算資源のメモリが少量の場合においても、同様にデータを削る処置が取られることがある。そのような状況において、特徴量の重要度を定量的に計測することは有効な特徴量を選別し高精度な学習済みモデルを作成する上で非常に役に立つ。本研究では予測に有効な特徴量を評価することを目的としているため、特徴量を削ってより高精度な学習済みモデルを作成することはしてない。しかし、実社会において予測の精度が重要な場合においてはデータの削減は大変有効である。

5.1.2　アンサンブル学習に用いるモデル選択の基準として

機械学習において交差検証を用いる場合やアンサンブル学習で複数の学習済みモデルを作成する場合に、モデル間の学習度を Explainable AI を使って比較することで、過学習を極力抑えたアンサンブル学習が可能になる。予測根拠の似たモデルを複数利用してしまうと、よりバイアスがかかった結果に陥りやすく、過学習にもつながる。なるべく、異なるアプローチでかつ精度の良いモデル同士を組み合わせることで、より高精度のアンサンブル学習が実現できる。

5.1.3　学習過程におけるミスの防止

Explainable AI では学習済みモデルの学習度合いについてより詳細に把握することができる。それによって、学習が想定通りに行えていなかった場合に、そのミスに気付ける利点がある。

5.2　本章における注意点

本章で記述した特徴量の書き方に関していくつか注意する点がある。初めに、図の関係上、いくつかの名称を以下のように省略した。

- 営業活動によるキャッシュ・フロー → 営業キャッシュ
- 投資活動によるキャッシュ・フロー → 投資キャッシュ
- 現金及び現金同等物の期末残高 → 期末残高

また、以下の表には「項目 4 の TF-IDF の SVD34 番目」などという特徴量名があるが、これは経営に関する項目 4 を TF-IDF をし、さらに SVD で次元削減したベクトルの 34 番目の要素を示す。項目 1-5 は、以下の内容を表す。

項目 1 経営方針，経営環境及び対処すべき課題等
項目 2 事業等のリスク
項目 3 経営者による財政状態，経営成績及びキャッシュ・フローの状況の分析
項目 4 経営上の重要な契約等
項目 5 研究開発活動

加えて、過去 4 年間などという特徴量名で注意することは、過去 4 年間は「2015 年度」から 2018 年度までの 4 年間の平均や分散や最大値を表しており、2014 年度から 2018 年度までの値を表しているわけではない。過去 3 年間は「2016 年度」から 2018 年度までの 3 年間を表す。

5.3　Feature Importance

本節では、Feature Importance を使ってどの特徴量が予測精度に寄与をしたのかを学習済みモデルを用いて定量的に評価する。LightGBM のような決定木構造のアルゴリズムでは、分岐の際にどの特徴量を用いたかが予測精度に寄与した特徴量を計測する上で役に立つ。単に分岐に用いられた特徴量をカウントすることでも特徴量の重要性を測ることはできるが、決定木の分岐の前半で用いられた特徴量と最後に用いられたものでは重要度が異なるので、それを反映した重要度を計測する。学習のために生成した全ての決定木で和を取ったものをソートし、特徴量を評価する。

Feature Importance による重要度の高い特徴量上位 15 位は以下の通りである。

表 5.1: 売上高の Feature Importance (LightGBM)

特徴量名	Feature Importance 値
2018 年度の売上高	61359
売上高の最大値 (過去 4 年間)	38452
売上高の最小値 (過去 4 年間)	13661
売上高の平均値 (過去 4 年間)	10650
総資産額の増加率 (過去 2 年間)	6283
2018 年度の総資産額	5899
2018 年度の 売上高/従業員数	5186
2018 年度の営業キャッシュ	4531
2018 年度の利益	4517
売上高の分散 (過去 4 年間)	4337
従業員数の分散 (過去 4 年間)	4263
項目 4 の TF-IDF の SVD34 番目	3888
利益の平均 (過去 4 年間)	3516
2018 年度の 売上高/利益	2939
項目 1 の TF-IDF の SVD9 番目	2516

表 5.2: 売上高の Feature Importance (XGBoost)

特徴量名	Feature Importance 値
売上高の最大値 (過去 4 年間)	92116
2018 年度の売上高	73347
売上高の最小値 (過去 4 年間)	30722
2018 年度の総資産額	28177
売上高の平均値 (過去 4 年間)	21210
2018 年度の営業キャッシュ	15188
従業員数の分散 (過去 4 年間)	14773
2018 年度の 売上高/従業員数	13088
利益の平均 (過去 4 年間)	12719
売上高の分散 (過去 4 年間)	11537
2018 年度の利益	10507
2018 年度 売上高/利益	9780
利益/売上高 の分散 (過去 4 年間)	8873
営業キャッシュ/総資産額 の分散 (過去 4 年間)	8418
過去 2 年間の総資産額の増加率	8334

売上高の予測では、過去の売上高に関する特徴量が際立って予測に寄与していることが分かる。次いで、純資産と負債を含めた総資産額やその伸び率が双方のモデルで上位に入った。また本業による資金の動きを示す、営業活動によるキャッシュ・フローや従業員数など、会社の規模を端的に示す特徴量が散見される。

表 5.3: 売上高伸び率の Feature Importance (LightGBM)

特徴量名	Feature Importance 値
項目 1 の TF-IDF の SVD10 番目	5603
項目 1 の TF-IDF の SVD9 番目	4755
期末残高の増加率 (過去 3 年間)	4347
経営方針，経営環境及び対処すべき課題等の文字数	4153
投資キャッシュ/利益 の最大値 (過去 4 年間)	4107
売上高/財務キャッシュ (2018 年度)	3552
項目 1 の TF-IDF の SVD1 番目	3304
株価の伸び率 (過去 4 年間)	3295
株価の伸び率 (過去 3 年間)	3174
総資産額の増加率 (過去 2 年間)	3048
財務キャッシュ/総資産額 の分散 (過去 4 年間)	3047
項目 3 の TF-IDF の SVD118 番目	2759
項目 1 の TF-IDF の SVD5 番目	2616
営業キャッシュ/総資産額の増加率 (過去 3 年間)	2603
項目 5 の TF-IDF の SVD83 番目	2516

表 5.4: 売上高伸び率の Feature Importance (XGBoost)

特徴量名	Feature Importance 値
期末残高の増加率 (過去 3 年間)	13211
売上高/総資産額 の増加率 (過去 2 年間)	13137
項目 1 の TF-IDF の SVD9 番目	11801
項目 5 の TF-IDF の SVD83 番目	10396
項目 3 の TF-IDF の SVD118 番目	10077
利益率の合計 (過去 4 年間)	7094
総資産額の増加率 (過去 3 年間)	6731
営業キャッシュ/総資産額 の分散 (過去 4 年間)	6709
株価の伸び率 (過去 3 年間)	6421
項目 3 の TF-IDF の SVD94 番目	6343
株価の伸び率 (過去 4 年間)	6337
項目 1 の TF-IDF の SVD10 番目	6209
項目 1 の TF-IDF の SVD3 番目	5906
項目 3 の TF-IDF の SVD101 番目	5503
項目 1 の TF-IDF の SVD48 番目	5177
2018 年度の利益	5174

続く、売上高伸び率の予測では、売上高予測の場合と比較し、LightGBM と XGBoost 双方が類似しない結果となった。双方のアルゴリズムによる結果を見ると、経営状況に関する記述から作成された特徴量が多く予測に貢献している。特に項目 1 や項目 3 から作成された特徴量が目立つ。また、売上高予測の結果とは異なり、利益率や増加率など割合に関する特徴量が多く予測に関わっていることがわかる。経営に関する項目を用いていない特徴量で最も目立つものが「現金及び現金同等物の期末残高」の伸び率だ。現金及び現金同等物の期末残高は、期首残高から期の営業/投資/財務のキャッシュフローが行われた後の残高と等しい。2016 年度の期末残高と比較し 2018 年度にどれほど残高が増減したかが重要な特徴量であると理解できる。また、双方のモデルにランクインしている特徴量が株価の伸び率である。過去 3 年間の伸びと、過去 4 年間の伸びに相関があることは否めないが、もし仮に過去 3 年間の株価の伸び率を削除して学習を実行していればより上位になった可能性がある。

5.4 Permutation Importance

Permutation Importance も「5.3 Fearure Importance」と同様、予測に寄与した特徴量を評価するためのアルゴリズムである。Permutation Importance では、学習済みモデルを利用し一つ一つの特徴量をシャッフルした場合としていない場合の精度を比較する。その差が大きいほど特徴量は重要であり、小さいほど重要ではないと判定される。

