

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7621608号
(P7621608)

(45)発行日 令和7年1月27日(2025.1.27)

(24)登録日 令和7年1月17日(2025.1.17)

(51)国際特許分類

F I

G 0 6 Q 10/0635(2023.01)

G 0 6 Q 10/0635

G 0 6 Q 40/12 (2023.01)

G 0 6 Q 40/12 4 2 0

請求項の数 12 (全37頁)

(21)出願番号	特願2024-55350(P2024-55350)	(73)特許権者	523162982
(22)出願日	令和6年3月29日(2024.3.29)		中島 真澄
(65)公開番号	特開2024-160043(P2024-160043 A)		東京都文京区向丘1丁目19-1 文京学院大学内
(43)公開日	令和6年11月11日(2024.11.11)	(73)特許権者	523169110
審査請求日	令和6年4月5日(2024.4.5)		藤田 邦彦
(31)優先権主張番号	特願2023-75347(P2023-75347)		東京都文京区向丘1丁目19-1 文京学院大学内
(32)優先日	令和5年4月29日(2023.4.29)	(73)特許権者	523162993
(33)優先権主張国・地域又は機関	日本国(JP)		吉田 啓佑
早期審査対象出願			東京都文京区向丘1丁目19-1 文京学院大学内
		(74)代理人	100175064
			弁理士 相澤 聡
		(72)発明者	中島 真澄

最終頁に続く

(54)【発明の名称】 財務諸表不正推定装置、財務諸表不正推定方法及びプログラム

(57)【特許請求の範囲】

【請求項1】

企業又はその経営者が開示するテキスト情報を分析して1以上の説明変数を生成するテキスト情報分析部と、

前記説明変数と、前記企業にかかる財務諸表不正の有無と、の相関関係を示す学習済みモデルを格納したモデル格納部と、

前記テキスト情報分析部が生成した前記説明変数を、前記学習済みモデルに入力し、前記企業にかかる財務諸表不正の有無を推定する不正推定部と、を有し、

前記説明変数として前記テキスト情報の数字率を含む

財務諸表不正推定装置。

10

【請求項2】

前記説明変数として前記テキスト情報の表現力をさらに含む、

前記表現力とは、形容詞の数と副詞の数の和を名詞の数と動詞の数の和で割った値である
請求項1記載の財務諸表不正推定装置。

【請求項3】

前記説明変数として前記テキスト情報の漢字含有率又は学年をさらに含む

請求項1記載の財務諸表不正推定装置。

【請求項4】

前記説明変数として前記テキスト情報の単語数をさらに含む

請求項1記載の財務諸表不正推定装置。

20

【請求項 5】

前記説明変数として前記テキスト情報の表現力、漢字含有率又は学年、単語数をさらに含み、

前記表現力とは、形容詞の数と副詞の数の和を名詞の数と動詞の数の和で割った値である

請求項 1 記載の財務諸表不正推定装置。

【請求項 6】

前記説明変数として前記テキスト情報の単語の多様性及びトーンをさらに含む

請求項 5 記載の財務諸表不正推定装置。

【請求項 7】

前記説明変数と、前記企業にかかる財務諸表不正の有無と、を入力して両者の相関関係を学習し、前記学習済みモデルを生成する機械学習部をさらに有し、

前記機械学習部はアンサンブル学習により前記学習を行う

請求項 1 乃至 6 いずれか 1 項記載の財務諸表不正推定装置。

【請求項 8】

前記機械学習部は勾配ブースティング決定木 (Gradient Boosting Decision Tree) により前記学習を行う

請求項 7 記載の財務諸表不正推定装置。

【請求項 9】

前記機械学習部はランダムフォレストにより前記学習を行う

請求項 7 記載の財務諸表不正推定装置。

【請求項 10】

前記テキスト情報として MD & A (Management ' s Discussion and Analysis of Financial Condition and Results of Operations) を使用する

請求項 1 乃至 6 いずれか 1 項記載の財務諸表不正推定装置。

【請求項 11】

コンピュータにおいて実行される財務諸表不正推定方法であって、

企業又はその経営者が開示するテキスト情報を分析して 1 以上の説明変数を生成するテキスト情報分析ステップと、

前記テキスト情報分析ステップにおいて生成された前記説明変数を、前記説明変数と前記企業にかかる財務諸表不正の有無との相関関係を示す学習済みモデルに入力し、前記企業にかかる財務諸表不正の有無を推定する不正推定ステップと、を有し、

前記説明変数として数字率を含む

財務諸表不正推定方法。

【請求項 12】

請求項 11 記載の方法をコンピュータに実行させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は財務諸表不正推定装置、財務諸表不正推定方法及びプログラムに関し、特に企業又はその経営者が開示するテキスト情報に基づいて財務諸表不正の有無を推定する技術に関する。

【背景技術】

【0002】

財務諸表に関する不正な情報が開示されると、投資者が正しい投資判断ができず、市場が機能しなくなり、公共財が棄損してしまう。投資者が不正企業を検出できない、非不正企業を選択できないことによって市場による適正な分配ができなくなるのである。かかる問題に対処するため、従来より様々な不正検出手法が提案されている。

【0003】

海外では、財務諸表データに対してロジスティック回帰を適用した手法（非特許文献 1

10

20

30

40

50

乃至４）、アニュアルレポートのテキスト情報に対して計量的分析を適用した手法（非特許文献５及び６）がある。

【０００４】

こうした計量統計的手法だけではなく、財務諸表データに対して機械学習を適用した手法として非特許文献７乃至１５がある。また、テキスト情報に対して機械学習を適用した手法として非特許文献１６乃至１９がある。

【０００５】

また、テキストと財務諸表データ両方を用いたハイブリッド型の機械学習手法として非特許文献２０乃至２４がある。

【０００６】

日本では、財務諸表不正に対してロジスティック回帰分析を適用した手法として非特許文献２５乃至２７がある。また、テキスト情報に対して統計的分析を適用した手法として非特許文献２８乃至非特許文献３０がある。

【０００７】

また、財務諸表データに対して機械学習を適用した手法として非特許文献３１がある。

【０００８】

特許文献１には、機械学習により、企業の財務諸表に含まれる複数の勘定科目に基づいて、異常な勘定科目を検出する装置が記載されている。

【０００９】

特許文献２には、機械学習により、個々の企業の財務諸表に含まれる各勘定科目や取引情報などに基づいて、複数の企業間で行われる異常な取引を検知する装置が記載されている。

【先行技術文献】

【非特許文献】

【００１０】

【文献】Beasley, M.1996. An empirical analysis of the relation between the board of director composition and financial statement fraud. The Accounting Review, 71 (4): 443-465.

【文献】Beneath, M.D. 1999. Incentives and penalties related to earnings over statements that violate GAAP, and post-Sarbanes Oxley period. The Accounting Review, 83 (3): 757-787.

【文献】Dechow, P. M., R. G. Sloan, and A. P. Sweeney. 1996. Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC. Contemporary Accounting Research 13 (1): 1-36.

【文献】Skousen, C. J, K. R. Smith, and C. J. Wright. 2009. Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99, in M. Hirschey, K., John, A. K. Makhija (ed.) Corporate Governance and Firm Performance (Advances in Financial Economics, 13), Emerald Group Publishing Limited, 53-A81.

【文献】Churyk, N. T., C. C. Lee, and B. D. Clinton. 2009. Early detection of fraud: Evidence from restatements. Advances in Accounting Behavioral Research 12: 25-40.

【文献】Lee, C.C., N. T. Churyk, and B. D. Clinton. 2013. Validating early fraud prediction using narrative disclosure. Journal of Forensic & Investigative Accounting, 5 (1): January-June.35-57.

【文献】Fanning, K.M, and K. O. Cogger.1998. Neural network detection of management fraud using published financial data, Intelligent Systems in Accounting, Finance and Management 7(1): 21-41.

【文献】Green, B.P., and J.H. Choi. 1997. Assessing the risk of management fraud through neural-network technology. Auditing: A Journal of Practice & The

10

20

30

40

50

ory 16:14-28.

【文献】Hoogs, B., T. Keihl, C. Lacombe, and D. Senturk. 2007. A genetic algorithm approach to detection temporal patterns indicative of financial statement fraud. *International Journal of Intelligent Systems in Accounting, Finance and Management* 15: 41-56.

【文献】Kirks, E., C. Spanthis, and Y. Manolopoulos. 2007. Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications* 32: 995-1003.

【文献】Ravisankar, P., V. Ravi., G.R. Rao, I. Bose. 2011. Detection of financial statement fraud and feature selection using data mining techniques, *Decision Support System* 50: 491-500.

10

【文献】Huang, S.Y. 2013. Fraud detection model by using support vector machine techniques. *International Journal of Digital Content Technology and its Applications* 7:32-42.

【文献】Omar, N. Z. A. Johari, M. Smith. 2017. Predicting fraudulent financial reporting using artificial neural network. *Journal of Financial Crime* 24(2):362-387.

【文献】Bao, Y., B. Ke, B. Li, Y. J. Yu, and J. Zhang. 2020. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research* 58 (1): 199-235. <https://doi.org/10.1111/1475-679X.12292>

20

【文献】EL-Bannany, M. A.H. Dehghan, and A.M. Khedr 2021. Prediction of financial statement fraud using machine learning techniques in UAE. 2021 18th International Multi-Conference on Systems, Signals & Devices.

【文献】Humpherys, S.L. K.C. Moffitt, M.B. Burns, J.K. Burgoon, W. F. Felix. 2011. Identification of fraudulent financial statements using linguistic credibility analysis, *Decision Support Systems* 50: 585-594.

【文献】Goel, S., J. Gongolly., S.R.Faerman. and O. Uzner. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting* 7: 25-46 .

30

【文献】Glancy, F.H. and S.B. Yadav. 2011. A computational model for financial reporting fraud detection, *Decision Support Systems* 50: 595-601.

【文献】Goel, S., and O.Uzuner. 2016. Do Sentiments Matter in Fraud Detection? Estimating Semantic Orientation of Annual Reports. *Special Issue in Accounting, Auditing and Finance* 23 (3) : 215-239

【文献】Brown, N. C., R. M. Crowley, and W. B. Elliott. 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research* 58 (1): 237-291.

【文献】Cecchini, M., H. Aytug, G.J. Koehler, P. Pathak. 2010. Making words work: Using financial text as a predictor of financial events, *Decision Support Systems* 50:164-175.

40

【文献】Chen, Y-J., C-H. Wu., Y-M, Chen, H-K Chen. 2017. Enhancement of fraud detection for narratives in annual reports. *International Journal of Accounting Information Systems* 26: 32-45.

【文献】Hajek, P. and R. Henriques. 2017. Mining corporate annual reports for intelligent detection of financial statement fraud -A comparative study of machine learning methods. *Knowledge-Based Systems* 128: 139-152.

【文献】Gleichman, T.C. 2022. The Detection of Fraudulent Financial Statements Using Textual and Financial data, *Techinish Univesitat Ilmenau, Dissertation.*

【文献】Song. M., N. Oshiro, and A. Shuto. 2016. Predicting accounting fraud:

50

Evidence from Japan, The Japanese Accounting Review 6: 17-63.

【文献】Nakashima, M. 2021. Can the Fraud Triangle Explain Fraudulent Financial Statements? Evidence from Japan, Journal of Forensic and Investigative Accounting 13 (1): 198-232.

【文献】佐藤夏輝・小村亜唯子・平井裕久．「事前不正確率の違いによるロジスティック回帰を用いた不正会計の見地モデルの検知精度」『高崎経済大学論集』64 (4):107-128．

【文献】Nakashima, M., H. Hirose, and H. Hirai. 2022. Fraud detection by focusing on readability Evidence from Japan. Journal of Forensic and Investigative Accounting 13 (1): 276-298.

【文献】Nakashima, M. 2022. Fraud Detection Method by Textual Analysis of CEO Letters in the Perspective of Obfuscation Hypothesis: Evidence from Japanese Firms Listed on the U.S. Stock Exchange . 『経営論集』 32 (1): 83-99.

【文献】中島真澄．2022．「CEOレターのトーンと財務ファンダメンタルズからの不正検出—米国上場日本企業における実証」『会計監査ジャーナル』19：104 - 114．

【文献】宇宿 哲平・近藤 聡・白木 研吾・菅 美希・宮川 大介．2019．「機械学習手法を用いた不正会計の検知と予測」『RIETI Discussion Paper Series』19-J-039

【文献】白田由香利，「機械学習回帰におけるShapley値の理論説明と事例紹介」，DEIM Forum 2022，2022年3月，[online]，インターネット URL: https://event.dbsj.org/deim2022/post/tutorial/deim2022_tutorial_T2.pdf，（検索日：2024年1月4日）

【文献】吉田秀穂ほか，「決定木ベースモデルの解釈におけるSHAP値の有用性の検証」，The 34th Annual Conference of the Japanese Society for Artificial Intelligence 2020，2020年6月，[online]，インターネット URL: https://www.jstage.jst.go.jp/article/pjsai/JSAI2020/0/JSAI2020_3E5GS204/_pdf/-char/ja，（検索日：2024年1月4日）

【特許文献】

【0011】

【文献】特許第6667865号

【文献】特許第7146218号

【発明の概要】

【発明が解決しようとする課題】

【0012】

しかしながら、財務諸表データを用いる手法（ハイブリッド型を含む）は、財務諸表不正の検出という目的を必ずしも十分に達成することができない。財務諸表不正の性質上、不正検出の際において財務諸表上の不正の兆候を識別することは困難であり、不正の兆候の一部は、不正が存在していなくとも、示されることもある。財務諸表が一般に認められた会計原則（Generally Accepted Accounting Principles：GAAP）から逸脱していても逸脱が適切に開示されている場合は不正の存在と言われるものでないし、一方、財務諸表がGAAPに準拠していても、その原則が完全でないことから不正という誤解を招くこともある。財務諸表データからの財務比率だけで不正検出するのは不十分である。

【0013】

また、過年度に財務諸表の虚偽記載などの不正が検出された場合、企業は訂正報告書が義務付けられている（金融商品取引法第24条2）。その訂正報告書が提出されると、有価証券報告書データベースのうち、財務諸表データは訂正報告書の数字で上書きされる。

【0014】

一方、企業またはその経営者が開示するテキスト情報は、定量的な情報に対して優位性があり、定量的な情報を補完する情報が含まれている。テキストの部分は、財務諸表に記載された情報以外の有用な情報あるいは監査済財務諸表には含まれない（非財務的な）説明や解釈を含んでいるため、融資や投資を決定する際に重要な役割を果たす可能性がある。

【 0 0 1 5 】

また、不正のトライアングル理論の合理化ファクターとして、経営者のテキスト情報には経営者の価値観や信条が直接的反映されているので、テキスト情報から不正企業経営者に共通する特性を発見できる可能性が高い。また、テキスト分析は、テキスト情報が経営者によってどのように作られたかを考察することを容易にし、企業開示の属性とその背後にある経営者や企業の特性との相互作用の分析を可能にする。不正を予測するためには、このような経営者の意思決定プロセスに対する視点が重要である。

【 0 0 1 6 】

また、財務諸表データとは異なり、記述（テキスト）情報に関しては訂正報告書が提出されたとしても変更されずそのままデータベース上に開示されている。したがって、記述情報の場合は、財務諸表不正が反映されたままの状態が維持されるため、記述（テキスト）情報による不正検出は性能分析が安定的である。

10

【 0 0 1 7 】

そこで、テキスト情報に焦点を合わせた不正検出手法の開発が望まれる。特に、財務諸表不正を検出することを目的とした場合、どのようなテキスト情報を対象として、どのような説明変数を生成し、どのようなモデルで推定を行うのが好適であるかは未だ明らかにされていない。

【 0 0 1 8 】

本発明はこれらの問題を解決するためになされたものであり、企業又はその経営者が開示するテキスト情報に基づいて財務諸表不正の有無を推定する財務諸表不正推定装置、財務諸表不正推定方法及びプログラムを提供することを目的とする。

