データサイエンス入門Ｂ
第２回　会計・金融におけるデータサイエンス・AI の活用

担当　坂上　学

１．金融分野での活用事例

金融市場等においては、実にさまざまな金融詐欺（financial fraud）が発生する。代表的なものは、（　　）詐欺、（　　）詐欺、（　　）詐欺（ネズミ講・マルチ商法)、粉飾決算、資産横領などである。

過去に史上最大級の巨額投資詐欺事件として有名なのがバーナード・（　　）事件である。公認（　　）のハリー・（　　）は早くから違法性に気付き、2000 年には米国（　　）委員会に告発をしていたが、9 年間放置され被害を拡大させてしまった。

１－１．保険金詐欺の検知技術

IAIS（　　国際機構）の ICP21 では、保険金詐欺の検知技術として、以下のものを挙げている。

- （　　）に基づく専門的な判断の活用
- （　　）信号 red flag のリストのチェック
- （　　）レビューの実施
- 社内外（　　）その他の情報源のチェック
- 音声（　　）分析、データ（　　)、（　　）ネットワーク、文章の信憑性チェックツール、IT ツールの活用
- 支払請求者に対する（　　）
- （　　）調査の実施

１－２．従来技術の特徴と問題点

保険金詐欺の発見に用いられていた従来の手法は、（　　）システムと呼ばれるものである。通常とは異なる（　　）や（　　）等の危険信号を発見した場合、これらを（　　）化して評価するルールを定め、一定の（　　）を超えた場合に、（　　）が必要な事案として抽出するものである。

ルールベースシステムは、その（　　）さが利点であるが、以下のような問題点もまた抱えていた。

- （　　）や検知漏れとなることが多い
- 追加調査に多くの（　　）的資源が必要
- （　　）的詐欺に対応できない
- （　　）なタイプの詐欺に対応できない

したがってこの手法を利用する保険会社は、定期的にルールの見直すことが重要とされている。

１－３．保険金詐欺の検知技術

次に示すような（　　）の分析技術を組み合わせ、得られた結果を（　　）モデリングによって（　　）評価することにより、保険金詐欺の傾向値を高い確実性で検出することに成功している。

なお予測モデリングとは、複数のデータ（　　）から収集された情報を相互に参照して（　　）マイニングを行い、データ相互の（　　）や（　　）性に内在する、不正である可能性が高いことを示す（　　）を識別して、詐欺の可能性を表すスコアを算出するものである。

技術	概要
（　　）	類似する保険金請求どうしを比較し、異常な点や金額の齟齬を検証し水増し請求等を検知する。システム内に蓄積されている多くの支払済保険金の情報を再利用する。
（　　）	すでに把握している詐欺犯罪者をデータベースから検索し、過去の不正請求の情報と照合して詐欺の常習者を検知するとともに、新たに提出された保険金請求の妥当性をより正確に評価する。
（　　）	自動車に設置されるテレマティクス装置は、保険料割引だけでなく事故状況の特定にも利用できる。複数の車両にテレマティクス装置が設置されていた場合、データを矛盾なく偽装することは困難である。
（　　）	行動分析、音声認識技術を基に、電話による会話から AI が不正請求を識別する技術も開発されている。保険会社による事故受付の電話の間、システムは特定の単語やフレーズ、そして声のトーンの形をとる潜在的な詐欺のシグナルを拾い出す。電話のオペレータは、システムによる警告を受けて応答を変更することができ、また、当該請求には疑わしいとのフラグが立てられる。
（　　）	画像、テキスト等の非構造化データは、保険金請求データの最大 80%を占める。テキストマイニングは、鑑定人の報告書、e メール、請求書類等を解析し、矛盾や状況を明らかにする上で役立つデータを抽出する。これらのデータは、文章の断片、書き間違い、表現の揺らぎ等を含む場合があり、整形、加工されてからシステムにより分析される。テキストマイニングは、Facebook、YouTube、その他のソーシャルメディアの膨大な量のデータの分析にも用いられるようになっている。
（　　）	ブログ、Twitter、Facebook、LinkedIn、YouTube、その他多くのソーシャルメディア投稿を検索して分析を行う。事故関係者どうしのつながりや、事故の際にどこにいたかも識別対象である。例えば、目を負傷した請求者が、しばらくしてから見たばかりの映画に関するフォーラムで発言したり、火災保険金の請求者が、火災の直前に不動産を売ることについての質問を投稿したりという事実が判明する場合もある。ソーシャルメディア分析の実行は 1 回限りの作業ではなく、定期的に、または請求の期間中に何度か実行する必要がある。
（　　）	一見無関係な大量の請求データを参照して、人、場所、アカウント、企業、電話番号、車両識別番号等の間の隠れた関係を明らかにする。例えば、自動車事故の被害者、車両の所有者、ドライバー、修理工場等の住所や電話番号が多くの保険金請求

