



日本銀行ワーキングペーパーシリーズ

格付け分類モデルにおける機械学習の応用： 機械学習の説明可能性を高める手法

橋本龍一郎*

ryuuichirou.hashimoto@boj.or.jp

三浦翔*

kakeru.miura@boj.or.jp

吉崎康則*

yasunori.yoshizaki@boj.or.jp

No.23-J-3
2023年3月

日本銀行
〒103-8660 日本郵便（株）日本橋郵便局私書箱30号

* 金融機構局

日本銀行ワーキングペーパーシリーズは、日本銀行員および外部研究者の研究成果をとりまとめたもので、内外の研究機関、研究者等の有識者から幅広くコメントを頂戴することを意図しています。ただし、論文の中で示された内容や意見は、日本銀行の公式見解を示すものではありません。

なお、ワーキングペーパーシリーズに対するご意見・ご質問や、掲載ファイルに関するお問い合わせは、執筆者までお寄せ下さい。

商用目的で転載・複製を行う場合は、予め日本銀行情報サービス局 (post.prd8@boj.or.jp) までご相談下さい。転載・複製を行う場合は、出所を明記して下さい。

格付け分類モデルにおける機械学習の応用： 機械学習の説明可能性を高める手法*

橋本龍一郎[†]・三浦翔[‡]・吉崎康則[§]

2023年3月

【要旨】

近年、金融機関における様々な業務において機械学習の活用が進んでいる。信用リスク管理の分野では、クレジットスコアリングモデルやデフォルト率モデルに機械学習を活用する事例がみられ始めている。本稿では、格付け分類モデルの推計に機械学習を応用した。まず、説明変数等の条件を揃えたうえで、機械学習と順序ロジット回帰によって格付け分類モデルを推計し、モデル構造の差異によって、どの程度予測精度に違いが生じるのかを確認した。次に、近年急速に研究が進んでいる機械学習の説明可能性を高めるための手法（「説明可能な AI」）を応用して、変数重要度の計測やモデル予測値の要因分解を行った。分析結果は以下の通りである。第一に、機械学習の応用により、財務指標と信用力の非線形的な関係が捕捉しやすくなり、順序ロジット対比で大幅に予測精度が改善した。第二に、SHAP（SHapley Additive exPlanations）や PDP（Partial Dependence Plot）を用いてモデル予測値の要因分解を行うことで、売上高、総資産回転率や ICR といった財務指標が企業の信用力に与える影響が大きいことが分かったほか、ICR が 2 倍以下に低下すると信用力が急激に減少するといった非線形性が観察された。これらの説明可能性を高めるために用いる手法の仮定や限界を正しく把握しつつ活用することで、機械学習の課題の一つである説明可能性の低さを補完することが可能となる。

JEL 分類番号：C49、C55、G32

キーワード：信用リスク管理、機械学習、説明可能性、説明可能な AI

* 本稿の作成に当たっては、多くの日本銀行スタッフから有益なコメントを頂いた。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者達個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者達個人に属する。

[†] 日本銀行金融機構局 (ryuuichirou.hashimoto@boj.or.jp)

[‡] 日本銀行金融機構局 (kakeru.miura@boj.or.jp)

[§] 日本銀行金融機構局 (yasunori.yoshizaki@boj.or.jp)

1. はじめに

近年、金融機関における様々な業務において機械学習の活用が進んでいる。Bank of England (2022)によれば、実務において機械学習を活用する英国の金融機関の割合は7割強を占めるに至っており、顧客との関係構築、マネー・ロンダリング防止対策や各種リスク管理、粉飾決算の検出など幅広い業務分野で活用されている¹。

信用リスク管理の分野では、European Banking Authority (2021) (以下、EBA (2021))が指摘するように、クレジットスコアリングモデルやデフォルト率モデルなど、個別企業の財務データやマクロ経済変数から、個別債務者の信用力を予測するモデルを中心に、機械学習を活用する事例がみられ始めている。

従来、こうしたデフォルトや格付けの決定要因の分析には、主にロジット回帰などパラメトリックな手法が用いられてきた。近年、デフォルト率モデルの分野においては、機械学習では、より複雑な非線形性を捕捉可能なことから、パラメトリックな手法対比、デフォルト率の予測精度が向上するとの実証研究が多くみられている。例えば、Alonso and Carbó (2021)は、スペインの個別行におけるリテールローンのデータを用いて、機械学習では、ロジット回帰対比、デフォルト率の予測精度が向上することを示した。また、三浦ほか (2019)は、入出金情報を用いて、中堅中小企業を中心とした非上場企業にも適用可能なデフォルト率モデルを構築し、機械学習では、ロジット回帰対比、予測精度が向上することを示した。

他方、格付け分類モデルの分野においては、小林 (2001)²など幾つかの先行研究がみられているが、パラメトリックな手法が中心であり、機械学習を応用した分析は少ない。格付け分類モデルは、企業の財務指標やマクロ変数から、個別企業の格付けを推定する枠組みである。一般に、機械学習は、複雑なモデル構造を取り得るため、財務指標（説明変数）と格付け（被説明変数）の間に非線形性が存在する場合、パラメトリックな手法に比して予測精度が高くなりやすく、機械学習を応用するメリットは大きいと考えられる。機械学習の応用により、格付け分類モデルの予測精度の向上が実現できれば、個別企業の信用リスク評価の精緻化にとどまらず、これを応用したシナリオ分析の有用性の向上も期待できる³。

¹ 機械学習は、民間金融機関だけでなく、中央銀行においても活用が広がっている。Araujo et al. (2022)は、データの収集、金融・経済分析、金融政策運営、プルーデンス業務など、中央銀行の様々な業務で機械学習が活用されている点を指摘している。

² わが国製造業のデータを用いて、財務指標を説明変数、社債格付けを被説明変数とする順序プロビット回帰と多項プロビット回帰を推計している。

³ 2022年10月の金融システムレポートでは、機械学習の手法に基づく格付け分類モデルを用いて、資金調達コスト上昇に伴うICRの悪化が、デフォルトカーブに及ぼす影響を推計している。詳細は日本銀行 (2022)を参照。

機械学習については、複雑な非線形性を捕捉可能なメリットの代償として、モデルの複雑さや、それに伴う説明可能性の低さがデメリットとして指摘されている。例えば、EBA(2021)は、機械学習では、モデルの予測値と説明変数の関係の把握が難しいことから、経営陣や監督当局に対する説明が困難である点を指摘している⁴。

こうした説明可能性の低さへの指摘に対応して、近年では機械学習の説明可能性を高める手法——いわゆる「説明可能な AI (eXplainable AI, XAI)」——の研究が急速に進展している⁵。そこで、本稿では、まず、機械学習を用いた格付け分類モデルの予測精度の評価を行った後、機械学習の説明可能性を高める手法を用いて、企業の財務指標と信用力の関係について定量的な分析を行う。そのうえで、これらの分析結果を踏まえて、格付け分類モデルの推計に機械学習を用いる際の留意点を整理する。

本稿の貢献は次のとおりである。第一に、格付け分類モデルの推計に機械学習を応用した点である。機械学習を応用し、ロジット回帰対比での予測精度の向上を報告した実証分析が多くみられるデフォルト率モデルとは異なり、格付け分類モデルに機械学習を応用した研究は少ない。本稿では、説明変数等の条件を揃えたうえで、機械学習と順序ロジット回帰によって格付け分類モデルを推計し、モデル構造の差異によって、どの程度予測精度に違いが生じるのかを確認した。その結果、デフォルト率モデルの推計に機械学習を応用した先行研究と同様に、格付け分類モデルにおいても、機械学習を用いた場合、順序ロジット回帰対比で予測精度が向上することが示された。

第二に、機械学習の説明可能性を高める手法を応用し、企業の財務指標と信用力の関係について定量的な分析を行った点である。具体的には、予測値に対する各説明変数の寄与を定量的に把握する SHAP (SHapley Additive exPlanations) や、ある変数を変化させた際の予測値の変化を可視化する Partial Dependence Plot (以下、PDP) を用いて、推計した格付け分類モデルにおける信用力の決定要因の把握を試みた。その結果、デフォルト率モデルについての先行研究と同様、売上高や ICR などの説明力が高いことに加え、ICR が信用力に与える影響において、非線形性が存在することなどが観察された。

本稿の構成は以下の通りである。2 章では、分析に使用するデータとモデルの概要を説明する。3 章では、従来用いられてきた順序ロジット回帰と機械学習のそれぞれで推計した格付け分類モデルの予測精度の比較を行う。4 章では、SHAP や PDP を用いて、企業の財務指標と信用力の関係を定量的に分析したのち、格付け分類モデルに機械学習を用いる際の留意点を考察する。

⁴ Alonso and Carbó (2022)は、モデルの複雑さや説明可能性の低さのほか、金融機関が機械学習を利用する際の留意点として、機械学習モデルの推計にクラウド・コンピューティングなどの外部サービスを利用した場合における、外部サービスと金融機関のシステムの連携の難しさや、顧客データの保護等を挙げている。

⁵ XAI の研究を包括的に網羅した文献としては、Molnar (2019)や森下 (2021)などがある。

2. データおよびモデルの概要

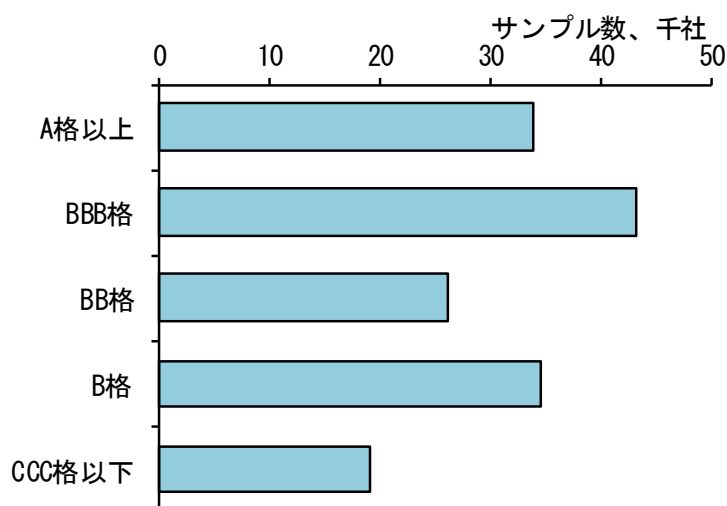
2-1. 分析に使用するデータ

(データの概要)

本稿では、海外企業約 6 千社の格付けデータおよび財務データを用いた。観測期間は 1991 年 1~3 月から 2022 年 4~6 月、データ頻度は四半期である。