20

【 0 0 1 9 】

本発明の一実施形態において、財務諸表不正推定装置は、企業又はその経営者が開示するテキスト情報を分析して1以上の説明変数を生成するテキスト情報分析部と、前記説明変数と、前記企業にかかる財務諸表不正の有無と、の相関関係を示す学習済みモデル又は統計モデルを格納したモデル格納部と、前記テキスト情報分析部が生成した前記説明変数を、前記学習済みモデル又は前記統計モデルに入力し、前記企業にかかる財務諸表不正の有無を推定する不正推定部と、を有し、前記説明変数として前記テキスト情報の数字率を含む。

本発明の一実施形態において、財務諸表不正推定装置は、前記説明変数として前記テキスト情報の表現力をさらに含む。

30

本発明の一実施形態において、財務諸表不正推定装置は、前記説明変数として前記テキスト情報の漢字含有率又は学年をさらに含む。

本発明の一実施形態において、財務諸表不正推定装置は、前記説明変数として前記テキスト情報の単語数をさらに含む。

本発明の一実施形態において、財務諸表不正推定装置は、前記説明変数として前記テキスト情報の表現力、漢字含有率又は学年、単語数をさらに含む。

本発明の一実施形態において、財務諸表不正推定装置は、前記説明変数として前記テキスト情報の単語の多様性及びトーンをさらに含む。

本発明の一実施形態において、財務諸表不正推定装置は、前記説明変数と、前記企業にかかる財務諸表不正の有無と、を入力して両者の相関関係を学習し、前記学習済みモデルを生成する機械学習部をさらに有し、前記機械学習部はアンサンブル学習により前記学習を行う。

40

本発明の一実施形態において、前記機械学習部は勾配ブースティング決定木（Gradient Boosting Decision Tree）により前記学習を行う。

本発明の一実施形態において、前記機械学習部はランダムフォレストにより前記学習を行う。

本発明の一実施形態において、財務諸表不正推定装置は、前記テキスト情報としてMD & A（Management's Discussion and Analysis of Financial Condition and Results of Operat

50

i o n s) を使用する。

本発明の一実施形態において、財務諸表不正推定方法は、コンピュータにおいて実行される財務諸表不正推定方法であって、企業又はその経営者が開示するテキスト情報を分析して1以上の説明変数を生成するテキスト情報分析ステップと、前記テキスト情報分析ステップにおいて生成された前記説明変数を、前記説明変数と前記企業にかかる財務諸表不正の有無との相関関係を示す学習済みモデル又は統計モデルに入力し、前記企業にかかる財務諸表不正の有無を推定する不正推定ステップと、を有し、前記説明変数として数字率を含む。

本発明の一実施形態において、プログラムは、方法をコンピュータに実行させる。

【発明の効果】

10

【0020】

本発明により、企業又はその経営者が開示するテキスト情報に基づいて財務諸表不正の有無を推定する財務諸表不正推定装置、財務諸表不正推定方法及びプログラムを提供することができる。

【図面の簡単な説明】

【0021】

【図1】不適切な会計・経理の開示企業数を示すグラフである。

【図2】財務諸表不正推定装置1の概略的なハードウェア構成を示すブロック図である。

【図3】財務諸表不正推定装置1の概略的な機能構成を示すブロック図である。

【図4】不正企業および非不正企業が開示するテキストの複雑さ(学年)を示す表である。

20

【図5】不正企業群および非不正企業群25社のt検定結果を示す表である。

【図6】財務諸表不正推定装置1の動作例を示すフローチャートである。

【図7】実験に使用した6つの説明変数を示す表である。

【図8】ROC曲線を示す図である。

【図9】AUROCを示す図である。

【図10】実験に使用した6個の説明変数について、不正企業群と非不正企業群との間の有意差の有無を示す表である。

【図11】6つの説明変数(漢字含有率、単語の多様性、数字率、単語数、トーン、表現力)を用いた場合における、学習済みモデルの評価指標を示す図である。

【図12A】決定木を用いた場合のSHAP値の絶対値の平均を示すグラフである。

30

【図12B】決定木を用いた場合のSHAP値の分布を示すグラフである。

【図12C】決定木を用いた場合のSHAP値の分布を示すグラフである。

【図12D】決定木を用いた場合のSHAP値の分布を示すグラフである。

【図12E】決定木を用いた場合のSHAP値の分布を示すグラフである。

【図12F】決定木を用いた場合のSHAP値の分布を示すグラフである。

【図12G】決定木を用いた場合のSHAP値の分布を示すグラフである。

【図13A】ランダムフォレストを用いた場合のSHAP値の絶対値の平均を示すグラフである。

【図13B】ランダムフォレストを用いた場合のSHAP値の分布を示すグラフである。

【図13C】ランダムフォレストを用いた場合のSHAP値の分布を示すグラフである。

40

【図13D】ランダムフォレストを用いた場合のSHAP値の分布を示すグラフである。

【図13E】ランダムフォレストを用いた場合のSHAP値の分布を示すグラフである。

【図13F】ランダムフォレストを用いた場合のSHAP値の分布を示すグラフである。

【図13G】ランダムフォレストを用いた場合のSHAP値の分布を示すグラフである。

【図14A】XGBoostを用いた場合のSHAP値の絶対値の平均を示すグラフである。

【図14B】XGBoostを用いた場合のSHAP値の分布を示すグラフである。

【図14C】XGBoostを用いた場合のSHAP値の分布を示すグラフである。

【図14D】XGBoostを用いた場合のSHAP値の分布を示すグラフである。

【図14E】XGBoostを用いた場合のSHAP値の分布を示すグラフである。

50

【図14F】X G B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図14G】X G B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図15A】L i g h t G B Mを用いた場合のS H A P値の絶対値の平均を示すグラフである。

【図15B】L i g h t G B Mを用いた場合のS H A P値の分布を示すグラフである。

【図15C】L i g h t G B Mを用いた場合のS H A P値の分布を示すグラフである。

【図15D】L i g h t G B Mを用いた場合のS H A P値の分布を示すグラフである。

【図15E】L i g h t G B Mを用いた場合のS H A P値の分布を示すグラフである。

【図15F】L i g h t G B Mを用いた場合のS H A P値の分布を示すグラフである。

【図15G】L i g h t G B Mを用いた場合のS H A P値の分布を示すグラフである。

10

【図16A】C a t B o o s tを用いた場合のS H A P値の絶対値の平均を示すグラフである。

【図16B】C a t B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図16C】C a t B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図16D】C a t B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図16E】C a t B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図16F】C a t B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図16G】C a t B o o s tを用いた場合のS H A P値の分布を示すグラフである。

【図17】4つの説明変数（漢字含有率、数字率、単語数、表現力）を用いた場合における、学習済みモデルの評価指標を示す図である。

20

【図18】2つの説明変数（数字率、単語数）を用いた場合における、学習済みモデルの評価指標を示す図である。

【図19】2つの説明変数（数字率、漢字含有率）を用いた場合における、学習済みモデルの評価指標を示す図である。

【図20】2つの説明変数（数字率、表現力）を用いた場合における、学習済みモデルの評価指標を示す図である。

【図21】1つの説明変数（数字率）のみを用いた場合における、学習済みモデルの評価指標を示す図である。

【図22】数字率の有効性を検証するためのサンプルデータのヒストグラムを示す図である。

30

【図23A】数字率の有効性を検証するための対比実験の結果を示す図である。

【図23B】数字率の有効性を検証するための対比実験の結果を示す図である。

【図23C】数字率の有効性を検証するための対比実験の結果を示す図である。

【図24D】数字率の有効性を検証するための対比実験の結果を示す図である。

【図24A】数字率の有効性を検証するための対比実験の結果を示す図である。

【図24B】数字率の有効性を検証するための対比実験の結果を示す図である。

【図24C】数字率の有効性を検証するための対比実験の結果を示す図である。

【図24D】数字率の有効性を検証するための対比実験の結果を示す図である。

【図24E】数字率の有効性を検証するための対比実験の結果を示す図である。

【図25A】数字率の有効性を検証するための対比実験の結果を示す図である。

40

【図25B】数字率の有効性を検証するための対比実験の結果を示す図である。

【図25C】数字率の有効性を検証するための対比実験の結果を示す図である。

【図25D】数字率の有効性を検証するための対比実験の結果を示す図である。

【図25E】数字率の有効性を検証するための対比実験の結果を示す図である。

【発明を実施するための形態】

【0022】

本発明の実施の形態にかかる財務諸表不正推定装置1は、企業の財務諸表に関する不正の有無を推定する情報処理装置である。ここでいう財務諸表には、連結企業の場合の連結財務諸表、連結子会社をもたない単体の企業の場合の個別財務諸表を含む。財務諸表不正は、「財務諸表をミスリードすることを通して投資者や債権者の信頼を損なう、経営者に

50

よって関与された恣意的な不正」(Elliott and Willingham 1980)や、「企業の財務上の結果を意図的に不正表示することであり、投資者や債権者のような財務諸表利用者をだます目的で財務諸表の開示や金額を省略したり、意図的に不正表示したりすることによって達成されること」(Ravisankar et al. 2011; ELBannay et al. 2021, 650)と定義されている。本実施の形態では、企業または経営者が関与したり、指示したりすることによって投資者や債権者のような財務諸表利用者をミスリードする目的で財務諸表上の開示や金額を意図的に省略したり、不正表示したりすることと定義する。そこには循環取引などによって結果的に財務諸表上の実際の金額と異なる数値になる、売上や利益の架空計上など、財務諸表上に反映された不正も、財務諸表不正に含める。なお、財務諸表不正には、経営者や社員個人による不正、子会社(連結子会社)・関連会社による不正等の、財務諸表に反映されない不正は含まれない。

10

【0023】

図1は、自社開示、金融庁・東京証券取引所などの公表資料に基づいて、東京商工リサーチが、上場企業、有価証券報告書の提出企業を対象に、「不適切な会計・経理」で過年度決算に影響が出た企業、今後影響が出る可能性を開示した企業数を集計したデータである(東京商工リサーチ、2021、2021年全上場企業「不適切な会計・経理の開示企業データ」調査に依拠して出願人が作成)。

【0024】

図1では、当事者を、会社、子会社・関連会社、従業員、役員の4つに分類している。このうち、会社が当事者であるものは、「会社ぐるみの会計不正」を意味し、上の定義でいう財務諸表不正を示している(図1において最も濃色で示されている成分に相当)。なぜなら、会社ぐるみの会計不正だけが、財務諸表不正に反映され、その意図的な財務諸表不正を隠べいするために、経営者はテキスト情報を作成、開示するからである。

20

【0025】

図2は、財務諸表不正推定装置1の概略的なハードウェア構成を示すブロック図である。財務諸表不正推定装置1は、CPU11、揮発性メモリ13、不揮発性メモリ14、インタフェース15、バス20、入出力装置70を有する情報処理装置である。

【0026】

CPU11(Central Processing Unit)は、不揮発性メモリ14に格納されたプログラムをバス20を介して読み出し、プログラムに従った情報処理を実行することにより特有の機能を実現する。

30

【0027】

不揮発性メモリ14は、財務諸表不正推定装置1の電源の状態にかかわらず記憶状態が保持される記憶装置であり、例えばハードディスクやSSD等である。一般に、不揮発性メモリ14に記憶されているプログラムやデータは、プログラム実行時に揮発性メモリ13に展開される。

【0028】

揮発性メモリ13には、不揮発性メモリ14から展開されたプログラムやデータをはじめ、一時的な計算データや入出力装置70を介して入力又は出力されるデータ等が格納される記憶装置である。

40

【0029】

入出力装置70はディスプレイ等のデータ出力装置、キーボードやポインティングデバイス等のデータ入力装置、外部との通信を制御する通信インタフェース等を含む。CPU11から出力された表示データは、インタフェース15を介してディスプレイに表示される。キーボードから入力された指令やデータは、インタフェース15を介してCPU11に渡される。通信インタフェースはCPU11が出力する送信データをインタフェース15により取得し、外部に対して出力する。また通信インタフェースは外部より受信データを取得し、インタフェース15を介してCPU11に引き渡す。

【0030】

50

図3は、本発明の実施の形態にかかる財務諸表不正推定装置1の概略的な機能構成を示すブロック図である。財務諸表不正推定装置1は、テキスト情報入力部101、テキスト情報分析部102、不正推定部103を有する。また、必要に応じ機械学習部104を有する。

【0031】

テキスト情報入力部101は、企業またはその経営者が開示するテキスト情報を取得する。以下にテキスト情報の一例を示すが、本発明はこれらの例に限定されるものでない。

【0032】

テキスト情報として特に好適なのはMD&A (management's discussion and analysis of financial condition and results of operations) である。有価証券報告書内「経営者による財政状態及び経営成績の検討と分析」において、経営者により開示されるこのテキスト情報には、経営成績に重要な影響を与える要因についての分析や、資本の財務および資金の流動性にかかわる財源などの説明が含まれる。

【0033】

なぜMD&A開示に焦点を合わせるのかというと、第1に、金融庁が内容に関する指針を出しており、上場企業にとって開示が義務付けられた情報であり、多くの人が財務諸表とともに読んでいるからである。第2に、現在、MD&A開示は外部監査の対象となっていないが、MD&A開示は上場企業の財務実績に関する「インサイダーの視点」を提供するもので、MD&A開示は、経営者の正直性や財務情報との整合性だけでなく、「経営陣の倫理的価値観 (tone at the top)」を評価する上で非常に重要な情報となっているからである。第3に、MD&A作成者 (経営者) が、MD&Aに財務報告不正が反映されていることを知っていた可能性があり、何らかのストレスにさらされ、テキスト (文章) に影響が反映されている。そのため、MD&A開示に焦点を合わせることは有意義である。

【0034】

加えて、MD&Aは形式が予め決まっているので情報の収集及び処理がしやすいこと、裁量が含まれる余地が小さいため定型的な処理に適していることが特徴である。

【0035】

あるいは、テキスト情報はCEOレターであっても良い。CEOレターは、アニュアルレポート、企業ウェブサイト等の電子媒体、又は株主通信等の紙媒体等において開示される投資家向けのメッセージであり、MD&Aに準ずる情報が含まれることが多い。CEOレターは、MD&Aに比べて裁量の余地が大きいため、経営者の考え、態度や価値観がより反映されやすいという特徴がある。

【0036】

あるいは、テキスト情報は有価証券報告書に記載されたリスク情報であっても良い。

【0037】

テキスト情報分析部102は、テキスト情報入力部101に入力されたテキスト情報を分析し、1以上の説明変数を生成する。以下に説明変数の一例を示すが、本発明はこれらに限定されるものではない。

【0038】

財務諸表不正を実施した経営者は、テキスト情報において、財務諸表不正を投資者から隠したいという意図をもって戦略的に財務諸表不正を隠ぺいしようとする。そのような、経営者による財務諸表不正の隠蔽という戦略をあばくための理論として、難解化仮説 (Obfuscation Hypothesis)、情報操作理論 (Information Manipulation Theory)、対人欺瞞理論 (Interpersonal Deception Theory) の3つの理論がある。本実施の形態では、これらの理論に基づいていくつかの説明変数を定義する。

【0039】

(1) 難解化仮説

10

20

30

40

50

言語上の複雑性は情報量と難解化の2つの構成要素に分解できる。このうち情報量は情報の非対称性と負の相関があり、難解化は情報の非対称性と正の相関がある。ある研究によれば、赤字企業は、文章を難解にして業績不振を隠蔽する一方で、情報量を増やす開示を行う傾向がある。難解化仮説 (Obfuscation Hypothesis) とは、経営者が悪いニュースについては隠蔽したり、より前向きな情報を示したりすることによって投資者をミスリードするという仮説である。情報量については、企業または経営者が関与する財務諸表不正の検出が目的であること、Bad news 自体が読みにくいということが別の研究で実証されていることから、本実施の形態では考慮しない。本実施の形態では、経営者の裁量がより働きやすいと考えられる難解化にのみ焦点を合わせる。

【0040】

10

不正企業は非不正企業よりもテキストの複雑性 (学年) が高く、投資者をミスリードさせる目的で難解化させる。例えば、不正なアニュアルレポートには、難解な単語、複雑な構文がより多く使用されていることがわかっている。不正企業は、発言の複雑性を高め、発言内容を難解にする。不正なアニュアルレポートには語彙の多様性が非不正のアニュアルレポートよりも高く、読みにくいという特徴がある。そこで、仮説 H1a 「不正企業経営者は、投資者をミスリードする目的で文章を難解化させるので、漢字含有率が高く、単語の多様性が高い。」を設定する。