	に関与しているような場合の検知が可能で、組織的な不正行為を特定するのに有効である。 システムは、新しい請求が発生する度に相互に関連するネットワークを継続的に更新して疑わしいネットワークにフラグ付けを行うため、調査担当者は顧客全体のネットワーク関係を瞬時に検証できる。

１－４．保険金詐欺に対する不正検知技術の採用割合

伝統的な危険信号とルールベースシステムが最も多いものの、ネットワークリンク分析、予測モデリング、アノマリー分析等のデータサイエンス的手法の利用が増えているが、2019 年時点では、AI の活用はまだ少ない。

種類	割合
危険信号＋ルールベース	84%
（　　）分析	60%
予測モデリング	55%
（　　）分析	55%
（　　）	33%
（　　）	6%
その他	4%

２．会計分野での活用事例

会計分野へのデータサイエンス・ＡＩの活用は、会計不正（accounting fraud）への対応として始まった。会計不正の典型的な例は、（　　）である。通常は、損失がでているのに利益が出ているかのように装うのが（　　）である。しかし膨大な利益が出ているのに、あまり出ていないように装う場合もある。その場合は、（　　）と呼ばれる。もう一つの例は、（　　）、すなわち従業員による回収金の着服などである。

２－１．さまざまな監査制度

会計という領域では、（　　）と呼ばれる実務が営まれており、証券取引所に上場している会社や、一定規模以上の会社は強制されている。企業内には（　　）が整備されており、内部監査人による（　　）や、公認会計士による（　　）を通じて、事前に不正や誤謬のチェックがおこなわれるため、実際に公表される決算書から会計不正が発見されることは、あまりない。

内部統制の有効性については、内部統制（　　）制度によって常にチェックが入るようになっている。この制度をもたらすきっかけとなったのは（　　）事件である。その後も不正経理事件が続き、多くの投資家が莫大な損失を被る事件が続いたため、企業責任、開示制度、監査人の独立性、監査法人の監視体制を強化するために、2002 年 7 月に（　　）法が成立した。法案を提出した、（　　）と（　　）の 2 人の議員の名前をと