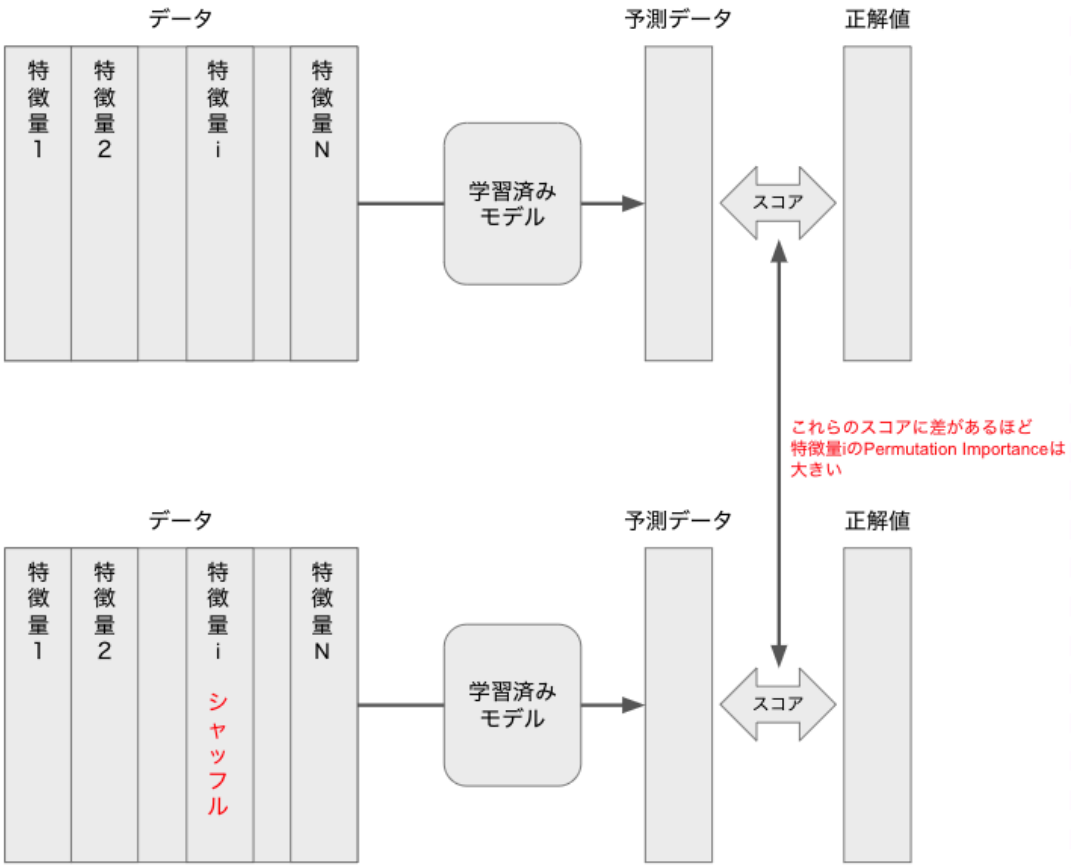

図 5.1: Permutation Importance の概略

以下の表 5.7 と表 5.8 は Permutation Importance の上位 15 位を表している。基本的に上位の特徴量は Feature Importance と変わらないが、Permutation Importance の場合は TF-IDF 以外の特徴量が多く確認できる。

表 5.5: 売上高の Permutation Importance (LightGBM)

特徴量名	Permutation Importance 値
2018 年度の売上高	187.69358
売上高の最大値 (過去 4 年間)	77.57066
売上高の平均値 (過去 4 年間)	10.95107
売上高の最小値 (過去 4 年間)	8.72992
2018 年度の営業キャッシュ	1.44423
利益の平均 (過去 4 年間)	1.17851
2018 年度の総資産額	1.12767
従業員数の分散 (過去 4 年間)	0.96271
項目 5 の TF-IDF の SVD93 番目	0.91597
売上高の分散 (過去 4 年間)	0.80863
項目 4 の TF-IDF の SVD32 番目	0.79260
2018 年度の利益	0.75416
2018 年度の 売上高/利益	0.72060
2018 年度の従業員数	0.59248
項目 4 の TF-IDF の SVD34 番目	0.40287

表 5.6: 売上高の Permutation Importance (XGBoost)

特徴量名	Permutation Importance 値
2018 年度の売上高	123.07533
売上高の最大値 (過去 4 年間)	108.87200
売上高の平均値 (過去 4 年間)	16.36349
売上高の最小値 (過去 4 年間)	9.05173
2018 年度の総資産額	2.65892
2018 年度の営業キャッシュ	1.95406
従業員数の分散 (過去 4 年間)	1.29746
利益の平均 (過去 4 年間)	1.10202
2018 年度の 売上高/利益	1.05807
営業キャッシュ/総資産額 の分散 (過去 4 年間)	0.70717
売上高の分散 (過去 4 年間)	0.68725
2018 年度の利益	0.63119
投資キャッシュ/総資産額 の平均値 (過去 4 年間)	0.57399
2018 年度の期末残高	0.52770
財務キャッシュ	0.52365

売上高予測モデルに関しては、全体として Feature Importance の結果と似通っている。特に上位５位の特徴量は順位は異なるものの、ほとんど同じ特徴量が上位になっている。全く異なった２つのアルゴリズムで類似した結果となったことから、特徴量の重要度の妥当性がわかる。

表 5.7: 売上高伸び率の Permutation Importance (LightGBM)

特徴量名	Permutation Importance 値
株価の伸び率 (過去 3 年間)	0.75785
株価の伸び率 (過去 4 年間)	0.74873
期末残高の増加率 (過去 4 年間)	0.51149
投資キャッシュ/利益 の最小値 (過去 4 年間)	0.27922
項目 1 の TF-IDF の SVD9 番目	0.22990
2018 年度の利益	0.19918
利益の分散 (過去 4 年間)	0.17136
営業キャッシュ/利益 の最大値 (過去 4 年間)	0.16235
項目 2 の TF-IDF の SVD26 番目	0.16080
項目 1 の TF-IDF の SVD10 番目	0.14243
項目 5 の TF-IDF の SVD74 番目	0.12992
投資キャッシュ/利益 の最大値 (過去 4 年間)	0.11429
項目 5 の TF-IDF の SVD83 番目	0.09910
総資産額の増加率 (過去 2 年間)	0.09421
項目 5 の TF-IDF の SVD64 番目	0.09168

表 5.8: 売上高伸び率の Permutation Importance (XGBoost)

特徴量名	Permutation Importance 値
期末残高の増加率 (過去 3 年間)	1.04622
株価の伸び率 (過去 3 年間)	0.54534
項目 1 の TF-IDF の SVD9 番目	0.53148
株価の伸び率 (過去 4 年間)	0.46065
利益の分散 (過去 4 年間)	0.30368
投資キャッシュ/利益 の最小値 (過去 4 年間)	0.28611
2018 年度の利益	0.28078
項目 3 の TF-IDF の SVD124 番目	0.26477
項目 3 の TF-IDF の SVD118 番目	0.26042
項目 5 の TF-IDF の SVD74 番目	0.25448
売上高/総資産額 の増加率 (過去 2 年間)	0.2224
項目 3 の TF-IDF の SVD64 番目	0.22178
項目 1 の TF-IDF の SVD3 番目	0.21857
項目 5 の TF-IDF の SVD83 番目	0.21638
項目 5 の TF-IDF の SVD64 番目	0.19735

Feature Importance の結果と異なる点が、特徴量の重要度の値のばらつきである。Feature Importance では上位の特徴量が拮抗しているのにも関わらず、Permutation Importance では上位の重要度の値が他を大きく離している。

5.5 PDP

5.5.1 概要

PDP は Partial Dependence Plot とも呼ばれ、特徴量と目的変数の関係を可視化するパッケージである [17]。PDP では調べたい特徴量の数値を一律で変化させていくことで予測結果の平均値がどのように変化するかを計算している。

数学的な PDP の定義は式 5.1 のようになる。

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C) \tag{5.1}$$

$\hat{f}$ は学習済みモデル、x_S は調べたい特徴量で、X_C は学習に用いた x_S 以外の特徴量である。x_S はスカラー値で X_C はサンプル数 N 行、特徴量-1 列の行列である。これを変形した式 (5.2) で推定する。

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{5.2}$$

$x_C^{(i)}$ は X_C の i 番目のサンプルである。調べたい特徴量の値が全て同じ x_S だった場合、それと i 番目の x_C の組み合わせから予測される値の平均値を式 (5.2) は意味する。

5.5.2 結果

PDP は一つの学習済みモデルに対し適用できるツールである。したがって、Seed 値の数×バリデーションの Fold 数の全ての学習済みモデルの中から、いくつかの PDP の結果を並べることで特徴量を評価する。なお、PDP においては LightGBM アルゴリズムで生成した学習済みモデルにのみ PDP を適用している。

まず例として、図 5.2 はシード値が 0 でバリデーション Fold 値が 1 の場合の学習済みモデルを用いた PDP である。図 5.2 は 2019 年度の売上高の予測に最も貢献した特徴量である「2018 年度の売上高」を全てのサンプルで一律に変化させた場合の 2019 年度の売上高の期待値の推移を可視化している。はじめに、縦軸が 2018 年度の売上高が全てのサンプルで 0 だった場合の値を基準とする、2019 年度の売上高の期待値の変化量を示している。なお、2019 年度の売上高は自然対数をとった値の変化量となる。続いて、横軸は一律に変化させた特徴量のその一律の値を示している。横軸の単位は 10 兆円 (1e7 百万円) である。この結果からわかることは、2018 年度の売上高が 2 兆 5000 億円までは期待値が上昇するが、それを超えたあたりから伸び幅は落ち着く。

