

機械学習を用いた倒産予知モデルの研究

片桐 英樹* 平井 裕久* 松丸 正延**

Bankruptcy prediction model using machine learning

Hideki KATAGIRI* Hirohisa HIRAI* Masanobu MATSUMRU**

1. 緒言

世界経済の不安定な状況の中、倒産予測の関心が高まっている。機械学習と人工知能テクノロジーの台頭により、倒産予測の精度を向上させるために多くの新しい方法論が提案されてきている。

平井[1]は特徴量に財務指標のみを用い、業種別に複数の機械学習手法による分析を行った。しかし、不均衡データへの対策は Synthetic Minority Oversampling Technique (以下: SMOTE) のみを用いており、他の対策との比較を行っていない。また、Zhou[2]は特徴量に財務指標のみを用い、複数の機械学習手法と不均衡データへの対策を使用しているが、業種別にモデルを構築していない。Zhao [3]は、データサンプルの不均衡など、倒産予測モデルの従来研究調査をしている。しかし、投融資ネットワーク指標 (以下: NW 指標) を用いた論文は見当たらない。

本研究では、財務指標のみを用いた倒産予測モデルに対して、企業間の長期的な信頼関係を捉えた NW 指標を追加して予測精度向上を図る。また、倒産企業と非倒産企業の不均衡データを扱う際に生じる倒産企業のデータパターンの学習困難を克服するために、3 種類のリサンプリング手法を用いる。さらに、業種ごとに倒産予測を行い、業種ごとの倒産要因を明らかにする。機械学習手法としては、LightGBM とランダムフォレストを用いて比較を行う。

3. 研究内容

3.1 分析データ

東京証券取引所の分類基準に基づく建設業・不動産業・サービス業・小売業・卸売業・電気機器業の 6 種類の業種についてそれぞれ 3 種のデータを作成する。

まず、日経 NEEDS-Financial QUEST を用いて、企業の財務指標及び投融資関係データを取得する。財務指標の数は 161 指標、NW 指標の数は 12 指標である。次に、3 種のデータを作成する。1 種目は財務指標のみのデータ (以下: 財務データ)、2 種目は財務指標と NW 指標を用いたデータ (以下: 投融資データ)、3 種目は投融資データから NW 指標を除いたデータ (以下: 比較データ) である。

比較データは、それぞれのデータとの比較を行うために用いる。

財務データとの比較では、有効データ数の差による比較を行う。投融資データとの比較では、使用する指標の差による比較を行う。各業種、各データの有効データを表 1 に示す。ただし、括弧内の数字は使用する指標数である。

表 1 業種別有効データ数

業種	倒産			非倒産		
	財務 (161)	投融資 (173)	比較 (161)	財務 (161)	投融資 (173)	比較 (161)
建設	42	25	25	6280	3848	3848
不動産	34	21	21	3603	2089	2089
サービス	25	5	5	11445	4480	4480
小売	23	10	10	11531	5808	5808
電気機器	23	11	11	9013	4722	4722
卸売	16	5	5	10935	5650	5650

3.2 分析手法と評価指標

業種 6 種類、特徴量 3 種類、不均衡データへの対策であるリサンプリング手法 3 種類、機械学習手法 2 種類を組合せの計 108 種のモデルを構築する。

リサンプリング手法には、k-means 法を用いたアンダーサンプリング (以下: k-means)、SMOTE、SMOTE + Edited Nearest Neighbor (以下: SMOTE+ENN) を用いる。機械学習手法には、Light Gradient Boosting Machine (以下: LightGBM) とランダムフォレストを用いる。モデルの評価では、倒産予知の研究であるため、Recall を重視する。

4. 実証実験

4.1 実験結果

紙面の制約上、建設業、不動産業、小売業の 3 業種に着目して、LightGBM における NW 指標の有用性と機械学習とリサンプリング手法の組合せの検証について記載する。ランダムフォレストの結果については簡単に触れる。なお、不動産業と小売業については、結果の一部抜粋を示す。サービス業、電気機器業、卸売業については結果も割愛する。建設業、不動産業、小売業の結果を表 2、表 3、

*教授 経営システム工学科
Professor, Dept. of Industrial Engineering and Management
**客員教授 工学研究所
Visiting Professor, Research Institute for Engineering

表 4 に示す。ただし、評価指標の Accuracy を Acc, Recall を Rec, Precision を Pre, F1-score を F1 と表記した。

また、リサンプリングにおいて SMOTE を用いたモデルを S, SMOTE+ENN を用いたモデルを S+E, k-means を用いたモデルを K と表記した。全ての業種において、財務モデルと比較モデルの比較を行うと、比較モデルの精度が下回った。この原因は有効な倒産データ数が減少し、精度に影響を及ぼしていると考えられる。

表 2 建設業の検証結果(単位:%)