【0041】

仮説 H1a に基づき、テキスト情報分析部 102 は以下の説明変数を生成できる。

・漢字含有率：テキスト情報に出現する漢字の数を、カタカナの数、ひらがなの数、漢字の数の和で割った値。

20

・単語の多様性：テキスト情報に出現する単語の種類の数を単語数で割った値。ここで単語の種類の数とは、重複を除いた単語数のことである。

・学年：テキストの複雑さを示す「学年 (Grade)」は、“世の中に存在する日本語文章の難易度分布の中で、ある文章がどのような位置にあるのか”を示す指標である (Sato 2008)。佐藤 (2011) に依拠して、OB I ソフトウェアで、バランスコーパスに基づく相対的な難易度を、日本の学校の学年に対応する13段階に分けて、複雑度を測定する方法を用いている。すなわち、小学校1年次生 (複雑さレベル1) から学年高等学校3年次生 (複雑さレベル12)、さらに大学 (複雑さレベル13) までの13段階のうちどれかの学年で表示される (Sato 2008)。

30

【0042】

図4は、不正企業および非不正企業それぞれのテキストについて、複雑さを示す指標である学年を測定した結果である。Nakashima (2023) の研究では、不正企業のテキストの複雑性は、非不正企業のテキストの複雑性よりも高いことがうかがえる。学年 (Grade) は、漢字含有率と相関が高いと思われ、難解化仮説に含まれることとなる。すなわち、漢字含有率の代替として、学年を説明変数として使用することが可能である。

【0043】

(2) 情報操作理論

情報操作理論によれば、受容者をミスリードする目的で提示する情報の量 (情報量のコントロール)、質 (情報の歪曲)、方法 (曖昧なスタイルでの提示)、関連性 (無関係な情報) を操作することによって、情報開示の一部または全部を利用することが可能である。不正検出研究として可読性を考察した研究 Nakashima et al. (2022) によれば、不正企業と非不正企業間で長さ (文字数) に有意差は観察されていない。不正企業は、財務諸表数値に虚偽があり、その虚偽、特に数字や単語数を隠ぺいすることが予想できる。そこで、仮説 H1b 「不正企業経営者は、虚偽の数字を示すことができず、隠蔽するので、非不正企業よりも数字率および単語数が低くなる。」を設定する。

40

【0044】

仮説 H1b に基づき、テキスト情報分析部 102 は以下の説明変数を生成できる。

・数字率：テキスト情報に出現する全ての文字に占める数字の割合。

50

・単語数：テキスト情報を単語に分割した際の単語の数。

【 0 0 4 5 】

(3) 対人欺瞞理論

ある研究によれば、不正な M D & A は、真実の M D & A に比べ、平均して 3 倍以上のポジティブなセンチメントと 4 倍以上のネガティブなセンチメントを含んでおり、このことは、不正な M D & A では、ポジティブとネガティブの両方のセンチメントがより顕著に使用されていることを示唆している。不正企業はより印象管理に従事し、だましてよりポジティブに企業を描写する。実際に、不正企業経営者はポジティブニュースを大げさにしたり、ネガティブなニュースを最小化したり隠蔽したりして誤った印象を伝えていることが発見されている。

10

【 0 0 4 6 】

対人欺瞞理論によれば、不正実行者は、特定性 (S p e c i f i c i t y) を低下させ、包括的な用語を用いることによって、文書に曖昧さを加え、責任を拡散しようとする。したがって、不正企業経営者は、M D & A を曖昧にし、責任追及から逃れようと、表現力を低くすることが予想できる。そこで、仮説 H 1 c 「不正企業経営者は、ポジティブな用語が少なく、ネガティブな用語が多く、また、表現力が低い。」を設定する。

【 0 0 4 7 】

仮説 H 1 c に基づき、テキスト情報分析部 1 0 2 は以下の説明変数を生成できる。

・トーン：テキスト情報に出現するポジティブ用語の数からネガティブ用語の数を引いたものを、ポジティブ用語の数とネガティブ用語の数の和で割った値。ここで、ポジティブ用語及びネガティブ用語の数は以下の手順で測定できる。まず、テキスト情報としての日本語の文章を単語ごとに切断してから、品詞を判断する。そして、辞書を参照し、ポジティブ用語、ネガティブ用語に相当する単語が文章中にいくつあるのかを算定する。辞書は、公知の単語感情極性対応表を使用できる。

20

・表現力：形容詞の数と副詞の数の和を名詞の数と動詞の数の和で割った値。

【 0 0 4 8 】

不正推定部 1 0 3 は、テキスト情報分析部 1 0 2 が生成した 1 以上の説明変数を取得し、これをモデル格納部 1 0 3 1 に格納された推定モデルに入力する。推定モデルは、説明変数と財務諸表不正の有無との相関関係を示す学習済みモデル 1 0 3 1 1 又は統計モデル 1 0 3 1 2 である。推定モデルは、説明変数を入力すると、予め定義された相関関係に照らし、財務諸表不正の有無を示す目的変数を出力する。本実施の形態では、目的変数は 2 値 (不正あり又はなし) とした。なお、本発明はこれに限定されるものではなく、例えば財務諸表に不正がある可能性を示す数値等を目的変数としても良い。

30

【 0 0 4 9 】

不正推定部 1 0 3 は、統計モデル 1 0 3 1 2 を用いて推定を行う場合は、ロジスティック回帰分析をはじめとする公知の統計分析手法を採用することができる。

【 0 0 5 0 】

一方、学習済みモデル 1 0 3 1 1 を用いて推定を行う場合、典型的には機械学習部 1 0 4 が予め学習済みモデルを生成する必要がある。

【 0 0 5 1 】

40

機械学習部 1 0 4 は、任意の機械学習アルゴリズムに従って、テキスト情報分析部 1 0 2 が生成した 1 以上の説明変数と、財務諸表不正の有無を示す目的変数との相関関係を学習する。機械学習部 1 0 4 は、大量の学習用データ (テキスト情報分析部 1 0 2 が生成した 1 以上の説明変数と、財務諸表不正の有無を示す目的変数とのセット) を繰り返し入力し、学習過程を繰り返し実行する。このように学習過程を繰り返し実行することにより、1 以上の説明変数と、財務諸表不正の有無を示す目的変数との相関性を示す学習済みモデルが構築される。学習済みモデルが示す相関性は、学習過程が進むにつれて徐々に信頼性を増してゆく。十分に信頼できる水準の学習済みモデルが構築されると、その学習モデルは、1 以上の説明変数に対応する目的変数の値を推定するために使用できるものとなる。

【 0 0 5 2 】

50

不正推定部 103 は、機械学習部 104 が構築した学習済みモデルに基づき、1 以上の説明変数に対応する目的変数すなわち財務諸表不正の有無を自動的かつ正確に求めることができる。すなわち不正推定部 103 は、テキスト情報分析部 102 が生成した 1 以上の説明変数を学習済みモデルに入力として与えることにより、学習済みモデルが財務諸表不正の有無を示す目的変数を自動的かつ正確に出力する。

【0053】

本実施の形態では、機械学習手法として、コンペティション等で多く用いられるアンサンブル型学習手法を採用した。具体的には、ランダムフォレスト、X G B o o s t、L i g h t G B M、C a t B o o s t が挙げられる。また、これらの手法に対する対比のため、決定木分析も行った。以降、目的変数は 0 又は 1 の 2 値であることを前提に説明する。

10

【0054】

(1) 決定木分析(分類)

単一の木構造を用いた教師あり機械学習による分析の総称である。本実施の形態で用いた他の手法(ランダムフォレスト、X G B o o s t、L i g h t G B M、C a t B o o s t)と比較して、得られるモデルの精度では劣る傾向がある。しかしながら、モデルの木構造の可視化の点において他手法に勝る。

【0055】

本実施の形態では、s c i k i t - l e a r n に標準的に実装されている決定木分析手法の一種である C A R T (C l a s s i f i c a t i o n a n d R e g r e s s i o n T r e e s) を用いた。C A R T は、各分岐点でデータを二分割する木(二分木)により分析を行う手法である。

20

【0056】

(2) ランダムフォレスト

アンサンブル型学習手法の一種である。元の訓練データから重複を許しつつランダムにいくつかのデータを抽出し、新しい訓練データを構築する。その後、構築された新しい訓練データを元に決定木分析をおこなう。上記の操作を並列に行うことで複数のモデル(弱学習器)を得る。目的変数の推定の際には、複数の弱学習器から得られる出力値に基づいた多数決を行う。より多くの弱学習器が出力した推定結果を最終的な出力とする。この方法により、単一のモデルに比べ、過学習の危険性の軽減が期待される。

【0057】

30

(3) X G B o o s t

以降の手法もランダムフォレストと同様にアンサンブル型学習手法である。各弱学習器が木の構造を持っている点においてもランダムフォレストと類似する。しかし、弱学習器の構成(決定)法・集約法は大きく異なる。ランダムフォレストにおいては、過学習の抑制が目的であったが、以降の手法の主な目的は予測精度の向上である。一方で、以降説明する手法は決定木分析やランダムフォレストよりも、過学習が生じやすい手法である。

【0058】

X G B o o s t は目的変数が 1 である確率を各弱学習器の推定値の総和を用いて推定する(厳密には、学習率と呼ばれるハイパーパラメータと各学習器の推定値の積の総和を用いるが、本説明では、学習率は常に 1 であると仮定する)。各学習器は逐次的に構築する。1 番目の弱学習器は目的変数が 1 である確率を推定するが、n 番目(n は 2 以上の整数)の弱学習器は(n - 1) 番目までの弱学習器の推定値の総和と真の値の間の誤差を推定する。木の構造を持つ各弱学習器における葉の分割については、情報利得を基準とする。ただし情報利得として、決定木分析においてよく用いられるエントロピーやジニ係数とは異なるものを用いる。詳細は割愛するが、真の値との誤差の最小化を達成するような情報利得を用いている。なお、葉の分割を検討する順番は l e v e l - w i s e t r e e g r o w t h に則る。すなわち、同じ深さの全ての葉について分割を検討しきった後に、1 つ深い枝の分割の検討に移るといった手順をとる。

40

【0059】

(4) L i g h t G B M

50

L i g h t G B Mは、学習アルゴリズムにおいて、X G B o o s tと類似する手法である。目的変数が1である確率を、木の構造を持つ弱学習器たちの推定値の総和で推定する点において、X G B o o s tと変わらない。X G B o o s tとの主な違いは各回帰木の構成方法にある。

【0060】

一点目の違いは各特徴量をヒストグラム化している点である。ヒストグラム化により厳密な枝分かれを探さずに済み、計算コストを抑えることが可能である。二点目の違いは「level-wise tree growth」ではなく「leaf-wise tree growth」により、木の構造を決定している点である。両者の違いは葉の分割を検討する順番にある。なお、木を完全に成長させた場合には同じ木が構成される。Leaf-wise tree growthは、情報利得の最も大きくなる枝から順に伸ばす構成方法である。leaf-wise tree growthを用いる場合は、「pre-pruning」と呼ばれる剪定アルゴリズムを適用することができる。pre-pruningとはさらなる分岐を行うことで予測誤差が下がるなら分岐を行い、そうではないのならそこで木の成長を止めるアルゴリズムである。Pre-pruningはpost-pruningに比較して精度が落ちる可能性があるものの、計算コストを抑えることができる。なお、post-pruningとは決定木を完全に成長させてから、最も予測誤差の小さい状態を選択する剪定方法である。

10

【0061】

ヒストグラム化、leaf-wise tree growth及びpre-pruningにより、lightGBMの計算コストはXGBoostよりも「軽く」なる傾向がある。ただし、leaf-wise tree growthによる決定木は複雑になりやすく、過学習には注意が必要である。

20

【0062】

(5) CatBoost

XGBoostやLightGBMと同様に、CatBoostも木の構造を持つ弱学習器たちの推定値の総和によって目的変数が1である確率を推定する手法である。勾配ブースティング決定木(Gradient Boosting Decision Tree)の一種である。各弱学習器を構築する際に、訓練データからランダムにサンプリングしたデータを用いるため、XGBoostやLightGBMよりも過学習を起こしにくい手法とされる。

30

【0063】

<実験>

本実施の形態にかかる財務諸表不正推定装置1を使用し、実在する企業が開示したテキスト情報を対象として、財務諸表不正の有無を推定する実験を行なった。その手法と結果を以下に示す。

【0064】

(1) サンプル選択とデータ

分析対象企業は、企業はすべて日本取引所に上場している企業3483社のうち、銀行業・保険業(165社)、米国基準およびIFRS基準その他基準(320社)を除外した、日本基準を適用している企業2998社である。不正企業群は、東京商工リサーチ(株)による『2020年不適切な会計・経理』(以下、不適切会計)を開示した上場企業』の2019年に不適切会計とされた会社73社のうち、会社ぐるみの不正31社、従業員関連の不正が12社、上級経営者個人の不正が3社、子会社関連不正が27社である。会社ぐるみの不正31社のうち、6社が上場廃止企業であり、最終サンプルは25社である。業種と総資産および売上高の類似した企業群をペアの非不正企業としている。図5は、不正企業群25社、非不正企業群25社のt検定結果を示している。図5から、不正企業、非不正企業の総資産、売上高に有意差がなく、ペアサンプルとして妥当であることがわかる。

40

【0065】

50

当該不正企業群、非不正企業群のMD & A開示データは、2013年から2019年の7年分のMD & Aを含めている。過去7年分のMD & A開示で分析するのは、東芝やオリンパスの不正事例から発覚した年度から遡って最短7年間は不正に関与しているからである。

【0066】

(2) 分析手法

(2-1) 機械学習による財務諸表不正推定の試行

図6に示すフローチャートに従って、財務諸表推定装置1を使用し、機械学習による不正検知モデルの生成及び当該モデルによる財務諸表不正推定を試行した。機械学習手法としては決定木、ランダムフォレスト、XGBoost、LightGBM、CatBoostを用いた。

10

【0067】

(2-1-1) 学習フェーズ

S101: テキスト情報の取得

テキスト情報入力部101が、企業またはその経営者が開示するテキスト情報を取得する。本実験では、(1) サンプル選択とデータにおいて示した企業50社の2013年度から2018年度までの6年分のMD & Aを学習用のテキスト情報として収集、入力した。

【0068】

S102: 説明変数の生成

テキスト情報分析部102は、テキスト情報入力部101に入力されたテキスト情報を分析し、説明変数を生成する。本実験では、図7に示す6つの説明変数を生成した。

20

【0069】

S103: 学習済みモデルの生成

機械学習部104は、S102で生成された6つの説明変数と、財務諸表不正の有無を示す目的変数(2値)と、のセットを学習用データとして繰り返し入力し、学習済みモデルを生成する。本実験では、機械学習手法として決定木、ランダムフォレスト、XGBoost、LightGBM、CatBoostの5種類を採用し、それぞれ学習済みモデルを生成した。

【0070】

(2-1-2) 推定フェーズ

S104: テキスト情報の取得

テキスト情報入力部101が、企業またはその経営者が開示するテキスト情報を取得する。本実験では、(1) サンプル選択とデータにおいて示した企業50社の2019年度のMD & Aを推定用のテキスト情報として収集、入力した。

30

【0071】

S105: 説明変数の生成

テキスト情報分析部102は、テキスト情報入力部101に入力されたテキスト情報を分析し、説明変数を生成する。本実験では、図7に示す6つの説明変数を生成した。

【0072】

S106: 財務諸表不正の有無の推定

不正推定部103は、S105で生成された6つの説明変数をS103で生成された学習済みモデルに入力し、財務諸表不正の有無を示す目的変数(2値)を推定結果として出力する。

40

【0073】

(2-2) 評価

複数の機械学習手法(決定木、ランダムフォレスト、XGBoost、LightGBM、CatBoost)により生成された分類器の有効性を評価するため、Accuracy、Precision、Recall、Fmeasure、AUROCを算出する。Accuracyは、正しく分類されたMD & A開示の総数を分析されたMD & A開示の総数で除したものである。Precisionは、不正と分類されたMD & A開示の総