モデルの被説明変数は、Moody's Investors Service から取得した発行体の長期格付⁶である。図表 1 は格付け別のサンプル企業数の分布である。これをみると、「BBB 格」のサンプルは 4.3 万社と多い一方で、「CCC 格以下」は 1.9 万社と他の格付け対比少なく、サンプル数の偏りが存在することが分かる。こうしたクラス間の偏り（クラス・インバランス）に対処せずに分類モデルを推計すると、サンプル数の多いクラスに分類しやすいモデルが推計されることで、サンプル数の少ないクラスに対する予測精度が悪化するといった問題が発生する。このため、Chawla et al. (2002) に従い SMOTE (Synthetic Minority Over-sampling Technique) を用いて全てのクラスのサンプル数を揃えてモデルを構築した⁷。

図表 1 格付け別のサンプル企業数



(注) 全観測期間における、延べ企業数を格付け別に示した。

⁶ 本稿では、便宜上、Moody's の格付けの A3 以上を A 格以上、Baa1~Baa3 を BBB 格、Ba1~Ba3 を BB 格、B1~B3 を B 格、Caa1 以下を CCC 格以下として分類した。

⁷ クラス・インバランスに対応する手法としては、SMOTE のほか、サンプル数の少ないクラスのサンプルをリサンプルするオーバーサンプリング、サンプル数の多いクラスをダウンサンプリングする手法が存在する。これらの手法を用いた場合でも、本稿とほぼ同様の結果が得られた。

説明変数は、S&P Global Market Intelligence から取得した個別企業の財務データをもとに算出した対象企業の事業規模、債務返済能力、収益性、財務レバレッジ、流動性、業種成長率の6項目に関する指標である(図表2)。具体的には、①事業規模を表す指標として売上高(対数値)、②債務返済能力を表す指標としてICR(EBITDA/利息支払額)とネットDER((有利子負債-現預金)/自己資本)、③収益性を表す指標としてEBITDAマージン(EBITDA/売上高)、純利益成長率と総資産回転率(売上高/総資産)、④財務レバレッジを表す指標としてレバレッジ(有利子負債/総資産)、⑤流動性を表す指標として流動比率(流動資産/流動負債)、⑥業種成長率を表す指標として業種全体の売上高成長率⁸の9変数を選択した。なお、これらの指標は、Moody's Investors Service (2021)や、S&P Global Market Intelligence (2020)を参考に、格付機関が重視する指標と概ね整合的になるように選択している。

図表2 分析に用いた財務指標

財務指標	変数
事業規模	売上高(対数値)
債務返済能力	ICR(倍)
	ネットDER(倍)
収益性	EBITDAマージン(%)
	純利益成長率(%)
	総資産回転率(倍)
財務レバレッジ	レバレッジ(%)
流動性	流動比率(倍)
業種成長率	業種全体の売上高成長率(%)

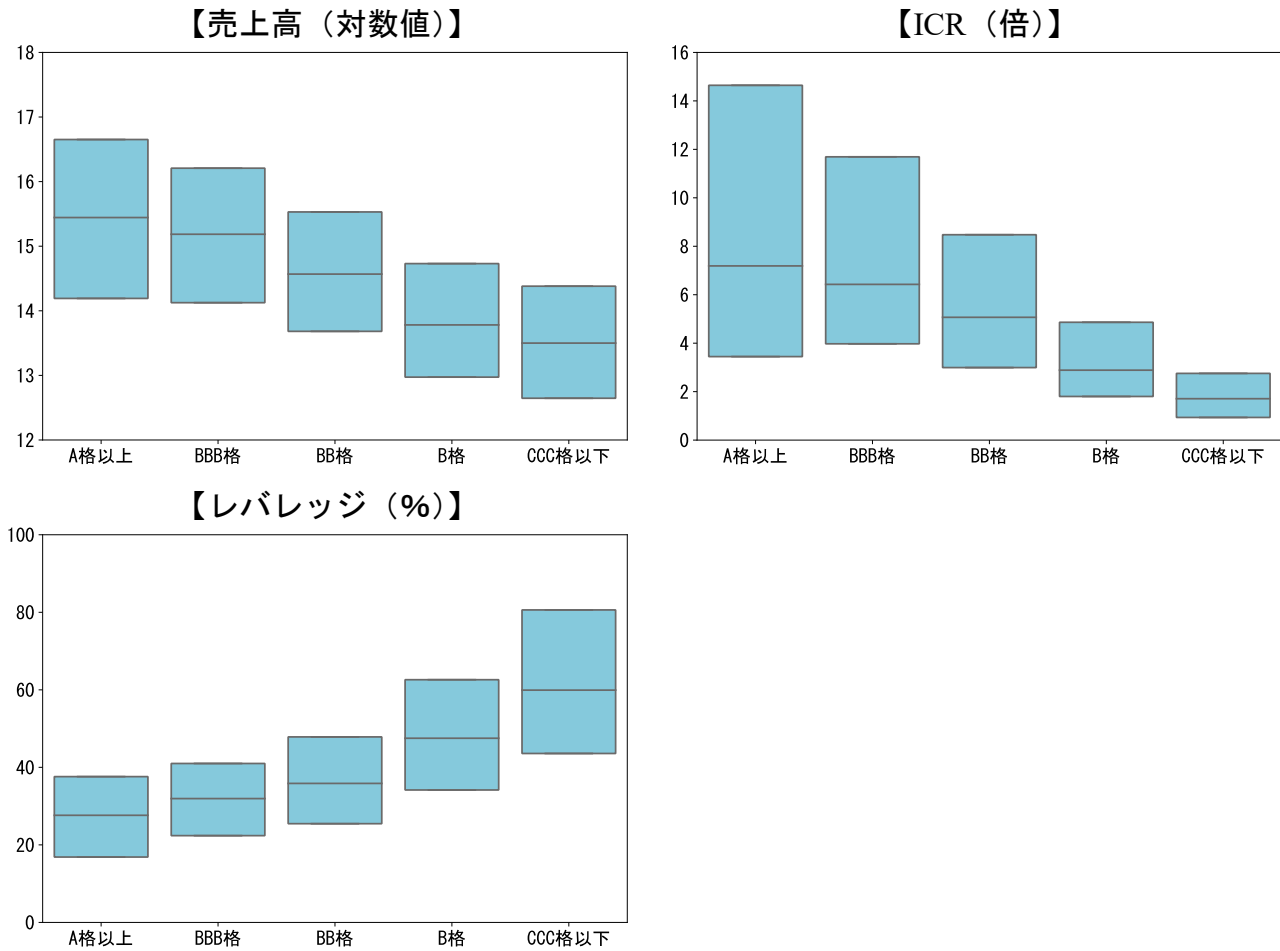
(格付けと財務指標の関係)

図表3~6では、説明変数に用いた財務指標について、格付け別の分布を示している。これらの分布からは、格付けの高低と財務指標の大小の関係について、幾つかのパターンがあることが分かる。

まず、格付けの高低と財務指標の大小の単調性である。図表3で売上高とICRの中央値をみると、格付けが低下するにつれて、中央値の水準が単調に減少している。レバレッジについては、格付けが低下するにつれて、中央値の単調な上昇が観察される。また、上昇幅に着目すると、A格以上からBB格にかけての中央値の上昇幅は比較的小さい一方で、B格以上では上昇幅が高まるなど、非線形的な関係も窺われる。

⁸ 業種分類は、S&Pの世界産業分類基準(Global Industry Classification Standard)の24産業グループを用いた。産業分類の詳細は<https://www.spglobal.com/spdji/jp/landing/topic/gics/>を参照。

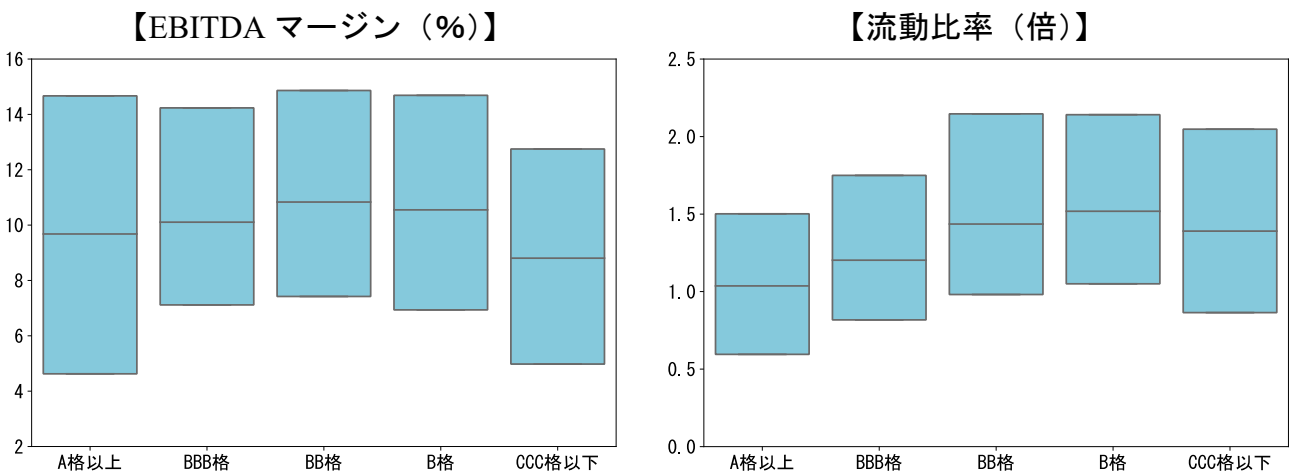
図表3 格付け別の財務指標①



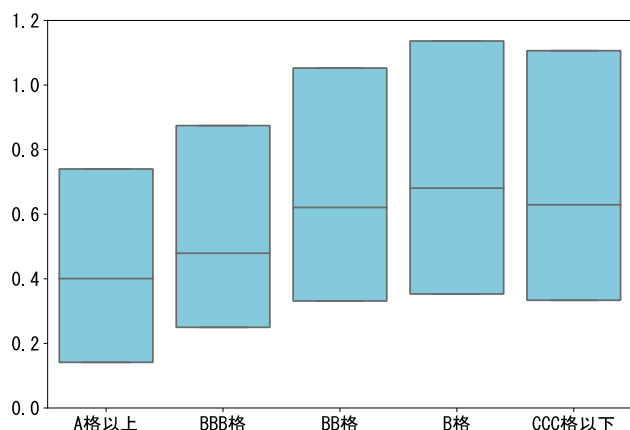
(注) 棒グラフ内の横線は中央値を、棒グラフの下端と上端は 25%-75%タイル値を示す (以下同様)。

