研究レポート　No.9　～決算書データに基づく非違予測～

2020 年 10 月 21 日　株式会社アイズファクトリー　https://bodais.com/company/

概要

決算書データを利用した解析の一例として非違予測を取り上げる。予測モデルを構築するには正解データが必要であり、効果的なモデル構築を行うためには正解データに求められる要件が存在する。非違行為の予測というあまり馴染みのない分析事例において、どのように正解データが検討されるのか、犯罪予測や病理検査の例に倣って論じる。

1.　はじめに

非違行為とは一般に違法行為のことを指し、違法であっても刑罰が科されるか行政処分かで法的効果は様々である。ここでは刑罰が科される犯罪よりも広い概念と捉えることにする。非違行為を事前に予測することは、災害の予測と同様に、国の統治を効率的におこなうために重要な役割を果たす。可能性の高い順に対策をとっていくことがリソース配分の最適化繋がるからである。本稿では、非違行為の予測モデルの構築の実例を紹介したい。どのようなモデルを考えるかは、正解データを規定することから始まる。一口に非違行為といっても概念が広すぎて、数理論的な定式化に乗せるのは困難だからである。

正解データを種類ごとに規定して予測モデルを検討することは、犯罪学の分野では先例がある。犯罪がいつどこで発生するかを知ることは、効率的な警察活動を行うために非常に有用である。犯罪発生の予測には、数理解析的な要素以外に、同じ場所や時間帯で同種の犯罪が繰り返される傾向（時間空間集積性）や、過去に成功した手口は効率が良いため再び犯罪を行いやすいという犯罪行動理論が考慮される。この傾向は罪種によって特徴が異なっており、例えば、財産犯では、犯罪合理性（捕まるリスク・労力の最小化、犯罪利益の最大化）について一定の意思決定がなされ、その計画性に起因して時間集積性が認められる。その一方で、非合理的で偶発的な粗暴犯には時間集積性は認められない。空間集積性についても、ひったくりは特定の地区で長期にわたり繰り返される傾向が見られたり、侵入犯は被害地点付近において短い時間で繰り返されるなど、各罪種の発生メカニズムに依存して集積性の傾向は異なる[1]。このような犯罪学の知見である近接反復性を活用した機械学習モデルの研究がなされている。[2,3]

2.　正解データ

本稿では、決算書データに基づいて、国税納付に関しての非違行為を予測することを検討してみたい。具体的には、国税の納付に関して税務調査の対象となるか否かを判定（法令順守法人と非違法人とに分類）することを考える。

判定問題における正解データの構築について定性的な理解を行うには、マルウェア感染や病理検査での陽性判定において、問題なし（ホワイト）か、問題あり（ブラック）かを判定する場合を考えると理解しやすい。モデル構築のための正解データ（調査対象）が、ブラックまたはグレーなデータのみからなる場合、対象間の特徴がうまく差別化できていないため、どんなモデルを作成してもブラックとグレーの峻別は難しい。ホワイトデータを一定量含めることで、モデルの精度を確保することができるのである。従って、モデル構築を考える際には、正解データの白黒比率を検討する必要がある。ダイレクトメールの反応率では3%程度の白黒比率（正解率）であるので、これに倣って白黒比率が3%程度に達するまで問題とならない法令順守企業を分析データに含めることにする。件数的にはおよそ、260 万社に対して 8 万社の割合の構成となる。

3.　分析粒度

犯罪分析では罪種ごとにモデルを検討しているが、非違法人の特徴は、業種だけでなく売上規模によっても異なっており、そうした業種に拠る特徴の違いを加味した予測モデルを構築する。4 桁からなる業種コードでは、273 業種に分割されており、その粒度も揃っていない。そのために、全法人データを業種・売上規模に応じて適当な粒度のクラスタに分割する必要がある。本稿では、予測モデルの統計的信頼性を担保するのに必要な法人数を算出し、これを基準として粒度調整を行い業種の統合を実施している。