50

数に対する、不正と正しく分類されたMD & A開示数の比率である。Recallは、実際の不正のMD & A開示の総数に対する、不正と正しく分類されたMD & A開示の数の比率である。RecallはTrue Positive Rate（真陽性率）とも呼ばれる。以下では、True Positive RateをTPRと略記する。対して、実際の非不正のMD & A開示の総数に対する、不正と誤って分類されたMD & A開示の数の比率をFalse Positive Rate（偽陽性率）と呼ぶ。以下では、False Positive RateをFPRと略記する。

【0074】

Fmeasureは、それぞれの加重調和平均を用いてPrecisionとRecallを1つのメトリックスに統合したものである。数値が高いほど推定性能が高いことを意味する（Humphery et al. 2011, 591）。

10

【0075】

AUROCの説明のために、本実験で用いる機械学習モデルたちの推定方法について述べる。本実験において、機械学習モデルたちは不正を行っている確率をまず算出する。予め閾値（通常は0.5）を設定しておき、確率が閾値以上の場合に不正と分類している。閾値を低くすればするほど、機械学習モデルたちによって不正と分類される傾向が強まる。その結果TPRは高まるものの、FPRも高まる。反対に、閾値を低くすればするほど、FPRは下がるがTPRも下がってしまう。閾値を0から1の間で変化させた際のTPRとFPRの関係をグラフで表したものをROC曲線と呼ぶ（図8、曲線a）。ランダムな分類（例えば、0以上1以下の間の値を取る乱数を不正の確率にした分類）を行った際には、図8の直線bのようなROC曲線になる。TPR = 1かつFPR = 0が成り立つような閾値が存在する場合はその閾値で分類を行うことで、AccuracyもPrecisionもRecallも1になるような理想的な分類を行うことができる。このような閾値が存在するような、理想的な場合は図8の線cのようなROC曲線が得られる。

20

【0076】

AUROC（Area Under the ROC curve）とはROC曲線がどれほど理想的なROC曲線（図8の線c）に近いかを測る尺度である。AUROCは図9において示す領域d（曲線bとTPR = 0に挟まれた領域）の面積で与えられる。ROC曲線が図8、図9の線cに近づくほど、AUROC曲線の値は大きくなり、理想的な分類が行われた際のAUROCは1である。すなわち、AUROC曲線の値が1に近いほど、機械学習モデルの推定性能が高いと示唆される。また、ランダムな分類が行われた場合、AUROCは0.5になる。すなわち、AUROCが0.5を超えていれば、機械学習モデルがランダムな分類より高い推定能力を持つことが示唆される。

30

【0077】

（3）分析結果

（3-1）統計学的検定の結果

本実験で用いる6個の説明変数に対して、不正企業の群と非不正企業の群の間に有意差があるかをマンホイットニーのU検定（有意水準5%）によって調べた。結果を図10に示す。不正企業は、非不正企業と比べて、漢字含有率が高く、単語の多様性が高く、数字率が低く、単語数が低く、トーンが低く、表現力が低いことがわかった。ただし、数字率については有意差が認められなかった。しかしながら、後述の（3-2）モデルの評価により、数字率は本実験の機械学習モデルの推定に強く寄与していることがわかった。

40

【0078】

（3-2）モデルの評価

本実験では、2013年度から2018年度までのデータを訓練データ、2019年度のデータをテストデータとした。また、説明変数として図7に示す6つの説明変数を使用した。

【0079】

決定木、ランダムフォレスト、XGBoost、LightGBM、CatBoostを用いた不正検知モデルの評価指標（Accuracy、Precision、Recall

50

11、F measure、AUROC)はそれぞれ図11のとおりであった。なお、各評価指標の値はテストデータに対して算出したものである。

【0080】

また、これらのモデルのSHAP値はそれぞれ図12乃至図16の通りであった。

【0081】

図12A乃至図12Gは決定木を使って生成したモデルのSHAP値を示す図である。図12Aは、6つの説明変数「数字率」「単語数」「漢字含有率」「表現力」「単語の多様性」「トーン」のSHAP値の絶対値の平均を比較したグラフである。図12B乃至図12Fは、上記6つの説明変数をそれぞれ使用した場合のSHAP値の分布を示したグラフである。図12Bは標準化された「数字率」に対するSHAP値の分布、図12Cは標準化された「単語数」に対するSHAP値の分布、図12Dは標準化された「漢字含有率」に対するSHAP値の分布、図12Eは標準化された「表現力」に対するSHAP値の分布、図12Fは標準化された「単語の多様性」に対するSHAP値の分布、図12Gは標準化された「トーン」に対するSHAP値の分布を示している。

10

【0082】

図13Aはランダムフォレストを使って生成したモデルのSHAP値を示す図である。図13Aは、6つの説明変数「数字率」「単語数」「漢字含有率」「表現力」「単語の多様性」「トーン」のSHAP値の絶対値の平均を比較したグラフである。図13B乃至図13Fは、上記6つの説明変数をそれぞれ使用した場合のSHAP値の分布を示したグラフである。図13Bは標準化された「数字率」に対するSHAP値の分布、図13Cは標準化された「単語数」に対するSHAP値の分布、図13Dは標準化された「漢字含有率」に対するSHAP値の分布、図13Eは標準化された「表現力」に対するSHAP値の分布、図13Fは標準化された「単語の多様性」に対するSHAP値の分布、図13Gは標準化された「トーン」に対するSHAP値の分布を示している。

20

【0083】

図14AはXGBoostを使って生成したモデルのSHAP値を示す図である。図14Aは、6つの説明変数「数字率」「単語数」「漢字含有率」「表現力」「単語の多様性」「トーン」のSHAP値の絶対値の平均を比較したグラフである。図14B乃至図14Fは、上記6つの説明変数をそれぞれ使用した場合のSHAP値の分布を示したグラフである。図14Bは標準化された「数字率」に対するSHAP値の分布、図14Cは標準化された「単語数」に対するSHAP値の分布、図14Dは標準化された「漢字含有率」に対するSHAP値の分布、図14Eは標準化された「表現力」に対するSHAP値の分布、図14Fは標準化された「単語の多様性」に対するSHAP値の分布、図14Gは標準化された「トーン」に対するSHAP値の分布を示している。

30

【0084】

図15AはLightGBMを使って生成したモデルのSHAP値を示す図である。図15Aは、6つの説明変数「数字率」「単語数」「漢字含有率」「表現力」「単語の多様性」「トーン」のSHAP値の絶対値の平均を比較したグラフである。図15B乃至図15Fは、上記6つの説明変数をそれぞれ使用した場合のSHAP値の分布を示したグラフである。図15Bは標準化された「数字率」に対するSHAP値の分布、図15Cは標準化された「単語数」に対するSHAP値の分布、図15Dは標準化された「漢字含有率」に対するSHAP値の分布、図15Eは標準化された「表現力」に対するSHAP値の分布、図15Fは標準化された「単語の多様性」に対するSHAP値の分布、図15Gは標準化された「トーン」に対するSHAP値の分布を示している。

40

【0085】

図16AはCatBoostを使って生成したモデルのSHAP値を示す図である。図16Aは、6つの説明変数「数字率」「単語数」「漢字含有率」「表現力」「単語の多様性」「トーン」のSHAP値の絶対値の平均を比較したグラフである。図16B乃至図16Fは、上記6つの説明変数をそれぞれ使用した場合のSHAP値の分布を示したグラフである。図16Bは標準化された「数字率」に対するSHAP値の分布、図16Cは標準

50

化された「単語数」に対するSHAP値の分布、図16Dは標準化された「漢字含有率」に対するSHAP値の分布、図16Eは標準化された「表現力」に対するSHAP値の分布、図16Fは標準化された「単語の多様性」に対するSHAP値の分布、図16Gは標準化された「トーン」に対するSHAP値の分布を示している。

【0086】

SHAP値は、各機械学習モデルが目的変数を出力するに際し、各説明変数の影響をどれだけ受けたか（寄与度）を示している。SHAP値の絶対値の平均について説明変数間の順位をつけると、「数字率」は5つのモデルのいずれにおいても1位であった。一方、「トーン」と「単語の多様性」は、5つのモデルのいずれについても、4位以下にしかならなかった。図17は、推定にあまり寄与していないと思われるこれらの説明変数（「トーン」及び「単語の多様性」）を除いた場合のモデルの評価指標を示す表である。図17に示すように、決定木を除いた手法については6変数の場合と遜色ない推定精度を示した。

10

【0087】

なお、説明変数の機械学習モデルへの貢献を示す指標としてSHAP値が有用であることは、例えば非特許文献31及び32に示すように当業者にとっての技術常識である。仮に、SHAP値が高くてその説明変数の寄与度が低くなる例外的なケースが存在するとしても、本実験がその例外にあたると推察できるような合理的な理由は見出せない。

【0088】

「トーン」と「単語の多様性」を除いた4つの説明変数のうち、「数字率」とその他の1つの説明変数を用いて、2つの説明変数のみから推定を行った。比較的少ない説明変数であっても多少の推定精度を示した。

20

図18は「数字率」及び「単語数」のみを用いた際のモデルの評価指標である。

図19は「数字率」及び「漢字含有率」のみを用いた際のモデルの評価指標である。

図20は「数字率」及び「表現力」のみを用いた際のモデルの評価指標である。

説明変数として、「数字率」のみを用いた際の評価は図21の通りであった。

【0089】

第1に、機械学習を用いた推定においては、説明変数のうち数字率の寄与度が最も高く、数字率のみを用いた場合でも一定の推定精度が得られることがわかった。従来の財務諸表不正推定手法では、数多くの説明変数を生成して推定に使用するものが主流であった。一方、本発明において新たに導入された説明変数である数字率を使用すれば、1つの説明変数で一定の精度を有する財務諸表不正の推定が可能である。すなわち、数字率という新たな概念を推定処理に使用することによって、従来よりも計算リソースを大幅に抑制しつつも、推定精度を確保することができるようになった。

30

【0090】

また、数字率に、寄与度が比較的高い表現力、漢字含有率、単語数の少なくともいずれか1つを組み合わせることで、数字率のみを用いた場合に比べ推定精度をさらに上げられることがわかった。一方、トーン、単語の多様性の寄与度は比較的低いことがわかった。

【0091】

第2に、不正検知モデル（決定木、ランダムフォレスト、XGBoost、LightGBM、CatBoost）の性能を比較したところ、6変数の場合はランダムフォレストが最も性能が高いことがわかった。なお、ハイパーパラメータをチューニングすればCatBoostの性能が最も高くなる可能性がある。

40

【0092】

このように、（3-1）に示した統計学的検定と、（3-2）に示した機械学習モデルの評価とでは、数字率に対する評価が異なる結果となった。しかしながら、この結果は説明変数としての数字率の有用性を否定するものではない。

【0093】

図22は、サンプルデータにおける非不正企業と不正企業の分布を示すヒストグラムである。淡色は非不正企業の件数、濃色は不正企業の件数を示す。このサンプルデータでは、数字率は0から0.175以下の範囲に分布している。数字率の範囲を細分化した各レ

50

ンジにおける非不正企業、不正企業の件数は0から35件の範囲に分布している。

【0094】

このヒストグラムでは、数字率の増加に対する非不正企業の増減傾向と、不正企業の増減傾向とが異なる。非不正企業では、0.075をピークとして増加から減少に転じる比較的美丽な分布が見られる。一方、不正企業では0.063をピークとして増加から減少に転じた後、0.100では再度増加に転じ、0.125では再度減少に転じるという特徴的な動きが見られる。

【0095】

統計的検定は、比較したい2群の増減傾向を全体的に捉えて比較する。そのため、局所的な増減傾向の相違は捨象されうる。このサンプルデータのように、一方の群が特徴的かつ局所的な増減傾向を有していたとしても、その相違は有意差としては顕現しにくい。

10

【0096】

一方、機械学習は、より局所的に数字率を捉えて推定を行う。例えば数字率0.100の場合と数字率0.113の場合とで異なる推定を行うことが可能である。したがって、このようなサンプルデータを用いた場合であっても比較的正確に判定を行うことができる。

【0097】

本実験において、統計学的検定においては有意差が認められなかったものの、機械学習による数字率を用いた場合に一定の推定精度が得られたのは、このような要因によるものと考えられる。

【0098】

20

図12B、図13B、図14B、図15B及び図16Bによれば、本実験にかかるサンプルデータは、数字率が高いほど不正である可能性は低く、数字率が低いほど不正である可能性が高いものであることがわかる。そして、数字率が中間レンジにある場合は不正、非不正のいずれかである可能性が入り乱れていることがわかる。これは、サンプルデータが上述のヒストグラムのような分布をしている場合にみられる特徴である。このような特徴を有するデータは、統計的検定では有意差が出なくても、機械学習による推定では一定の精度を示しうるケースにあたる。

【0099】

このように、ある説明変数について統計的検定では有意差が出ない場合であっても、機械学習では一定の推定精度が得られる場合がある。上述のように、本実験では、数字率を用いた場合の機械学習モデルのAUROCの値は0.5を超え、ランダムな分類より高い推定能力を持つことを示唆している。また、機械学習モデルが目的変数を出力するに際し、各説明変数の影響をどれだけ受けたか（寄与度）を示すSHAP値は、数字率が最も高い。したがって、本実験において使用したような数字率を説明変数とする機械学習モデルは、財務諸表不正の有無を実用的な精度で推定できると結論づけることができる。

30

【0100】

<対比実験1>

財務諸表不正推定における数字率の有用性をさらに検証するため、数字率を用いずに他の1以上の説明変数を用いた場合の機械学習モデルのAUROC値と、数字率を併用した場合のAUROC値との対比実験を行った（対比実験1）。対比実験1の結果を図23に示す。なお、検証に用いた機械学習アルゴリズムに内在するランダム性により、図17乃至図22において示されている値とは多少の差異がある。

40

【0101】

図23A乃至図23Dにおいて、「変数リスト」は機械学習モデルが用いる説明変数を示している。ここには数字率が含まれていない。

【0102】

図23Aにおいて、「RandomF__AUROC」は、ランダムフォレストを用い、「変数リスト」に示す説明変数（数字率は含まれていない）のみを用いて生成したモデルのAUROC値を示している。「RandomF__AUROCwith数字率」は、同じくランダムフォレストを用い、「変数リスト」に示す説明変数と数字率とを用いて生成し

50

たモデルの A U R O C 値を示している。

【 0 1 0 3 】

図 2 3 B において、「X G B C l a s _ A U R O C」は、X G B o o s t を用い、「変数リスト」に示す説明変数（数字率は含まれていない）のみを用いて生成したモデルの A U R O C 値を示している。「X G B C l a s _ A U R O C w i t h 数字率」は、同じく X G B o o s t を用い、「変数リスト」に示す説明変数と数字率とを用いて生成したモデルの A U R O C 値を示している。

【 0 1 0 4 】

図 2 3 C において、「L G B M C l a _ A U R O C」は、L i g h t G B M を用い、「変数リスト」に示す説明変数（数字率は含まれていない）のみを用いて生成したモデルの A U R O C 値を示している。「L G B M C l a _ A U R O C w i t h 数字率」は、同じく L i g h t G B M を用い、「変数リスト」に示す説明変数と数字率とを用いて生成したモデルの A U R O C 値を示している。

10

【 0 1 0 5 】

図 2 3 D において、「< c a t b o o _ A U R O C」は、C a t B o o s t を用い、「変数リスト」に示す説明変数（数字率は含まれていない）のみを用いて生成したモデルの A U R O C 値を示している。「< c a t b o o _ A U R O C w i t h 数字率」は、同じく C a t B o o s t を用い、「変数リスト」に示す説明変数と数字率とを用いて生成したモデルの A U R O C 値を示している。

【 0 1 0 6 】

20

図 2 3 A 乃至図 2 3 D のいずれにおいても、他の説明変数に数字率を加えると A U R O C 値が増加し、推定精度が向上することがわかる。

【 0 1 0 7 】

< 対比実験 2 >

さらなる検証のため、説明変数として数字率を含むグループと、数字率を含まないグループとの A U R O C 値を比較する対比実験を行った（対比実験 2）。対比実験 2 では、2 つのグループの説明変数の数は同じになるようにした。対比実験 2 の結果を図 2 4 及び図 2 5 に示す。なお、検証に用いた機械学習アルゴリズムに内在するランダム性により、図 1 7 乃至図 2 2 において示されている値とは多少の差異がある。