一方で、格付けの高低と財務指標の大小が単調に変化しない指標も存在する。図表4で EBITDA マージン、流動比率、総資産回転率の中央値と格付けの高低をみると、山なり (上に凸) の関係が確認できる。EBITDA マージンの中央値をみると、A 格以上から、最大値である BB 格にかけて上昇した後、緩やかな低下に転じている。流動比率と総資産回転率の中央値も、EBITDA マージンと同様、山なりの関係にあるが、最大値は B 格となっている。

図表4 格付け別の財務指標②



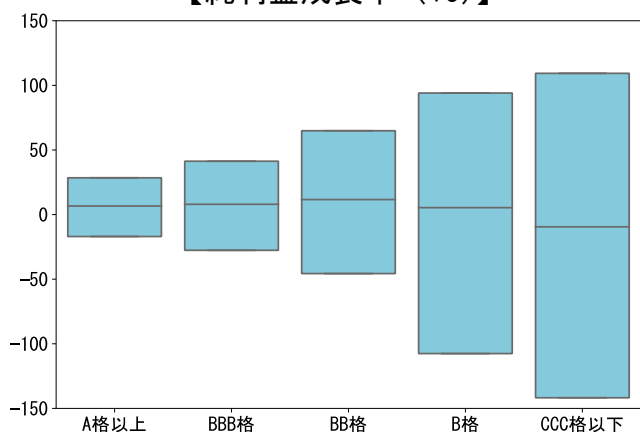
【総資産回転率（倍）】



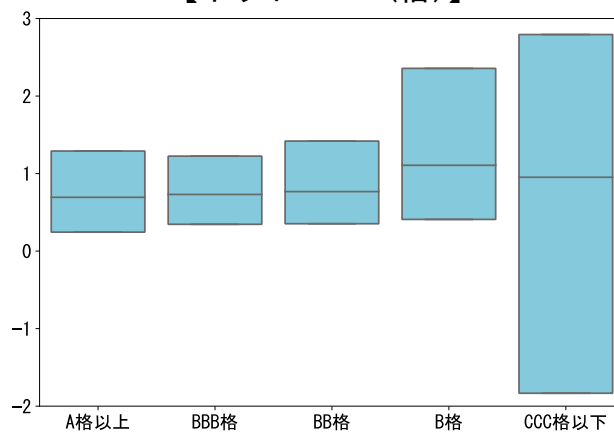
また、ICR、レバレッジ、純利益成長率、ネット DER といった財務指標では、サンプル毎のばらつきが格付けの高低と相関している。ICR をみると、各格付け区分における ICR の 25%–75% タイル区間は、CCC 格以下では 0.9~2.8 倍程度に止まっている一方で、A 格以上では 3.4~14.6 倍と広く、格付けが高いほど ICR の水準にバラツキが生じている。ICR とは対照的に、レバレッジの 25%–75% タイル区間は、A 格以上と BBB 格は概ね 20~40% だが、CCC 格以下では 44~81% と急激に上昇するなど、格付けが低いほど、25%–75% タイル区間が拡大している。図表 5 に示した純利益成長率とネット DER の中央値は格付け毎にさほど大きな違いはないものの、低格付けになるにつれて、25%–75% タイル区間の幅が拡大しており、外れ値を取りやすくなる傾向が確認される。低格付けグループには規模の小さい企業が相対的に多いもとの、両指標の分母である純利益や自己資本も小さい値を取りやすいことから、格付けが低くなるほど絶対値が大きい水準を取りやすくなることが理由として考えられる。

図表 5 格付け別の財務指標③

【純利益成長率（%）】



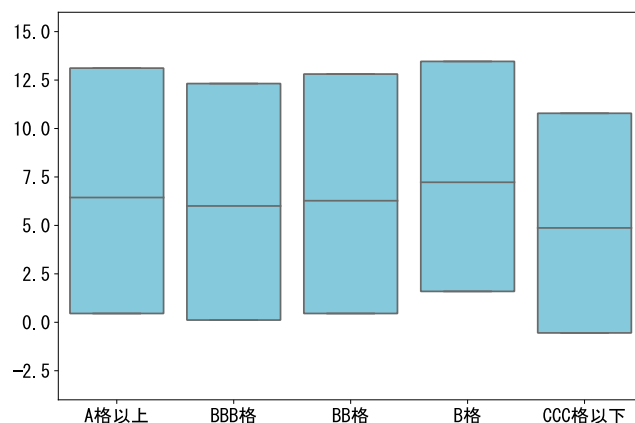
【ネット DER（倍）】



最後に、図表 6 に示した業種全体の売上高成長率については、中央値の水準が、A 格以上から BBB 格以下にかけて低下するが、B 格にかけて上昇に転じた後、CCC 格にかけて減少するなど、格付けの高低と指標の大小に、明確な関係が窺われない。

図表 6 格付け別の財務指標④

【業種全体の売上高成長率 (%)】

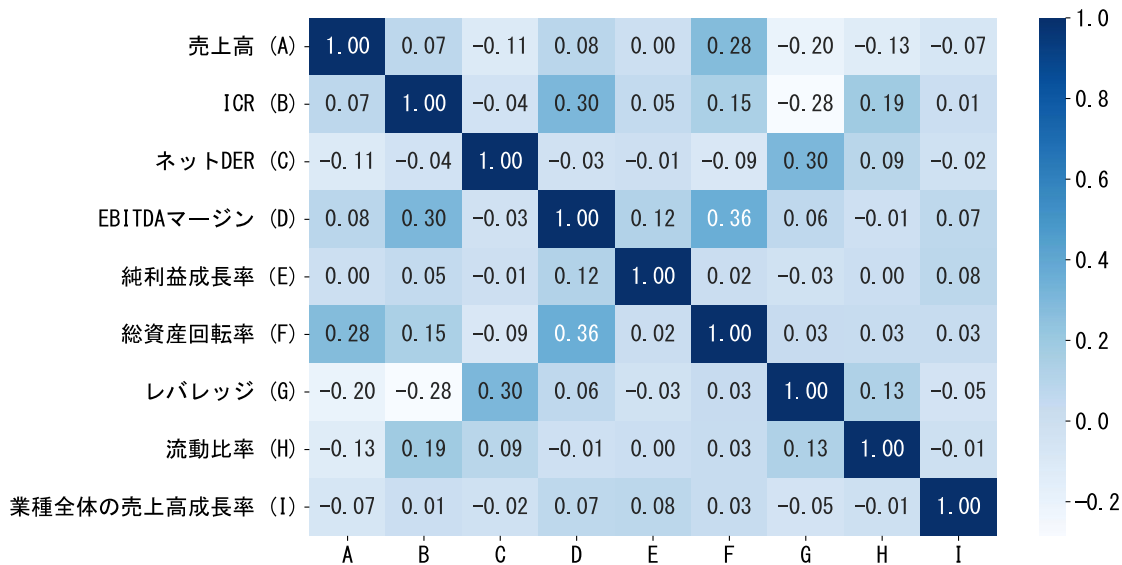


このように、財務指標の大小と格付けの高低は、必ずしも単調な線形の関係にあるわけではない。このため、予測精度の高い格付け分類モデルを推計する際には、財務指標と格付けの非線形的な関係を取り込むことや、指標の平均的な水準ではなく、分散や外れ値の情報を柔軟に扱う必要があることが分かる。すなわち、被説明変数と説明変数間の複雑な関係を捕捉可能である機械学習では、こうした財務指標の大小と格付けの高低の関係を取り込みやすいものの、線形モデルであるロジット回帰では、指標の平均的な水準を評価するため、予測精度が相対的に低い可能性が予見される。

(財務指標の相関行列)

ロジット回帰のような伝統的なパラメトリックモデルを推計する際には、説明変数間に多重共線性が存在する場合、パラメータの安定性が下がるといった問題が発生することが知られている。また、森下 (2021) は、機械学習についても、多重共線性がある場合にはモデルの評価が困難になる可能性を指摘している。もっとも、図表 7 の財務指標の相関行列をみると、2 指標間の相関は、最も高い EBITDA マージンと総資産回転率の組み合わせでも 0.36 とそれほど高くはない。このため、本稿の分析に用いたデータセットにおいては、多重共線性がパラメータの安定性やモデル評価に与える影響は大きくないと考えられる。

図表 7 財務指標の相関行列



(注) 各セルは相関の大きさに応じて色付けしている。

2-2. 推計するモデル

格付け分類モデルは、順序ロジット回帰、機械学習を用いて、A 格以上、BBB 格、BB 格、B 格、CCC 格以下の 5 クラスに分類するモデルを推計した。機械学習モデルには、高い予測精度と計算時間の速さを両立可能な勾配ブースティングツリー (LightGBM⁹、以下 LGBM) を選択した¹⁰。推計に際しては、データセットを学習データ、検証データ、テストデータの 3 つに分割し、学習・検証データを用いて推計を行い、テストデータでモデルの予測精度を評価した。

なお、4 章で紹介する SHAP や PDP の計算や解釈を容易にするため¹¹、「投資適格 (以下、IG)」「非投資適格 (以下、Non-IG)」の 2 クラスに分類するモデルも推計した。両クラスのサンプル数はほぼ同じ (IG : 7.7 万社、Non-IG : 8.0 万社) であるため、SMOTE によるクラス・インバランスへの対応は行っていない。

⁹ LightGBM (Light Gradient Boosting Machine) は、2016 年に Microsoft Research が開発した勾配ブースティングツリーの一つ。"Light"の名称の通り、パラメータのチューニングに要する推計時間を削減しつつ、高い精度を維持している。LGBM の詳細は Ke et al. (2017)を参照。

¹⁰ 代表的な他の機械学習モデル (Random Forest や XGBoost) では、モデルの予測精度に多少の差異がみられたものの、LGBM と概ね同様の結果が得られた。

¹¹ 多クラス分類問題における SHAP 値や PDP は、分類するクラスの数だけ作成される。例えば、SHAP 値であれば、「ある格付けと判別される確率に対するある説明変数の SHAP 値」を格付け毎に算出する必要があり、計算や解釈が難しくなる。

(順序ロジット回帰)