具体的には、非違法人比率（正解率）のサンプリング誤差が 95%信頼水準で 10%以内に収まるのに必要な法人数を算出すると、売上が中・低階級の場合で 5,000 法人、高階級の場合で 3,000 法人が必要となる。一方、業種コードの階層で粒度を考慮すると、業種コードの上 2 桁の階層では、中・低階級でおよそ 2 万法人、高階級でおよそ 5,000 千法人の粒度となる。そこで、業種コードの上 2 桁の階層を基本統合業種とし、業種による粒度のばらつきや業種の定性的な特性も加味して、基本統合業種の分割・統合を行い最終的な統合業種（中・低階級 94 業種、高階級 92 業種）を決定した。

4.　変数選択

分析用データは、法人申告事績データであり、売上、営業損益、個人換算所得金額などの勘定科目金額が含まれている。一般に、世帯年収の分布が対数正規分布に従うことが知られており、上述の勘定科目金額についても対数正規分布に従うと予想される。実際、勘定科目金額の常用対数を取ってヒストグラムを描くと、概ね正規分布の形状を示しており、対数正規分布に従うことが明らかになった。そこで、勘定科目金額は全て常用対数で対数化した値を使用することにした。その際に注意すべき点は、勘定科目金額の数値には 0 や負値も許されることである。そのため、対数化を次のように定義した。（Log は常用対数）

勘定科目数値>0 の時　　Log (勘定科目数値)
勘定科目数値=0 の時　　0
勘定科目数値<0 の時　－Log (勘定科目数値)

前年からの変化を考慮するため、各勘定科目の前年からの差分も説明変数に取り込む。対数化した変数の前年度との差分は、対前年度比を扱うことと同等である。

因みに、対数化した変数による線形回帰モデルは、計量経済学で言うところの乗法モデルに対応しており、その回帰係数は価格弾力性に相当する（経済時系列データを傾向変動、循環変動、季節変動、不規則変動に因子分解された形をとる）。線形回帰式の両辺の指数関数を取ると線形和で記述された回帰式が、各変数の冪乗積に変換されるので乗法モデルと呼ぶ。[4] このとき、冪関数の冪に

相当するのが回帰係数であり、この値が 1 より小さい正値を取る場合には、乗法モデルは説明変数の増加に伴い収穫逓減の振る舞いを示す。

説明変数の説明力を測る指標として AIC を採用した。AIC（赤池情報量基準）とは、本来、回帰モデルの適合度を測る指標として開発されたもので、所与のデータの下での予測値の対数尤度とモデルのパラメータ数を組み合わせた指標である。予測残差が小さく、モデルのパラメータ数が少ないほど AIC は小さな値をとる。[5]

各説明変数について当該変数だけで非違法人を予測する単変数モデルを作成し、その AIC に拠り当該変数の効果を測定することとした。マルチコ対応としては、変数間の相関係数を算出して強相関の変数ペアについては一方を削除するとともに、変数ごとに VIF を計算して VIF が高い変数は除外するという操作を行った。その際の判定閾値は、相関係数が 0.9 以上、VIF 値が 10 を超える場合とした。

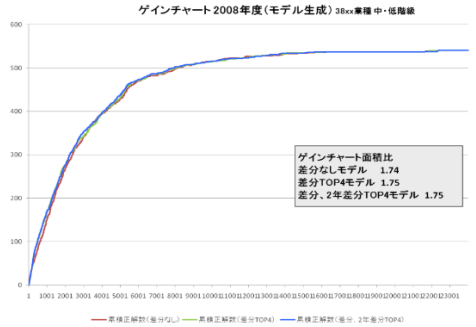
5. ロジスティック回帰分析による非違確率予測

今回は、分析モデルの単純化と比較の容易さを考慮して、全業種で同じ変数を使用することにした。その選択には業種コード 380* における相関係数、VIF による変数選択および AIC による変数順位を参考とした。結果として、73 変数（勘定科目 43 変数、差分変数 30 変数）が選択された。統合業種ごとにモデル構築と予測検証を実施して、面積比の値を評価した結果を表 1. に示す。

表 1. は、モデル精度を業種ごとに面積比で算出し、業種コード大分類で平均値を求めたものである。「予測との差」はモデルゲインチャートの面積比と予測ゲインチャートの面積比の差分の平均値である。中・低階級においては、全般的に精度が高く、面積比は全業種平均で 1. 72、モデルと予測との面積比の乖離も 0. 10 であり、ロバストかつ高精度のモデルが実現できている。