【 0 1 0 8 】

30

図 2 4 A は、1 つの説明変数を使用した場合において観測された A U R O C 値の一覧である。「変数リスト」は、機械学習モデルの生成に使用された説明変数（6 パターン）を示している。本実験では、6 つの説明変数「トーン」「単語数」「漢字含有率」「表現力」「数字率」「多様性」のうち、いずれか 1 つの説明変数のみを使用した。モデル生成にあたって使用した機械学習手法は、ランダムフォレスト（R a n d o m F _ A U R O C）、X G B o o s t（X G B C l a s _ A U R O C）、L i g h t G B M（L G B M C l a _ A U R O C）、C a t B o o s t（< c a t b o o _ A U R O C）の 4 種類である。これらの組み合わせ（2 4 パターン）についてそれぞれ A U R O C 値を測定した。

【 0 1 0 9 】

図 2 5 A は、図 2 4 A の「変数リスト」に挙げられた 6 パターンを、数字率を含むグループと含まないグループとに分別し、それぞれのグループにおける A U R O C 値の統計量を算出、比較した表である。数字率を含むグループは、数字率を含まないグループと比較して少なくとも A U R O C 値の平均値及び中央値が高い。このことから、全体的な傾向として数字率を含むグループの方が推定精度が高いと評価できる。

40

【 0 1 1 0 】

図 2 4 B は、2 つの説明変数を使用した場合において観測された A U R O C 値の一覧である。「変数リスト」は、機械学習モデルの生成に使用された説明変数の組合せ（1 5 パターン）を示している。本実験では、6 つの説明変数「トーン」「単語数」「漢字含有率」「表現力」「数字率」「多様性」のうち、異なる 2 つの説明変数を組合せて使用した。モデル生成にあたって使用した機械学習手法は、ランダムフォレスト（R a n d o m F _

50

A U R O C)、X G B o o s t (X G B C l a s _ A U R O C)、L i g h t G B M (L G B M C l a _ A U R O C)、C a t B o o s t (< c a t b o o _ A U R O C) の 4 種類である。これらの組み合わせ (6 0 パターン) についてそれぞれ A U R O C 値を測定した。

【 0 1 1 1 】

図 2 5 B は、図 2 4 B の「変数リスト」に挙げられた 1 5 パターンを、数字率を含むグループと含まないグループとに分別し、それぞれのグループにおける A U R O C 値の統計量を算出、比較した表である。数字率を含むグループは、数字率を含まないグループと比較して A U R O C 値の平均値、中央値、最大値のいずれもが大きい。このことから、全体的な傾向としてもピーク性能としても、数字率を含むグループの方が推定精度が高いと評価できる。

10

【 0 1 1 2 】

図 2 4 C は、3 つの説明変数を使用した場合において観測された A U R O C 値の一覧である。「変数リスト」は、機械学習モデルの生成に使用された説明変数の組合せ (2 0 パターン) を示している。本実験では、6 つの説明変数「トーン」「単語数」「漢字含有率」「表現力」「数字率」「多様性」のうち、異なる 3 つの説明変数を組合せて使用した。モデル生成にあたって使用した機械学習手法は、ランダムフォレスト (R a n d o m F _ A U R O C)、X G B o o s t (X G B C l a s _ A U R O C)、L i g h t G B M (L G B M C l a _ A U R O C)、C a t B o o s t (< c a t b o o _ A U R O C) の 4 種類である。これらの組み合わせ (1 2 0 パターン) についてそれぞれ A U R O C 値を測定した。

20

【 0 1 1 3 】

図 2 5 C は、図 2 4 C の「変数リスト」に挙げられた 2 0 パターンを、数字率を含むグループと含まないグループとに分別し、それぞれのグループにおける A U R O C 値の統計量を算出、比較した表である。数字率を含むグループは、数字率を含まないグループと比較して少なくとも A U R O C 値の平均値及び中央値が高い。このことから、全体的な傾向として数字率を含むグループの方が推定精度が高いと評価できる。

【 0 1 1 4 】

図 2 4 D は、4 つの説明変数を使用した場合において観測された A U R O C 値の一覧である。「変数リスト」は、機械学習モデルの生成に使用された説明変数の組合せ (1 5 パターン) を示している。本実験では、6 つの説明変数「トーン」「単語数」「漢字含有率」「表現力」「数字率」「多様性」のうち、異なる 4 つの説明変数を組合せて使用した。モデル生成にあたって使用した機械学習手法は、ランダムフォレスト (R a n d o m F _ A U R O C)、X G B o o s t (X G B C l a s _ A U R O C)、L i g h t G B M (L G B M C l a _ A U R O C)、C a t B o o s t (< c a t b o o _ A U R O C) の 4 種類である。これらの組み合わせ (6 0 パターン) についてそれぞれ A U R O C 値を測定した。

30

【 0 1 1 5 】

図 2 5 D は、図 2 4 D の「変数リスト」に挙げられた 1 5 パターンを、数字率を含むグループと含まないグループとに分別し、それぞれのグループにおける A U R O C 値の統計量を算出、比較した表である。数字率を含むグループは、数字率を含まないグループと比較して A U R O C 値の平均値、中央値、最大値のいずれもが大きい。このことから、全体的な傾向としてもピーク性能としても、数字率を含むグループの方が推定精度が高いと評価できる。

40

【 0 1 1 6 】

図 2 4 E は、5 つの説明変数を使用した場合において観測された A U R O C 値の一覧である。「変数リスト」は、機械学習モデルの生成に使用された説明変数の組合せ (6 パターン) を示している。本実験では、6 つの説明変数「トーン」「単語数」「漢字含有率」「表現力」「数字率」「多様性」のうち、異なる 5 つの説明変数を組合せて使用した。モデル生成にあたって使用した機械学習手法は、ランダムフォレスト (R a n d o m F _ A

50

U R O C)、X G B o o s t (X G B C l a s _ A U R O C)、L i g h t G B M (L G B M C l a _ A U R O C)、C a t B o o s t (< c a t b o o _ A U R O C) の 4 種 類 である。これらの組み合わせ (2 4 パターン) についてそれぞれ A U R O C 値を測定した。

【 0 1 1 7 】

図 2 5 E は、図 2 4 E の「変数リスト」に挙げられた 6 パターンを、数字率を含むグループと含まないグループとに分別し、それぞれのグループにおける A U R O C 値の統計量を算出、比較した表である。数字率を含むグループは、数字率を含まないグループと比較して A U R O C 値の平均値、中央値、最大値のいずれもが大きい。このことから、全体的な傾向としてもピーク性能としても、数字率を含むグループの方が推定精度が高いと評価できる。

10

【 0 1 1 8 】

対比実験 1 及び 2 によれば、数字率を含む説明変数を使用した機械学習モデルにおいては、数字率を含まない場合と比較して、財務諸表不正の有無を高精度にかつ一定の反復可能性をもって推定することができる。

【 0 1 1 9 】

本実施の形態によれば、テキスト情報入力部 1 0 1 が企業またはその経営者が開示する M D & A や C E O レターをはじめとするテキスト情報を取得し、テキスト情報分析部 1 0 2 がテキスト情報を分析して数字率を含む 1 以上の説明変数を生成し、不正推定部 1 0 3 がランダムフォレストや C a t B o o s t をはじめとするメソッドにより推定モデルを構築する。これにより、精度の高い財務諸表不正推定を行うことが可能となるため、投資者の投資損失やリスクを回避するとともに、投資者が非不正企業に投資する場合の投資効果の向上を図ることができる。

20

【 0 1 2 0 】

以上、本発明の実施の形態について説明したが、本発明は上記実施形態に限られたものではなく、趣旨を逸脱しない範囲で適宜変更することが可能である。本発明はその発明の範囲内において、実施の形態の任意の構成要素の変形、もしくは実施の形態の任意の構成要素の省略が可能である。

【 0 1 2 1 】

例えば、上述の実施例ではテキスト情報として日本語のテキストを想定したが、本発明はこれに限定されるものではなく、英語をはじめとする外国語のテキストを対象としても良い。この場合、漢字含有率に代わる説明変数として、例えば単語の長さ (単語を構成する文字数) の平均値を用いることができる。単語の多様性、数字率、単語数、トーン、表現力等の説明変数は日本語の場合と同様に算出できる。なお、トーンを算出する際に参照する辞書については対象言語向けにあらかじめ準備されたものを使用する。

30

【 0 1 2 2 】

また、本発明の情報処理はハードウェアにより実現されても良く、C P U がコンピュータプログラムを実行することにより実現されても良い。コンピュータプログラムは、様々なタイプの非一時的なコンピュータ可読媒体 (n o n - t r a n s i t o r y c o m p u t e r r e a d a b l e m e d i u m) 又は一時的なコンピュータ可読媒体 (t r a n s i t o r y c o m p u t e r r e a d a b l e m e d i u m) によりコンピュータに供給され得る。

40

【符号の説明】

【 0 1 2 3 】

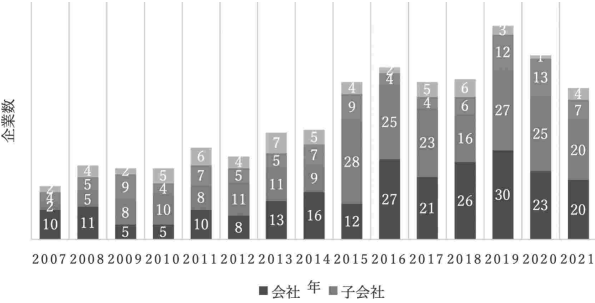
- 1 財務諸表不正推定装置
- 1 1 C P U
- 1 3 揮発性メモリ
- 1 4 不揮発性メモリ
- 1 5 インタフェース
- 1 6 インタフェース
- 2 0 バス

50

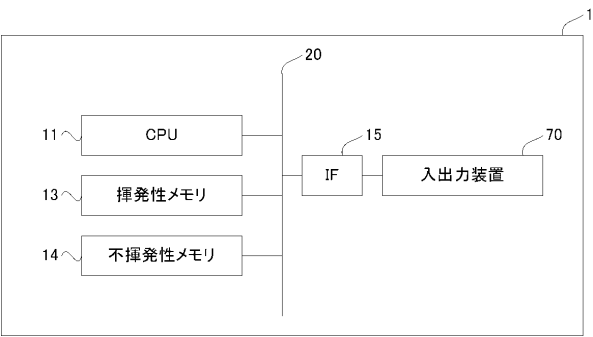
- 7 0 入出力装置
- 1 0 1 テキスト情報入力部
- 1 0 2 テキスト情報分析部
- 1 0 3 不正推定部
- 1 0 3 1 モデル格納部
- 1 0 3 1 1 学習済みモデル
- 1 0 3 1 2 統計モデル
- 1 0 4 機械学習部

【図面】

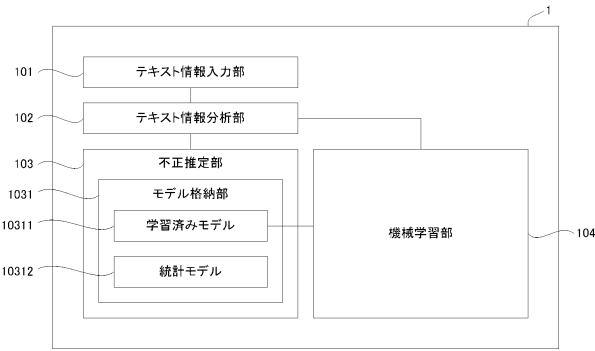
【図 1】



【図 2】



【図 3】



【図 4】

	Textual Complexity (School Grade)				
	In the Pre-COVID19				
	MEAN	MEDIAN	MAX	MIN	S.D.
Non-Fraudulent Firms	12.389	13.000	13.000	8.000	1.607
Fraudulent Firms	12.558	13.000	13.000	8.000	1.378

10

20

30

40

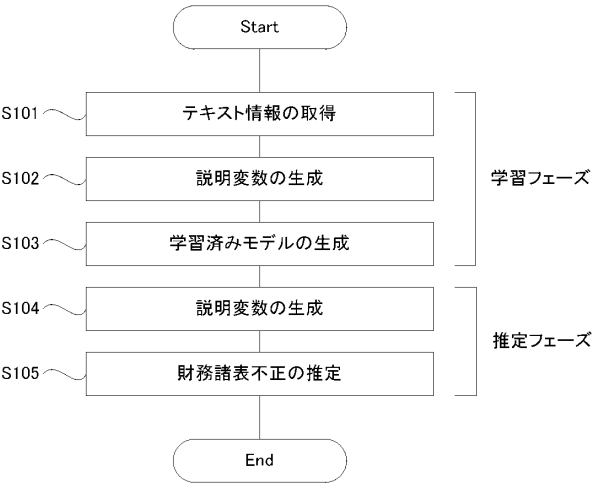
50

【図 5】

	不正企業群		非不正企業群		t-検定	
	平均値	標準偏差	平均値	標準偏差	t value	p value
Total Assets	794,232	3,467,175	718,737	3,396,914	0.205	0.838
OCF	44,517	206,057	41,482	195,923	0.141	0.888
SALES	566,990	2,301,039	574,202	2,590,879	-0.027	0.978
EARN	23,516	105,109	23,797	118,240	-0.023	0.981

変数の定義:
変数 定義
Total Assets Total assets at the ending period, Data from aol database.
OCF Operating cash flows at the ending period, Data from aol database.
SALES Sales at the ending period, Data from aol database.
EARN Net income at the ending period, Data from aol database.

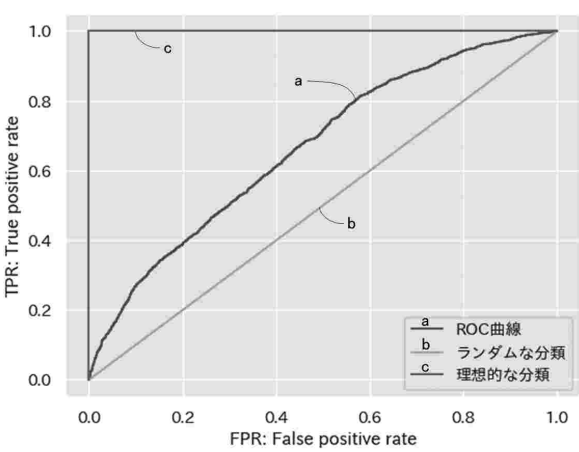
【図 6】



【図 7】

	理論	変数	符号	
H1a	難解化仮説	漢字含有率	不正>非不正	漢字の数をカタカナの数、ひらがなの数、漢字の数の和で割った値
		単語の多様性	不正>非不正	単語の種類数を単語数で割った値
H1b	情報操作理論	数字率	不正<非不正	MD&A 中の全ての文字に占める数字の割合
		単語数	不正<非不正	MD&A を単語に分割した際の単語の数
H1c	対人欺瞞理論	トーン	不正<非不正	ポジティブ用語の数からネガティブ用語の数を引いたものをポジティブ用語の数とネガティブ用語の数の和で割った値
		表現力	不正<非不正	形容詞の数と副詞の数の和を名詞の数と動詞の数の和で割った値

【図 8】



10

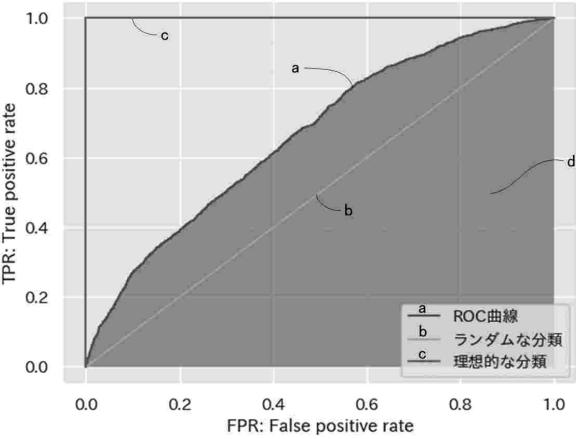
20

30

40

50

【図 9】



【図 10】

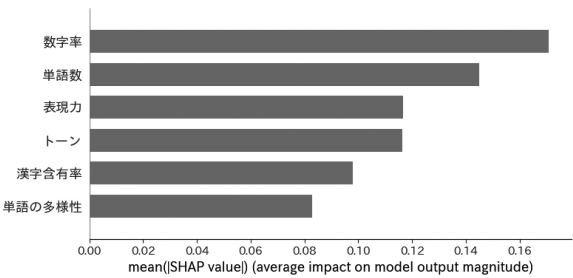
変数	不正企業の平均値	非不正企業の平均値	p 値	有意差
漢字含有率	0.571086033	0.562060636	0.0394751	認められる
単語の多様性	0.312558315	0.283698629	0.00817588	認められる
数字率	0.072464557	0.075511517	0.196544608	認められない
単語数	1368.657718	1933.7	0.000373165	認められる
トーン	-0.929814764	-0.926002555	0.039353201	認められる
表現力	0.004702222	0.006754342	0.000502744	認められる

