F-044

# 複数顧客に向けた共通の予測モデルの精度向上手法の検討
# A method for performance improvement of common prediction model to multiple customers

濱本敬大† 田中剛†
Keita Hamamoto Tsuyoshi Tanaka

## 1. はじめに

機械学習を用いた予測に基づく意思決定はその客観性や正確性から、融資判断や株価の予測など、金融領域の多くの場面において採用されている。特に、融資業務では、融資対象者が将来返済できなくなる確率を予測する貸し倒れ(デフォルト)予測モデルを用いて融資の審査が行われることが多い。融資業務で収益をあげるため、デフォルト予測モデルの予測精度の向上が求められており、機械学習を用いた予測モデルを導入するケースが増大している。

融資審査用のデフォルト予測モデルを銀行などの金融機関に提供するベンダは、銀行から受け取ったデータをもとに機械学習を用いた予測モデルを学習・チューニングしたものを納品し、そのエラー対処及び性能モニタリングなどの運用管理や定期的な再学習などのサービスを併せて提供することが多い。このときベンダ側では銀行や融資商品ごとに特化した固有の予測モデルを構築し提供するのではなく、複数の銀行やローン商品からデータを集積し共通の予測モデルを提供することで収益性を向上することができると考えられる。これは事業の拡大、すなわち予測モデルを提供する銀行数の増大に比例してモデル構築や運用管理等のベンダ側で負担するコストが増大するのを防ぐことができるからである。

しかし、銀行ごとの客層の違いに起因して、学習するデータセットの統計的な分布に差異が生じる。統計的特性の異なるデータを単純に混ぜたデータセットで学習したモデルでは予測精度が低下してしまう問題が発生する。そのため、複数銀行データから単一予測モデルを構築しても精度が低下しない方法を開発することが課題である。

そこで、本研究では複数銀行から集めたデータを用いて単一の予測モデルを構築するタスクにおいて、生じうる予測精度低下の要因を整理し、精度向上の手法を検討する。

## 2. 共通予測モデルの概要及び設定

通常の機械学習を用いたデフォルト予測では、過去の融資実績データを用いて、年収や借入期間などの説明変数から、デフォルトか否かを表す二値の目的変数(0:ノンデフォルト、1:デフォルト)を予測する分類モデルを構築する。

本研究で着目する複数銀行共通予測モデルの構築にあたっては複数の銀行から過去の融資実績データを集めて単一のモデルを学習し、各銀行に予測モデルを提供する。学習に際して、機械学習モデルは全銀行のデータを混ぜた学習データの統計的性質を認識して学習を進めるため、各銀行の個別の傾向を正しく捉えられない可能性がある。そこで本研究では各説明変数の大小とデフォルト有無の相関関係をその変数のデフォルト傾向と呼び、銀行ごとのデフォルト傾向の違いに着目して精度低下の要因を整理し、精度向上手法の検討を行う。

問題の簡単化のため銀行数は 2 とし、またデフォルト予測においてニューラルネットワークなどのブラックボックスモデルの利用は可監査性を考慮すると困難であること[1]を鑑み、ホワイトボックスモデルの一つであるロジスティック回帰のみを用いることとする。

## 3. 予測精度低下要因の分析と解決策の提案

ある説明変数$x$に対し、$x$と目的変数の間の相関係数を$\mathrm{corr}(x;b)$と表すことにし、単に相関係数というときはこの目的変数との相関係数を意味することとする。二つの銀行を A, B とし、$b \in \{A, B, A \cup B\}$は相関係数を計算するデータセットを表す。$b = A$の時は銀行 A のデータのみを用いて計算し$b = A \cup B$の時は両銀行のデータを用いて計算する。場合に応じて順位相関など他の相関指標を用いてもよい。共通予測モデルは合算したデータのデフォルト傾向である$\mathrm{corr}(x; A \cup B)$をもとに学習を進めてしまうことに注意し、以後$\mathrm{corr}(x; b)$ 間の関係をもとに精度低下要因を整理する。

### 3.1 傾向の埋没による精度低下

銀行間で変数$x$に対するデフォルト傾向が反転している場合、すなわち

$$\mathrm{sign}\big(\mathrm{corr}(x; A)\big) \neq \mathrm{sign}\big(\mathrm{corr}(x; B)\big)$$