業種（大分類）	平均	標準偏差	予測との差
製造業	1.81	0.07	0.05
卸売業	1.75	0.05	0.06
小売業	1.73	0.08	0.10
建設業	1.79	0.00	0.03
運送業	1.70	0.21	0.18
サービス業1	1.64	0.23	0.17
サービス業2	1.78	0.01	0.01
料理・飲食店業	1.68	0.08	0.15
その他の業_金融	1.75	0.01	0.03
その他の業	1.68	0.07	0.07
総計	1.72	0.15	0.10

【表 1】 ゲインチャート面積比（中・低階級）



【図 1】 ゲインチャートのイメージ

モデルに有効な変数について AIC ランキングを集計し、表 2. にまとめた。表 2. には、業種ごとに AIC ランキング上位 5 変数にランクインした回数を集計した。中・低階級では、繰越欠損金額、申告所得控除前、売上、売上総利益、が上位を占めている。業種による細かな違いは有るものの概ね同様の傾向を示している。

項目名	サービス業1	サービス業2	その他の業	運送業	卸売業	建設業	小売業	製造業	料理・飲食店業	総計
繰越欠損金額	21	2	10	9	12	2	16	18	4	94
申告所得控除前	21	2	10	9	12	2	15	18	3	92
売上	19	2	10	7	12	2	12	18	4	86
売上総利益	16	2	4	8	10	1	14	10	1	66
付加価値金額	7	1	9	1	10	2	12	13	3	58
個人換算所得金額	6		6	8	3		2	9	1	35
資本の部計	2			2			5	1	1	11
営業損益	6					1		3		10
その他販管費	1	1	1		1		3		1	8
現金預金	1						1			2
建物	1									1
減価償却費	1									1
資産合計				1						1
資本の部計 差分	1									1
資本金額	1									1
代表者貸借料 差分									1	1
地代家賃・租税公課	1									1
売上原価									1	1

【表 2】 AIC 上位ランク回数（中・低階級）

表には示さなかったが、高階級においても、面積比は全業種平均で 1. 52、モデルと予測との乖離も 0. 08 でロバストかつ高精度なモデルを実現できた。有効変数については、繰越欠損金額、申告所得控除前、個人換算所得金額、付加価値金額、が上位を占めた。

6. まとめ

本稿では、犯罪予測を例にとり、正解データの特徴が罪種ごとにモデルに反映されるという背景事情を概説した。その考え方を基にして、非違予測の場合には、業種と売上規模という 2 軸での分類が必要となることを、具体的な手順に従って説明した。正解データを規定する要件には、サンプル数や正解率といった単純な数値的要件だけではなく、予測モデルを当てはめる集団をどのように規定するのかという分析粒度の要件も関係してくる。更には、予測の課題をどう設定するか（非違の判定とするか、非違の大きさ（不正金額）とするか）という設計的要件でもモデルの構築設計に大きく関わってくる。分析精度を向上させるためには、データクレンジングや変数選択などの解析精度に係る重要な処理も然ることながら、正解データの構成をこれらの規定要件に照らし合わせて検討することも重要である。

7. 参考文献

[1] 菊池, 雨宮, 島田, 齋藤, 原田, “近接反復被害の罪種間比較 -時空間 K 関数の応用-”, GIS-理論と応用, Vol.18, No.2,(2010) pp.21-30.

[2] 大山,雨宮,島田,中谷,”地理的犯罪予測研究の潮流”,GIS-理論と応用, Vol.25, No.1 (2017).

[3] 中川, 小西, 宮野, “犯罪発生履歴データの機械学習による時空間カーネル密度推定型犯罪予測の最適化”, FIT2018（第 17 回情報科学技術フォーラム）.

[4] 中村,新家,美添,豊田「経済統計入門（第 2 版）」東京大学出版会.

[5] 坂元,石黒,北川, 情報科学講座 A・5・4「情報量統計学」, 共立出版（1983）.