日本オペレーションズ・リサーチ学会
2022年 秋季研究発表会

Integrated Gradients による格付 AR 値変化の寄与度分解

賛助会員　（株）日経金融工学研究所　*木村和央　KIMURA Kazuo

1. はじめに

ロジットモデルに代表されるクラス分類を目的としたモデルの精度評価指標として，AUC，AR 値といった順序統計量（序列精度）が重要視されている (山下・三浦 [1])．このうち，金融機関においては伝統的に AR 値の利用が多く，その時系列変化に大きな関心が寄せられている．木村ほか [2] では，格付 AR 値（以下 AR 値）を対象に，格付別の構成比と相対デフォルト率の変化差分を用いた寄与度分解を提案したが，相対デフォルト率には全体のデフォルト率の影響が入り込むため結果の解釈に困難な部分が伴った．この困難をクリアするべく，本稿では説明可能な AI(XAI) 技術の 1 つである Integrated Gradients(IG) を利用して，AR 値の変化要因をセグメント・格付別の構成比変化とデフォルト率変化に寄与度分解する手法を提案し，実際の債務者データに対して適用した．

2. Integrated Gradients(IG) の概略

詳細は Sundararajan et al.[3] を参照．基準となるサンプルの入力値を $\boldsymbol{x'}$，評価したいサンプルの入力値を $\boldsymbol{x}$ とする．入力値が $\boldsymbol{x'}$ から $\boldsymbol{x}$ まで変化したとき，出力値は $F(\boldsymbol{x'})$ から $F(\boldsymbol{x})$ まで変化するが，その差である $F(\boldsymbol{x})-F(\boldsymbol{x'})$ を次式により入力値の各要素 x_i の変化からの寄与度に分解する．

$$IG_i(\boldsymbol{x}) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(\boldsymbol{x'} + \alpha \times (\boldsymbol{x} - \boldsymbol{x'}))}{\partial x_i} d\alpha.$$

通常，解析的に計算するのは困難であるため，m 個の区間に分割して近似した区分求積法にて行う．

$$IG_i^{\text{approx}}(\boldsymbol{x}) ::= (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial F(\boldsymbol{x'} + (k/m) \times (\boldsymbol{x} - \boldsymbol{x'}))}{\partial x_i} \cdot \frac{1}{m}.$$

なお，F がほとんど至るところ微分可能であれば，寄与度の合計が出力値の差分に一致する．

$$\sum_i IG_i(\boldsymbol{x}) = F(\boldsymbol{x}) - F(\boldsymbol{x'}).$$

*本稿の内容は筆者に属し，所属組織の見解ではない．

3. 格付 AR 値の定義

業種等のセグメント $s(=1,\ldots,S)$，格付 $k(=1,\ldots,K)$ の債務者数の全債務者数に対する構成比を a_{sk}，デフォルト率を p_{sk} とすると，全体のデフォルト率は $P = \sum a_{sk}p_{sk}$．このとき，AR 値 (AR) は以下のとおり．

$$AR = \frac{1/2}{(1-P)P} \sum_{u,v}^{S} \sum_{i,j}^{K} a_{ui}a_{vj}(p_{vj} - p_{ui}) \operatorname{sgn}(j-i).$$

4. IG に必要な微分計算とその解釈

構成比 a_{sk} とデフォルト率 p_{sk} が入力値 $\boldsymbol{x}$，AR 値 (AR) が出力値 $F(\boldsymbol{x})$ と考えて，IG に必要な微分計算を行う．

$$\frac{\partial AR}{\partial a_{sk}} = AR \cdot \frac{2P-1}{(1-P)P} \cdot p_{sk} + \frac{1}{(1-P)P} \sum_{v}^{S} \sum_{j}^{K} a_{vj}(p_{vj} - p_{sk}) \operatorname{sgn}(j-k).$$

$$\frac{\partial AR}{\partial p_{sk}} = AR \cdot \frac{2P-1}{(1-P)P} \cdot a_{sk} - \frac{1}{(1-P)P} \sum_{v}^{S} \sum_{j}^{K} a_{sk}a_{vj} \operatorname{sgn}(j-k).$$

右辺第 1 項は，全体デフォルト率 P の変化を通じて AR 値に与える影響と解釈できる．通常、$P < 0.5$ より，$2P-1 < 0$ であるから，a_{sk} と p_{sk} の増加/減少は AR の減少/増加に寄与する．

右辺第 2 項は，格付間の整序性の変化を通じて AR 値に与える影響と解釈できる．格付別デフォルト率が $j \lessgtr k$ のとき $p_{vj} \lessgtr p_{sk}$ であれば，a_{sk} 微分の式の右辺第 2 項は正となり，a_{sk} の増加/減少は AR の増加/減少に寄与する．他方，p_{sk} 微分の式の右辺第 2 項は，$\sum$ 記号内の項別に，$k > j$ ならば正，$k < j$ ならば負となる．つまり，格付 k のデフォルト率上昇/下落は，上位格付とのデフォルト率差を拡大/縮小させるため AR の増加/減少となり，下位格付とのデフォルト率差を縮小/拡大させるため AR の減少/増加となる．このバランスにより右辺第 2 項の正負が決まる．