10

【図 11】

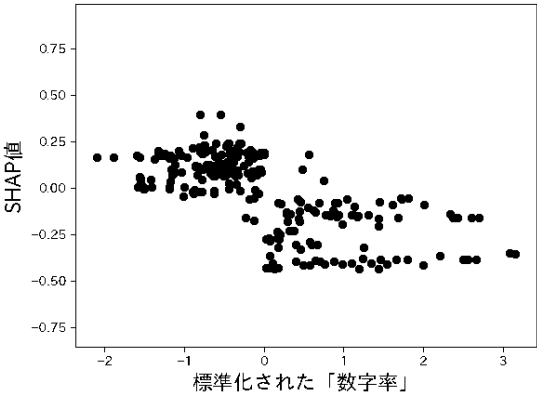
手法	Accuracy	Precision	Recall	F measure	AUROC
決定木	0.660	0.650	0.591	0.619	0.655
ランダムフォレスト	0.723	0.714	0.682	0.698	0.731
XGBoost	0.702	0.667	0.727	0.696	0.715
LightGBM	0.681	0.682	0.652	0.667	0.672
CatBoost	0.617	0.611	0.500	0.550	0.727

【図 12 A】

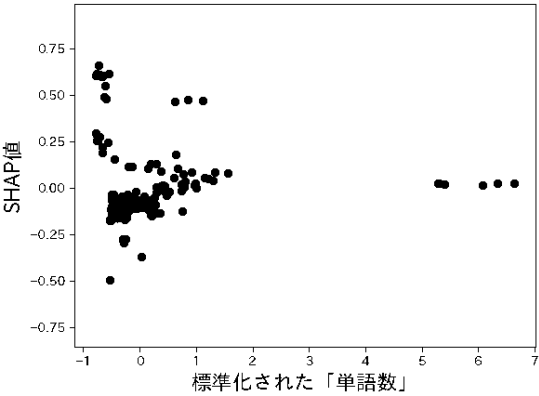


20

【図 12 B】



【図 12 C】

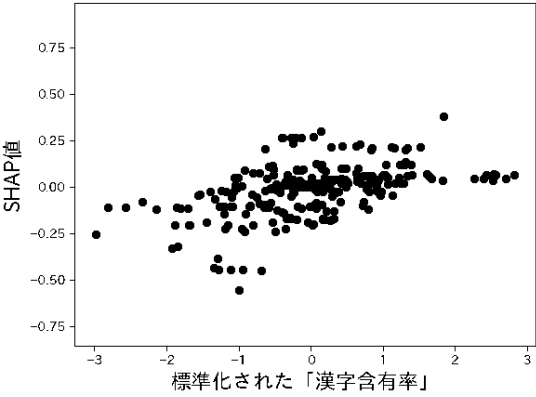


30

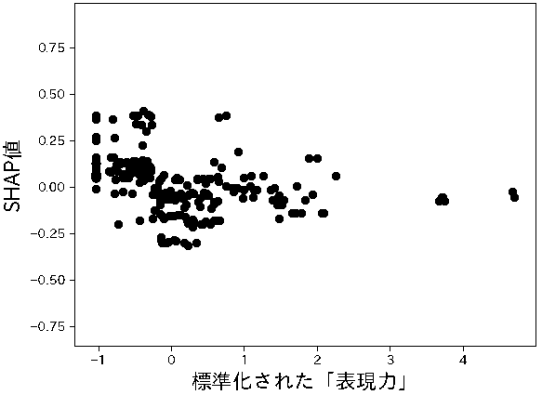
40

50

【図 1 2 D】

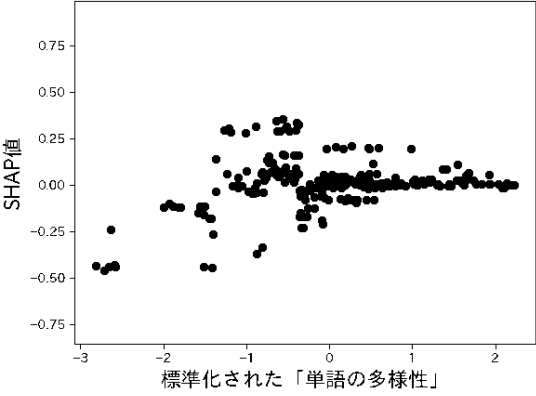


【図 1 2 E】

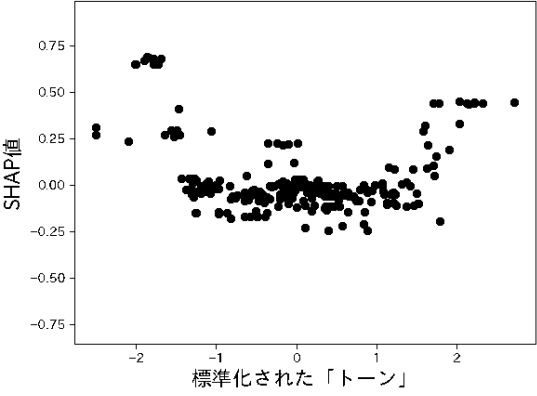


10

【図 1 2 F】



【図 1 2 G】



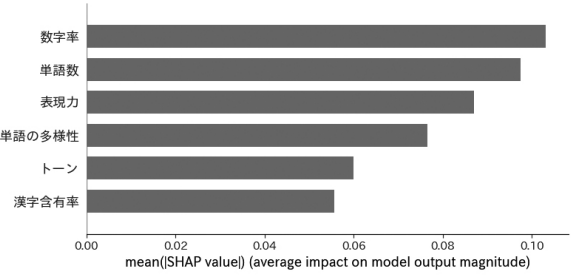
20

30

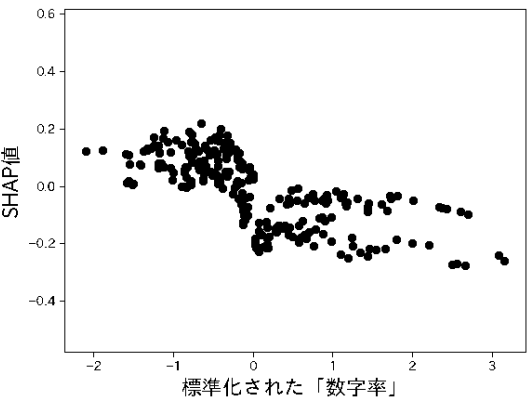
40

50

【図 1 3 A】

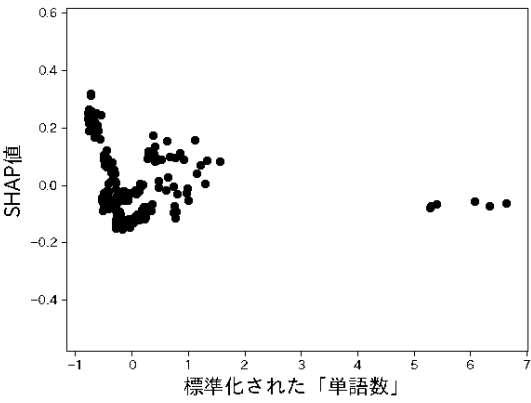


【図 1 3 B】

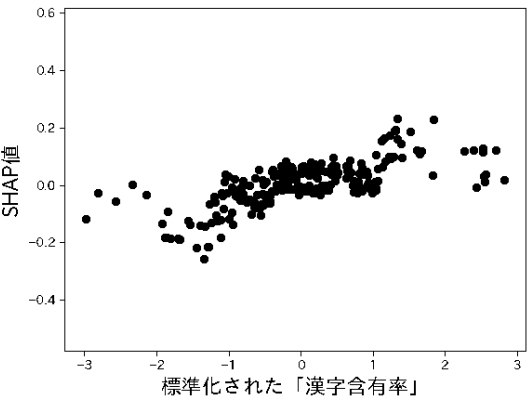


10

【図 1 3 C】



【図 1 3 D】



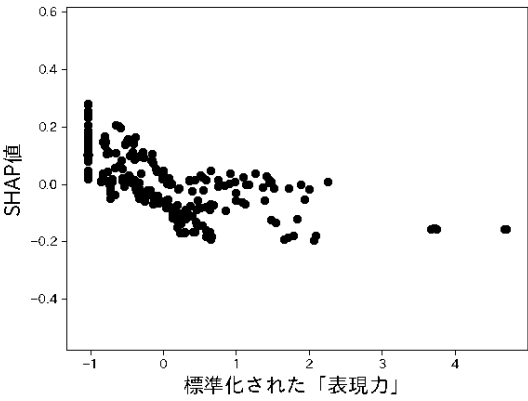
20

30

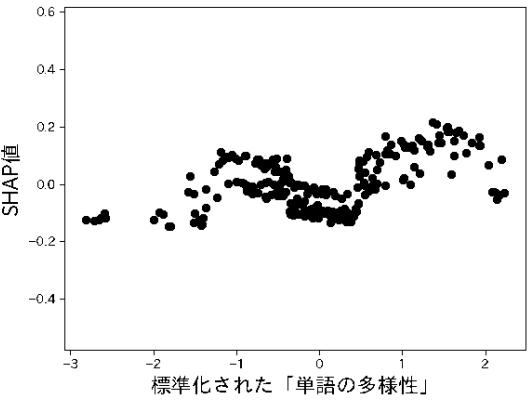
40

50

【図 1 3 E】

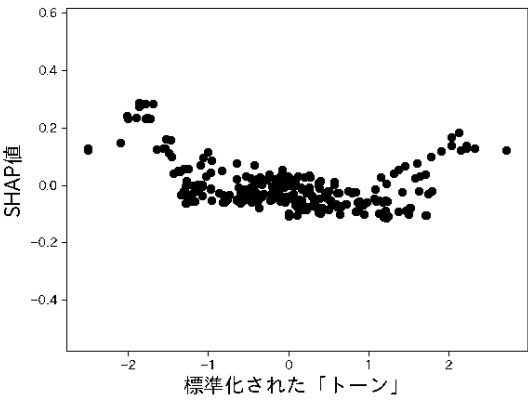


【図 1 3 F】

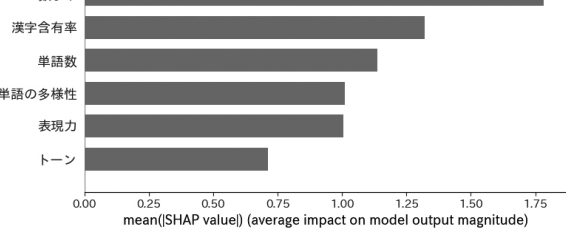


10

【図 1 3 G】



【図 1 4 A】



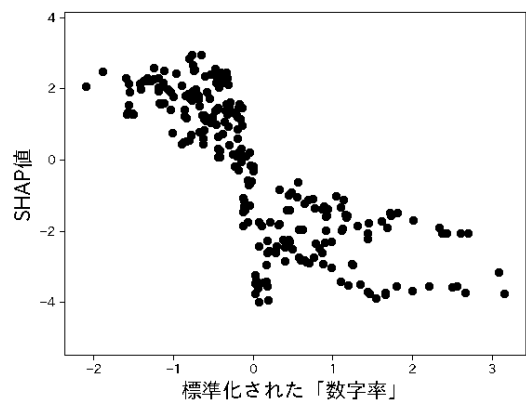
20

30

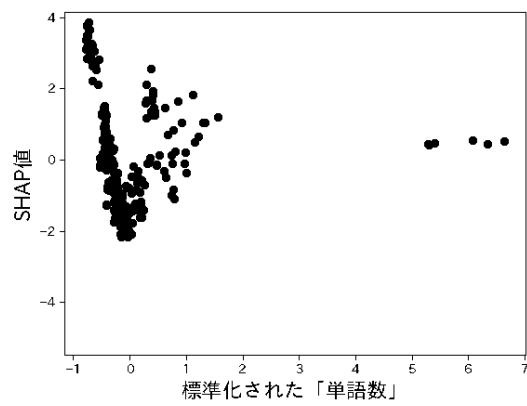
40

50

【図 1 4 B】

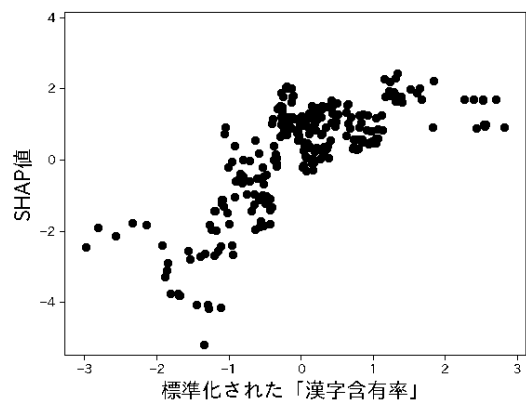


【図 1 4 C】

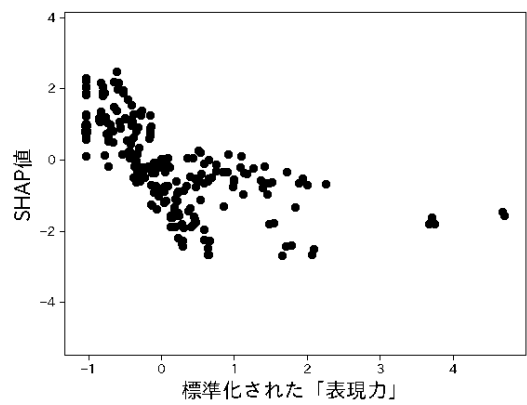


10

【図 1 4 D】



【図 1 4 E】



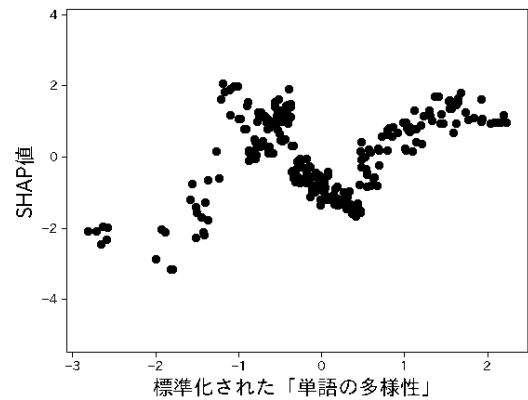
20

30

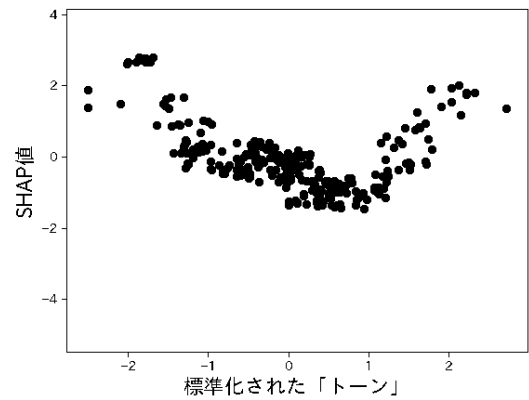
40

50

【図 1 4 F】

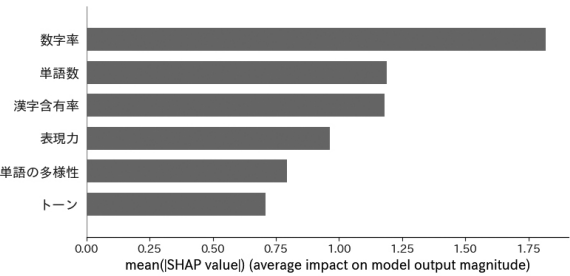


【図 1 4 G】

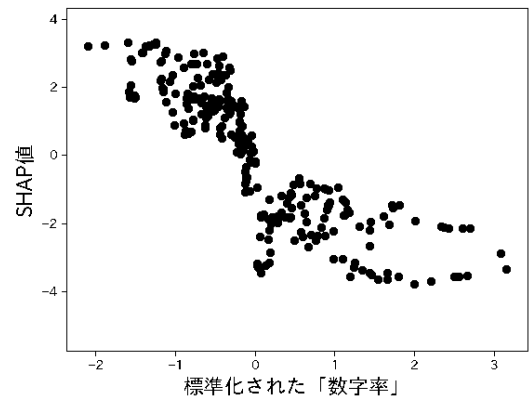


10

【図 1 5 A】



【図 1 5 B】



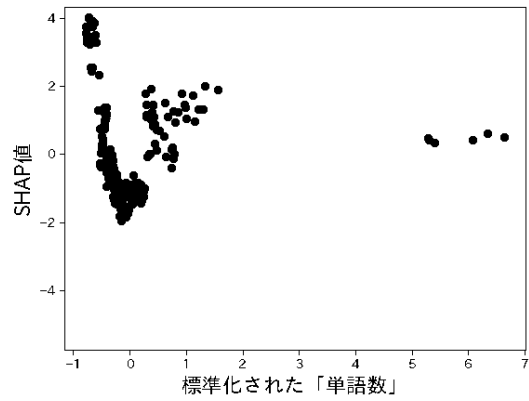
20

30

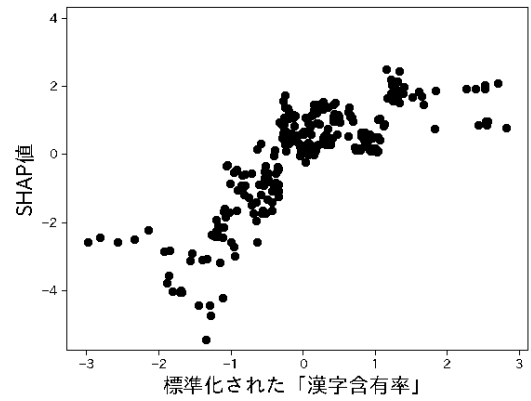
40

50

【図 1 5 C】

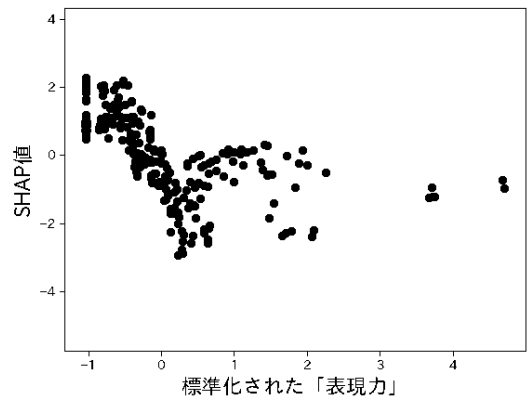


【図 1 5 D】

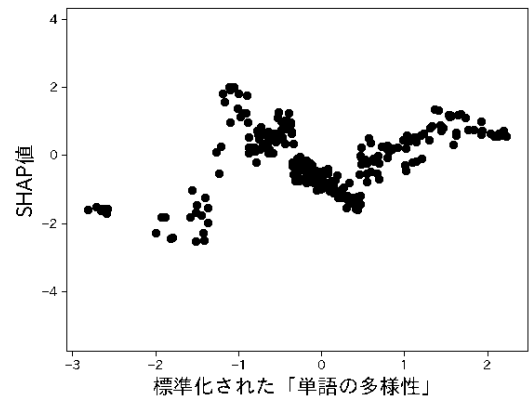


10

【図 1 5 E】



【図 1 5 F】



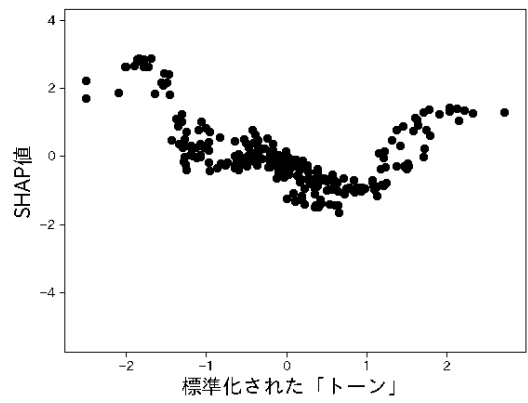
20

30

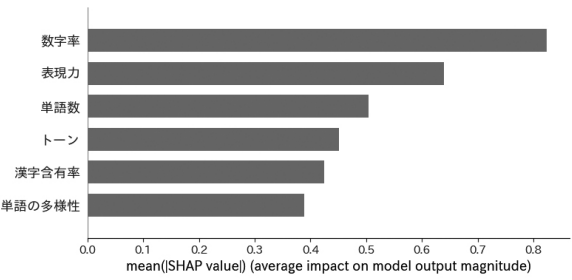
40

50

【図 15 G】

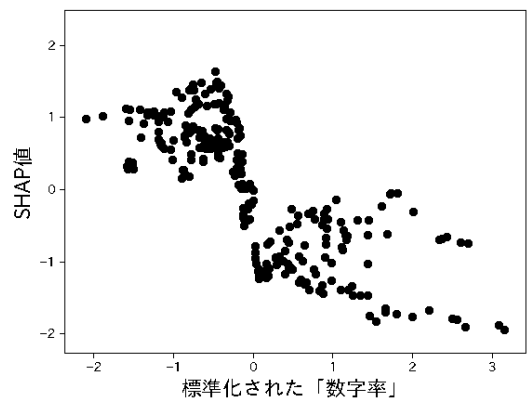


【図 16 A】

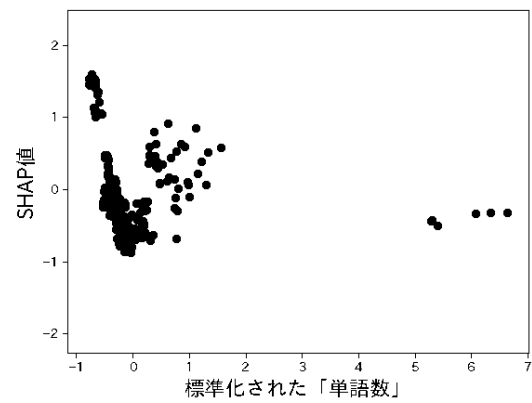


10

【図 16 B】



【図 16 C】



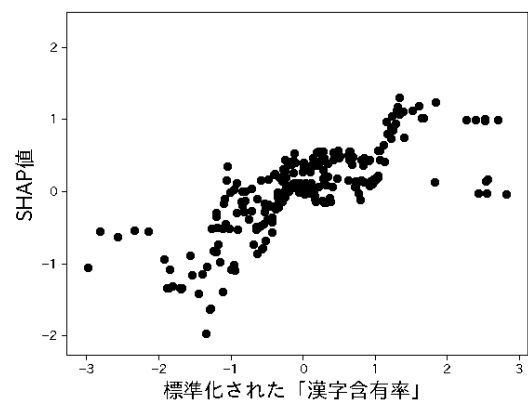
20

30

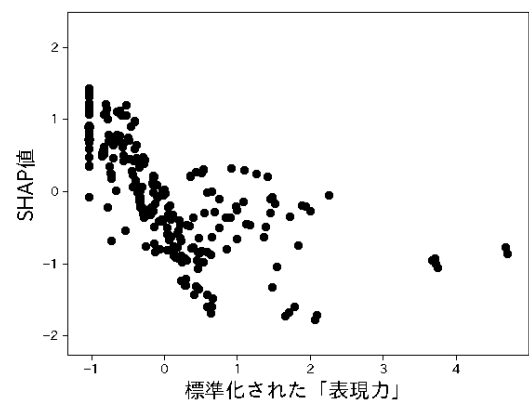
40

50

【図 1 6 D】

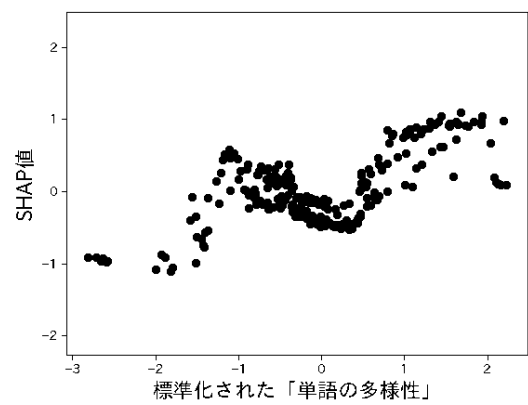


【図 1 6 E】

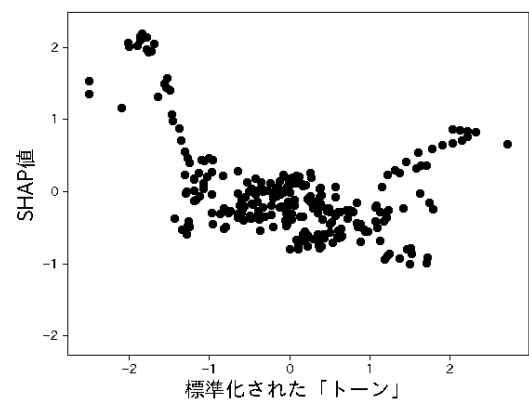


10

【図 1 6 F】



【図 1 6 G】



20

【図 1 7】

手法 \ 評価指標	Accuracy	Precision	Recall	F measure	AUROC
決定木	0.617	0.642	0.409	0.500	0.605
ランダムフォレスト	0.766	0.789	0.682	0.732	0.793
XGBoost	0.787	0.800	0.727	0.762	0.778
LightGBM	0.745	0.778	0.636	0.700	0.785
CatBoost	0.638	0.745	0.500	0.564	0.564

【図 1 8】

手法 \ 評価指標	Accuracy	Precision	Recall	F measure	AUROC
決定木	0.511	0.480	0.545	0.511	0.513
ランダムフォレスト	0.617	0.600	0.545	0.571	0.630
XGBoost	0.553	0.524	0.500	0.512	0.553
LightGBM	0.617	0.620	0.455	0.526	0.599
CatBoost	0.553	0.526	0.455	0.488	0.673

30

40

50

【図 19】

評価指標	Accuracy	Precision	Recall	F measure	AUROC
決定木	0.574	0.550	0.500	0.524	0.570
ランダムフォレスト	0.638	0.647	0.500	0.564	0.688
XGBoost	0.702	0.667	0.727	0.696	0.697
LightGBM	0.638	0.609	0.636	0.622	0.635
CatBoost	0.660	0.667	0.545	0.600	0.682

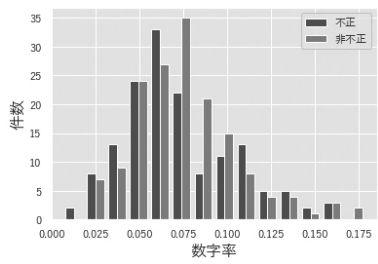