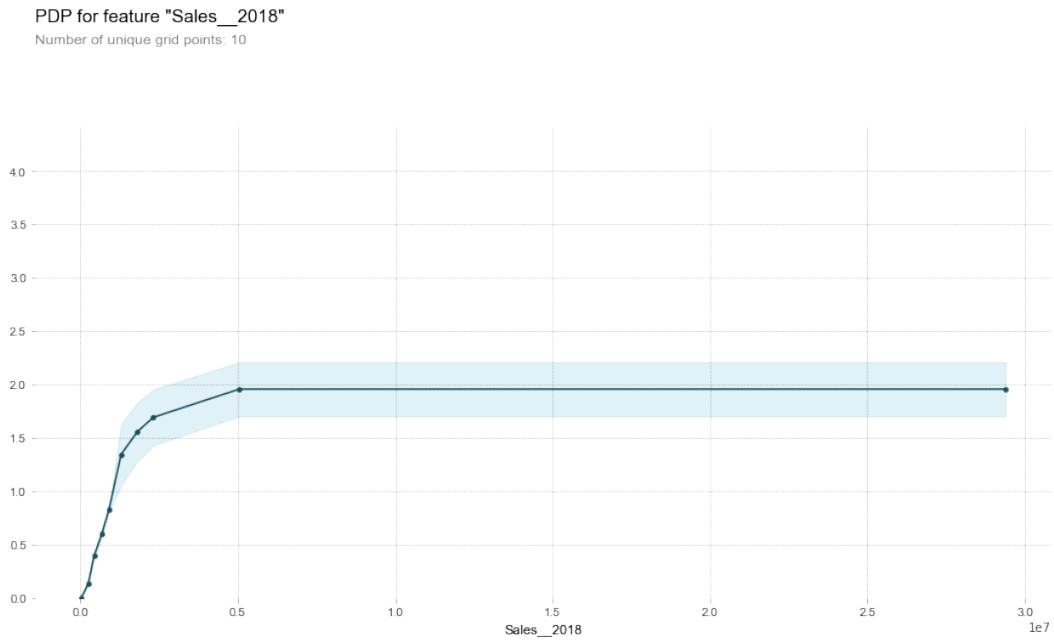

図 5.2: 2018 年度の売上高と 2019 年度の売上の予測結果の関係

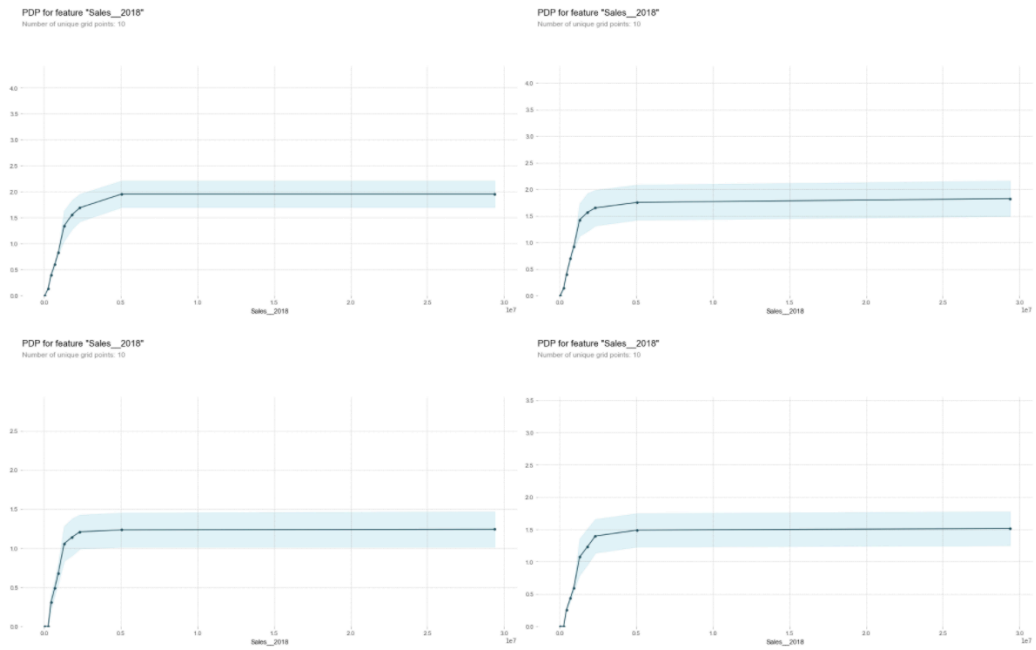

図 5.3: 2018 年度の売上高と 2019 年度の売上の予測結果の関係

同じシード値で別のバリデーション Fold の結果が図 5.3 である。学習の際、15 のバリデーション Fold で学習モデルを作成し検証を行ったが、そのうちの 4 つの学習モデルに対して PDP を計算し表にまとめた。Fold 値によって期待値の振れ幅に多少違いがあるが、同じような曲線を描くことがわかる。

はじめに、図 5.4 は 2018 年度の 2016 年度に対する株価の比率と 2019 年度の売上比率の推移である。興味深いことに全てのシード値において株価の比率が 1.5 を超えたあたりから予測値の期待値が跳ね上がることが見て取れる。

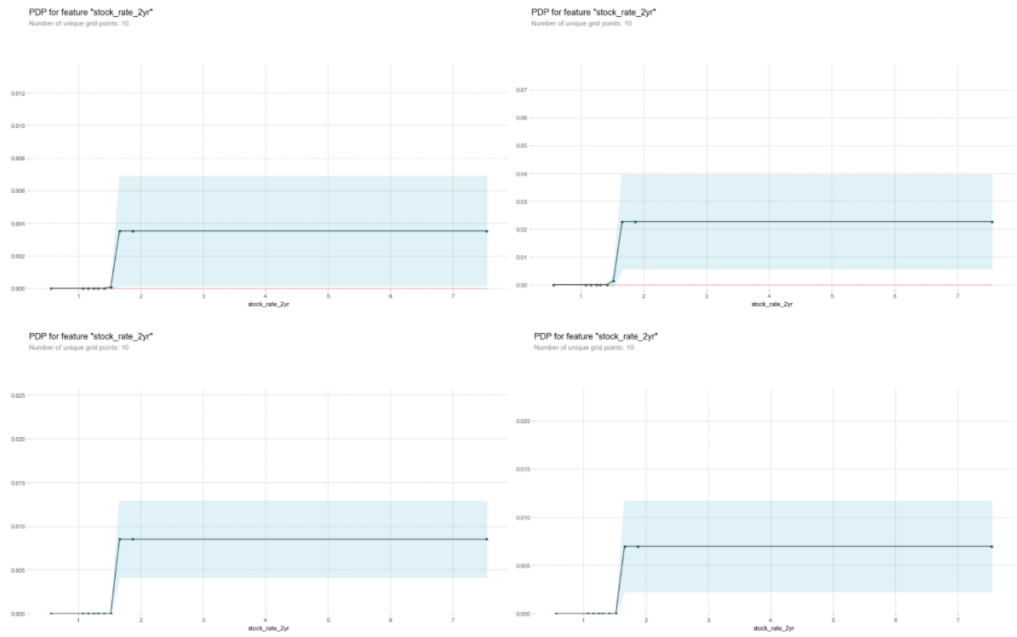

図 5.4: 「2018 年度の 2016 年度に対する株価の比率」と「2019 年度の売上比率」の予測結果の関係

最後に図 5.5 では、「経営方針，経営環境及び対処すべき課題等」に関する記述量と 2019 年度の売上比率を表した図である。Fold の値によって曲線は若干異なるが、全体として経営方針に関する記述量は 0 字から 3000-4000 字に増加するに従い予測値の期待値は大きく下がるが、10000 字を超えたあたりから期待値が若干増加する傾向があることが見て取れる。

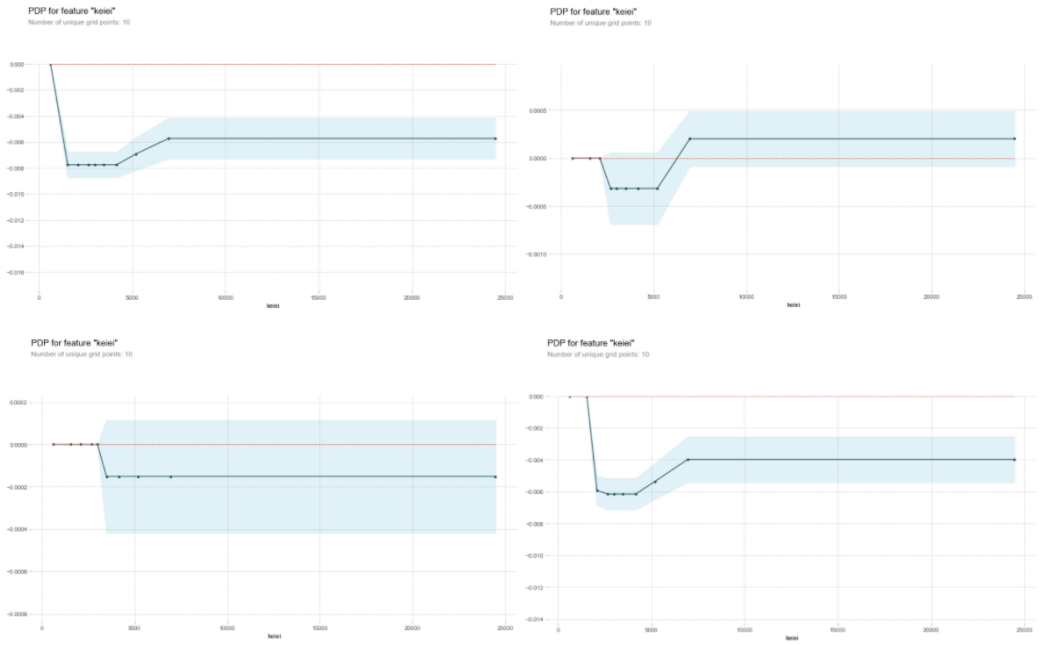

図 5.5: 項目 1 に関する記述量と 2019 年度の売上比率の期待値

以上のように、PDP を用いて各特徴量を一律で変化させることで予測値の期待値がどのように変動するのか例を用いて紹介した。経営方針に関する記述量と売上高伸び率の関係で示したように、一見予測と関係のない特徴量でも興味深い振る舞いをすることがあった。今回はサンプル数が日経 225 社のみと少ないデータで分析を行ったので、この振る舞いが一般的であるということは難しいが、データを増やし、さらにサンプルのバイアスを減らすことで、より一般的な興味深い知見が得られる可能性がある。