の場合、データを混ぜることによってそれぞれの傾向が埋没してしまう。ここで$\mathrm{sign}(z)$は実数$z$の符号を表す。

銀行 A のデータ件数が多いまたはデフォルト傾向が強い場合には$\mathrm{sign}\big(\mathrm{corr}(x; A \cup B)\big) = \mathrm{sign}\big(\mathrm{corr}(x; A)\big)$となり、銀行 A に対しては有利な学習が進行することとなる。例えば$\mathrm{corr}(x; A \cup B) > 0$であれば線形モデルにおいて変数$x$の重みが正の値をとるように学習され、$\mathrm{corr}(x; A) > 0$であるためこの学習結果は銀行 A にとって好ましい。一方、少数派の銀行 B にとっては本来の傾向$\mathrm{corr}(x; B) < 0$とは逆の判別を受けてしまうため、銀行 B に対する予測精度が低下する。

両銀行のデータ件数及びデフォルト傾向の強さが拮抗している場合には$\mathrm{corr}(x; A \cup B) \simeq 0$となり、変数$x$が予測に有用であるとみなされず、学習されない。結果、各銀行に存在した変数$x$のデフォルト傾向が見逃されてしまい、両銀行の精度が低下してしまう。

この現象は例えば長期・短期のローン商品のデータを合算して共通予測モデルを構築する際、長期ローン商品においては高齢ほどデフォルトリスクが高く(定年退職が近いため)、短期ローン商品においては若齢ほどリスクが高い(収入が少ないため)などのケースで発生しうる。

### 3.2 有害な変数による予測精度低下

稀なケースであるが、両銀行で変数$x$に対するデフォルト傾向が一致しているにもかかわらず、合算後にデフォルト傾向が反転してしまう場合、すなわち

$$\mathrm{sign}\big(\mathrm{corr}(x; A \cup B)\big) \neq \mathrm{sign}\big(\mathrm{corr}(x; A)\big) = \mathrm{sign}\big(\mathrm{corr}(x; B)\big)$$

† （株）日立製作所 研究開発グループ
Hitachi Ltd., Research and Development Group

の場合には、両銀行にとって不利益な学習が進行してしまう。こうした変数を有害な変数と呼ぶこととする。

この現象は、両銀行でデフォルト比率が大きく異なる場合に発生する可能性がある。例えば銀行 A では平均年収が低いがデフォルト比率が低く、一方銀行 B では平均年収が高いもののデフォルト比率が高い、といった場合である。また人為的に作成した特徴量が期せずして有害な変数となってしまう可能性もある。

### 3.3 精度低下を抑制する手法の検討

上述の 2 つの現象による精度低下を防ぐため当該変数に対する学習を抑制する必要があり、また予測モデルの透明性を損なってもならない。

そこで本研究では精度向上手法の最も簡便な方法として①当該変数を削除する手法を検討した。加えて、単に削除するだけでは当該変数の情報がすべて失われてしまうため②当該変数を削除し、代わりに当該変数と銀行フラグの交互作用項を追加する手法を検討した。

## 4. 実験とその結果及び考察

3 章で述べた現象によって精度低下が生じていること、及び提案手法による精度向上効果を確かめるため、数値実験を行った。実験には台湾での融資実績データ[2](25200 件、23 変数) を基に、2 銀行からのデータを模して適切に分割して用いた。精度指標には roc-auc スコアを用いた。

### 4.1 実験 a : 傾向の埋没による精度低下

3.1 節に記した内容を再現するため、EDUCATION 変数の値が{0,1,2,3}である多数のデータを銀行 A,それ以外少数のデータを銀行 B として分割した。一か月前の返済額を表す数値変数$x$ ="PAY_AMT1"に着目し、図 1(a)にその分布と相関係数を示す。銀行 A の相関係数$\mathrm{corr}(x;A)$は両銀行データを混ぜて計算した相関係数$\mathrm{corr}(x;A\cup B)$とほぼ等しく、少数派である銀行 B の相関係数$\mathrm{corr}(x;B)$はこれらと符号が異なる。3.1 節で説明した通りこの変数は銀行 B の精度低下を引き起こしていることが示唆される。

変数$x$による精度低下及び提案手法①による精度向上効果を調べるため変数$x$を削除する前後の精度を表 1(a)にまとめた。削除によって銀行 B の精度が 0.035 ポイント上昇し、銀行間の精度バランスも改善していることがわかる。

