HITOTSUBASHI UNIVERSITY HUB BUSINESS SCHOOL

FSMBA PROGRAM FINANCIAL STRATEGY

HUB-FS Working Paper Series

FS-2021-J-001

機械学習手法を用いた不正会計予測：非上場企業データを用いた検討

宇宿哲平、近藤聡、白木研吾

有限責任 あずさ監査法人

宮川大介

一橋大学大学院経営管理研究科

柳岡優希

株式会社東京商工リサーチ

First version: June 22, 2021

機械学習手法を用いた不正会計予測：

非上場企業データを用いた検討*

2021 年 6 月

宇宿哲平、近藤聡、白木研吾、宮川大介、柳岡優希**

＜要旨＞

本研究は、宇宿ほか (2019)で構築された機械学習手法に基づく上場企業向け不正会計予測モデルを、非上場企業を含む広範な企業群に適用したうえで、その予測精度を検証したものである。具体的には、2014 年から 2017 年の期間における詳細な財務情報を伴う各年 30 万社程度のデータセットに対して、株式会社東京商工リサーチ（TSR）が保有する企業レベルのネガティブイベントに関するテキスト情報から構築した不正会計イベントの発生を示すフラグを付与した上で、不正会計の予測変数として実証会計分野で一般的に用いられている変数のほか、企業属性、実物及び金融面における取引関係情報、同業他社情報、周辺立地情報、株主関係情報を用いた予測モデルを構築した。Out of sample における AUC 指標に基づく精度評価から、第一に、非上場企業の不正会計予測について 0.8 を上回る予測精度が実現された。第二に、こうした高い予測精度を実現するために、極端に高次元の予測変数は必要とされず、各企業の信用評点及び基本的な企業属性から構築された予測変数群で実務的観点から見て十分な予測精度が実現されることを確認した。これらの結果は、利用可能な情報が上場企業に比して乏しいことが想定される非上場企業についても、機械学習手法の利用による適切なモデル構築を行うことで不正会計予測が実行可能であることを示唆するものである。

JEL Classification: M42, C53, C14

Key Words: 不正会計、機械学習、未上場企業、非上場企業

* 本稿は、国立大学法人一橋大学と㈱東京商工リサーチとの共同研究契約に基づき、有限責任あずさ監査法人を研究協力者として実施された研究プロジェクトの成果である。

** 宇宿、近藤、白木：有限責任 あずさ監査法人 Digital Innovation 部、宮川（corresponding author）：一橋大学大学院経営管理研究科 准教授 〒101-8439 東京都千代田区一ツ橋 2-1-2 E-mail: dmiyakawa@hub.hit-u.ac.jp、柳岡：株式会社東京商工リサーチ 経営企画室 〒100-6810 東京都千代田区大手町 1-3-1 E-mail: masaki.yanaoka_2@tsr-net.co.jp。

１．はじめに

上場企業にとって「開示情報の誤り」は、取引金融機関や実物取引における相手先（販売先、仕入先）に加えて市場からの信認を失いかねない重大なインシデントである。特に、こうした開示情報の誤りが企業による何らかの意図の下で行われた場合、会計情報に関する不正行為（不正会計）とみなされ、当局からペナルティが課される場合もある。

不正会計に関するこうした実務上の重要性を踏まえて、多くの既存研究が不正会計の発生パターンに関する理論的な検討を進め、検知や予測を目的とした実証モデルの構築を試みてきた（Dechow et al. 1996, 2010, 2011; Song et al. 2016）。これらの文献では、企業がどのような状況において不正会計を行うインセンティブを有するかを演繹的に記述したうえで、こうした理論的想定の下で構築された変数群（例：裁量的会計発生高）を予測変数として用いることで、不正会計の予測モデルを構築している。

これらの先行研究が、明示的な理論的想定の下で解釈可能な予測モデルの構築に成功している一方で、実務的な観点からは、少なくとも二つの課題が認識される。第一に、予測モデルの構築に当たって利用されていない膨大な情報が存在する。例えば、企業の財務情報からは無数の財務指標を構築することができるほか、それらのラグ値や階差といった処理から得られる変数に加えて、こうしたデータが欠損しているという情報にも予測に当たって有益なシグナルが含まれている可能性がある。また、金融機関や実物取引におけるカウンターパート（販売先、仕入先）の属性情報や株主の情報、また、同業他社や周辺に立地する地理的に近接した企業の属性からもこうしたシグナルを抽出する余地があると考えられる。これらの膨大な情報から、不正会計の発生を帰納的にモデル化するためには、高次元の情報を潜在的な予測変数として取り扱う必要があり、こうした目的にとって機械学習手法は有力な選択肢となる。