5.6 SHAP

5.6.1 概要

SHAP(SHapley Additive exPlanations) とは、各サンプルの特徴量がどの程度予測に貢献したかを可視化するアルゴリズムである。[18]

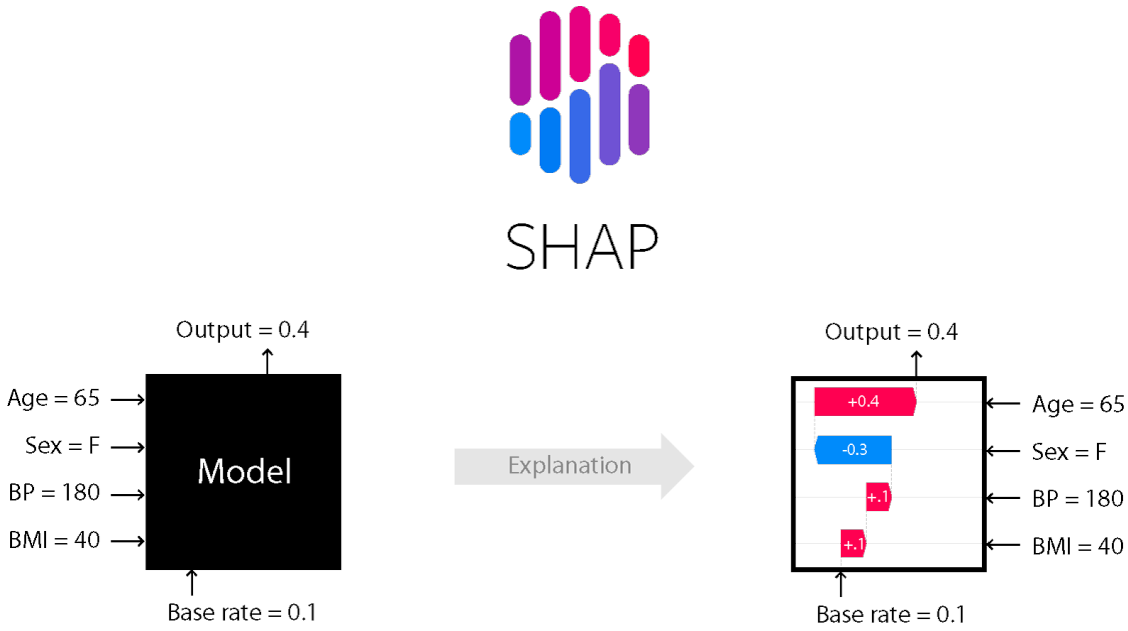

図 5.6: SHAP の概略図

結果を計算するにあたり、計算コストを削減するために、SHAP では式 (5.3) を用いて近似をする。

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5.3}$$

ここで $\phi_j (\in \mathbb{R})$ は j 番目の特徴量の予測値への寄与度を表し、$z' (\in \{0,1\}^M)$ は特徴量ベクトル x を下の式のように簡略化したものである。また M は z' のサイズである。

$$z'_j = \begin{cases} 1 & (x_j > 0) \\ 0 & (x_j = 0) \end{cases}$$

上式の ϕ の計算には、ゲーム理論の中でも協力ゲームと呼ばれる、複数人のチームで利得を得た場合の各メンバーの寄与度を算出するアルゴリズムを用いている。これを SHAP に応用する場合、チームのメンバーを一つひとつの特徴量、チームで得た利得を予測値と置き換える。協力ゲーム理論におけるメンバーの寄与度の組み合わせは Shapley Value として知られ、式 (5.6) で表現される。

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} Prob(S)(v(S \cup \{i\}) - v(S)) \tag{5.4}$$

N はメンバー全体の集合で、S はメンバー i を除いた N の部分集合である。$v(S)$ はメンバーの部分集合 S の結果への寄与分で、$v(S \cup \{i\}) - v(S)$ は「メンバー i が S と共に寄

与した分」と「S のみが寄与した分」の差を表している。$Prob(S)$ は N の中で S が組まれる確率であり、チーム全員のメンバー数 n を使って式 (5.7) で表される。

$$Prob(S) = \frac{|S|!\,(n - |S| - 1)!}{n!} \tag{5.5}$$

これを応用し、メンバーを特徴量 x、チームの寄与 $v(S)$ を学習モデルとデータ $f_S(x_S)$ の二つに置き換えると、SHAP 値 ϕ_i は

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} Prob(S)(f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \tag{5.6}$$

$$Prob(S) = \frac{|S|!\,(n - |S| - 1)!}{n!} \tag{5.7}$$

と書き換えられる。

5.6.2 結果

SHAP では、異なる学習済みモデルによる結果を加算することができる。本研究の SHAP では LightGBM モデルで学習した Fold:15、Seed:15 パターンの計 225 の学習済みモデルに対して平均した結果を表示する。

売上高予測の SHAP

結論として、694 の特徴量のうち予測に大きく寄与した特徴量は過去 4 年間の売上高による特徴量となった。その他、キャッシュフローや利益など金額に関する特徴量が比較的高く寄与していた。また、上位 20 特徴量には従業員数も含まれている。以下では、学習済みモデルから計算された SHAP の結果の例をいくつか紹介する。

図 5.7 は (株) ディー・エヌ・エーの SHAP の結果である。図の中央部にある base value は 225 社の正解値 (2019 年度の売上高の対数) の平均値である。$f(x)$ はこのサンプルの予測値を示しており、その値は 12.03($log(X) = 12.03$ より X =167,711 百万円、なお正解値は 124,116 百万円) である。その平均値からどの特徴量が予測値を押し上げたのかまた押し下げたのかが見て取れる。赤の特徴量はその予測値に正の向き寄与した特徴量であり、一方青の特徴量はその予測値を負の向きに寄与した特徴量である。(株) ディー・エヌ・エーの場合、予測値の負の方向に最も寄与した特徴量は「2018 年度の売上高 (Sales_2018)」である。また 1.394e+5 という数はそのサンプルの特徴量の値を表す。この結果はあくまで、2019 年度の日経 225 社の売上の平均値よりも下であると予測した根拠を示しており、(株) ディー・エヌ・エーの売上高が前年度と比較し下落した要因を可視化した図ではないことには注意したい。

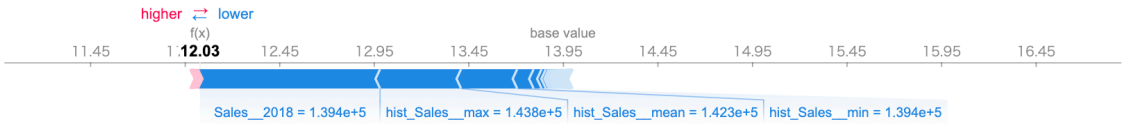

図 5.7: (株) ディー・エヌ・エーの売上高予測の SHAP の結果

続いて、図 5.8 はトヨタ自動車 (株) の SHAP の結果である。トヨタ自動車 (株) は日経 225 社の中でも最も売上高の高い企業であり、様々な特徴量が正の方向に寄与していることがわかる。(株) ディー・エヌ・エーと同様に、もっとも正に寄与した特徴量は「2018 年度の売上高 (Sales__2018)」であり、続いて「過去 4 年間の売上高の最大値 (hist_Sales__max)」となった。

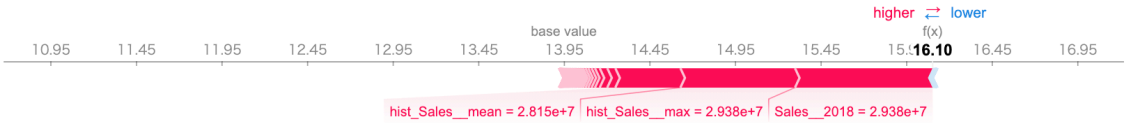

図 5.8: トヨタ自動車 (株) の売上高予測の SHAP の結果

図 5.9 は日経 225 社の全てのサンプルの SHAP 値を縦に並べて予測値で並べ替えた表である。図 5.9 の最も左が図 5.8 と同じ値になっている。694 つある特徴量でも 2018 年度の売上高が際立って寄与していることがわかる。このように SHAP では PDP や Permutation Importance では不可能だった、サンプルごとの各特徴量の寄与度を可視化することができる。