山下・三浦 (2011)によれば、格付けのような順序付き離散データの分類モデルを推計する際には、順序ロジット回帰を用いることが多い。順序ロジット回帰では、 J 個のクラスが存在する順序付き離散データ Y は、潜在変数 y^* と閾値によって以下の通り決定されると仮定する¹²。ここで、 t_1, \dots, t_{j-1} はクラスを決定するための閾値であり、 y^* の大きさに応じてクラスに分類される。

$$Y = \begin{cases} 1 & -\infty \leq y^* \leq t_1 \\ 2 & t_1 < y^* \leq t_2 \\ \vdots & \\ J & t_{j-1} < y^* \leq \infty \end{cases}$$

ここで、潜在変数 y^* は M 個の説明変数の各成分 x_i ($i = 1, 2, \dots, M$)を対応する重み w_i で足し合わせた重み付き線形和として定義される。また、 $Y = j$ となる確率は、 y^* がクラス j の閾値の間に収まる確率に対応する。

$$y^* = \sum_{i=1}^M w_i x_i = \mathbf{w}^T \mathbf{x}$$

$$P(Y = j | \mathbf{x}) = P(t_{j-1} < y^* \leq t_j | \mathbf{x})$$

上式の通り、同モデルは潜在変数が説明変数に対して線形に定式化されており、一般化線形モデルの一つである¹³。順序ロジット回帰の定式化の特徴は、重み w_i がクラスに依存しない、すなわち、分類においてどの説明変数を重視するかは、クラス間で一定と仮定していることである（平行性の仮定）。この仮定は強い仮定であり、クラス間で説明変数の重要度が大きく異なるような場合には、順序ロジット回帰を利用することが適切ではなく、モデルによる予測精度が低下する可能性が高い。

本稿では、前述の9変数の財務指標と業種ダミーのうち、AICを用いて選択した7項目を説明変数、格付けを被説明変数として、順序ロジット回帰を行った。なお、2クラス分類問題では、順序ロジットは通常の二項ロジットとなる。推計結果の詳細は参考図表1に記載している。

(LGBM)

本稿で使用する機械学習モデルである LGBM は、機械学習の分野において最も多

¹² 順序ロジット回帰の理論的な背景については McCullagh (1980)などに詳しい。

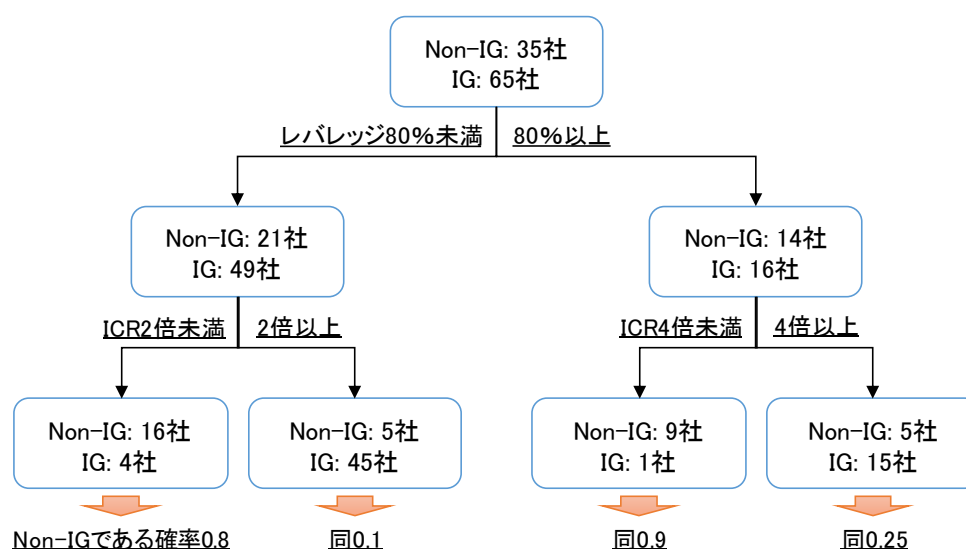
¹³ 一般に、順序ロジット回帰では、説明変数の重み w_i とクラス間の閾値 t_j は2項ロジット回帰と同様に最尤法を用いて算出される。

く用いられるモデルの一つである。LGBMなどの勾配ブースティングツリーは、多くの決定木を直列に連結するモデル構造を取っており、決定木が進む毎に実際のデータとモデルの予測との誤差を徐々に縮小させる仕組みとなっている¹⁴。

勾配ブースティングツリーにおける一つ一つの決定木は、学習データのサンプルを樹形図状に段階的に分岐させるものであり、分岐点毎に説明変数を用いた分割条件（ICRが2倍未満、など）が設定される。こうした決定木の構造により、LGBMが説明変数と被説明変数の間の非線形的な関係を捉えることが可能となる。

図表8は、決定木による分割のイメージを図示したものである。それぞれの分岐点において、条件を満たすサンプルは左側、それ以外のサンプルは右側、といった形で分割され、次の分岐ポイントにそれぞれ進む。決定木内の分割が終了したあとの到達点（ターミナルノード）には、それぞれ予測値（Non-IGである確率80%/IGである確率20%、など）が設定される。それぞれの決定木では、ツリーの最大の深さ¹⁵などの事前に設定されたパラメータを所与として、学習データとモデルの予測結果の乖離を最小化するように、分割に用いる説明変数の選択やその閾値およびターミナルノードの予測値が決定される。

図表8 ツリーによる分割のイメージ



もっとも、ツリーの分割を必要以上に行うなど、モデル構造を複雑にしすぎると、モデルが学習データに過剰に適合する結果、アウトオブサンプルの予測精度が極端に

¹⁴ ツリーベースのモデルには、決定木を直列結合するブースティングと呼ばれる手法を用いたモデルと、ランダムフォレストに代表されるようなツリーを並列に学習させるバギングと呼ばれる手法を用いたモデルなどが存在する。

¹⁵ 学習データの分割を何段階行うかを意味するパラメータ。より多くの回数だけ分割を行えば、学習データとモデルの予測値の誤差をそれだけ縮小させることが可能だが、テストデータのような学習に用いていないデータへのフィットが悪化する可能性がある。

悪化する現象（いわゆる過学習）が起こりやすくなるほか、推計にかかる時間が大幅に増加する。LGBM は、決定木の分割方法や閾値の選択方法を改善したり、一つ前の決定木で誤差の大きかったサンプルを中心に損失の改善を試みるといった工夫を導入することで、勾配ブースティングツリーのモデリングの過程における課題の一つである推計速度の遅さに対応している。その結果、LGBM を用いることで、相対的に短い推計時間で高い予測精度を得やすいという利点がある。

こうした利点を踏まえ、本稿では、前述の 9 変数の財務指標と業種(カテゴリ変数)を説明変数、格付けを目的変数、Log-loss を損失関数とする LGBM を推計した¹⁶。

（推計結果の解釈と留意点）

格付機関は、一般的に、企業の財務情報や企業の属する国のマクロ経済指標といった定量情報に加えて、企業の事業環境や先行きのリスクなどの定性情報も勘案して格付けを付与している¹⁷。このうち、定性情報については必ずしもデータセットの企業全てについて利用可能ではない。このため、本稿では、小林 (2001)など他の先行研究と同様に、財務指標を用いて格付け分類モデルを推計している。また、推計の都合上、格付けを IG/Non-IG や 5 クラスに集約している。

このため、財務情報のみを勘案した本稿のモデルでは、格付機関が実際に決定した格付けを完全に予測することはできない点には留意が必要である。3 章で述べるとおり、LGBM は高い予測精度を実現できる一方で、実際の格付けとモデルの予測には依然として乖離が存在しており、これらの要因が影響している可能性がある。もっとも、財務情報を用いてどの程度格付けを適切に推計できるかといったモデル精度の比較の観点から、同一のデータセットを用いて推計した順序ロジット回帰と LGBM で、予測精度を比較することは有益と考えられる。

3. モデルの予測精度

本章では、分類モデルの予測精度を評価する際に一般に用いられる、正答率や AUC (Area Under the Curve) といった指標により、推計したモデルの予測精度を比較する。

¹⁶ 推計に際しては、過学習のリスクを低減するため、検証データで損失関数に改善が認められなくなった時点で推計を停止する Early stopping の手法を用いた。また、パラメータの調整は、グリッド・サーチで行った。

¹⁷ Moody's Investors Service (2021)や S&P Global Market Intelligence (2020)によれば、実際の格付けプロセスでは、財務情報や企業の属する国のリスクスコアやインフレ率といった定量情報に加え、格付け機関の担当者が独自に収集・分析した企業の事業環境（業種内の立ち位置・業種に対する規制・ESG への取り組み度合い）や企業の資本政策の動向（金融子会社の有無・経営層による良好な財務状況・資金調達の維持へのコミットメント）に関する定性情報を考慮している。

3-1. 正答率による予測精度の評価

まず、正答率を確認する。図表9は、予測された格付けを横軸に、実際の格付けを縦軸に集計した混合行列（Confusion matrix）である。正答率は、モデルが正しく格付けを判別できた企業の割合であり、ここでは、混合行列の対角成分の和をサンプル数で割ることによって、全サンプルの正答率を定義する。全サンプルの正答率は、5クラス問題では、順序ロジット回帰が43.1%、LGBMが79.4%となっており、いずれのモデルでも実際の格付けとの乖離はみられるが、LGBMの正答率の方が高い結果となった。順序ロジット回帰の混合行列を仔細にみると、両端（A格以上、CCC格以下）の正答率は55~60%と相応に高い一方で、それ以外の格付けでは正答率が30%台と低いことが分かる。この背景として、順序ロジット回帰では、2章で述べた財務指標と格付け間の非線形性や、財務指標の分散と格付け間の相関といった複雑な関係性を捕捉することが難しいことなどが考えられる。なお、2クラス問題では、ロジット回帰が77.7%、LGBMが89.6%と、5クラス問題同様、LGBMの方が正答率が高い。

図表9 混合行列

①5クラス問題

【順序ロジット回帰】

	予測					正答率(%)	
	A格以上	BBB格	BB格	B格	CCC格以下		
実際	A格以上	4,049	1,716	710	183	105	59.9
	BBB格	2,593	2,953	2,240	743	138	34.1
	BB格	626	1,015	1,947	1,308	347	37.1
	B格	260	556	1,473	2,450	2,195	35.3
	CCC格以下	57	127	376	1,108	2,131	56.1

全サンプル正答率: 43.1

【LGBM】

	予測					正答率(%)	
	A格以上	BBB格	BB格	B格	CCC格以下		
実際	A格以上	5,992	569	139	48	15	88.6
	BBB格	681	6,965	723	223	75	80.4
	BB格	132	528	3,796	628	159	72.4
	B格	53	234	645	5,171	831	74.6
	CCC格以下	7	49	129	590	3,024	79.6