第二に、大規模かつ高次元のデータと機械学習手法を用いた予測モデル構築が上記の理由から効果的である一方で、データ収集に係る有形無形のコストを勘案すると、可能な限り少ない情報を基にして高精度の予測モデルを構築することも同時に期待される。勿論、基本的には大規模で高次元のデータを用いることで、将来時点におけるイベントの発生と相関する事象を精緻に表現することが出来るというメリットがあると考えられる。しかし、高い柔軟性を許容された予測モデルに付随する問題として、モデル構築用のデータに対する過剰適合によって out of sample の予測精度が損なわれるという問題もあり、必ずしも高次元のデータが望ましいとは言えない側面もある。また、そもそも、高次元の変数間における相関関係の存在を踏まえると、強く相関している変数群の中で代表的な一変数を予測変数として用いることにより「低コスト」で目的が果たされる可能性もある[1]。

本研究では、こうした問題意識を踏まえて、第一に、非上場企業を含む広範な企業群を対象として、機械学習手法を用いた不正会計予測モデルを構築した上で、その out of sample での予測精度を検証した。具体的には、まず、2014 年から 2017 年の期間において詳細な財務情報を伴う各年 30 万社程度のデータセットに対して、株式会社東京商工リサーチ（TSR）が保有する企業レベルのネガティブイベントに関するテキスト情報から構築した不正会計イベントの発生を示すフラグを付与した。次に、不正会計の予測変数として実証会計分野で一般的に用いられている変数のほか、企業属性、実物及び金融面における取引関係情報、同業他社情報、周辺立地情報、株主関係情報を用いた予測モデルを構築した。

上場企業を対象として不正会計の予測を目的とした機械学習ベースの予測モデル構築を行った既存研究としては、宇宿ほか (2019)が挙げられる。本研究の特徴は、こう

[1] ここでの議論は予測変数の次元（種類）を念頭に置いているが、データ収集コストと予測精度のバランスを検討する文脈では、観測数についても同種の議論が存在する。所与の予測精度を実現するためのデータサイズに関する議論としては、例えば、Bajari et al. (2018)を参照のこと。

した既存研究が対象としてきた上場企業に加えて、非上場企業を対象としたモデル構築を行い、その予測精度を検証した点にある。非上場企業はその定義から株式が金融市場で取引されることはないため、不正会計の発生が直接的に金融市場の機能不全をもたらす可能性は低い。しかし、第一に、非上場企業における不正会計の発生は当該企業と取引関係を有する上場企業に関する重大なリスク要因として捉えることができる。この事実は、金融市場の観点からも非上場企業を対象としたモデル構築には一定の意義があることを意味する。第二に、銀行を中心とする金融機関の視点から、主たる業務である融資の実行に当たって非上場企業の不正会計リスクを見積もる十分な動機が存在する。この意味で、上場企業で確認されている不正会計予測の文脈における機械学習モデルの有効性が非上場企業でも確認されるか否かには実務上の高い関心が認められると言えるだろう。

第二に、変数の次元と out of sample での予測精度に関して明示的な分析を行う。既述の通り、情報収集に要するコストと達成される予測精度との間にトレードオフが存在するという点について特段の異論は生じ得ないものの、期待される一定水準の予測精度を達成するためにどの程度のサイズのデータセットが必要となるかは実証的な問題となる。そこで本稿では、膨大な予測変数群を幾つかのカテゴリに分割したうえで、個々のカテゴリに含まれる変数群がどの程度 out of sample の予測精度に貢献するかを記述し、実務的に許容できる予測精度を達成するために必要となる情報のサイズを検討する。

予測モデルの精度指標として一般的に参照される AUC を用いて、上記の要領で構築した予測モデルの out of sample における予測精度を確認した結果から、以下の二点が確認された。第一に、非上場企業を対象とする不正会計予測について AUC の水準として 0.8 を上回る予測精度が実現された。潜在的な予測変数として用いた変数群の選択によってこの予測精度の水準は変化するものの、最も低いケースにおいても 0.76 を上

回る水準が達成されており、実務的に見ても十分な予測精度が実現されている。第二に、こうした高い予測精度を実現するために必要となる予測変数の次元を検討したところ、様々な変数群をフルに用いた予測モデルが最も高い予測精度を実現しておらず、信用調査会社の TSR が各企業に付与した総合評価を示す信用評点と標準的な企業属性を潜在的な予測変数として用いたモデルが最も高い予測精度を実現していることが分かった。特に、利用可能な変数群を全て投入した予測モデルは概ね最も低い予測精度となっており、信用評点と標準的な企業属性を機械学習手法の下で柔軟な形で参照することにより高精度の予測モデルを構築できることが確認される一方で、「極めて」高次元の情報がもたらすベネフィットが過剰適合などのコストによって相殺されていることが窺える。