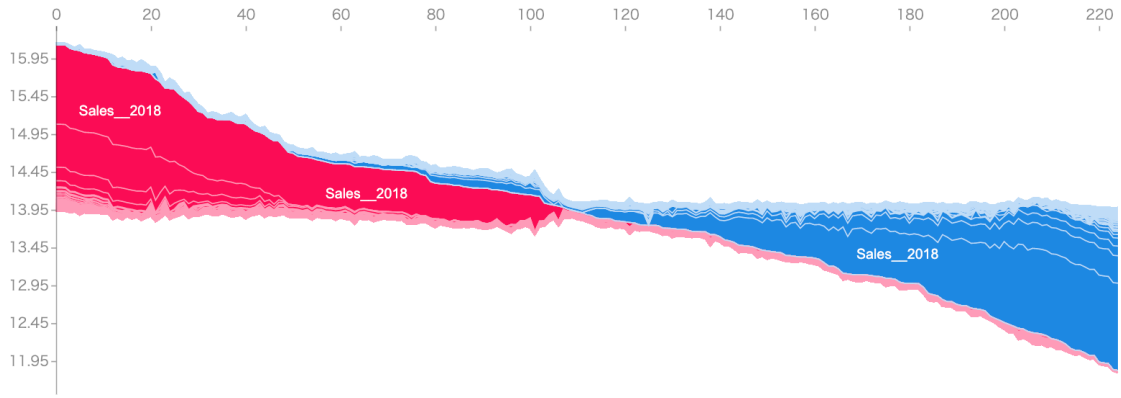

図 5.9: 日経 225 社全ての売上高予測の SHAP

特徴量ごとの寄与度を加算した図が、図 5.10 である。本項の冒頭で述べたように、売上高に関する 4 つの特徴量（2018 年度の売上高・過去 4 年間の売上高の最大値・過去 4 年間の売上高の平均値・過去 4 年間の売上高の最小値）が予測に大きく寄与していることがわかる。

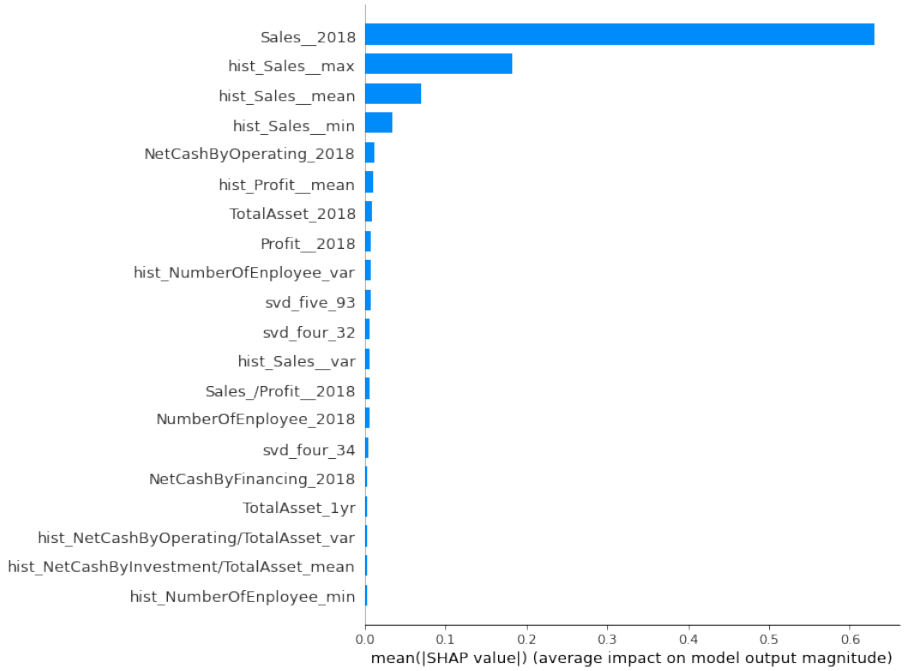

図 5.10: 694 特徴量のうち特に売上高予測に寄与度の高い 20 の特徴量

売上率予測の SHAP

本節では売上率予測の学習済みモデルから得られた SHAP の結果を示す。結論として、株価の伸び率が最も予測に寄与しており、さらに経営理念などの項目から生成された特徴量の寄与度が比較的高いことがわかった。さらに、キャッシュフロー値を利益で割った値から生成された特徴量も、最も寄与した 20 の特徴量に 3 つ含まれている。前項と同様にサンプルごとの例を紹介しながら全体の結果を紹介する。

まず、図 5.11 の base value(=1.042) は日経 225 社の前年度からの 2019 年度の売上率の平均値を表しており、$f(x) = 1.04$ は (株) ディー・エヌ・エーの売上率の予測値を表している。(株) ディー・エヌ・エーの売上率の予測の際に最も負に寄与した特徴量は、2018 年度までの株価の伸び率（2016 年度比較）であり、最も正に貢献した特徴量は項目 1 に関する記述を TF-IDF し SVD で次元削減したベクトルの 10 番目の要素である。

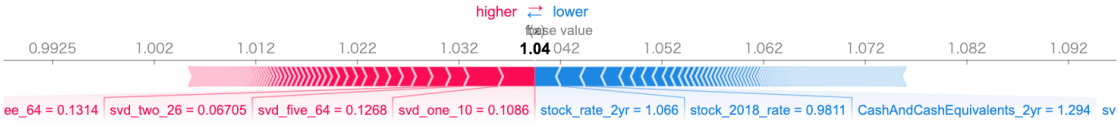

図 5.11: (株) ディー・エヌ・エーの売上率の SHAP の結果

図 5.12 は 225 社の中で最も予測値が高かった伊藤忠商事 (株) の SHAP の結果である。最も正に寄与した特徴量は現金及び現金同等物の期末残高の伸び率 (2016 年度比較) となっ

た。Feature Importance や Permutation Imporance と同様、売上高予測で少ない特徴量が大きく寄与していたことと反して、売上率予測では様々な特徴量が予測に寄与していることが見て取れる。

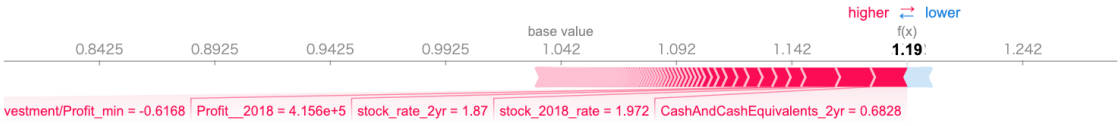

図 5.12: 伊藤忠商事 (株) の売上率の SHAP の結果

図 5.13 は日経 225 社の全てのサンプルの SHAP 値を縦に並べて予測値で並べ替えたものである。その下に続く、図 5.14 は日経 225 社全ての売上率の正解値を並べたものであり、図 5.13 と比較して正解値の分布を学習済みモデルがうまく捉えていることがわかる。

s

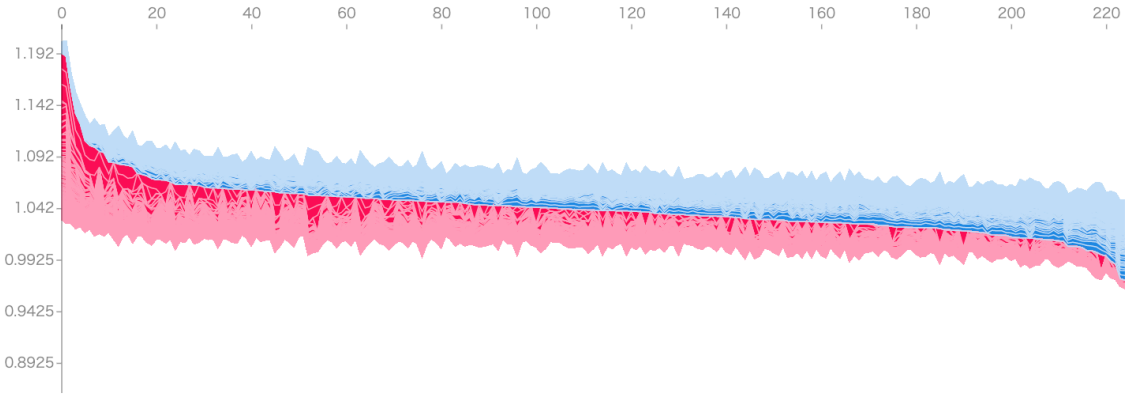

図 5.13: 日経 225 社全ての売上率の SHAP を予測値で並べた図

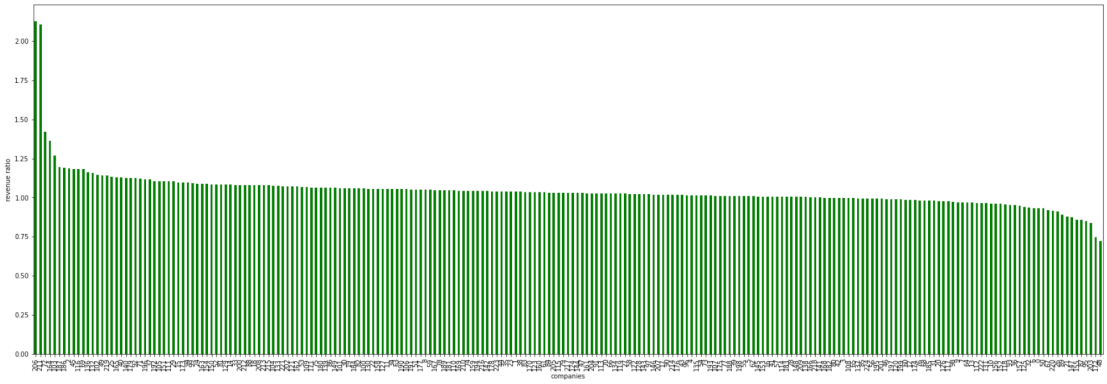

図 5.14: 日経 225 社全ての売上率の正解値を並べた図

図 5.15 は売上率予測に最も寄与した 20 の特徴量である。上位 3 特徴量は、表 5.7:売上高伸び率の Permutation Importance (LightGBM) の特徴量と全く同じ結果となった。

