

令和 5 年 6 月 9 日現在

機関番号：62603
研究種目：若手研究
研究期間：2020～2022
課題番号：20K13581
研究課題名（和文）実務利用可能な中小企業の期待損失率推計手法の開発：地銀統合ビッグデータを用いて

研究課題名（英文）Development of a Practically Applicable Estimation Method for Expected Loss Ratio of SMEs: Using Integrated Big Data of Regional Banks

研究代表者
長幡 英明（Nagahata, Hideaki）
統計数理研究所・リスク解析戦略研究センター・外来研究員

研究者番号：00815128
交付決定額（研究期間全体）：（直接経費） 3,200,000円

研究成果の概要（和文）：デフォルト時貸出残高損失率(LGD)推計用のデータベースを用いて、LGD推計精度向上に寄与するようなデータ構造化の実現を目指した。データ構造化について、具体的にはデフォルト時点からt時点前・デフォルト時点からt時点後も含めたデータ構造化を行った。推定精度の向上のため、デフォルト関連情報をより深く・広範囲に取り込めることから、統計モデルや機械学習の学習による推計に有効であることが確認された。統計モデルや機械学習の高度化によるLGD推計精度の向上だけでなく、推計に用いるデータベースから見直した点が推計精度向上や新たな知見の獲得を重視するようにピボットした結果、明確な進捗を得られたと考えている。

研究成果の学術的意義や社会的意義

本研究は一般では公開されていない担保・保証・債務回収の情報を含む、地銀5行統合データベースを用いた期待損失率推計を扱い、汎用的で高精度な推計手法を開発・提案します。これによって、期待損失率推計に対し統計的・機械学習的接近法を用いた推計手法が確立・実務利用され、その成果は信用リスク研究の発展、銀行の融資審査の高度化、金融行政の合理化、中小企業金融の円滑化に貢献することを旨とします。

研究成果の概要（英文）：Using the database for estimating loan loss ratio (LGD) at default, we aimed to realize data structuring that would contribute to improving the accuracy of LGD estimation. Specifically, we structured the data to include the time point t before and t after the default point. The data structure was confirmed to be effective for estimation by learning statistical models and machine learning, since it can incorporate default-related information more deeply and extensively in order to improve estimation accuracy. We believe that we have made clear progress not only in improving the accuracy of LGD estimation through the sophistication of statistical models and machine learning, but also as a result of our review from the database used for estimation, which pivoted to focus on improving estimation accuracy and gaining new knowledge.

研究分野：信用リスク

キーワード：信用リスク 機械学習 データ構造化 データ結合 デフォルト時損失率 ビッグデータ

様式 C - 19、F - 19 - 1、Z - 19 (共通)

1. 研究開始当初の背景

2008年の金融危機から分かるように、社会における金融・経済の安定化のためには、銀行のリスク管理の精緻化が不可欠となっています。銀行の抱えるリスクには、主に信用リスク・市場リスク・オペレーショナルリスクの3つがあります。この3つのリスクの中で最も大きいリスクが信用リスクで、さらに銀行が抱える信用リスクでは、貸出リスクが一番大きいことが知られています。

そして、貸出リスクは主に貸出先企業の期待損失率によって補足され、それは(1) デフォルト確率(PD)と(2) デフォルト時の貸出残高損失率(LGD)に分解することができます。PDとは、一定の期間内(通常は1年間)に当該貸出先からの返済が滞る、つまり債務不履行状態に陥ってしまう確率のことです。

LGDとは債務不履行となった貸出先について、債務不履行発生時点の貸出残高のうち実際の損失に至る金額の割合のことを指しています。銀行はこれらの要素を精緻に推計することで、貸出債権のリスクを正確に把握する必要があります。

国際的に業務を展開している銀行の健全性を維持するための新たな自己資本規制であるバーゼルでは、(A) 標準的手法(外部格付を利用)と(B) 内部格付手法で信用リスクを推定するよう記載されており、(B)の内部格付手法は(B-1) PDのみ推計する「基礎的内部格付手法」と(B-2) PDとLGDの両方を推計する「先進的内部格付手法」に分類することができます。

しかし、このように銀行債権の信用リスクを推計しようとしたとき、銀行データの高い秘匿性により統合データベースの欠如が課題となります。特に、LGDの分析に不可欠な「貸出の回収実績(銀行の貸出によってどれだけ回収できたか(=損失が発生したか)の実績値)を含むLGD推計用のデータベースの不足が顕著になっています。

さらに、高精度なLGD推計手法を実務利用するためには「担保」や「保証」などの有効だと思われる変数も含み、結果が特定の銀行に依存しないよう複数の銀行のデータが統合されたデータベースが必要となります。この～をすべて満たすような日本唯一の信用リスク計量化用データベースのために、高度信用リスク統合データベースコンソーシアムが組成され、本研究で利用する期待損失率推計のための地銀5行統合データベースが構築されました。

2. 研究の目的

- ・上記～をすべて満たすような日本唯一の信用リスク計量化用データベース構築
- ・構築されたデータベースを用いて期待損失率推計モデル開発

3. 研究の方法

統合データベースに対し、統計モデル・機械学習による最適な推計モデルを構築するため、下記サブテーマ(A)～(ウ)の解明に取り組みました。

(ア) 最高の推計精度をもつモデル

推計精度を基準として、統計モデル・機械学習モデルの最先端モデルを網羅的に検討する。また、(イ)の土台となる調査として、モデルの背後にある数理統計学・統計的学習理論についても研究する。