【図 20】

評価指標	Accuracy	Precision	Recall	F measure	AUROC
決定木	0.702	0.682	0.682	0.682	0.701
ランダムフォレスト	0.638	0.647	0.500	0.564	0.635
XGBoost	0.596	0.644	0.600	0.409	0.644
LightGBM	0.617	0.611	0.500	0.550	0.665
CatBoost	0.574	0.550	0.500	0.524	0.635

【図 21】

評価指標	Accuracy	Precision	Recall	F measure	AUROC
決定木	0.553	0.524	0.500	0.512	0.550
ランダムフォレスト	0.553	0.524	0.500	0.512	0.625
XGBoost	0.553	0.524	0.500	0.512	0.617
LightGBM	0.532	0.500	0.455	0.476	0.587
CatBoost	0.553	0.522	0.545	0.533	0.595

【図 22】



【図 23 A】

変数リスト	RandomF_AUROC	RandomF_AUROCwith数字率
[トーン]	0.341818182	0.659090909
[単語数]	0.528181818	0.664545455
[トーン, 単語数]	0.479090909	0.606363636
[漢字含有率]	0.511818182	0.678181818
[トーン, 漢字含有率]	0.550909091	0.693636364
[単語数, 漢字含有率]	0.664545455	0.71
[トーン, 単語数, 漢字含有率]	0.600909091	0.697272727
[表現力]	0.402727273	0.647272727
[トーン, 表現力]	0.586363636	0.624545455
[単語数, 表現力]	0.636363636	0.75
[トーン, 単語数, 表現力]	0.611818182	0.64
[漢字含有率, 表現力]	0.475454545	0.624545455
[トーン, 漢字含有率, 表現力]	0.650909091	0.675454545
[単語数, 漢字含有率, 表現力]	0.658181818	0.766363636
[トーン, 単語数, 漢字含有率, 表現力]	0.637272727	0.724545455
[単語の多様性]	0.52	0.570909091
[トーン, 単語の多様性]	0.482727273	0.628181818
[単語の多様性, 単語数]	0.61	0.644545455
[トーン, 単語の多様性, 単語数]	0.536363636	0.585454545
[単語の多様性, 漢字含有率]	0.596363636	0.686363636
[トーン, 単語の多様性, 漢字含有率]	0.579090909	0.658181818
[単語の多様性, 単語数, 漢字含有率]	0.649090909	0.74
[トーン, 単語の多様性, 単語数, 漢字含有率]	0.585454545	0.674545455
[単語の多様性, 表現力]	0.539090909	0.738181818
[トーン, 単語の多様性, 表現力]	0.563636364	0.652727273
[単語の多様性, 単語数, 表現力]	0.636363636	0.739090909
[トーン, 単語の多様性, 単語数, 表現力]	0.589090909	0.714545455
[単語の多様性, 漢字含有率, 表現力]	0.626363636	0.776363636
[トーン, 単語の多様性, 漢字含有率, 表現力]	0.666363636	0.746363636
[単語の多様性, 単語数, 漢字含有率, 表現力]	0.663636364	0.733636364
[トーン, 単語の多様性, 単語数, 漢字含有率, 表現力]	0.602727273	0.748181818

【図 23 B】

変数リスト	XGBClass_AUROC	XGBClass_AUROCwith数字率
[トーン]	0.369090909	0.694545455
[単語数]	0.53	0.604545455
[トーン, 単語数]	0.478181818	0.577272727
[漢字含有率]	0.571818182	0.644545455
[トーン, 漢字含有率]	0.581818182	0.665454545
[単語数, 漢字含有率]	0.624545455	0.736363636
[トーン, 単語数, 漢字含有率]	0.561818182	0.641818182
[表現力]	0.455454545	0.641818182
[トーン, 表現力]	0.605454545	0.618181818
[単語数, 表現力]	0.709090909	0.714545455
[トーン, 単語数, 表現力]	0.578181818	0.663636364
[漢字含有率, 表現力]	0.443636364	0.614545455
[トーン, 漢字含有率, 表現力]	0.718181818	0.641818182
[単語数, 漢字含有率, 表現力]	0.678181818	0.767272727
[トーン, 単語数, 漢字含有率, 表現力]	0.616363636	0.705454545
[単語の多様性]	0.567272727	0.661818182
[トーン, 単語の多様性]	0.510909091	0.656363636
[単語の多様性, 単語数]	0.62	0.612727273
[トーン, 単語の多様性, 単語数]	0.543636364	0.567272727
[単語の多様性, 漢字含有率]	0.594545455	0.727272727
[トーン, 単語の多様性, 漢字含有率]	0.490909091	0.696363636
[単語の多様性, 単語数, 漢字含有率]	0.596363636	0.685454545
[トーン, 単語の多様性, 単語数, 漢字含有率]	0.550909091	0.610909091
[単語の多様性, 表現力]	0.513636364	0.689090909
[トーン, 単語の多様性, 表現力]	0.571818182	0.709090909
[単語の多様性, 単語数, 表現力]	0.578181818	0.645454545
[トーン, 単語の多様性, 単語数, 表現力]	0.569090909	0.656363636
[単語の多様性, 漢字含有率, 表現力]	0.558181818	0.765454545
[トーン, 単語の多様性, 漢字含有率, 表現力]	0.62	0.729090909
[単語の多様性, 単語数, 漢字含有率, 表現力]	0.610909091	0.763636364
[トーン, 単語の多様性, 単語数, 漢字含有率, 表現力]	0.598181818	0.687272727