モデル	Acc	Rec	Pre	F1
S 財務	93.89	69.05	7.21	13.06
S+E 財務	91.68	76.19	5.84	10.85
K 財務	75.81	76.19	2.06	4.02
S 比較	91.73	52.00	4.23	7.83
S+E 比較	94.17	44.00	5.16	9.24
K 比較	72.35	64.00	1.56	3.03
S 投融資	95.00	52.00	6.99	12.32
S+E 投融資	95.41	48.00	7.10	12.37
K 投融資	81.96	84.00	3.07	5.92

表 3 不動産業の検証結果(単位:%)

モデル	Acc	Rec	Pre	F1
S 財務	94.36	70.59	10.96	18.97
S 投融資	94.37	52.38	9.74	16.42
S+E 投融資	95.13	61.91	12.75	21.14
K 投融資	71.78	52.38	1.95	3.77

表 4 小売業の検証結果(単位:%)

モデル	Acc	Rec	Pre	F1
S 財務	93.74	60.87	1.92	3.73
S+E 財務	95.11	60.87	2.46	4.72
K 財務	94.57	52.17	1.91	3.69
S 投融資	93.60	10.00	0.28	0.54
S+E 投融資	93.40	20.00	0.53	1.04
K 投融資	99.83	0.00	0.00	0.00

4.2 建設業の検討・考察

建設業の機械学習手法において Recall の値が高いのは LightGBM においては、K 投融資モデルであった。ランダムフォレストにおいても、K 投融資モデルであった。

特徴量重要度を算出すると、次数中心性、ページランク、オーソリティ度の NW 指標が上位にランクされていた。建設業は一つのプロジェクトに対し、複数の企業が協力し合う構造をもち、企業間の協力関係が不可欠であり NW 指標は重要な指標であると考えられる。一方、財務指標では、自己資本比率や預借率の安全性指標が上位にランクされていた。依頼から納品まで時間がかかることから財務安定性が求められていると考えられる。

機械学習とリサンプリングの組み合わせでは、Light GBM と K 投融資モデル、ランダムフォレストにおいても K 投融資モデルとの組み合わせが有用である。

4.3 不動産業の検討・考察

機械学習手法において Recall の値が高いモデルは S 財務モデルであり、2 番目は S+E 投融資モデルであった。ランダムフォレストにおいては、SMOTE+ENN または k-means との組み合わせであり、NW 指標を含んだ投融資モデルは高い値を示さなかった。

特徴量重要度を算出すると、売上高設備投資比率やキャッシュフローに関する財務指標が上位にランクされていた。売上高設備投資比率は、売上高に対する設備投資額の割合を示すことから、売上高で回収した資金で、次の販売に向けての多くの土地や建物を取得していることを表していると考えられる。キャッシュフローは現金の流れを示す。不動産業においては、取引を有利に進める上で、土地や建物の取引額を自社の現金で賄える状態にしておくことが重要である。顧客を対象とする不動産事業では、企業同士は競争相手であることから、資金調達の多様性を表す NW 指標が、予測の結果に与える影響は小さいと考えられる。

機械学習とリサンプリングの組み合わせでは LightGBM と SMOTE モデル、ランダムフォレストにおいても k-means モデルとの組み合わせが有用である。

4.4 小売業の検討・考察

小売業の機械学習手法において Recall が高いモデルは、S 財務モデル、S+E 財務モデルであった。ランダムフォレストにおいても S 財務モデル、S+E 財務モデルで高い値を示し、投融資ネットワーク指標は含まれていなかった。投融資ネットワーク指標を用いることで、倒産データが減少し、学習が困難になるためと考えられる。倒産データ数が少ない小売業、卸売業、電気機器業も同様の結果となっている。

特徴量重要度では、売上高純金利負担率が上位にランクされた。小売業において需要変動や新商品の導入など発生する資金ニーズを売上高で賄うことが重要であることが示唆された。

機械学習手法ではランダムフォレストの k-means がどのデータセットでも精度が高く有用であり、LightGBM の精度は低かった。

5. 結言

本研究では、財務指標に加え、NW 指標を用いて、業種ごとに複数の機械学習手法と不均衡データへの対策を考慮した予測モデルを構築し、NW 指標の有用性と機械学習とリサンプリング手法の組み合わせの検証を行い、精度向上を図った。一部の業種モデルで NW 指標の有用性があることを示した。加えて、業種によって有用な機械学習手法とリサンプリング手法の組み合わせを明らかにした。

参考文献

- [1] 平井江利香, 財務状況の経年変化を考慮した業界別倒産予知モデルの構築, 神奈川大学工学部経営工学科 2019 年度卒業論文 (2020)
- [2] Zhou Ligang, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods", Knowledge-Based Systems, Vol.41, pp.16-25 (2013)
- [3] Zhao Jinxian et al, "Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction", Machine Learning with Applications, Vol.15, 100527(2024)