これらの結果は、利用可能な情報が上場企業に比して乏しいことが想定される非上場企業についても、機械学習手法の利用による適切なモデル構築を行うことで、一定の費用対効果の制約を満たす不正会計予測が実行可能であることを示唆するものである。

次節以降の構成は以下の通りである。第二節では、本研究で使用する予測モデルの構築方法を概観する。第三節では、モデルの構築と予測精度の検証に用いるデータセットを、企業データとネガティブデータに分けて説明した後に変数の定義を示す。第四節では、予測精度の検証結果を示すと共に、今後の研究に向けたディスカッションを行う。第五節では本稿のまとめを示す。

２．予測モデル

本研究では、不正会計の有無のように二値の変数を対象とした予測モデル構築に際して標準的に用いられる Random Forest（Breiman 2001）を用いる。同手法は、決定木

ベースの予測モデルをサブサンプル毎に構築した結果を合算することで分類器を構築するものである。また、不正会計の発生頻度が少ないことを踏まえて、Chen et al. (2004)によって提案された、Random Forest の拡張版である Weighted Random Forest を用いることで、個々の決定木構築及びアンサンブルに際してレアイベントへ相対的に大きなウェイトを付与した上での予測モデル構築を行う[2]。構築済みモデルの out of sample での予測精度評価には、ROC 曲線に基づく AUC を参照する。

３．データ

3.1　企業データ

本研究では、個々の企業に関する変数を構築するために、TSR が保有する以下のデータセットを用いる。第一に、個社の設立年月、住所、従業員数といった基本情報を格納した TSR 企業情報ファイルである。このデータセットには、TSR が個社に付与した信用評点、取引銀行リスト、業種、売上高や利益などの簡易的な財務データ、年齢などの経営者情報が含まれている。第二に、販売先や仕入先といった取引関係に加えて、株主リストを格納した TSR 企業相関ファイルである。第三に、詳細な財務データを格納した TSR 財務情報ファイルである。以上の標準的なデータセットに加えて、本研究では、企業の詳細な親子関係を記録した TSR 企業グループ情報ファイルも用いている。

[2] 次節で概観する TSR データを用いた企業レベルのダイナミクス（例：倒産、休廃業・解散、被合併、成長）予測モデルを構築した事例として、一橋大学と TSR による共同研究成果である特許「企業情報処理装置、企業のイベント予測方法及び予測プログラム」（特許番号：第 6611068 号、特許取得日：令和 1 年 11 月 8 日）が挙げられる。本研究で構築した不正会計予測モデルはこの技術に基づいている。

3.2 ネガティブ情報

不正会計イベントの計測に当たっては、TSR が保有する企業単位のネガティブ情報を格納したデータベースを用いる。具体的には、不正会計の発生を示すとみなし得る複数のキーワードからなるリストを用いて、個々のテキスト情報に対して不正会計フラグを付与する。また、併せて有価証券報告書の虚偽記載に対する金融庁の課徴金勧告、証券取引委員会の告発、適時開示情報を用いた不正フラグの構築も行う。

3.3 変数

前節で記載したデータセットを用いて、まず、過去の会計不正の実績に対応した変数を構築する。具体的には、架空、粉飾、循環取引、虚偽、課徴金などのキーワードが一つでも含まれる場合に 1 を取り、そうでない場合に 0 を取るダミー変数を構築し予測対象の被説明変数として用いる。次に、予測変数として以下の変数群を構築する。全ての予測変数は被説明変数の計測ウィンドウの始点において観測されているもののみを用いる。なお、対応する変数が欠損している場合に対応した欠損ダミー変数も構築する。

第一に、TSR が付与した信用評点（fscore）である。第二に、簡易的な決算数値、継続年数、従業員数、社長情報などを含む企業自身の情報（firmown）である。第三に、財務諸表から構築された詳細な財務指標（kessan）である。第四に、予測対象企業と同一市区町村に立地している企業の売上高成長率の平均値や同業他社に関して同種の値を計測したもの（geoind）である。第五に、取引銀行数やメイバンクの変更を計測した（bank）である。第六に、販売先及び仕入れ先との取引関係についてネットワーク統計量や平均的な信用評点の水準などを用いて計測した（network）である。第七に、network と同様の方法で株主との関係を既述した（shareholder）である。