5. 実際の債務者統合データへの適用

債務者統合データ[1]へ本手法を適用し，2004〜17年度の AR 値の時系列変化につき寄与度分解を実施した．セグメント（業種）は，1. 製造，2. 建設，3. 卸売，4. 小売，5. 不動産，6. サービス，7. その他である．格付は，決算データのスコアリング結果から付与したもので，最上位格が 1，最下位格が 7 である．なお，区間分割数を 1,000 とした台形公式にて区分求積計算を実施した．以下，構成比変化とデフォルト率変化の寄与度は合算して扱った．

図 1 は業種別，図 2 は格付別の寄与度分解である（以下，数値を 1,000 倍して表示していることに留意）．全体の AR 値は，2006 年度とリーマンショックの 08 年度に大幅低下したが，その後は回復基調であった．データ秘匿のため図は省略したが，2008 年度の業種別に計算された AR 値は，不動産業が他の業種の 2 倍程度低下しており，一般に，全体の AR 値低下は不動産業の要因であると解釈されているが，図 1 によれば製造業の寄与度が上回った．これは債務者数に占める製造業の割合が不動産業よりも多いことに起因するものと考えられるが，本手法により定量的な比較･解釈が可能となった．また，前回の相対デフォルト率を用いた寄与度分解では，最下位格付である Rank7 の寄与が重要と報告していたが，図 2 によれば中位格付である Rank3,4 からの寄与度が顕著であり，これも債務者数割合で説明が可能と考えられる．表 1 は 2008 年度における業種×格付別の寄与度であるが，製造業は Rank1 の寄与度が大きく，Rank2〜4 も同様の傾向であるのに対し，不動産業は Rank3 と 4 の寄与度が大きいことがわかった．図 3 は時系列分散の寄与度分解であり，業種ではその他，格付では低格付である Rank6,7 が全体の動きとは逆相関となった．

6. おわりに

本稿では，Integrated Gradients を用いて格付 AR 値の変化につき寄与度分解する手法を提案し，実際の債務者統合データに適用した．今回採用した手法により，相対デフォルト率から通常のデフォルト率に変更したことで結果解釈がわかりやすくなり，また，業種セグメントに関する寄与度分解も合わせて可能となった．

[1]データは所属機関が独自に金融機関から許可を得て収集.

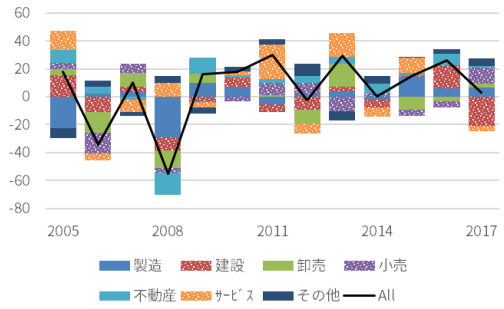

図 1: 格付 AR 値変化の業種別寄与度分解

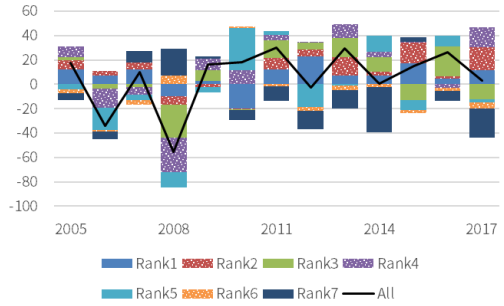

図 2: 格付 AR 値変化の格付別寄与度分解

表 1: 2008 年度の業種・格付別寄与度分解

	製造	建設	卸売	小売	不動	サー	その他
Rank1	**-8.5**	1.5	-2.4	-0.3	-0.2	0.1	0.0
Rank2	**-6.7**	-0.2	-2.0	-1.0	-1.0	1.6	2.6
Rank3	**-7.5**	-2.3	-3.1	**-4.4**	**-7.8**	-3.2	1.2
Rank4	**-6.6**	**-9.1**	**-4.0**	0.1	**-9.0**	2.0	-1.6
Rank5	-3.1	-3.6	**-4.3**	1.0	1.0	**-4.7**	1.0
Rank6	0.3	1.7	1.2	1.5	0.5	1.8	0.3
Rank7	2.8	2.9	1.9	0.0	0.3	**12.5**	1.3

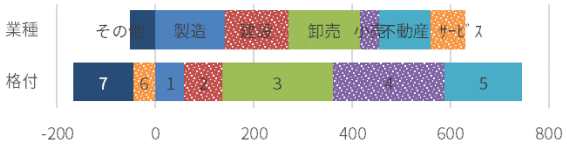

図 3: 時系列分散の業種/格付別寄与度分解

参考文献

[1] 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.

[2] 木村和央, 宋明子, 友添峻希 (2019). 格付 AR 値変化の寄与度分解. 日本オペレーションズ・リサーチ学会 2019 年秋季研究発表会予稿集.

[3] M Sundararajan, A Taly, Q Yan(2017). Axiomatic attribution for deep networks. International conference on machine learning, 3319-3328.