創業企業の信用リスクモデルにおける人的要因の有効性
― 機械学習モデルとの比較 ―

日本政策金融公庫 国民生活事業本部 *内海 裕一 UTSUMI Yuichi
日本政策金融公庫 国民生活事業本部 峰下 正博 MINESHITA Masahiro
05000227 日本政策金融公庫 国民生活事業本部 尾木 研三 OGI Kenzo
01505910 慶應義塾大学理工学部 枇々木 規雄 HIBIKI Norio

1. はじめに

創業前の企業は決算書の情報がない。尾木ら[1]は、創業前の限られた情報だけで信用リスクモデル（ロジスティック回帰モデル）を構築し、実務で利用可能な精度であることを明らかにした。分析の過程で「創業時の年齢」や「斯業経験年数」といった創業者の人的要因が有効であることを確認したが、データ数が少なく、オーバーフィッティングの可能性があるため、尾木ら[1]はダミー変数を用いている。

その後、データの蓄積が進んだことから、本研究では「創業時の年齢」「斯業経験年数」とデフォルトとの関係を見直して、モデルの精度向上を目指す。さらに、最近注目されている機械学習の手法を用いて精度が向上できないか試みる。

分析の結果、「創業時の年齢」は4次関数、「斯業経験年数」は2次関数で近似したスコアを変数として投入すると、AR値が48.9%から50.9%に2%ポイント上昇した。また、同じ変数を用いて機械学習の手法を使ったモデルを複数構築して比較した結果、同程度のAR値を得られた。

2. データの概要

本研究の使用データは、日本政策金融公庫が2011年度から2017年度までに融資した創業企業92,638社のデータDB1と、2017年以降に取得を開始した18,075社のデータDB2である。データの概要を表1に示す。＊印は尾木ら[1]のモデルで採用された変数である。

表1　使用データ

データ	融資時期	データ数	主な変数
DB1	2011年4月～2018年3月	92,638	＊創業時の年齢 ＊負債情報（金額、用途、履行状況） ＊業種 ・創業時の従業者数 ・融資金額・資本金 など
DB2	2017年1月～2018年3月	18,075	＊斯業経験年数 ＊創業者の手持ち資金

3. ロジスティック回帰モデルの構築

3.1 モデル1の構築（創業時の年齢の関数近似）

データベース(DB)によって使用できる変数が異なるため、尾木ら[1]と同様にDBごとにモデルを構築し、それぞれのモデルから算出されるスコアを加重する統合モデルを構築する。まず、DB1のK個の変数からモデル1を構築する。創業企業iが融資後2年以内にデフォルトする確率 $p_{1,i}$ は以下のとおりである。

$$p_{1,i}=\frac{1}{1+e^{z_{1,i}}},\quad z_{1,i}=ln\left(\frac{1-p_{1,i}}{p_{1,i}}\right)=\alpha_1+\sum_{k=1}^{K}\beta_{1,k}f_{1,i,k}$$

尾木ら[1]は人的要因として「創業時の年齢が40歳未満」か否かのダミー変数を作成し、説明変数とした。本研究のDB1で創業時の年齢別のデフォルト率をみると、図1のような曲線となった。鈴木[2]は創業時の年齢と経済的パフォーマンスとの関係について、先行研究などを踏まえると体力と知力がアンバランスな若年層と高齢層のパフォーマンスが低いと予想されると述べている。図1はこの指摘とおおむね合致しており、年齢を変数とするには、ダミー変数ではなく、年齢別デフォルト率を関数近似することが望ましい。

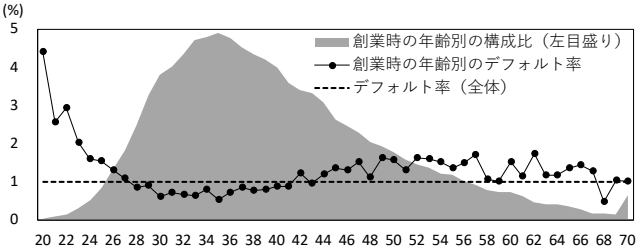

図1　創業時の年齢別の構成比とデフォルト率

そこで、年齢とデフォルト率の関係について、①年齢別デフォルト率の3次関数近似、②4次関数近似、③5次関数近似、の3パターンを作成し、単変数AR値を④40歳未満ダミー(尾木ら[1])と比較した。結果は表2のとおりで、③と同水準で次数の低い②の4次関数近似を採用した。