これらの変数群に加えて、実証会計分野で一般的に使用される変数として、ソフト

資産比率、裁量的会計発生高、異常裁量費用を内容とする変数群（theoretical）を用いるほか、上場ステータスを示す（listed）、親会社が保有する子会社数、親子で産業分類が異なるか否か、売上高の親子比、資本金の親子比などからなる（parent）も用いる[3]。

分析に当たっては、2014 年から 2017 年に亘る各年約 100 万レコードの中から、詳細な財務情報が付与されている各年約 30 万レコードを用いた。

４．精度検証結果

4.1　予測精度

前節までに示したフレームワーク及びデータに基づいて構築した不正会計の予測モデル（一年間の予測ウィンドウ）の精度検証結果を以下で示す。精度の検証に当たっては、不正会計イベントが連続した年において観察されることから生じる information leakage を回避するために、2014 年から 2017 年にかけてのデータセットについて企業方向でデータを二分割した上で、一方をモデル構築用、他方を精度検証用として用いた。

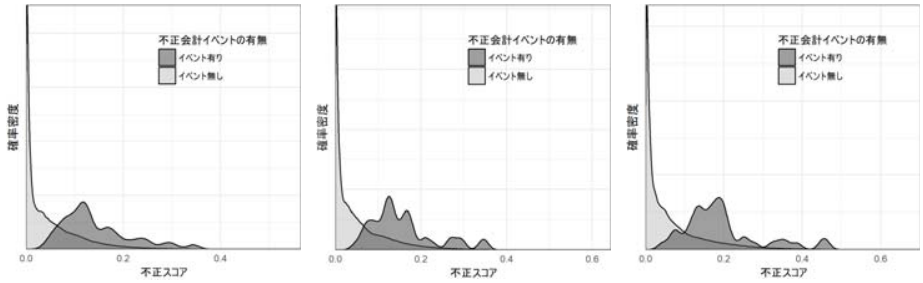
第一に、図 1 a は非上場企業（上場企業子会社を除く）を対象とするサンプルに対して fscore 及び firmown のみを用いて構築したモデルの予測精度について、out of sample でのイベント有りサンプルと無しサンプルのスコア分布を AUC と共に示したものである。同様に図 1 b は fscore、firmown、listed のみを、図 1 c は fscore、firmown、listed、parent のみを用いた場合の結果を AUC と共に示している。何れのケースも out of sample で 0.8 を超える AUC が達成されている。一方で、listed や parent の追加による AUC の改善はごく僅かであり、fscore 及び firmown から構成される予測モデルが、非線形性を許容するそのモデル特性などを背景として高い予測精度を実現していることが確認さ

[3] 代表的な変数については補論を参照。

れる。

図 1 a: AUC=0.882　　図 1 b: AUC=0.886　　図 1 c: AUC=0.887

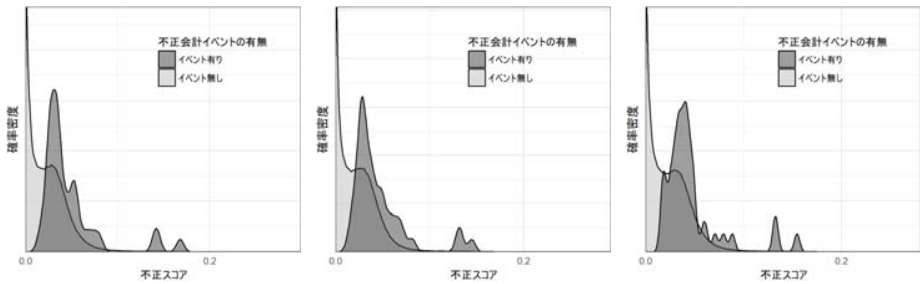
s.e.=0.015　　s.e.=0.014　　s.e.=0.014



第二に、図２a～c は、図１の三枚のパネルに対応するモデルへ、kessan、geoind、bank、network、theoretical を追加したときのスコア分布を AUC と併せて示したものである。何れのケースにおいても 0.8 弱の AUC が達成されており実務的には相応の予測精度が得られている。しかし、予測変数を拡充したにもかかわらず図１記載の各ケースに比して予測精度が低下しているという結果は、「極めて」高次元の情報がもたらすベネフィットが過剰適合などのコストによって相殺されていることを示唆するものである。

図 2 a: AUC=0.768　　図 2 b: AUC=0.769　　図 2 c: AUC=0.762

s.e.=0.025　　s.e.=0.027　　s.e.=0.026



最後に、図３の三枚のパネルは図１の三枚のパネルにおける予測変数の選択パター

ンを維持したうえで、分析対象企業へ上場企業と上場企業子会社を含めた上で再度予測モデルの構築を行い、out of sample での予測精度を確認したものである。比較的コンパクトの予測変数群で十分な予測精度が実現されているという上記の結果が、予測対象を拡充した場合でも引き続き確認されている