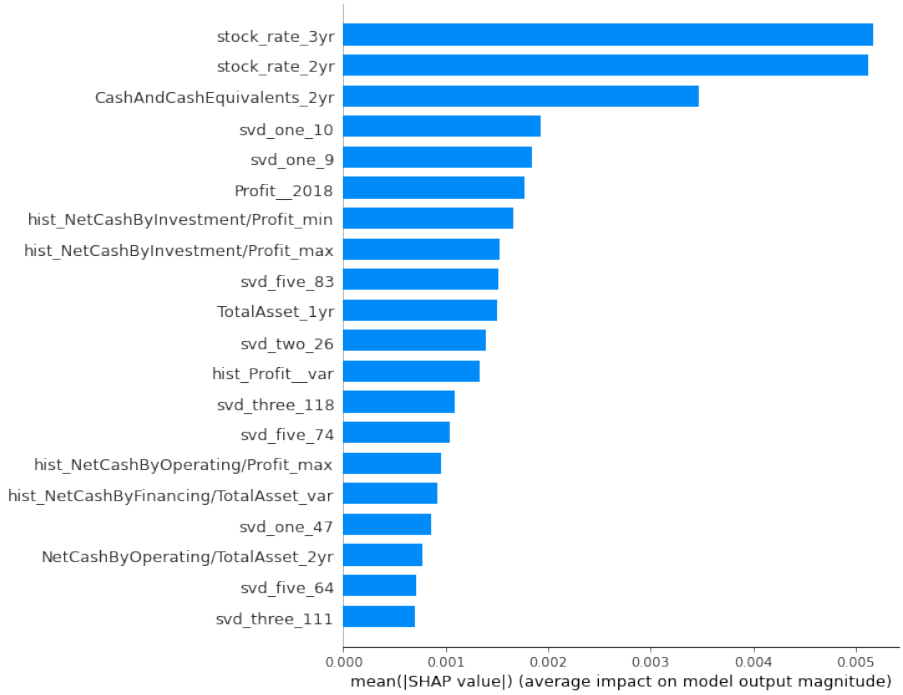

図 5.15: 694 特徴量のうち売上率予測に特に寄与度の高い 20 の特徴量

また、SHAP を用いることで各特徴量の中での値の大小と、SHAP 値がどう関係しているかを可視化することができる。図 5.16 がそれを表している。例えば、過去 3 年間の株価の伸び率は、伸び率が高い一部のサンプルで SHAP 値が大きい値を取っており、低い値の SHAP は 0.003 付近で似たような値をしていることがわかる。3 位の過去 2 年間期末残高の伸び率と SHAP の関係は一部の低い伸び率のサンプルにおいて高い SHAP 値を取っていることがわかる。また全体として SHAP 値の範囲が正に偏っている。

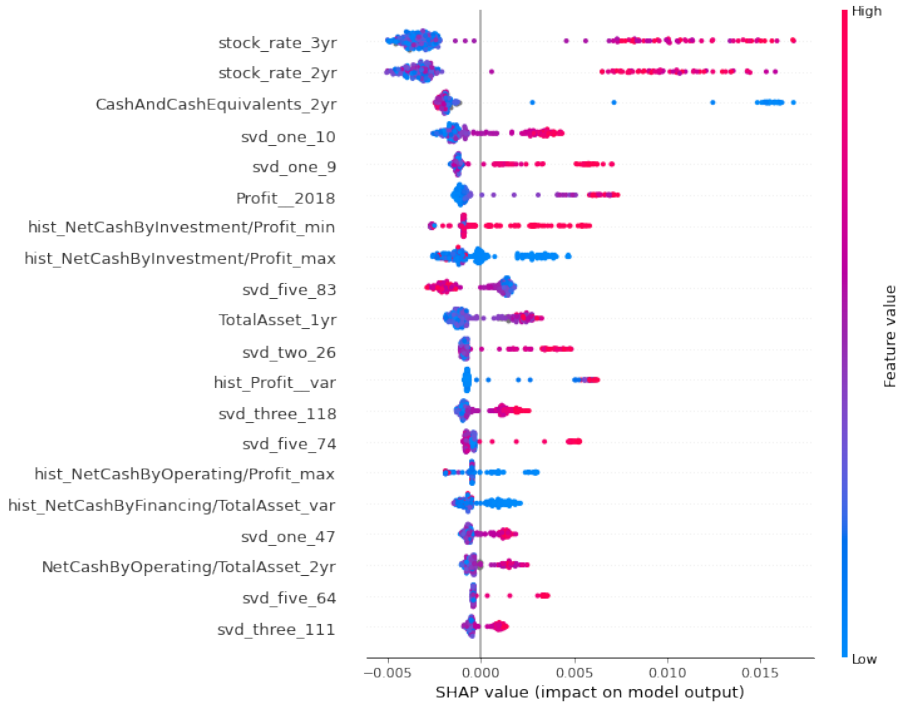

図 5.16: 694 特徴量のうち売上率予測に特に寄与度の高い 20 の特徴量

第6章　総括

6.1　総括

本研究を通して、日経225社の貸借対照表を一つのファイルにまとめ、それを用いて売上高とその売上率について予測モデルを作成した。さらに、複数のExplainable AIを使って予測に寄与した特徴量を定量的に評価し、同時にExplainable AIの活用例を紹介した。日経225社の決算データを一つにまとめたデータは公開されていないことから、今後の決算データの研究において誰でも有効活用でいるようにGitHubに公開している(以下のリンク参照)。一方、Explainable AIを使った分析では予測する対象が売上高と売上率を予測することに留まったが、利益やキャッシュフローなど、予測の対象を変更し活用することでさらに興味深い知見が得られるのではないだろうか。

今回の分析を通して得られた結論として、売上高予測においては前年度の売上高が最も予測に寄与し、さらに過去の売上高の最大値・最小値・平均値が、他のどの特徴量よりも強く予測に寄与している結果となった。さらに、営業活動によるキャッシュ・フローや総資産額など、会社の体力と言える得量量が次いで予測に強く寄与している結果となった。一方、売上率予測に関しては用いたExplainable AIフレームワークによって結果が異なったが、過去3年の株価の変化率や期末残高の増加率、また自然言語から生成した特徴量が大きく予測に寄与している結果が得られた。有効な特徴量として期待していた、業種による区分やMicrosoft Azureによる印象分析で得られた特徴量は、いずれの予測においても上位の重要度となることはなかった。

本研究で用いた日経225社の決算書関連のデータ
`https://github.com/hodaka0714/XAI_annualreport`

6.2　今後の展望

今後の展望として、会社のサンプル数をさらに増やすことでサンプルに偏りの少ない一般的な結果が得られる可能性がある。さらに、売上率予測で有効であった経営に関する記述から得られた特徴量がなぜ有効だったのか、単語ベースに特徴量を作りここでもExplainable AIを活用することで、詳細な理由が判明するのではないだろうか。

謝辞

最後に、指導教員としてご指導をいただいた、名古屋大学大学院情報学研究科複雑系科学専攻の時田恵一郎教授には、学部時代から 3 年間、ご多忙にも関わらず様々な場面でご指導をいただきました。心から感謝を申し上げます。また、私が所属している多自由度システム情報論講座の杉山雄規教授、谷村省吾教授、中村泰之准教授には中間発表等を通して研究に関する助言をいただきました。感謝を申し上げます。さらに、多自由度システム情報論講座の先輩、後輩、同期の皆様にも、論文の校閲やその他論文に関する議論にお付き合いいただきました。ありがとうございました。最後に、大学院の 2 年間を支えてくださった、家族に感謝の意を表します。

参考文献

[1] A. Kinski: “Google trends as complementary tool for new car sales forecasting : a cross-country comparison along the customer journey” (2016).

[2] 税所篤大力 菅愛子：“感情分析を通じて見える経営者の特性と企業の関係 －日経 225 社を対象にした分析－”, 経営課題に AI を!ビジネスインフォマティクス研究会 (第 10 回).

[3] 田村浩一郎上野山勝也 飯塚修平：“深層学習を用いたアンサンブルモデルによる企業価値推定モデルの提案”.

[4] “＜政策オープンラボの取組＞「有価証券報告書等の審査業務等における ai 等利用の検討」実証実験の結果の概要について：金融庁”, https://www.fsa.go.jp/news/r1/openlab/20190927/01.pdf (2019).

[5] “「家計の安定的な資産形成に関する有識者会議」（第 1 回）議事次第：金融庁”, https://www.fsa.go.jp/singi/kakei/siryou/20170203.html. (Accessed on 08/17/2020).