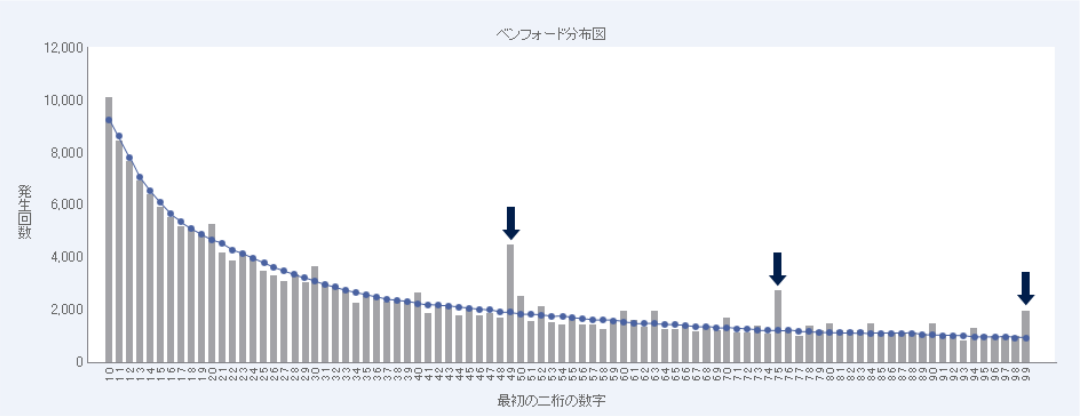
って SOX 法とも呼ばれている。

日本でも、（　　）取引法において内部統制報告制度が規定されたが、これがいわゆる日本版（　　）法であり、2008 年 4 月 1 日以降に開始される事業年度から適用となった。しかしながら、その後も（　　）事件や（　　）製紙事件が 2011 年に発覚し、2015 年には（　　）の不適切会計が 2015 年に発覚するなど、会計不正事件は後を絶たない。

２－２．監査データアナリティクス

企業の日々の取引は、膨大な量の（　　）データとして蓄積されていくが、これらのデータに対し、データサイエンスの技法を用いた監査データ（　　）が、今日隆盛を極めている。

たとえば仕訳データの最初の 2 桁の数字の出現頻度は（　　）の法則に従うので、異常な（　　）が見られた場合、不正が行われている可能性が高いということになる。



２－３．安全性分析が経営分析の基本

経営分析は、（　　）比率（流動資産÷流動負債）が（　　）%を超える企業は倒産をしないことを、ある銀行家が発見したことから始まったとされている。企業の安全性の分析には、流動比率の他に以下のような財務指標もしばしば用いられている。

- （　　）：当座資産と流動資産との比率
- （　　）：純資産に対する負債の比率
- （　　）：総資本に対する自己資本の比率
- （　　）：自己資本に対する固定資産の比率
- （　　）：事業利益が金融費用の何倍あるかを示す指標

２－４．アルトマンの Z スコアと AI 的意義

倒産予知の高度化は（　　）の Z スコアから始まった。彼、米国の 66 社の倒産企業と非倒産企業のデータを用いて（　　）モデルを構築し、以下の式で Z スコアを算出

し、（　　）を超えると安全、（　　）未満では倒産、その間の値がグレーであるとされる。

Z = 0.012X1 +0.014X2 +0.033X3 +0.006X4 +0.999X5

X1=運転資本÷総資産

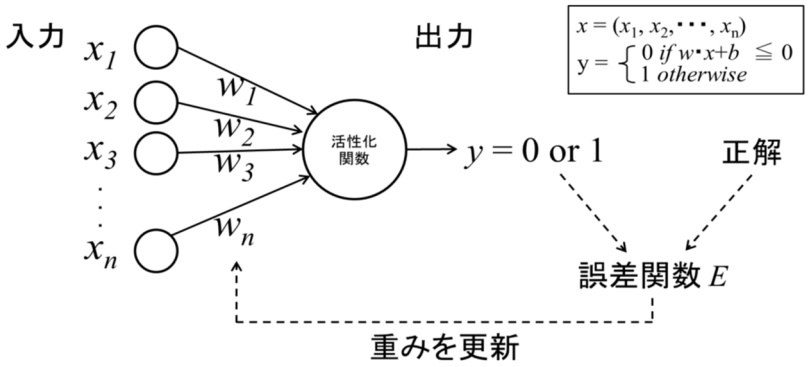
X2=利益剰余金÷総資産

X3=（支払利息・税引前当期利益）÷総資産

X4=株式時価総額÷負債簿価

X5=売上高÷総資産

Z スコアで用いられている判別式は、AI におけるニューラル・ネットワークの初期段階において開発された、（　　）と構造が類似していることが分かる。



しかしながら（　　）と（　　）によって、（　　）の問題を解けないことが指摘された。AND（　　）、OR（　　）、NAND（　　）、XOR（　　）という 4 つの基本回路のうち、XOR 回路については線形式では分離できない問題をいう。