図３a: AUC=0.886 s.e.=0.012

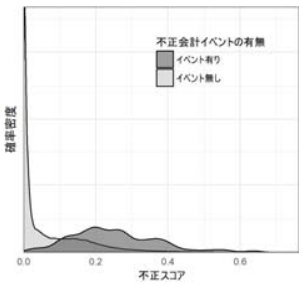


図３b: AUC=0.897 s.e.=0.017

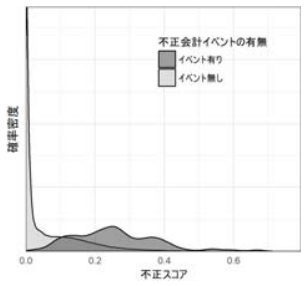
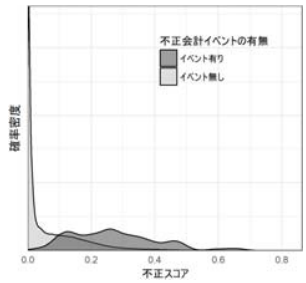


図３c: AUC=0.895 s.e.=0.011



4.2　ディスカッション

前節の結果は、上場企業で確認されている機械学習手法を用いた不正会計予測モデルの有効性が非上場企業でも確認されることを意味している。このように、不正会計フラグと企業レベルのパネルデータを利用することができれば、実務的に有効な取り組みが可能となるが、幾つかの留意点もある。

第一に、本稿では明示的に示していないが、上場企業「のみ」を対象とした不正会計の予測精度という点について、我々が構築したモデルは、膨大な財務指標を用いて予測モデルを構築した宇宿ほか (2019)の結果に必ずしも及ばない。この点に関連して、上場企業のみを対象として本稿で用いたフレームワークの下で上場企業向けの不正会計予測モデルを構築した場合には、前節の結果とは異なり、予測変数へ財務指標を含めることで予測精度が大幅に改善している。これらの結果は、上場企業を対象とした

不正会計予測において、宇宿ほか (2019)が用いた高次元の財務指標が重要な役割を果たしていたことを示唆している。

第二に、予測モデルの構築に際して常に生じる問題として、モデルの構造変化（concept drift）の問題が挙げられるが、不正会計を含む malpractice の予測に当たっては、不正の手口に関する変遷もあることから、この問題がより深刻となる可能性がある。この点に関して、本稿で用いた機械学習手法は迅速なモデルの改訂が可能であるという点で、裁量的な判断を必要とする局面を多く含む伝統的なパラメトリックモデルに比して優位性があると考えられるが、何れにしてもモデルの陳腐化による予測精度の低下が恒常的な問題として存在していることは認識すべきだろう。

５．まとめ

本研究は、非上場企業を含む広範な企業群を対象として機械学習手法に基づく不正会計予測モデルを構築し、その予測精度を検証したものである。年間数十万社に及ぶ企業レベルパネルデータと独自に構築した不正会計フラグ情報を用いたモデルの精度検証結果から、利用可能な情報が上場企業に比して乏しいことが想定される非上場企業についても、機械学習手法の利用による適切なモデル構築を行うことで不正会計予測が実行可能であることが確認された。本稿で構築した予測モデルの出力結果は、各企業に関する不正会計のリスクを示す指標として他の分析にも利用可能である。例えば、企業レベルの不正会計検知・予測とは異なる粒度の分析として、取引関係で接続されたクラスター単位での不正（例：循環取引）の検知・予測を行う際に、クラスター単位の予測変数として本研究で用いたスコアを参照するということも考えられるだろう。

補論：代表的な予測変数

変数区分	変数名
信用評点（fscore）	信用評点
企業自身の情報（firmown）	売上高関連変数 配当関連変数 利益率関連変数 創業年数 従業員数 資本金 その他
財務指標（kessan）	貸借対照表関連変数 損益計算書関連変数 その他
地理・業種情報（geoind）	近隣所在企業情報 同業他社情報 その他
銀行関連情報（bank）	取引銀行関連変数 その他
取引関係情報（network）	仕入先関連変数 販売先関連変数 その他
資本関係情報（shareholder）	株主関連変数 その他
実証会計分野の情報（theoretical）	ソフト資産比率 裁量的会計発生高 異常裁量費用
上場ステータス（listed）	上場関連変数 その他
親子関係情報（parent）	親子関係変数 その他