全サンプル正答率: 79.4

(注) テストデータを用いて算出。各セルは企業数に応じて色付けしている。

②2 クラス問題

【ロジット回帰】

		予測		正答率(%)
		IG	Non-IG	
実際	IG	12,268	3,162	79.5
	Non-IG	3,832	12,144	76.0
全サンプル正答率:				77.7

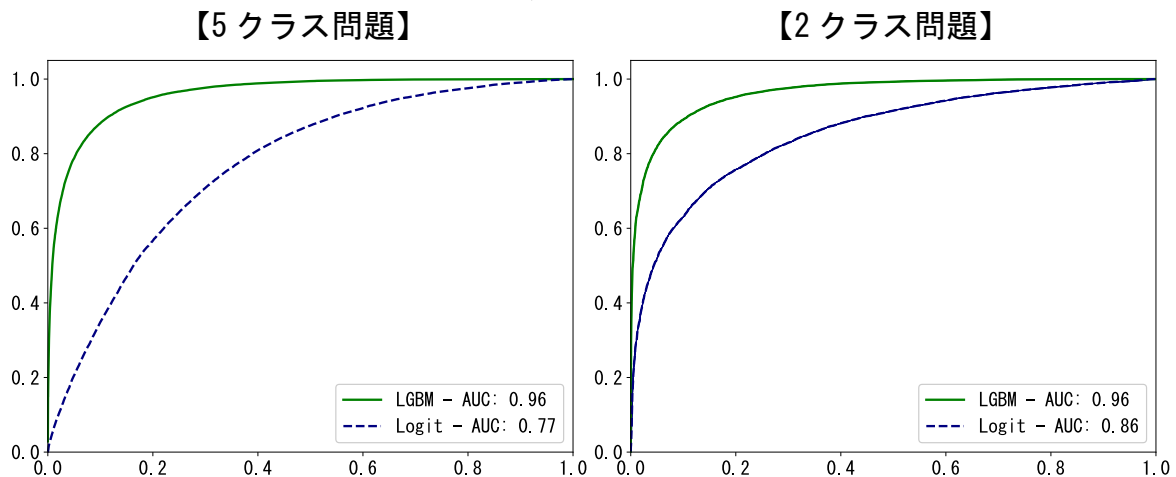
【LGBM】

		予測		正答率(%)
		IG	Non-IG	
実際	IG	13,966	1,464	90.5
	Non-IG	1,817	14,159	88.6
全サンプル正答率:				89.6

3-2. AUC による予測精度の評価

次に、Alonso and Carbó (2021)と同様に、分類モデルの予測性能を測定する代表的な評価指標の一つである AUC (Area Under the Curve) によって、モデルの予測精度を比較する。AUC は、縦軸に TPR (True Positive Rate¹⁸)、横軸に FPR (False Positive Rate¹⁹) をプロットした ROC (Receiver Operating Characteristic) 曲線²⁰の下部の面積を表し、AUC が大きいほど分類モデルの予測精度が高いことを意味する²¹。

図表 10 AUC



(注) テストデータを用いて算出。

図表 10 では、LGBM、順序ロジット回帰のそれぞれについて算出した AUC を示し

¹⁸ 真陽性率。Non-IG を陽性とみなした IG/Non-IG の 2 クラス問題では、Non-IG の企業を Non-IG と正しく予測した割合。5 クラス問題では、OvR (One-versus-Rest) の定義に沿って、ある格付け (例: A 格以上) を陽性、それ以外の格付けを陰性と定義したうえで、正しく当該格付けを予測した割合として計算。

¹⁹ 偽陽性率。上記 2 クラス問題では、実際には陰性である IG の企業を Non-IG と誤って予測した割合。5 クラス問題では、TPR と同様、OvR の定義に沿って集計。

²⁰ 分類モデルにおいて、クラスの判定基準 (閾値) を変化させた際に、TPR と FPR がどのように変化するかをプロットした曲線。

²¹ 具体的には、100% 正確に予測する完全な分類モデルでは AUC が 1 となる一方、ランダムな予測を行う分類モデルでは、正しく予測する確率は 50% となるため、AUC は 0.5 となる。

た。左図は5クラス問題、右図は2クラス問題のAUCである。これをみると、正答率と同様に、いずれのモデルでも実際の格付けとの乖離はみられるが、5クラス問題²²、2クラス問題の両方で、LGBMのAUCの方が、曲線の下面積が大きく、予測精度が高い結果となった。ただし、順序ロジット回帰のAUCをみると、5クラス問題では0.77、2クラス問題では0.86と、相応の予測精度²³が確保されている。

3-3. モデル間の予測精度の差異の背景

前節まででみたように、順序ロジット回帰、LGBMのいずれについても、相応の予測精度は確保したものの、実際の格付けとの乖離が残る結果となった。前述したように、本稿のモデルでは、格付機関の格付け判定とは異なり、定性情報である企業の将来性を勘案していない点が影響したものと考えられる。

両モデルの評価指標を比較すると、正答率とAUCいずれも、順序ロジット回帰対比で、LGBMの予測精度が高い結果となった。LGBMの予測精度が高い背景としては、2章で確認した説明変数と被説明変数の間の非線形的な関係や、説明変数の分散と被説明変数の相関などが、順序ロジット回帰では捉えられていないことが考えられる。例えば、ICRが2(倍)低下したとしても、10(倍)から8(倍)に低下した場合と、2(倍)からゼロに低下した場合には、後者の方が信用力は大きく悪化すると考えられるが、こうした非線形性は、順序ロジット回帰では、モデルの定式化上、捕捉できない²⁴。また、純利益成長率のように、格付け間で平均的な水準は大きく異ならないとしても、格付けが低下すると分散が大きくなるような関係は、分布や関数形を先験的に仮定するパラメトリックモデルでは捕捉できない。一方、LGBMのようなツリーベースの機械学習モデルは、その構造上、説明変数と被説明変数の非線形的な構造をうまく捉えることができるため、より高い予測精度が実現できたと考えられる。

4章では、こうしたパラメトリックモデルでは捕捉が難しい変数間の非線形的な構造が、実際に、機械学習モデルでどのように捉えられているかを含め、近年研究が進められている機械学習の説明可能性を高める枠組みを用いて、企業の財務指標と信用力の関係を確認していく。

²² 通常、AUCは2クラス問題の評価で用いられることが多いが、図表10の5クラスのROC曲線は、各クラスについてOvRの定義に沿って算出したROCを単純平均している。

²³ 尾木(2017)は、おおよその水準として、大企業を対象として財務情報を用いたデフォルト率モデルではAR値が0.8程度(AUCでは0.9程度)としている。

²⁴ 順序ロジット回帰でも、説明変数のべき乗項・交差項等や、説明変数の平均・分散で基準化したZ値を追加的に採用するなど、変数変換を通じた対応によって、こうした非線形性を取り込むことは可能となる。ただし、適切な変数の加工方法を選択するには相応の知見が求められるほか、これらの項を増やすにつれて、パラメトリックモデルの利点である説明可能性は低下する。

4. 企業の財務指標と信用力の関係

4-1. 機械学習の説明可能性を高める手法 (XAI)

関数形や分布を先験的に仮定し、推計された回帰係数によって説明変数の予測値への影響を推測することが可能なパラメトリックモデルに対し、機械学習はモデル構造が複雑で、説明可能性の低さが課題として指摘されることが多い。実務上は、予測値がどの説明変数の寄与によるものか把握することが重要であるが、近年、機械学習のモデルに対してもこうしたニーズに応えるべく、モデルの説明可能性を高める手法 (XAI) の研究が進められている。主な指標の概要は図表 11 の通りで、モデルとしての解釈を行いたい (グローバル)、予測結果の背景をサンプル毎に解釈したい (ローカル)、など、分析目的に合わせて、様々な指標が存在する。これらの指標の多くはモデル構造に依存せずに算出できるため (model-agnostic)、多くの機械学習モデルに適用可能である²⁵ほか、異なるモデル間で分析結果を比較することも可能である²⁶。

図表 11 主な指標の概要

	モデルレベル(グローバル)	サンプルレベル(ローカル)
概要・用途	<ul style="list-style-type: none"> モデルがどの説明変数を重要視して(学習して)いるか。 どの変数が予測値に対してどのような影響を与えているか。 	<ul style="list-style-type: none"> サンプルの出力結果に対し、それぞれの説明変数がどの程度寄与しているか。
名称 ²⁷	<ul style="list-style-type: none"> 変数重要度 (PFI, Permutation Feature Importance) PDP (Partial Dependence Plot) 	<ul style="list-style-type: none"> LIME (Local Interpretable Model-agnostic Explanations²⁸) ICE (Individual Conditional Expectation²⁹)
	<ul style="list-style-type: none"> SHAP (モデルレベル・サンプルレベルいずれも適用可) 	

²⁵ 金田ほか (2022)は、SHAP により原油価格の変動を可視化し、近年の原油価格の変動には需給およびマーケット要因に加え、金融政策要因 (FRB の BS サイズ) が寄与している点を指摘した。

²⁶ これらの手法は、幾つかの仮定に基づいており、その解釈には留意を要する。例えば、PDP は、説明変数間の依存関係を勘案せずに予測値への影響を算出するが、相関が存在する場合、実際には起こりにくい説明変数の組み合わせにも一定のウェイトを与えてしまうことがある。

²⁷ ツリー系のモデルでは Gain (分割したときの誤差の減少量) や Split (分割回数) などの変数重要度を用いることも可能だが、ここではモデルに依存しない変数重要度を記載している。なお、Gain 等の変数重要度については、Nembrini et al. (2018)に詳しい。

²⁸ LIME は、調べたいサンプルの周辺のデータ空間について線形回帰モデルを作成し、各変数が予測値にどの程度貢献しているかを測定する手法。詳細は Ribeiro et al. (2016a, 2016b)を参照。もっとも、LIME は「サンプルの周辺」をどのように定義するかで測定結果が変動しやすいといった問題があるため、SHAP の方がより利用される傾向がある。

²⁹ ICE は、調べたいサンプルについて、ある特徴量が変化したときにサンプルの予測値がどのように変化するかを 1 本の線で可視化する手法。ICE の詳細は Goldstein et al. (2015)を参照。なお、PDP は ICE を全サンプルで平均したものと一致する。

本章では、モデル内の各説明変数が予測にどれだけ寄与したかをサンプルレベルで評価する SHAP や、他の説明変数を一定としてある説明変数を変化させた際の予測値の変化を可視化する PDP を用いて、推計した LGBM における企業の財務指標と信用力の関係を確認する(図表 11 下線部)。なお、以下の議論は全て 2 クラス問題の LGBM で行うため、本章で紹介する SHAP 値や PDP は、モデルが Non-IG と判別する確率や、その寄与度を表している。