表2　年齢関数のスコアの単変数AR値

	①	②	③	④
AR値	14.8%	18.6%	18.6%	11.9%

他の変数を含めてステップワイズで変数選択したモデル1の結果を表3に示す。いずれもp値が0.1%未満で有意となった。標準化回帰係数でみると、年齢スコアは3番目に高い結果となった。また、尾木ら[1]のモデルに比べて、負債情報の変数が多く採用された。

表3　モデル1の標準化回帰係数

変数名	推計値	標準化回帰係数	変数名	推計値	標準化回帰係数
定数項	▲ 3.79	－	負債情報4	▲ 0.16	▲ 0.12
年齢スコア	0.63	0.13	負債情報5	▲ 0.10	▲ 0.06
業種スコア	0.92	0.46	負債情報6	▲ 0.03	▲ 0.04
負債情報1	0.78	0.09	負債情報7	▲ 0.09	▲ 0.11
負債情報2	0.06	0.08	負債情報8	▲ 1.67	▲ 0.11
負債情報3	▲ 0.10	▲ 0.04	負債情報9	▲ 3.00	▲ 0.14

3.2 モデル2の構築（斯業経験年数の関数近似）

次に、DB2の「斯業経験年数」と「創業者の手持ち資金」の2変数からモデル2を構築する。3.1節と同様に、モデル2は以下のとおりである。

$$p_{2,i}=\frac{1}{1+e^{z_{2,i}}},\quad z_{2,i}=ln\left(\frac{1-p_{2,i}}{p_{2,i}}\right)=\alpha_2+\beta_{2,1}g_{1,i}+\beta_{2,2}g_{2,i}$$

尾木ら[1]は人的要因として「斯業経験年数5年以内」か否かのダミー変数を作成し、説明変数とした。本研究のDB2で斯業経験年数別のデフォルト率をみると、図2のとおり、斯業経験年数が長くなるにつれてデフォルト率が低下するが、長すぎると上昇に転じるという曲線の関係が明らかとなった。鈴木[2]は創業者の斯業経験年数が長いほどデフォルトしにくいという結果を示している。ただ、斯業経験が長すぎる層は、勤務先の倒産や解雇といった予期せぬ事情でやむをえず創業した結果、デフォルト率が高くなっている可能性がある。

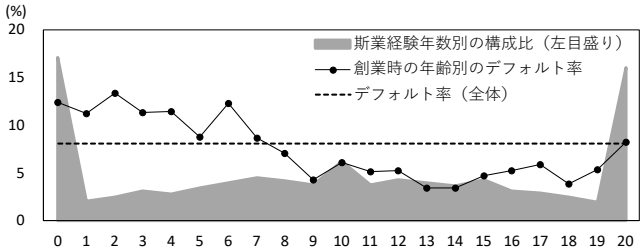

図2　斯業経験年数別の構成比とデフォルト率

この点を反映するため、斯業経験年数とデフォルト率の関係を関数近似した。具体的には①原数値（上限20年）、②2次関数近似、の2パターンを作成し、単変数AR値を③経験5年以内ダミー(尾木ら[1])と比較した。結果は表4のとおりで、AR値が最も高い②の2次関数近似を採用した。

表4　斯業経験年数関数のスコアの単変数AR値

	①	②	③
AR値	12.8%	24.5%	12.2%

モデル2の結果を表5に示す。標準化回帰係数は、斯業経験年数スコアの方が高い結果となった。

表5　モデル2の標準化回帰係数

変数名	推計値	標準化回帰係数	p値
定数項	▲ 1.45	―	―
斯業経験年数スコア	0.92	0.21	＜ 0.1%
創業者の手持ち資金	0.22	0.15	＜ 0.1%

