

機械学習を用いた教育ローンのクレジットスコアリングモデル

申請中 日本政策金融公庫 国民生活事業本部 *引寺 佑輔 HIKIDERA Yusuke
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

金融機関は融資業務における信用リスク評価に、クレジットスコアリングモデルを活用している。モデルは、融資先の財務や属性情報等から統計手法を用いて信用リスクをスコア化する仕組みで、融資の審査や融資後のモニタリングで用いている。

近年、機械学習を活用し、スコアリングモデルの精度を向上させようとする研究や技術開発が活発になってきている。とりわけ、教育ローンは、融資金額が小口で採算がとりにくい商品であることから、審査コストを低減させるためにクレジットスコアリングモデルの精度向上への期待が高まっている。

枇々木・尾木・戸城 (2011) は、約 35 万件の教育ローンの融資データを用いて、ロジットモデルでスコアリングモデルを構築し、AR 値による評価とその有用性について実証した。

本研究では、日本政策金融公庫国民生活事業本部 (以下、公庫という) が保有する教育ローンの融資データ 674,547 件を使用して、機械学習を用いたクレジットスコアリングモデルを構築し、アルゴリズムの違いによるモデルの精度を比較検証する。検証の結果、ロジットモデルよりも概ね AR 値が向上することを確認した。

2. データの概要

公庫は、年間約 10 万件の教育ローンを融資している。本研究で用いたデータは、2011 年度から 2017 年度までの公庫の教育ローンの融資データ 674,547 件を用いる。分析に用いた変数は、融資先の属性、過去の返済実績等の情報から絞り込んだ 18 変数を用いている。

変数選択のプロセスは以下の手順で行った。

- ①融資審査時等に得られた 233 の候補変数からステップワイズ (有意水準 1%) を用いて 63 変数に絞り込んだ。
- ②ランダムフォレストを使って変数重要度の高い 37 変数に絞り込んだ。
- ③ ②のうち相関の高い変数を削除し 18 変数とした。

説明変数の概要を表 1 に示す。この 18 変数に、性別および地域等の情報は含まれてはいない。

本分析では、融資後の 2 年後の月末までに「3 か月以上延滞」となったものをデフォルトと定義した。判定時点でのデフォルトを 1、非デフォルトを 0 としてデータセットを構築する。

表 1 説明変数の概要

変数	変数の数
融資先の属性情報等	7
過去の返済実績等	11

3. モデルの構築および評価

本研究で用いた分析手法と各モデルのハイパーパラメータを表 2 に示す。勾配ブースティング、ランダムフォレストおよび決定木においては、過学習を防ぐために最大深さを設けている。

表 2 各モデルのハイパーパラメータ

分析手法	主なハイパーパラメータ
ロジットモデル	正則化:L2, C=100
勾配ブースティング (注)	学習率:0.1, 最大深さ:11
ランダムフォレスト	最大深さ:11
決定木	最大深さ:9

(注) 勾配ブースティングは、Python の XGBoost を採用している。

モデルの評価は、AR 値により行う。AR 値は、0 から 100% で表され、モデルの精度が高いほど 100% に近づく評価指標である。

検証は、インサンプルの AR 値と 5 分割交差検証の AR 値を用いて性能評価する。交差検証とは、インサンプルを用いて未知のデータに対する性能を検証する手法である。

本分析では、5 分割交差検証で行う。具体的には、インサンプルのデータセットを 5 個のサブセットに分割し、分割したサブセットの 4 個でトレーニングを行い、残りの 1 個のサブセットで AR 値を測定する。この過程を 5 回繰り返し得られた AR 値の平均値で性能評価する。

4. 分析結果

分析結果を表3に示す。3節で示した各モデルについて、インサンプルと5分割交差検証で得られたAR値を示す。検証の結果、インサンプルと5分割交差検証のいずれの場合にも勾配ブースティングが最も高い精度を示した。図1および図2でも示すとおり、機械学習では未知のデータに対する判別精度が大きく低下した。