4-2. SHAP

SHAP (SHapley Additive exPlanations) は、それぞれの説明変数が予測値への程度寄与しているかを、サンプル毎に評価する手法であり、説明変数の寄与度を SHAP 値と呼ぶ³⁰。具体的には、SHAP は、以下の数式の通り、あるサンプルの予測値を、全サンプルの予測値の平均と各説明変数の SHAP 値に加法的に分解する手法である。

$$f(x) = E_x[\hat{f}(X)] + \sum_{j=1}^M \phi_j$$

あるサンプル x の予測値
全サンプルの予測値の平均
当該サンプルにおける M 個の説明変数の SHAP 値 ϕ_j の合計

この SHAP 値の計算方法の詳細は Lundberg and Lee (2017)などに詳しいが、ここでは、ICR とレバレッジの 2 つの説明変数で Non-IG と判別される確率を予測する、単純なモデルを例に説明する。まず、このモデルが、ICR 2 倍、レバレッジ 80%の企業が Non-IG である確率を 0.7 と予測し、全サンプルの予測値の平均が 0.5 である場合、両者の差 (0.2) を、ICR とレバレッジの寄与度 (SHAP 値) の和と考える。そのうえで、ICR の SHAP 値を算出する際には、シャープレイ値におけるプレイヤーの限界的な貢献度を求める手法に準じて、「ICR の情報を考慮せずに算出された予測値」から「追加的に ICR の情報を考慮して算出した予測値」への変化を、当該サンプルにおける ICR の SHAP 値とする³¹。レバレッジの SHAP 値も、同様に算出する。以下では、SHAP の利用法について、具体例を用いて解説する。

³⁰ ゲーム理論におけるシャープレイ値の考え方をベースに Lundberg and Lee (2017)や Lundberg et al. (2018)によって提唱された。なお、協力ゲーム理論におけるシャープレイ値は、各プレイヤーの協力によって得られた利得を、プレイヤー間で公正に分配するために、各プレイヤーの限界的な貢献度を測定する指標である。詳細は Shapley (1953)を参照。

³¹ 説明変数を加える順番によってその変数の限界的な寄与度が異なるため、シャープレイ値を厳密に算出する場合には、全ての組み合わせについて算出した寄与度を平均する必要がある。説明変数の数に応じて指数関数的に計算コストが増大することになる。Lundberg et al. (2018)は、条件付き期待値の考え方を採用して、ツリーベースの機械学習モデルにおいて SHAP 値の計算コストを大幅に削減する TreeSHAP アルゴリズムを開発した。このアルゴリズムによって、SHAP は、サンプルの予測値を説明変数の貢献度の和として表現するシャープレイ値の特徴を維持しつつ、より効率的に SHAP 値を算出することが可能となっている。

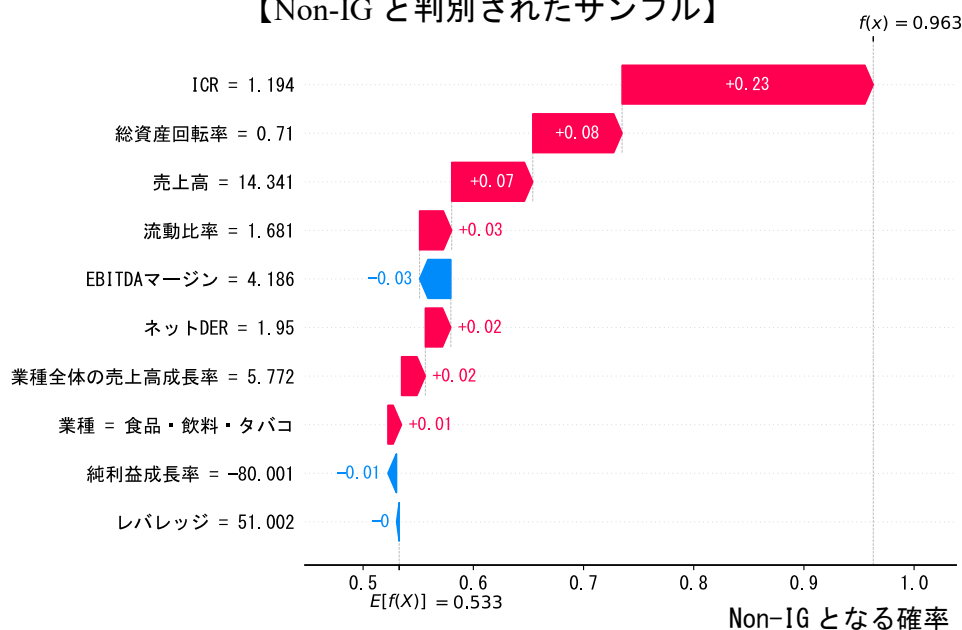
(サンプルレベルの評価)

実務上は、信用力の判定に際して重要度の高い説明変数を把握したいといった格付け分類モデル全体の評価に加えて、「ある企業が Non-IG である確率は 90%と推計されたが、どの財務指標が見劣りする結果として、こうした推計結果に繋がったのか」など、サンプルレベルで説明変数の寄与度を確認したいケースも想定される。

SHAP を用いることで、こうしたケースにも対応できる。図表 12、13 では、データセットからランダムに抽出した 2 社について、SHAP を用いて、説明変数の寄与度を示した。図表 12 は Non-IG と判別されたサンプル、図表 13 は IG と判別されたサンプルである。いずれも、横軸にはモデルの予測値 ($f(x)$ 、Non-IG である確率) を、縦軸には説明変数を並べ、モデルの予測値に対する寄与度 (SHAP 値) を、データ全体の期待値 ($E[f(x)] = 0.533$) との差として示した。

図表 12 のサンプルをみると、モデルの予測値が、 $f(x) = 0.963$ と大きく、Non-IG と判別された。財務指標毎にみると、ICR (1.194) の低さが、Non-IG である確率を 0.23 (23%pt) 押し上げているほか、総資産回転率 (0.71) や売上高 (14.341) の低さも、それぞれ 0.08、0.07 押し上げていることが分かる。

図表 12 SHAP によるサンプルレベルの評価①
【Non-IG と判別されたサンプル】



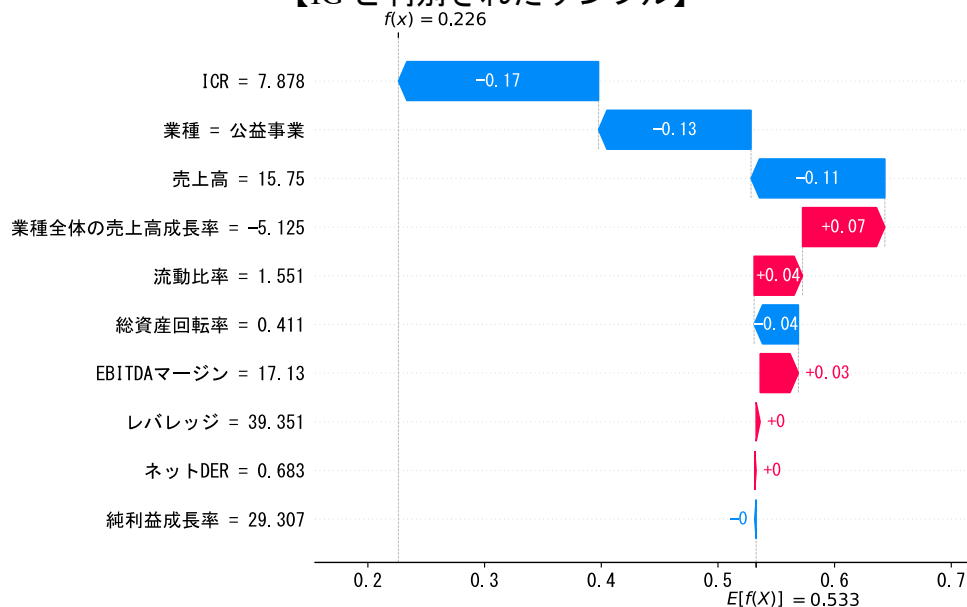
(注) テストデータからランダムにサンプルを抽出して SHAP 値を算出。
グラフの $E[f(x)]$ は、全サンプルの予測値 (Non-IG である確率) の平均を示す (以下同様)。

一方で、図表 13 のサンプルでは、モデルの予測値は $f(x) = 0.226$ と低く、IG と判別された。SHAP 値をみると、ICR (7.878)、売上高 (15.75) の高さが、Non-IG であ

る確率を、それぞれ▲0.17、▲0.11 引き下げているほか、業種（公益事業）も▲0.13 引き下げていることが分かる³²。

図表 13 SHAP によるサンプルレベルの評価②

【IG と判別されたサンプル】



（モデルレベルの評価）

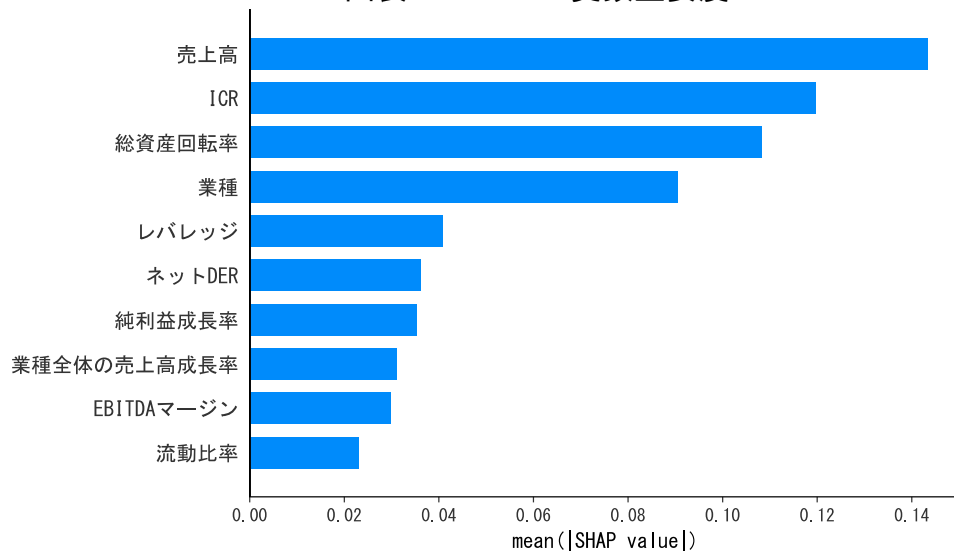
次に、SHAP を用いた代表的なモデルレベルの評価手法である、SHAP 変数重要度と SHAP Dependence Plot を紹介する。