[6] 航, 和孝：“専門家記事と機械学習に基づく web ニュースからの日経平均株価予測 (言語理解とコミュニケーション) – (第 10 回テキストマイニング・シンポジウム)”, 電子情報通信学会技術研究報告 = IEICE technical report : 信学技報, **116**, 451, pp. 19–24 (2017).

[7] 田中：“経済物理学とその周辺”, 横幹, **7**, 2, pp. 79–82 (2013).

[8] “Edinet”, https://disclosure.edinet-fsa.go.jp/. (Accessed on 07/16/2020).

[9] “その他統計資料 — 日本取引所グループ”, https://www.jpx.co.jp/markets/statistics-equities/misc/01.html. (Accessed on 07/16/2020).

[10] “pandas - python data analysis library”, https://pandas.pydata.org/. (Accessed on 07/16/2020).

[11] “Singular value decomposition (svd) tutorial”, https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm. (Accessed on 07/30/2020).

[12] “Stooq”, https://stooq.com/. (Accessed on 09/07/2020).

[13] T. Chen and C. Guestrin: “Xgboost: A scalable tree boosting system”, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794 (2016).

[14] L. B. Statistics and L. Breiman: "Random forests", Machine Learning, pp. 5–32 (2001).

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T. Liu: "Lightgbm: A highly efficient gradient boosting decision tree", NIPS (2017).

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama: "Optuna: A next-generation hyperparameter optimization framework", Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019).

[17] "5.1 partial dependence plot (pdp) — interpretable machine learning", `https://christophm.github.io/interpretable-ml-book/pdp.html`. (Accessed on 08/06/2020).

[18] S. M. Lundberg and S.-I. Lee: "A unified approach to interpreting model predictions", Advances in Neural Information Processing Systems 30 (Eds. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett), Curran Associates, Inc., pp. 4765–4774 (2017).

付 録A　付録

A.1　日経225社リスト (銘柄名 [17業種区分別])

- エネルギー資源
  国際石油開発帝石/ＪＸＴＧホールディングス
- 不動産
  東急不動産ホールディングス/住友不動産/東京建物/三菱地所
- 医薬品
  中外製薬/大塚ホールディングス/第一三共/エーザイ/武田薬品工業/アステラス製薬/大日本住友製薬/協和キリン
- 商社・卸売
  豊田通商/丸紅/伊藤忠商事/三菱商事/住友商事/双日
- 小売
  三越伊勢丹ホールディングス/高島屋/セブン＆アイ・ホールディングス/イオン/ファミリーマート/ファーストリテイリング/丸井グループ
- 建設・資材
  太平洋セメント/東海カーボン/ＴＯＴＯ/日本碍子/日本電気硝子/日本板硝子/ＡＧＣ/コムシスホールディングス/大成建設/東洋製罐グループホールディングス/大林組/清水建設/日揮ホールディングス/積水ハウス/大和ハウス工業/ＳＵＭＣＯ/長谷工コーポレーション/鹿島建設
- 機械
  オークマ/アマダ/小松製作所/日本製鋼所/住友重機械工業/日立建機/クボタ/日本精工/ダイキン工業/ジェイテクト/ＩＨＩ/三菱重工業/日立造船/ＮＴＮ
- 素材・化学
  花王/ＤＩＣ/宇部興産/日東電工/日本化薬/三菱ケミカルホールディングス/信越化学工業/日産化学/日本曹達/住友化学/昭和電工/日本製紙/王子ホールディングス/トクヤマ/クラレ/旭化成/帝人/ユニチカ/東洋紡/富士フイルムホールディングス/三井化学/デンカ/東レ/東ソー
- 自動車・輸送機
  横浜ゴム/ブリヂストン/三井Ｅ＆Ｓホールディングス/ＳＵＢＡＲＵ/本田技研工業/日産自動車/いすゞ自動車/トヨタ自動車/日野自動車/スズキ/ヤマハ発動機/マツダ/川崎重工業/三菱自動車工業

- 運輸・物流
  東急/商船三井/日本郵船/ヤマトホールディングス/東武鉄道/日本通運/東海旅客鉄道/川崎汽船/西日本旅客鉄道/京成電鉄/京王電鉄/ＡＮＡホールディングス/小田急電鉄/東日本旅客鉄道

- 銀行
  三井住友フィナンシャルグループ/三井住友トラスト・ホールディングス/りそなホールディングス/三菱ＵＦＪフィナンシャル・グループ/あおぞら銀行/新生銀行/静岡銀行/ふくおかフィナンシャルグループ/コンコルディア・フィナンシャルグループ/みずほフィナンシャルグループ

- 金融（除く銀行）
  ソニーフィナンシャルホールディングス/野村ホールディングス/ＳＯＭＰＯホールディングス/クレディセゾン/第一生命ホールディングス/東京海上ホールディングス/Ｔ＆Ｄホールディングス/松井証券/ＭＳ＆ＡＤインシュアランスグループホールディングス

- 鉄鋼・非鉄
  フジクラ/神戸製鋼所/ジェイ　エフ　イー　ホールディングス/大平洋金属/日本軽金属ホールディングス/三井金属鉱業/日本製鉄/三菱マテリアル/住友金属鉱山/ＤＯＷＡホールディングス/古河電気工業/住友電気工業

- 電力・ガス
  中部電力/東京電力ホールディングス/大阪瓦斯/東京瓦斯

- 電機・精密
  シチズン時計/日清紡ホールディングス/テルモ/コニカミノルタ/ミネベアミツミ/三菱電機/富士電機/安川電機/東京エレクトロン/ジーエス・ユアサ　コーポレーション/日本電気/富士通/沖電気工業/セイコーエプソン/パナソニック/オムロン/ＴＤＫ/ソニー/リコー/キヤノン/ＳＣＲＥＥＮホールディングス/ニコン/太陽誘電/オリンパス/ファナック/カシオ計算機/アドバンテスト/横河電機/アルプスアルパイン/京セラ

- 食品
  マルハニチロ/日清製粉グループ本社/明治ホールディングス/サッポロホールディングス/日本たばこ産業/キリンホールディングス/宝ホールディングス/キッコーマン/味の素/ニチレイ/アサヒグループホールディングス/日本水産

- 情報通信・サービスその他
  サイバーエージェント/トレンドマイクロ/日本郵政/リクルートホールディングス/バンダイナムコホールディングス/凸版印刷/Ｚホールディングス/ヤマハ/ソフトバンクグループ/ＫＤＤＩ/セコム/コナミホールディングス/エムスリー/ディー・エヌ・エー/ＮＴＴドコモ/東宝/日本電信電話/スカパーＪＳＡＴホールディングス/エヌ・ティ・ティ・データ/電通グループ