提案手法②の効果を調べるため変数$x$と銀行フラグの交互作用項を追加した場合の精度を表 1(a)の右端の列に記載した。変数削除によって低下していた銀行 A の精度が向上していることが確かめられる。

### 4.2 実験 b : 有害な変数による予測精度低下

次に 3.2 節に記した内容を再現するためデータの加工を実施した。まず、全データを等しい件数でランダムに銀行 A と B の二つに分割する。この時銀行 B のデフォルト件数が A の 3 倍になるような層化サンプリングを行う。次にデフォルト比率の高い銀行 B の変数$x$ ="LIMIT_BAL"の値に一律 20000 を加算する。この変数はクレジットの限度額を表しており、一般に値が小さいとデフォルトリスクが高い。図 1(b)にその分布と相関係数を示す。銀行 A、銀行 B ともに相関係数は負であるが、デフォルト比率の高い銀行 B データに大きな値を付与したため、両銀行データを混ぜて計

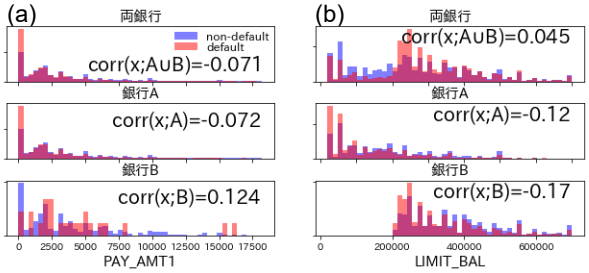


図 1　実験 a,b で注目した変数の分布

(a)

| 銀行 | 適用前 | ①削除後 | ②追加後 |
|---|---|---|---|
| 両銀行 | 0.719 | 0.718 | 0.719 |
| 銀行 A | 0.718 | 0.717 | 0.720 |
| 銀行 B | 0.577 | 0.612 | 0.600 |

(b)

| 銀行 | 適用前 | ①削除後 | ②追加後 |
|---|---|---|---|
| 両銀行 | 0.735 | 0.717 | 0.759 |
| 銀行 A | 0.681 | 0.713 | 0.708 |
| 銀行 B | 0.690 | 0.719 | 0.711 |

表 1　実験 a,b における提案手法適用前後の予測精度

算した相関係数は正の値を持つ。このことから 3.2 節で説明した通り$x$は有害な変数であり、両銀行に精度低下をもたらしていることが示唆される。

変数$x$に提案手法を適用する前後の精度を表 1(b)にまとめた。提案手法①の変数削除によって両銀行の精度がともに 0.03 ポイント程度上昇しており、変数$x$が両銀行の精度低下を引き起こしていること及び提案手法①による精度向上効果が確かめられた。2 銀行合算後の精度は 0.018 ポイント低下しているが、これは削除前のモデルが「(変数$x$の情報をもとに)銀行 B の人はリスクが高い」と判断していたことを示唆しており、銀行ごとの精度のみが関心である今回のタスクでは問題にならない。また提案手法②によって全銀行の精度は大きく上昇しているが、銀行別精度に関しては変数削除のみによる精度を上回ることはなかった。

## 5. まとめと今後の課題

本研究では、複数顧客に向けた共通予測モデルの精度向上手法を検討した。銀行間のデータ傾向の差異に起因した共通モデルの予測精度低下要因を整理し、変数削除や交互作用項追加という、簡便かつモデルの透明性を損なわない性能向上手法を検討した。融資実績データを用いた数値実験によって提案手法により共通モデルの予測精度が向上することを確認した。特に、少数派銀行の傾向が埋没してしまうケースでは精度バランスの向上が期待でき、有害な変数が存在するケースでは変数削除がとくに大きな精度向上効果を持つことが明らかとなった。

一方で強い多重共線性が存在する場合の検討や線形モデル以外への適用、交互作用項に代わる有用な特徴用の作成手法などに関しては、今後より詳細な検討が必要である。

### 参考文献

[1] 大久保豊, 尾藤剛, "【究解】信用リスク管理," 金融財政事情研究, pp.168-170, 2018

[2] Yeh, I. C., & Lien, C. H. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", Expert Systems with Applications, Vol.36, No.2 (2009)