先述の通り、SHAP 値は説明変数の寄与度をサンプル毎に確認するものであるが、説明変数の SHAP 値をデータセット全体で積み上げることによって（SHAP 変数重要度）、モデルレベルで評価を行うことも可能である。SHAP 変数重要度は、予測値に対する各説明変数の寄与度を表す指標であり、各サンプルの SHAP 値の絶対値を平均して算出する。図表 14 は、推計した格付け分類モデルの SHAP 変数重要度である。ここからは、売上高、ICR、総資産回転率といった財務指標が、IG/Non-IG の判別に与える影響が大きいことが分かる³³。

³² 業種などカテゴリ変数の SHAP 値は、当該企業の属する業種の情報を勘案すると、こうした高い ICR や売上高をもつ企業が Non-IG である確率が低下する、と解釈する。実際、参考図表 2 にあるように、公益事業は他業種と比較して高格付け先の比率が高い。

³³ 変数重要度には、サンプルレベルで算出した SHAP 値を積み上げて算出する SHAP 変数重要度のほかに、ある説明変数の値をデータセット内でランダムに入れ替えることによって、モデルの予測誤差がどの程度悪化するか（どの説明変数が最も予測誤差を小さくするのに寄与していたか）をモデルレベルで評価する Permutation Feature Importance という手法もある。同手法を用いた場合でも、図表 14 と概ね同様の結果となった。

図表 14 SHAP 変数重要度



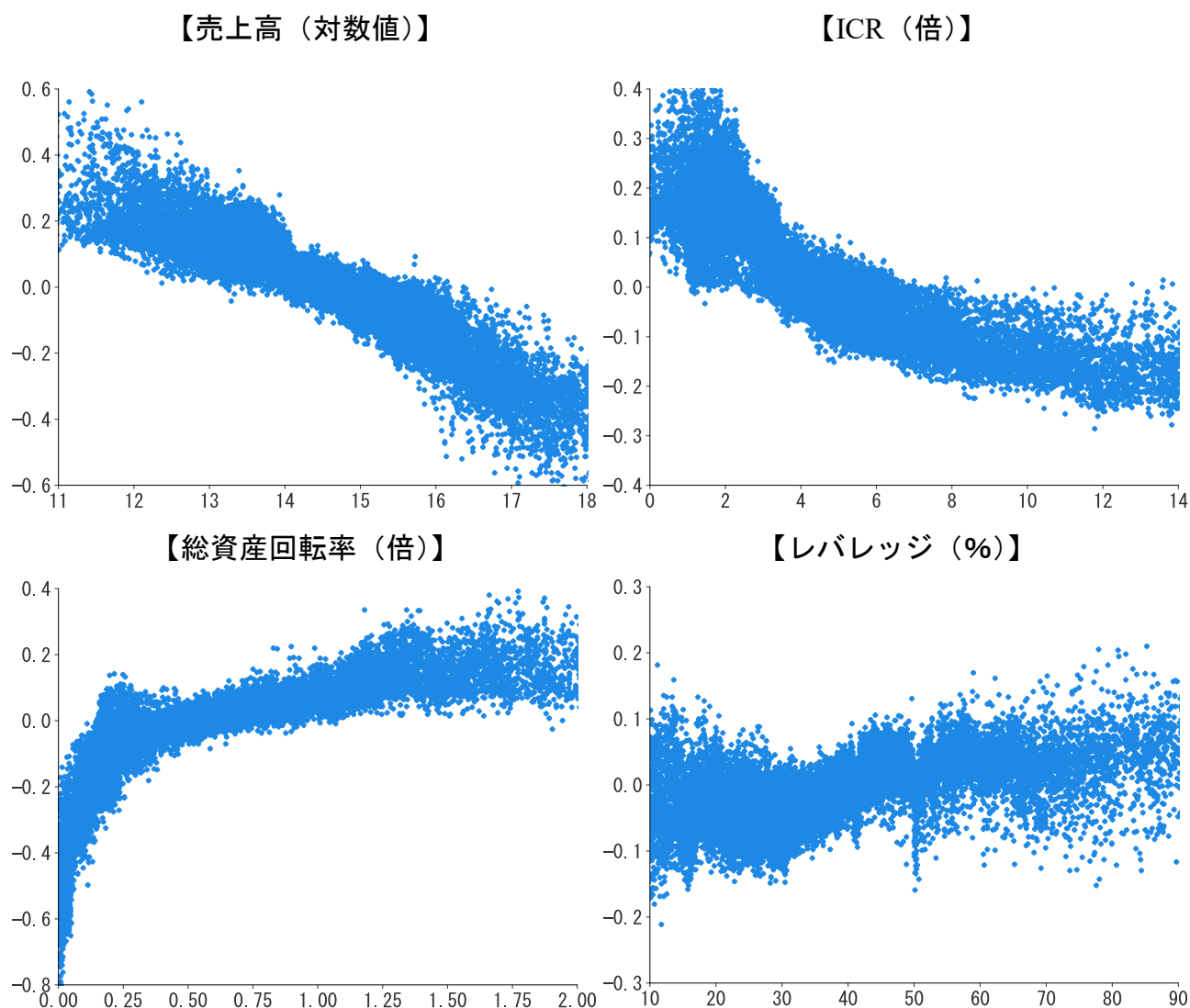
(注) テストデータを用いて算出。

SHAP Dependence Plot は、横軸に説明変数の値、縦軸に対応する SHAP 値をプロットすることで、説明変数の水準に応じて、SHAP 値やその分散の変化を可視化する手法である。図表 15 では、SHAP 変数重要度の高い売上高、総資産回転率、ICR、レバレッジについて、SHAP Dependence Plot を示した。

これをみると、売上高については、売上高が増加すると Non-IG である確率は低下する、単純な負の関係がある。ICR についても、売上高と同様、SHAP 値の間に負の関係があるほか、ICR が 2 倍以下に低下すると、SHAP 値が急激に上昇するという非線形性も窺える。また、レバレッジについては、レバレッジが上昇すると、SHAP 値は緩やかに増加するが、散布図の両端では SHAP 値の分散が大きいことが分かる。このことは、レバレッジが極端に高い、ないし低い場合には、他の説明変数との相互作用が IG/Non-IG の判別に与える影響が強まる点を示唆している。これらの結果は、2 章で確認したデータの特徴と整合的である。なお、総資産回転率については、直感に反し、同比率が高いほど SHAP 値は増加 (Non-IG である確率が上昇) した。これには、学習データ内の Non-IG 企業には、事業規模の小ささから、総資産回転率の分母である総資産が少なく、計算上、総資産回転率が高く算出されてしまう企業が多いことが理由として考えられる³⁴。

³⁴ 図表 4 でも、格付けが低くなるほど総資産回転率が高くなる傾向が確認できる。

図表 15 SHAP Dependence Plot



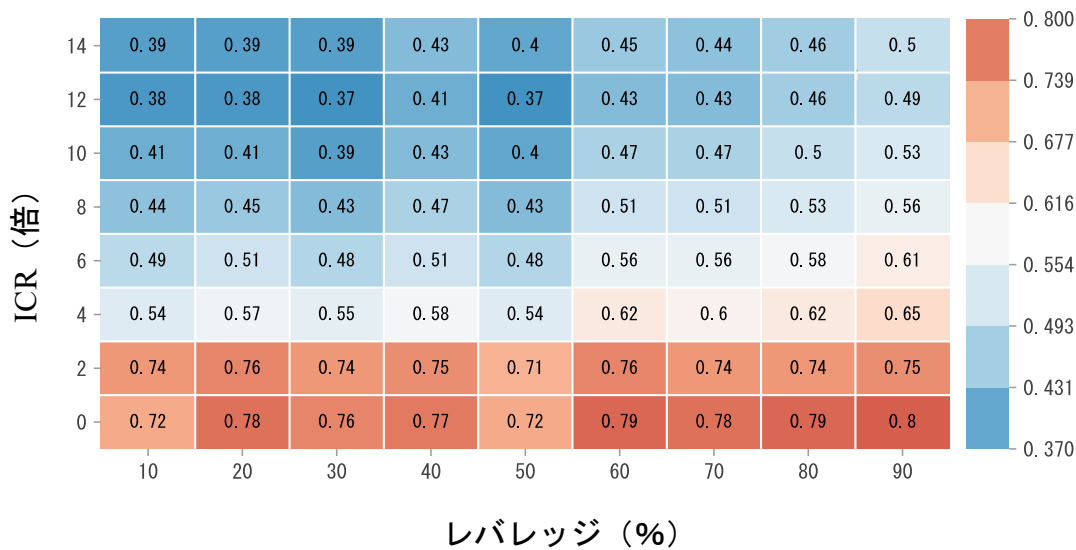
(注) テストデータの SHAP 値の散布図。グラフの横軸は各変数の水準、縦軸は SHAP 値 (Non-IG である確率への寄与度) を示す。

4-3. Partial Dependence Plots

PDP は、予測モデル全体において、説明変数と予測値の平均的な関係を可視化するモデルレベルの代表的な評価手法で、説明変数がモデルの予測結果に与える周辺効果を示したものである³⁵。具体的には、全サンプルについて、対象とする説明変数以外の値は所与としたうえで説明変数の水準を変化させ、当該説明変数の水準毎に算出したモデル予測値を単純平均することで算出する。

³⁵ PDP の理論的背景は Friedman (2001)を参照。

図表 16 Partial Dependence Plot (2 変数)



(注) グラフの横軸・縦軸は説明変数の水準、グラフ内の計数はモデルの予測値を示す。
テストデータを用いて算出。

前節の SHAP Dependence Plot では、一つの説明変数と予測値の関係を可視化した
が、PDP を用いることで、複数の説明変数が同時に変化した際の予測値への影響を可
視化できる³⁶。図表 16 では、PDP を用いて、推計した格付け分類モデルについて、
ICR およびレバレッジの組み合わせ毎に、Non-IG である確率を可視化した。

これをみると、図表 15 と同様、ICR (縦軸) が 2 倍以下になった際に、Non-IG で
ある確率が急激に高まるのが分かる。レバレッジ (横軸) については、ICR ほど強
い関係性は確認されないが、レバレッジが高い企業ほど Non-IG である確率が高い。