ロジットモデルと比較すると、インサンプルにおいては、機械学習がいずれも精度が高くなった。また、5分割交差検証においても、決定木を除いて機械学習のAR値が高かった。

金融機関がスコアリングモデルを構築するとき、その多くにロジットモデルが用いられているが、本研究の分析結果は機械学習を用いることでより精度の高いモデルを構築できる可能性があることを示している。

また、機械学習を用いて変数選択を行った結果、従来行ってきたロジットモデル構築時のステップワイズとモデル構築者の経験則に基づく変数選択においては採用されなかった変数が見つかった。これはロジットモデルのステップワイズと構築者の経験則だけでは捉えることのできないデフォルトと変数との関係性を、機械学習では捉えることができることを示唆するものであり、ロジットモデルの精度向上にも寄与する可能性がある。

表3 分析結果

分析手法	AR値	
	インサンプル	5分割交差検証
ロジットモデル	0.59	0.59
勾配ブースティング	0.91	0.66
ランダムフォレスト	0.69	0.62
決定木	0.60	0.55

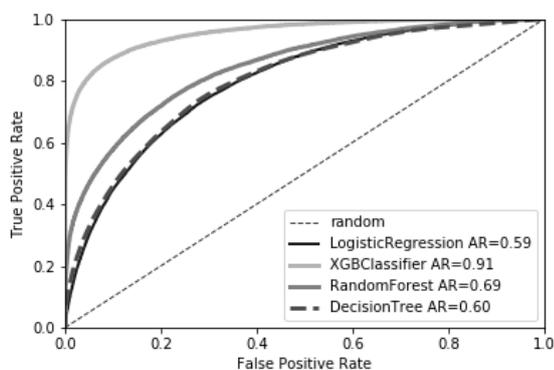


図1 ROC曲線(インサンプル)

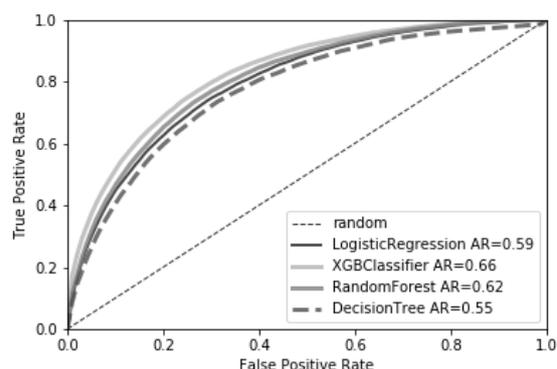


図2 ROC曲線(5分割交差検証)

5. おわりに

本研究では、公庫が保有する教育ローン融資データ 674,547 件を用いて、機械学習によるスコアリングモデルを構築し、アルゴリズムの違いによるモデルの精度を比較検証した。

検証の結果、機械学習は、一部を除いてロジットモデルよりもAR値が高いことを確認した。要因として、ロジットモデルでは捉えることのできない①有効な変数の抽出、②被説明変数と説明変数との非線形関係を機械学習では捉えることができていることが考えられる。この点の検証については今後の課題とし、今回構築したモデルを用いてアウトサンプルによる検証を行うとともに、ロジットモデルに有効な変数の抽出および変数の加工における機械学習の効果や活用方法等について分析を行いたい。

(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] 枇々木規雄, 尾木研三, 戸城正浩(2011), 教育ローンの信用スコアリングモデル, ジャフイー・ジャーナル「金融工学と市場計量分析」, 136-165, 朝倉書店.
- [2] 白田佳子(2019), 企業倒産のメカニズムと格付けとの関係: AI 技術の活用について. Journal of the Institute of Internal Auditors-Japan, 45(6), 1-7.
- [3] 山下智志, 三浦翔(2011), 信用リスクモデルの予測精度 AR 値と評価指標, 朝倉書店
- [4] A. Bandyopadhyay(2016), Studying borrower level risk characteristic of education loan in India, IIMB Management Review, 28, 126-135.