3.3 モデル3の構築（統合モデル）

モデル1とモデル2のスコアを用いて統合モデルのモデル3を構築する。まず、DB1とDB2の両方のデータを保有する18,075社のデータを用いて、創業企業iごとにモデル1のスコア$z^*_{1,i}$とモデル2のスコア$z^*_{2,i}$を算出する。次に、各スコアを説明変数としたロジスティック回帰モデル(モデル3)を構築する。モデル3の式は以下のとおりである。

$$p_{3,i}=\frac{1}{1+e^{z_{3,i}}}\ ,\ z_{3,i}=ln\left(\frac{1-p_{3,i}}{p_{3,i}}\right)=\alpha_3+\omega_1 z^*_{1,i}+\omega_2 z^*_{2,i}$$

モデル3の結果を表6に示す。モデルの精度を表すAR値は50.2%（インサンプル）となった。

表6　モデル3の標準化回帰係数

変数名	推計値	標準化回帰係数	p値
定数項	▲ 0.67	―	―
モデル1のスコア	0.54	0.35	＜ 0.1%
モデル2のスコア	0.66	0.17	＜ 0.1%

4. 機械学習モデルの構築

ロジスティック回帰モデルの説明変数を用いてランダムフォレストおよび勾配ブースティング（XGBoost）の機械学習モデルを構築する。ハイパーパラメータを調整して構築した結果を表7に示す。

表7　機械学習モデルのAR値

モデル	トレーニング(75%)	テスト(25%)
ランダムフォレスト	83.8%	54.2%
勾配ブースティング	78.9%	49.2%

5. アウトオブサンプルによる検証

モデルの頑健性を確認するため、アウトオブサンプルテストを行う。2018年4月～9月の融資データ6,549件を用いて、①尾木ら[1]のモデル、②モデル3の関数近似した変数にダミー変数を用いたモデルの各AR値をモデル3と比較した。また、③ランダムフォレスト、④勾配ブースティングの結果もあわせて、表8に示す。

表8　アウトオブサンプルによるAR値

	①	②	モデル3	③	④
AR値	46.0%	48.9%	50.9%	51.9%	50.6%

この結果から、①から②のAR値の上昇は、説明変数および回帰係数の見直しの効果、②からモデル3のAR値の上昇は、変数の関数近似の効果とわかる。また、モデル3、③、④はほぼ同水準となり、関数近似した変数を用いたモデルの精度は、機械学習モデルと同程度となった。

6. おわりに

本研究では、創業企業の信用リスクモデルにおいて、重要な人的要因とデフォルトとの関係を、理論との整合性を踏まえて見直すことによって、ロジスティック回帰モデルの精度が向上した。また、選択された変数を使って、機械学習モデルを構築した結果、同程度の精度を得られた。

今回の分析結果は、①同じ変数でも関数近似などの加工によってモデルの精度向上の余地があること、②ロジスティック回帰モデルでも変数加工によって機械学習モデルと同程度の精度を確保できる可能性があることを示唆するものである。ただ、実務的には、精度に差がない場合は、モデルの解釈性や理論との整合性の観点からロジスティック回帰モデルの採用が合理的と考えられる。

今後は、人的要因だけでなく、他の要因についても関数近似など、新たな変数を見いだすことで、さらなるモデルの精度の向上を目指したい。

（本稿で示されている内容は、筆者たちに属し、日本政策金融公庫としての見解をいかなる意味でも表さない。）

参考文献

[1] 尾木研三,内海裕一,枇々木規雄(2017)「創業企業の信用リスクモデル」『リスク管理・保険とヘッジ（ジャフィー・ジャーナル：金融工学と市場計量分析）』, 133-156, 朝倉書店.

[2] 鈴木正明(2012)「新規開業企業の軌跡　パネルデータにみる業績,資源,意識の変化」,勁草書房.