このように、2 変数の PDP では、ICR とレバレッジが変化した際の予測値への影響
などを可視化できる。変数の組み合わせによっては、一つの変数の挙動を観察するだけ
では、変数間の相互作用を捕捉できないケースも想定されるが、そうしたケースでも、
2 変数の PDP を用いることで説明可能性を高めることができる。

4-4. 格付け分類モデルに機械学習を活用する際の留意点

本節では、これまでの分析結果を踏まえつつ、格付け分類モデルに機械学習を用い
る際の留意点を整理する。

まず、PDP が説明変数間の依存関係を踏まえずに値を算出しているなど、XAI がモ

³⁶ 1 変数の PDP は SHAP Dependence Plot とほぼ同様の結果であったため、本稿では割愛する。

デルの予測結果に与える解釈は、幾つかの重要な仮定に依拠している点には留意が必要である。こうした仮定を把握せずに XAI を機械学習モデルに適用することで、モデルから誤った帰結を得てしまう可能性がある。XAI を用いる際には、当該手法の仮定や前提を把握し検証を行うことが重要と考えられる。

また、本章では説明を簡便にするために IG/Non-IG の 2 クラス問題について XAI を適用したが、多クラス分類問題の場合には、SHAP や PDP はクラスの数だけ作成される。すなわち、5 クラス問題であれば、あるサンプルにおける特定の説明変数の SHAP 値は 5 つ作成される。多クラスの SHAP や PDP の算出にはより多くの時間的コストを要するのみならず、算出結果の解釈が容易でない可能性がある。これに対し、ロジット回帰など従来のパラメトリックな手法は、関数形や分布を明示的に仮定するため、各説明変数が予測値に与える影響を回帰係数から容易に把握できるという点で、安定的に高い説明可能性が担保されているといえる。こうした点を踏まえると、XAI による機械学習モデルの解釈が困難である場合などには、パラメトリックな手法を用いて推計したモデルも併用し、予測精度とモデルの説明可能性を相互補完することも有用である。

5. おわりに

本稿では、格付け分類モデルの推計に機械学習を応用し、従来広く用いられてきた順序ロジット回帰との予測精度の比較を行った。機械学習は、モデルの構造上、説明変数と被説明変数の間の複雑な非線形性を取り込むことが可能である。その結果、デフォルト率の推計に機械学習を応用した先行研究と同様、従来のパラメトリックな手法対比、予測精度の大幅な改善がみられた。

また、SHAP や PDP といった機械学習の説明可能性を高める枠組み (XAI) を用いて、企業の財務指標と信用力の関係について定量的な分析を行った。その結果、ICR やレバレッジ比率などの財務指標と信用力との関係において、非線形性が存在することなどが分かった。こうした枠組みを用いることで、これまで機械学習の課題の一つとして捉えられることが多かった説明可能性の低さに一定程度対処することができることが確認された。

機械学習の説明可能性を高める手法は近年研究が著しく進展している分野であり、今後も新たな指標が開発されることが予想される。金融機関の業務における機械学習の活用が浸透していくもとの、こうした手法の重要性や有用性は今後も高まっていくと考えられる。

参考文献

- [1] 尾木研三 (2017) 「スコアリングモデルの基礎知識—中小企業融資における見方・使い方—」、金融財政事情研究会.
- [2] 金田規靖・木全友則・平木一浩・松栄共紘 (2022) 「SHAPを用いた機械学習モデルの解釈—原油価格の変動要因分析を例に—」、日本銀行金融研究所ファイナンス・ワークショップ「機械学習のファイナンス分析への応用」、2022年11月.
- [3] 小林正人 (2001) 「順序プロビット・モデルのテストと社債格付データへの応用」、金融研究、20(1)、1-18.
- [4] 日本銀行 (2022) 「金融システムレポート (2022年10月号)」、2022年10月.
- [5] 三浦翔・井實康幸・竹川正浩 (2019) 「入出金情報を用いた信用リスク評価—機械学習による実証分析—」、日本銀行ワーキングペーパーシリーズ、No. 19-J-4.
- [6] 森下光之助 (2021) 「機械学習を解釈する技術：予測力と説明力を両立する実践テクニック」、技術評論社.
- [7] 山下智志・三浦翔 (2011) 「信用リスクモデルの予測精度—AR値と評価指標—」、朝倉書店.
- [8] Alonso, R., J. M. Carbó. (2021) "Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation," Banco de España Working Paper Series, No. 2015.
- [9] Alonso, R., J. M. Carbó. (2022) "Measuring the Model Risk-adjusted Performance of Machine Learning Algorithm in Credit Default Prediction," *Financial Innovation*, 8(1), 1-35. <https://doi.org/10.1186/s40854-022-00366-1>
- [10] Araujo, D., G. Bruno, J. Marcucci, R. Schmidt, B. Tissot. (2022) "Machine Learning Applications in Central Banking," IFC Bulletin, 57. Bank for International Settlements.
- [11] Bank of England. (2022) "Machine Learning in UK Financial Services," Bank of England and FCA Joint Report.
- [12] Chawla, N. V., K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. (2002) "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- [13] European Banking Authority. (2021) "EBA Discussion Paper on Machine Learning for IRB Models," EBA Discussion Paper, No.4.
- [14] Friedman, J. H. (2001) "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>

- [15] Goldstein, A., A. Kapelner, J. Bleich, E. Pitkin. (2015) "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics*, 24(1), 44-65. <https://doi.org/10.1080/10618600.2014.907095>
- [16] Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu. (2017) "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Advances in Neural Information Processing Systems*, 30.
- [17] Lundberg, S. M., S. Lee. (2017) "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, December 2017, 4768-4777.
- [18] Lundberg, S. M., G. G. Erion, S. Lee. (2018) "Consistent individualized feature attribution for tree ensembles," *arXiv*. <https://doi.org/10.48550/arXiv.1802.03888>
- [19] McCullagh, P. (1980) "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society*, 42(2), 109-142. <https://www.jstor.org/stable/2984952>
- [20] Molnar, C., (2019) "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," 2nd edition.
- [21] Moody's Investors Service. (2021) "Rating Methodology: Manufacturing."
- [22] Nembrini, S., I. R. König, M. N. Wright. (2018) "The Revival of the Gini Importance?," *Bioinformatics*, 34(21), 3711-3718. <https://doi.org/10.1093/bioinformatics/bty373>
- [23] Ribeiro, M. T., S. Singh, C. Guestrin (2016a) "Model-Agnostic Interpretability of Machine Learning," *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, July 2016. <https://doi.org/10.48550/arXiv.1606.05386>
- [24] Ribeiro, M. T., S. Singh, C. Guestrin (2016b) "Why Should I Trust You?: Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2016, 1135-1144. <https://doi.org/10.1145/2939672.2939778>
- [25] Shapley, L. S. (1953) "A Value for N-person Games," *Contributions to the Theory of Games*, 2, 307-317.
- [26] S&P Global Market Intelligence. (2020) "Machine Learning and Credit Risk Modelling," S&P Global Inc.

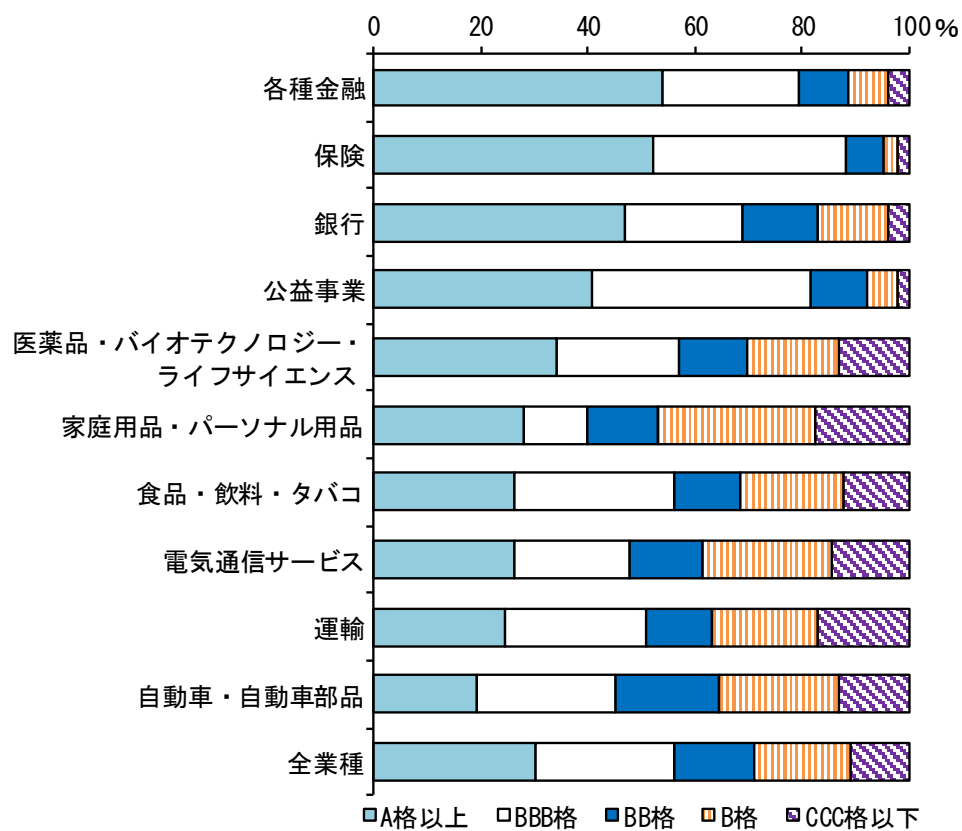
参考図表 1 順序ロジットの回帰結果

	回帰係数	標準誤差	Z値	P値
売上高(対数値)	-0.638	0.004	-150.080	0.000
ICR(倍)	-0.008	0.000	-23.925	0.000
EBITDAマージン(%)	-0.061	0.001	-74.840	0.000
総資産回転率(倍)	0.858	0.012	69.970	0.000
有利子負債・総資産比率(%)	0.035	0.000	117.832	0.000
流動比率(倍)	0.027	0.004	6.049	0.000
業種全体の売上高成長率(%)	-0.020	0.000	-43.967	0.000

	閾値
A格以上 - BBB格	-11.09
BBB格 - BB格	-9.54
BB格 - B格	-8.14
B格 - CCC格以下	-6.54

(注) 順序ロジット回帰の係数表とクラス間の閾値。推計に当たっては、AICをもとに純利益成長率とネットDERを除外したほか、業種固定効果を加えた。

参考図表 2 業種別の格付け構成比



(注) S&P Global Industry Classification Standard の 24 産業グループのうち、A 格以上の比率が高い上位 10 業種と全業種の格付け構成比。