A.2　日経 225 社に関する表

以下の図は全て 2019 年度の損益計算書から取得した値である。

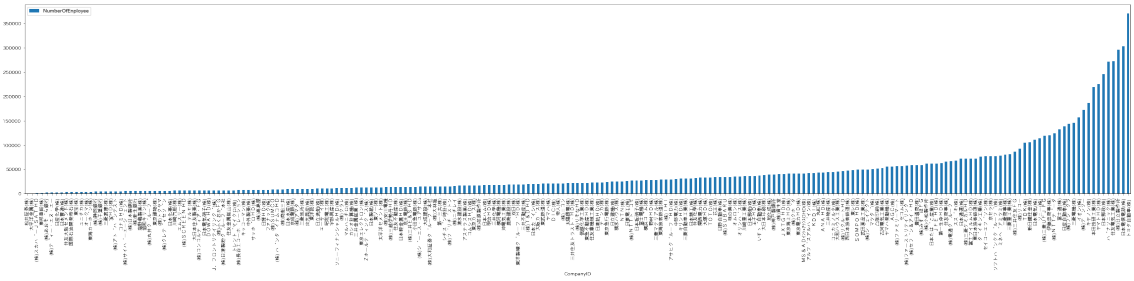

図 A.1: 225 社の売上高

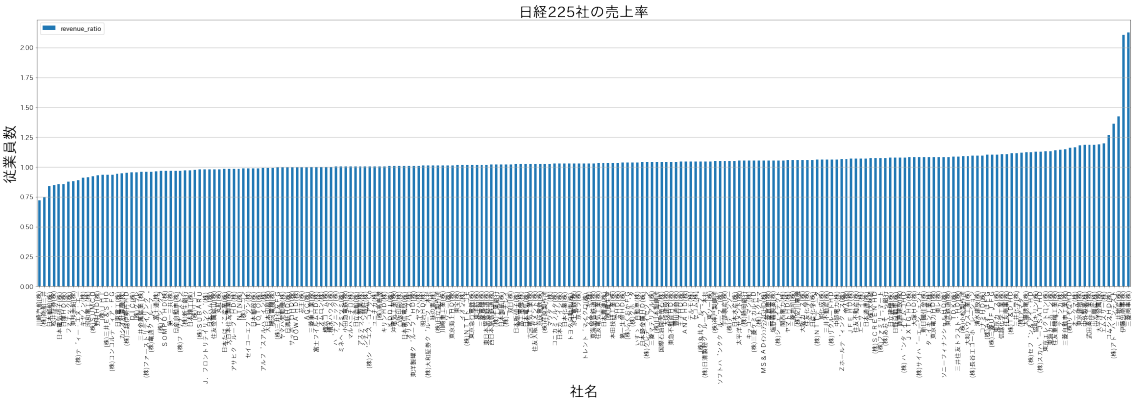

図 A.2: 225 社の売上率

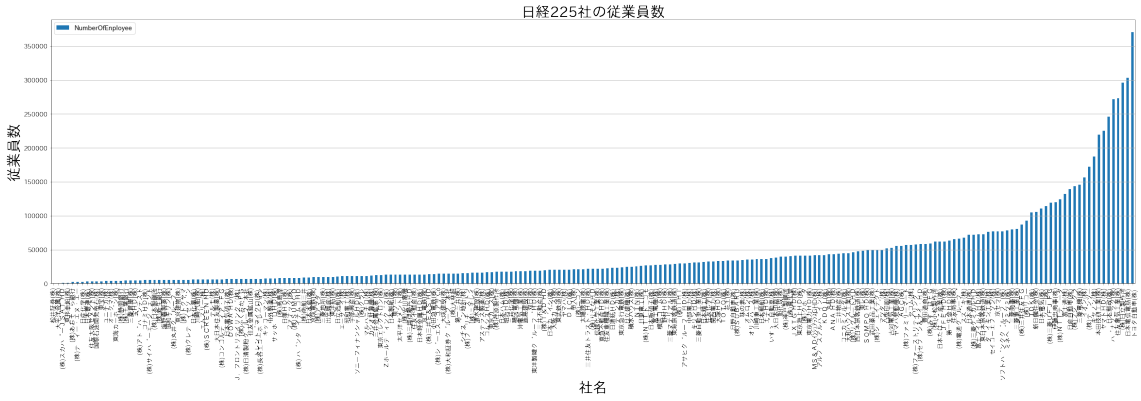

図 A.3: 225 社の従業員数

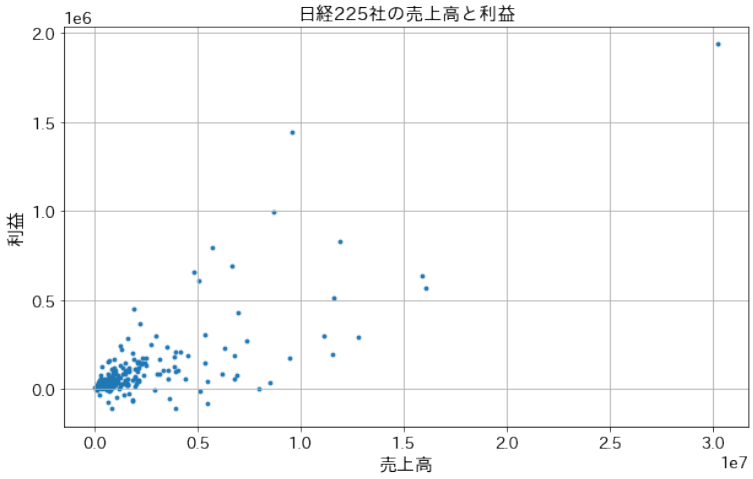

図 A.4: 売上高と利益の分布

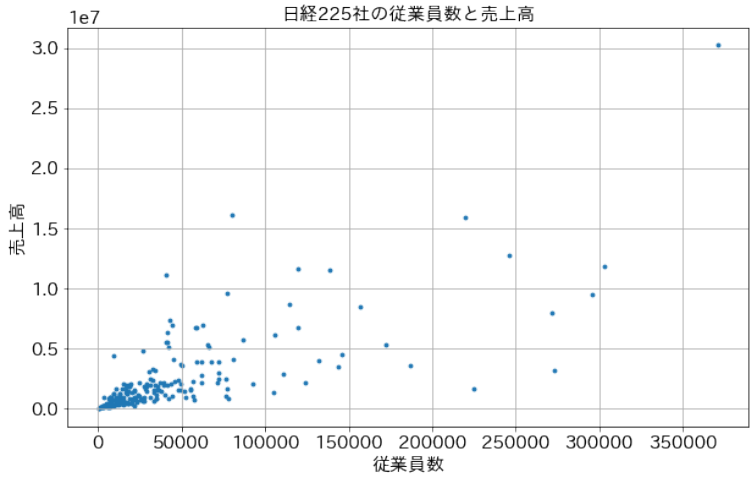

図 A.5: 従業員数と売上高の分布

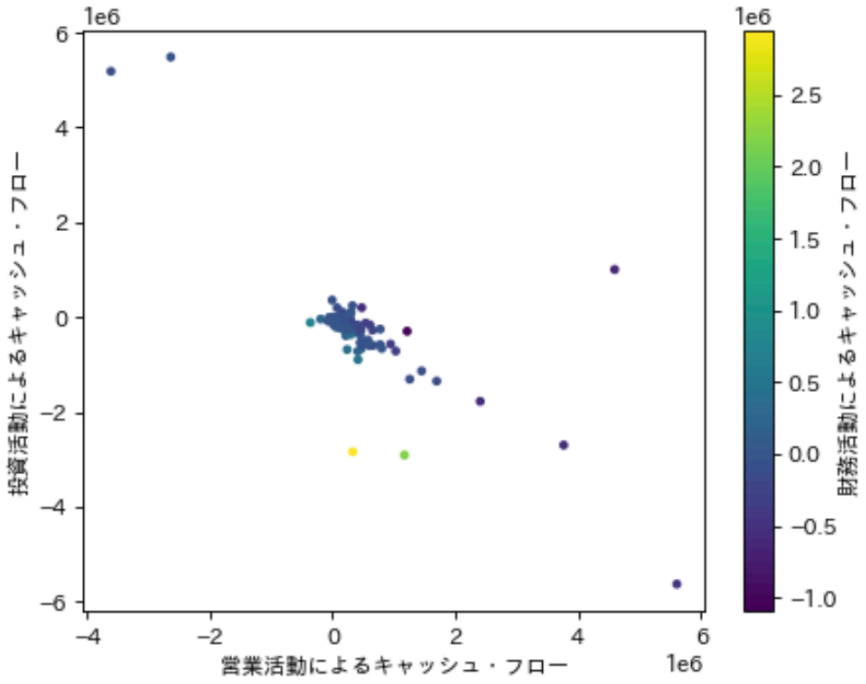

図 A.6: キャッシュフローの分布

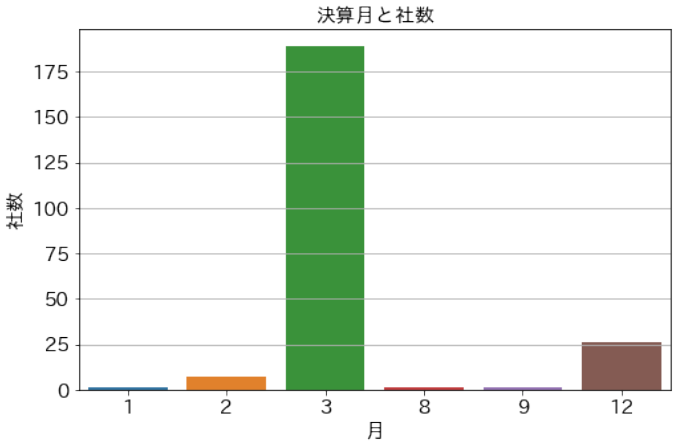

図 A.7: 決算月と社数

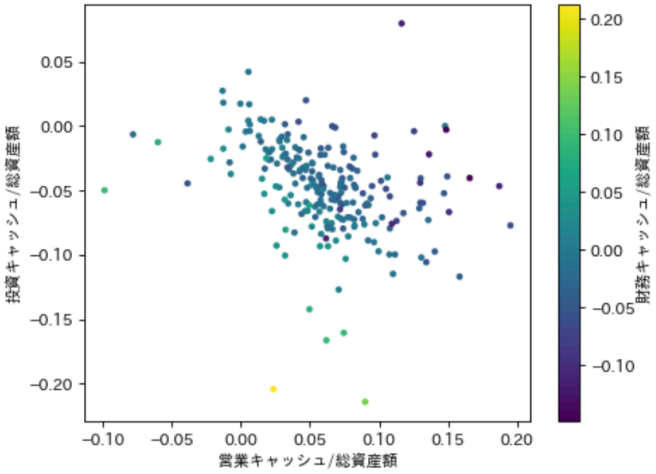

図 A.8: キャッシュフロー/総資産額の分布

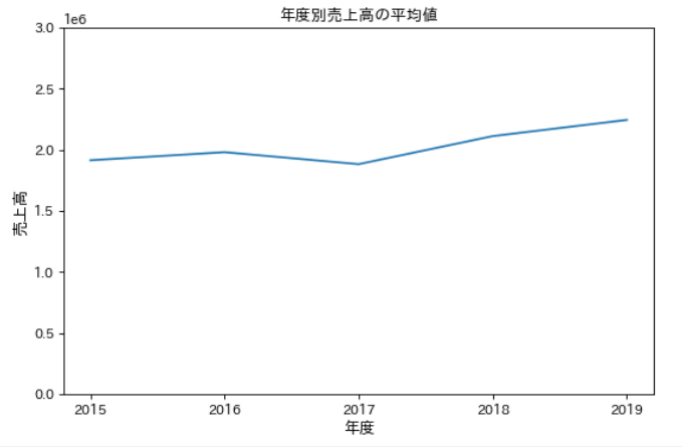

図 A.9: 年度別売上高の平均値