線形非分離問題の解決方法は大きく 2 つのアプローチがある。線形式で分離できないのであれば、（　　）を用いるというアプローチと、平面上では分離できないのであれば立体化して（　　化）するというアプローチである。後者のアプローチをさらに発展させたものが、（　　）ネットワークという AI 手法である。

３．定量・定性データの両方を用いた研究事例の紹介

定量データ（財務数値）と定性データ（文字情報）の両方を使った倒産分析の事例の概要を紹介する。

2000 年 4 月以降にわが国において、破綻した上場企業の倒産直前期と同時期に（　　）していた上場企業の有価証券報告書の非財務情報、すなわち（　　）報告、（　　）、（　　）情報などのテキスト部分を比較し、出現単語の差分を解析した。

サンプル企業の有価証券報告書から抽出されたサンプルデータをもとに、以下の 4 つのパターンにより解析をおこなった。

- 解析 1：範囲はファイル全体で、助詞・記号・数字をフィルタリング
- 解析 2：範囲は業績部分のみに限定し、助詞・記号・数字をフィルタリング
- 解析 3：範囲は業績部分のみに限定し、動詞・形容詞のみをフィルタリング
- 解析 4：範囲は業績部分のみに限定し、一般名詞のみをフィルタリング

また語句の重要度の評価方法としては、(1)「倒産企業の有価証券報告書に出現した平均頻度」から「非倒産企業の有価証券報告書に出現した平均頻度」を差し引いた値を用いて評価を行ったもの、(2)「倒産企業の有価証券報告書に出現した平均頻度」から「非倒産企業の有価証券報告書に出現した平均頻度」をさらに「合計頻度」で割った値を用いて評価を行ったもの、(3)「倒産企業の有価証券報告書に出現したファイルの割合」から「非倒産企業の有価証券報告書に出現したファイルの割合」を差し引いた値を用いて評価を行ったもの、(4) 機械学習（線形判別学習アルゴリズム）による評価を行ったもの、という 4 つの評価方法を用いた。

種々の解析結果を慎重に観察した結果、①データ範囲はファイル全体を用い、助詞・記号・数字をフィルタリングしたもの（解析 1）から、②「倒産企業の有価証券報告書に出現したファイルの割合」から「非倒産企業の有価証券報告告に出現したファイルの割合」を差し引いた出現頻度値について評価を行った（評価方法 3）ランキングに、興味深い結果が見られた。具体的には、出現頻度に顕著な差の見られる言葉（上位にある項目）には「配当」「利益処分」「留保」といったものが並び、なかでも「配当金」「配当性向」といった言葉の出現頻度は 2 群間で顕著な差が見られた。つまり、継続企業（非倒産企業）においては、常に株主を意識したメッセージが有価証券報告書の中に織り込まれていることが確認できる。また、倒産企業群では 1 社もこの言葉が出現していないことがわかる。

この分析結果からの結論をまとめると、倒産企業群と非倒産企業群とにおいて顕著とまでいえる出現率の差異がみられる文言は確認されなかったので、倒産企業を特定する文言を抽出するのは困難である一方で、継続企業を特定することは可能であることである。

４．テキストマイニングの基本的な手法

４－１．テキストマイニングの基本手法

テキストマイニング技術は、以下のような技術を組み合わせた複合技術である。

- 対象となる文書データを扱うための（　　）
- 膨大なテキストデータを扱うための文書（　　）の技術
- 膨大なデータから有用なパターンを発見するための（　　）
- テキスト分類のための（　　）のアルゴリズム

テキストマイニングにおける自然言語処理では、大きく 2 つの段階で処理がなされる。第 1 のステップは、文書中のテキストデータの基本的な情報を解析するために、(　　) 解析や (　　) 解析といった手法が用いる。第 2 のステップは、(　　) を抽出したり、語句内容の (　　) を認識させたり、(　　) を認識させたりする。