10

20

30

40

50

【図 2 3 C】

変数リスト	LGBMCla_AUROC	LGBMCla_AUROCwith数字字
「トーン」	0.426363636	0.685454545
「単語数」	0.509090909	0.599090909
「トーン」,「単語数」	0.407272727	0.567272727
「漢字含有率」	0.61	0.634545455
「トーン」,「漢字含有率」	0.618181818	0.64
「単語数」,「漢字含有率」	0.581818182	0.725454545
「トーン」,「単語数」,「漢字含有率」	0.56	0.641818182
「表現力」	0.431818182	0.664545455
「トーン」,「表現力」	0.652727273	0.683636364
「単語数」,「表現力」	0.528181818	0.709090909
「トーン」,「単語数」,「表現力」	0.538181818	0.658181818
「漢字含有率」,「表現力」	0.515454545	0.641818182
「トーン」,「漢字含有率」,「表現力」	0.74	0.647272727
「単語数」,「漢字含有率」,「表現力」	0.659090909	0.785454545
「トーン」,「単語数」,「漢字含有率」,「表現力」	0.605454545	0.698181818
「単語の多様性」	0.498181818	0.590909091
「トーン」,「単語の多様性」	0.409090909	0.62
「単語の多様性」,「単語数」	0.497272727	0.568181818
「トーン」,「単語の多様性」,「単語数」	0.442727273	0.552727273
「単語の多様性」,「漢字含有率」	0.583636364	0.654545455
「トーン」,「単語の多様性」,「漢字含有率」	0.556363636	0.629090909
「単語の多様性」,「単語数」,「漢字含有率」	0.561818182	0.678181818
「トーン」,「単語の多様性」,「単語数」,「漢字含有率」	0.529090909	0.636363636
「単語の多様性」,「表現力」	0.552727273	0.674545455
「トーン」,「単語の多様性」,「表現力」	0.496363636	0.703636364
「単語の多様性」,「単語数」,「表現力」	0.518181818	0.661818182
「トーン」,「単語の多様性」,「単語数」,「表現力」	0.543636364	0.663636364
「単語の多様性」,「漢字含有率」,「表現力」	0.569090909	0.741818182
「トーン」,「単語の多様性」,「漢字含有率」,「表現力」	0.594545455	0.72
「単語の多様性」,「単語数」,「漢字含有率」,「表現力」	0.59	0.711818182
「トーン」,「単語の多様性」,「単語数」,「漢字含有率」,「表現力」	0.596363636	0.672727273

【図 2 3 D】

変数リスト	<catboo_AUROC	<catboo_AUROCwith数字字
「トーン」	0.432727273	0.703636364
「単語数」	0.500909091	0.685454545
「トーン」,「単語数」	0.441818182	0.610909091
「漢字含有率」	0.615454545	0.701818182
「トーン」,「漢字含有率」	0.623636364	0.692727273
「単語数」,「漢字含有率」	0.578181818	0.696363636
「トーン」,「単語数」,「漢字含有率」	0.563636364	0.641818182
「表現力」	0.494545455	0.627272727
「トーン」,「表現力」	0.621818182	0.710909091
「単語数」,「表現力」	0.601818182	0.749090909
「トーン」,「単語数」,「表現力」	0.521818182	0.672727273
「漢字含有率」,「表現力」	0.485454545	0.638181818
「トーン」,「漢字含有率」,「表現力」	0.705454545	0.732727273
「単語数」,「漢字含有率」,「表現力」	0.636363636	0.743636364
「トーン」,「単語数」,「漢字含有率」,「表現力」	0.610909091	0.732727273
「単語の多様性」	0.517272727	0.647272727
「トーン」,「単語の多様性」	0.48	0.649090909
「単語の多様性」,「単語数」	0.534545455	0.679090909
「トーン」,「単語の多様性」,「単語数」	0.505454545	0.594545455
「単語の多様性」,「漢字含有率」	0.603636364	0.712727273
「トーン」,「単語の多様性」,「漢字含有率」	0.623636364	0.687272727
「単語の多様性」,「単語数」,「漢字含有率」	0.6	0.68
「トーン」,「単語の多様性」,「単語数」,「漢字含有率」	0.556363636	0.643636364
「単語の多様性」,「表現力」	0.545454545	0.685454545
「トーン」,「単語の多様性」,「表現力」	0.54	0.701818182
「単語の多様性」,「単語数」,「表現力」	0.6	0.734545455
「トーン」,「単語の多様性」,「単語数」,「表現力」	0.52	0.66
「単語の多様性」,「漢字含有率」,「表現力」	0.612727273	0.774545455
「トーン」,「単語の多様性」,「漢字含有率」,「表現力」	0.652727273	0.752727273
「単語の多様性」,「単語数」,「漢字含有率」,「表現力」	0.654545455	0.756363636
「トーン」,「単語の多様性」,「単語数」,「漢字含有率」,「表現力」	0.603636364	0.730909091

10

【図 2 4 A】

変数リスト	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
「トーン」	0.341818182	0.369090909	0.426363636	0.432727273
「単語数」	0.528181818	0.53	0.509090909	0.500909091
「漢字含有率」	0.511818182	0.571818182	0.61	0.615454545
「トーン」,「漢字含有率」	0.402727273	0.455454545	0.431818182	0.494545455
「単語数」,「漢字含有率」	0.603636364	0.564545455	0.587272727	0.605454545
「単語の多様性」	0.52	0.567272727	0.498181818	0.517272727

【図 2 4 B】

変数リスト	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
「トーン」,「単語数」	0.479090909	0.478181818	0.407272727	0.441818182
「トーン」,「漢字含有率」	0.550909091	0.581818182	0.618181818	0.623636364
「単語数」,「漢字含有率」	0.664545455	0.624545455	0.581818182	0.578181818
「トーン」,「表現力」	0.586363636	0.605454545	0.652727273	0.621818182
「単語数」,「表現力」	0.636363636	0.570909091	0.528181818	0.601818182
「漢字含有率」,「表現力」	0.475454545	0.443636364	0.515454545	0.485454545
「トーン」,「単語の多様性」	0.482727273	0.510909091	0.409090909	0.48
「単語数」,「単語の多様性」	0.61	0.62	0.497272727	0.534545455
「漢字含有率」,「単語の多様性」	0.596363636	0.594545455	0.583636364	0.603636364
「表現力」,「単語の多様性」	0.539090909	0.513636364	0.552727273	0.545454545
「トーン」,「数字字」	0.659090909	0.694545455	0.685454545	0.703636364
「単語数」,「数字字」	0.664545455	0.604545455	0.599090909	0.685454545
「漢字含有率」,「数字字」	0.678181818	0.644545455	0.634545455	0.701818182
「表現力」,「数字字」	0.647272727	0.641818182	0.664545455	0.627272727
「数字字」,「単語の多様性」	0.570909091	0.661818182	0.590909091	0.647272727

20

【図 2 4 C】

変数リスト	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
「トーン」,「単語数」,「漢字含有率」	0.600909091	0.561818182	0.56	0.563636364
「トーン」,「単語数」,「表現力」	0.611818182	0.578181818	0.538181818	0.521818182
「トーン」,「漢字含有率」,「表現力」	0.650909091	0.718181818	0.74	0.705454545
「単語数」,「漢字含有率」,「表現力」	0.658181818	0.678181818	0.659090909	0.636363636
「トーン」,「単語数」,「単語の多様性」	0.536363636	0.543636364	0.442727273	0.505454545
「トーン」,「漢字含有率」,「単語の多様性」	0.579090909	0.490909091	0.556363636	0.623636364
「単語数」,「漢字含有率」,「単語の多様性」	0.649090909	0.596363636	0.561818182	0.6
「トーン」,「表現力」,「単語の多様性」	0.563636364	0.571818182	0.496363636	0.54
「単語数」,「表現力」,「単語の多様性」	0.636363636	0.578181818	0.518181818	0.6
「漢字含有率」,「表現力」,「単語の多様性」	0.626363636	0.558181818	0.569090909	0.612727273
「トーン」,「単語数」,「数字字」	0.606363636	0.572727272	0.567272727	0.610909091
「トーン」,「漢字含有率」,「数字字」	0.693636364	0.665454545	0.64	0.692727273
「単語数」,「漢字含有率」,「数字字」	0.71	0.736363636	0.725454545	0.696363636
「トーン」,「表現力」,「数字字」	0.624545455	0.618181818	0.683636364	0.710909091
「単語数」,「表現力」,「数字字」	0.75	0.714545455	0.709090909	0.749090909
「漢字含有率」,「表現力」,「数字字」	0.624545455	0.614545455	0.641818182	0.638181818
「トーン」,「数字字」,「単語の多様性」	0.628181818	0.656363636	0.62	0.649090909
「単語数」,「数字字」,「単語の多様性」	0.644545455	0.612727273	0.568181818	0.670909091
「漢字含有率」,「数字字」,「単語の多様性」	0.686363636	0.727272727	0.654545455	0.712727273
「表現力」,「数字字」,「単語の多様性」	0.738181818	0.689090909	0.674545455	0.685454545

【図 2 4 D】

変数リスト	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
「トーン」,「単語数」,「漢字含有率」,「表現力」	0.637272727	0.616363636	0.605454545	0.610909091
「トーン」,「単語数」,「漢字含有率」,「数字字」	0.697272727	0.641818182	0.641818182	0.641818182
「トーン」,「単語数」,「表現力」,「数字字」	0.64	0.663636364	0.658181818	0.672727273
「トーン」,「漢字含有率」,「表現力」,「数字字」	0.675454545	0.641818182	0.647272727	0.732727273
「単語数」,「漢字含有率」,「表現力」,「数字字」	0.766363636	0.767272727	0.785454545	0.743636364
「トーン」,「単語数」,「漢字含有率」,「単語の多様性」	0.585454545	0.550909091	0.529090909	0.556363636
「トーン」,「単語数」,「表現力」,「単語の多様性」	0.589090909	0.569090909	0.543636364	0.52
「トーン」,「漢字含有率」,「表現力」,「単語の多様性」	0.666363636	0.62	0.594545455	0.652727273
「単語数」,「漢字含有率」,「表現力」,「単語の多様性」	0.663636364	0.610909091	0.59	0.654545455
「トーン」,「単語数」,「数字字」,「単語の多様性」	0.585454545	0.567272727	0.552727273	0.594545455
「トーン」,「表現力」,「数字字」,「単語の多様性」	0.658181818	0.696363636	0.629090909	0.687272727
「単語数」,「漢字含有率」,「数字字」,「単語の多様性」	0.74	0.685454545	0.678181818	0.68
「トーン」,「漢字含有率」,「数字字」,「単語の多様性」	0.652727273	0.709090909	0.703636364	0.701818182
「単語数」,「表現力」,「数字字」,「単語の多様性」	0.739090909	0.645454545	0.661818182	0.734545455
「漢字含有率」,「表現力」,「数字字」,「単語の多様性」	0.776363636	0.765454545	0.741818182	0.774545455

30

40

50

【 図 2 4 E 】

変数リスト	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
[トーン,'単語数','漢字含有率','表現力','数字率']	0.724545455	0.705454545	0.698181818	0.732727273
[トーン,'単語数','漢字含有率','表現力','単語の多様性']	0.602727273	0.598181818	0.596363636	0.603636364
[トーン,'単語数','漢字含有率','数字率','単語の多様性']	0.674545455	0.610909091	0.636363636	0.643636364
[トーン,'単語数','表現力','数字率','単語の多様性']	0.714545455	0.656363636	0.663636364	0.66
[トーン,'漢字含有率','表現力','数字率','単語の多様性']	0.746363636	0.729090909	0.72	0.752727273
[単語数,'漢字含有率','表現力','数字率','単語の多様性']	0.733636364	0.763636364	0.711818182	0.756363636

【 図 2 5 A 】

統計量及び数字率の有無	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
平均値_数字率あり	0.603636364	0.564545455	0.587272727	0.605454545
平均値_数字率なし	0.460909091	0.498727273	0.495090909	0.512181818
中央値_数字率あり	0.603636364	0.564545455	0.587272727	0.605454545
中央値_数字率なし	0.511818182	0.53	0.498181818	0.500909091
最大値_数字率あり	0.603636364	0.564545455	0.587272727	0.605454545
最大値_数字率なし	0.528181818	0.571818182	0.61	0.615454545

【 図 2 5 B 】

統計量及び数字率の有無	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
平均値_数字率あり	0.644	0.649454545	0.634909091	0.673090909
平均値_数字率なし	0.562090909	0.554363636	0.534636364	0.551636364
中央値_数字率あり	0.659090909	0.644545455	0.634545455	0.685454545
中央値_数字率なし	0.568636364	0.576363636	0.540454545	0.561818182
最大値_数字率あり	0.678181818	0.694545455	0.685454545	0.703636364
最大値_数字率なし	0.664545455	0.624545455	0.652727273	0.623636364

【 図 2 5 C 】

統計量及び数字率の有無	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
平均値_数字率あり	0.670636364	0.661181818	0.648454545	0.681636364
平均値_数字率なし	0.611272727	0.587545455	0.564181818	0.590909091
中央値_数字率あり	0.665454545	0.660909091	0.648181818	0.689090909
中央値_数字率なし	0.619090909	0.575	0.558181818	0.6
最大値_数字率あり	0.75	0.736363636	0.725454545	0.749090909
最大値_数字率なし	0.658181818	0.718181818	0.74	0.705454545

【 図 2 5 D 】

統計量及び数字率の有無	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
平均値_数字率あり	0.693090909	0.678363636	0.67	0.696363636
平均値_数字率なし	0.628363636	0.593454545	0.572545455	0.598909091
中央値_数字率あり	0.686363636	0.674545455	0.66	0.694545455
中央値_数字率なし	0.637272727	0.610909091	0.59	0.610909091
最大値_数字率あり	0.776363636	0.767272727	0.785454545	0.774545455
最大値_数字率なし	0.666363636	0.62	0.605454545	0.654545455

【 図 2 5 E 】

統計量及び数字率の有無	RandomF_AUROC	XGBClas_AUROC	LGBMCla_AUROC	<catboo_AUROC
平均値_数字率あり	0.718727273	0.693090909	0.686	0.709090909
平均値_数字率なし	0.602727273	0.598181818	0.596363636	0.603636364
中央値_数字率あり	0.724545455	0.705454545	0.698181818	0.732727273
中央値_数字率なし	0.602727273	0.598181818	0.596363636	0.603636364
最大値_数字率あり	0.746363636	0.763636364	0.72	0.756363636
最大値_数字率なし	0.602727273	0.598181818	0.596363636	0.603636364

10

20

30

40

50

フロントページの続き

- 東京都文京区向丘 1 丁目 1 9 - 1 文京学院大学内
- (72)発明者 藤田 邦彦
- 東京都文京区向丘 1 丁目 1 9 - 1 文京学院大学内
- (72)発明者 吉田 啓佑
- 東京都文京区向丘 1 丁目 1 9 - 1 文京学院大学内
- 審査官 山崎 雄司
- (56)参考文献 米国特許出願公開第 2 0 0 5 / 0 0 0 4 8 6 2 (U S , A 1)
- 首藤昭信, テキスト分析と会計学研究, 情報センサー, 2019年5月号, 日本, EY Japan, 2019年04月30日, [online], インターネット<URL: https://ey.com/ja_jp/library/info-sensor/2019/info-sensor-2019-05-02>, [令和6年9月10日検索]
- 中島真澄, ナラティブ情報(非財務情報)からの会計不正検出, 文京学院大学 オピニオンレター, Vol.32, 文京学院大学, 2022年09月16日, [online], インターネット<URL: http://bgu.ac.jp/about/activity/opinion_letter/>, [令和6年9月10日検索]
- 市原 直通, 外1名, F i n T e c hで変わる会計の世界, 企業会計, 日本, 株式会社中央経済社, 2017年06月01日, 第 6 9 巻 第 6 号, 55-63頁
- (58)調査した分野 (Int.Cl., D B 名)
- G 0 6 Q 1 0 / 0 0 - 9 9 / 0 0