(イ) そのモデルの最適なパラメータ

最小二乗推定・最尤推定・ベイズ推定などを用いて、最適推測論やときには網羅的・全探索的な方法を用い、最適なパラメータを探索する。

(ウ) モデル評価方法の検討

モデル選択やパラメータ推計だけでなく、それらの良し悪しを決定づけるモデル評価方法についても検討する。

4. 研究成果

デフォルト時貸出残高損失率(LGD)推計用のデータベースを用いて、LGD推計精度向上に寄与するようなデータ構造化の実現を目指しました。データ構造化について、具体的にはデフォルト時点からt時点前・デフォルト時点からt時点後も含めたデータ構造化を行いました。推定精度の向上のため、デフォルト関連情報をより深く・広範囲に取り込めることから、統計モデルや機械

学習の学習による推計に有効であることが確認されました。統計モデルや機械学習の高度化による LGD 推計精度の向上だけでなく、推計に用いるデータベースから見直した点が推計精度向上や新たな知見の獲得を重視するようにピボットした結果、明確な進捗を得られたと考えています。なお、時間情報を取り入れる発想の原点として、従属誤差に対する ANOVA の理論構築に触れていました。このような、より自然な仮定を課したモデルに対して理論構築した経験が、実データに対して自然な特徴量生成とデータ構造化の成功を促した点も添えさせていただきます。

また、各サブテーマについて、研究結果を記載いたします。

(ア) 最高の推計精度をもつモデル

推計モデルとしては、ゼロ過剰な連続データのための 2 段階モデル・ニューラルネット・ニューラルネットを用いた転移学習・ガウス過程・ランダムフォレスト・ガウス過程による特徴量設計を経由したランダムフォレストなどの独自の機械学習を用いた実験を実施し、有効なモデルを発見した。

(イ) そのモデルの最適なパラメータ

最小二乗推定・最尤推定・ベイズ推定など多様な推計手法を用いて、(ア) のモデル群のパラメータ推計を実施した。しかし、(ア) にあるような特徴量設計や、さらに上流に位置するデータ構造化を吟味した上での特徴量設計と比較して、際立った研究結果は得られなかった。

(ウ) モデル評価方法の検討

アドホックなモデル評価だが、「より頑健なモデル」を得られるよう以下のように工夫し、良好な結果を得た。具体的には、機械学習手法を用いる際のハイパーパラメータが安定的に得られるようになった。

* (1) Cross Validation と (2)時間方向へのバックテスト を組み合わせる

* (1) Cross Validation

* ハイパラ空間はアドホックに決定

* 設定された criteria に最適化するよう grid search

* しかし、時間方向で言えば、「未来データを学習し、過去データでの検証」も含まれているため、「時間方向に対しても頑健なモデル評価」を別途実施したい

* (2) 時間方向への バックテスト

* (1)と同時に、時間方向への n 回バックテストを実施

* Step 1 :

* t_1 年データを対象に (1) を実施し、 t_1 年データに(1)の意味で最適化されたモデル₁ を構築

* モデル₁ を用いて、 t_2 年データに対してバックテスト

* Step 2 :

* t_1, t_2 年データを対象に (1) を実施し、 t_1, t_2 年データに(1)の意味で最適化されたモデル₂ を構築

* モデル₂ を用いて、 t_3 年データに対してバックテスト

* Step n :

* t_1, \dots, t_n 年データを対象に (1) を実施し、 t_1, \dots, t_n 年データに(1)の意味で最適化されたモデル_n を構築

* モデル_n を用いて、 $t_{(n+1)}$ 年データに対してバックテスト

なお、今後の展望としては、本研究で構築されたモデルを実務利用していただくことを目指し、本研究結果の具体的な実装先となる As is 業務フロー と To Be 業務フローの整理、そのリターンとリスクの整理、実装する際の要件定義について、ご協力くださった地銀様との合意形成を目指し、内容を詰めていきたいと考えています。

5. 主な発表論文等

〔雑誌論文〕 計1件（うち査読付論文 1件 / うち国際共著 0件 / うちオープンアクセス 0件）

1. 著者名 Tanoue Yuta, Yamashita Satoshi, Nagahata Hideaki	4. 巻 22
2. 論文標題 Comparison study of two-step LGD estimation model with probability machines	5. 発行年 2020年
3. 雑誌名 Risk Management	6. 最初と最後の頁 155 ~ 177
掲載論文のDOI（デジタルオブジェクト識別子） 10.1057/s41283-020-00059-y	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

〔学会発表〕 計0件

〔図書〕 計1件

1. 著者名 Yuichi Goto, Hideaki Nagahata, Masanobu Taniguchi, Anna Clara Monti, Xiaofei Xu	4. 発行年 2023年
2. 出版社 Springer Nature	5. 総ページ数 89
3. 書名 ANOVA with Dependent Errors	

〔産業財産権〕

〔その他〕

-

6. 研究組織

氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
---------------------------	-----------------------	----

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------