４－２．形態素解析

形態素解析とは、簡単に言えば意味をもつ最小の言語単位、すなわち（　　）の単位で語句を区切って（　　）を割り当てることである。

たとえば「継続企業の前提が棄却される条件を探る」という文を形態素解析すると、次のような結果が得られる。

表　層　語	基　本　形	品　　詞
継続	継続	名詞・サ変接続
企業	企業	名詞・一般
の	の	助詞・連体化
前提	前提	名詞・一般
が	が	助詞・格助詞・一般
棄却	棄却	名詞・サ変接続
さ	する	動詞・自立
れる	れる	動詞・接尾
条件	条件	名詞・一般
を	を	助詞・格助詞・一般
探る	探る	動詞・自立

４－３．テキストマイニングのためのツール

先ほど紹介した研究では、ChaSen と呼ばれるツールを使って形態素解析を行っていた。ChaSen をはじめ形態素解析ツールには以下のようなものがある。

- ChaSen（可変長マルコフモデル）https://chasen-legacy.osdn.jp/
- JUMAN++（bi-gram マルコフモデル）https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN%2B%2B
- MeCab（bi-gram マルコフモデル）https://taku910.github.io/mecab/
- KAKASHI（最長一致）http://kakasi.namazu.org/index.html.ja

４－４．重要語の抽出

テキストマイニングでは、どのような内容が多いのか少ないのか、どのような内容が増えているのか減っているのか、どのような内容とどのような内容が関連性が高いか、といった点について分析を行っている。

これらの統計的な情報をもとにして、重要語句とはどのようなものであるかを考えることになるが、この場合の「内容」とは語句の意味を示しているだけでなく、（　　）あるいは文字コード列であったり、（　　）の意味であったりする点は留意する必要がある。

多くの場合、重要詩句・キーワードとなるのは（　　）であり、なかでも（　　）が重要とされ、助詞や助動詞がキーワードになることは滅多にない。

ここで代表的な重要語アルゴリズムである TF-IDF 法を紹介する。TF-IDF 法は,「ある語句の単文書内での（　　）と、その語句を含む（　　）を全文書数で割った数の（　　）との積をとった値を測定する。

出現頻度×（log（総文書数÷出現文書数）＋1）

TF-IDF 法の例示

- A 社の有価証券報告書には、「財政状態」という語が 1 回出現し、「資金繰り」という語が 3 回出現している。
- B 社の有価証券報告書には「財政状態」という語が 2 回出現し、「内部留保」という語が 2 回出現している。
- C 社の有価証券報告書には「財政状態」という語が 3 回出現し、「自己資本比率」という語が 1 回出現している。

各社の有価証券報告書における各語句の重要度の評価は以下のように計算される。

- A 社の有報における「資金繰り」という語句の評価：3×（log（3÷1）＋1）＝3.4771
- A 社の有報における「財政状態」という語句の評価：1×（log（3÷3）＋1）＝1
- B 社の有報における「内部留保」という語句の評価：2×（log（3÷1）＋1）＝2.4771
- B 社の有報における「財政状態」という語句の評価：2×（log（3÷3）＋1）＝2
- C 社の有報における「自己資本比率」という語句の評価：1×（log（3÷1）＋1）＝1.4771
- C 社の有報における「財政状態」という語句の評価：3×（log（3÷3）＋1）＝3

「資金繰り」という語は、A 社の有価証券報告書にしか出現せず、かつ 3 度も出現していることから評価が（　　）なる。「自己資本比率」という語は、C 社の有価証券報告書にしか出現しないが、1 度しか出現しないので、A 社の「資金繰り」に比べ評価が（　　）なる。C 社の有価証券報告書には「財政状態」が 3 度出現するが、すべての有価証券報告書に出現するので、A 社の有価証券報告書にしか出現しない「資金繰り」よりも評価が（　　）なる。

以上