書　評

入門 R による予測モデリング
―機械学習を用いたリスク管理のために―

岩沢宏和，平松雄司 著，東京図書，2019 年

統計数理研究所
川崎　能典

本書を手に取ろうという読者は, 表題のどのキーワードに反応するであろうか. R, 予測モデリング，機械学習，リスク管理のいずれかないし複数であろうから，これらを切り口に本書の特色を紹介しよう.

まずは「予測モデリング」である．本書はデータ解析において機械学習アプローチを取るという姿勢を冒頭第 1 章で明確に打ち出す．その際引用されるのが，レオ・ブライマンによる所謂“Two Cultures 論文”である. ここでいう 2 つの文化とは，「予測モデリング文化」と「生成モデリング文化」である．ここでは評者なりにこの 2 つの文化について改めて整理しておきたい.

統計的モデリングとは，現象の不確実性を確率分布によって表現し，その平均構造や分散構造等に通例複数の定式化を試みながら，データを生成する一番尤もらしい確率構造（Data Generating Process, DGP）を探索することである．ポイントはこの「生成」をどう捉えるか，である.

モデリングの対象となるデータが出てきているところの個別科学領域の知見を反映しながら，「これらの変数群の影響で応答変数が出てきている」という形で，ある種因果的あるいは説明的に現象を記述する立場があり，これが「生成モデリング文化」である．説明変数の変化に応答変数がどの程度影響するかは，例えば回帰モデルにおける係数に示され，何が現象の決定要因なのかについて一定の解釈を許すものとなる.

これに対し，応答変数の挙動を，例えば一期先予測の観点からもっとも良く記述する「入出力関係」を得ることを最優先し，場合によっては（そして実際多くの場合）モデルの解釈性・可読性を犠牲にすることもいとわない立場がある．平たく言えば，予測が当たるなら推論装置がブラックボックスでも受け入れる，という立場であり，これが「予測モデリング文化」である.

では，本書が例えば現代におけるブラックボックスの最先端である深層学習にま

っしぐらに進んでいるのかというと，決してそういうわけではない．むしろ本書には，2 つの文化双方を尊重する分別ある姿勢が垣間見えるのだが，それは著者らなりの実務経験を反映したものと推察する．

そこで本書では「機械学習」として何を取り上げているか，に議論を移そう．評者の見るところ，第 7 章，特に 7.4 節以降に分析例を伴って登場する統計モデル・機械学習手法が本書のスコープを端的に示している．それは，一般化線形モデル（GLM），正則化 GLM（ただし実例は正則化線形回帰），一般化加法モデル（GAM）である．実例においてはランダムフォレスト，勾配ブースティング法の予測結果も比較対象として示されるが，それはあくまで解釈性を手放せば達成できる予測精度の限界，というような意味合いで紹介される．

GLM や GAM は，見方によっては「生成モデリング文化」の範疇にあるとも言える．多数の説明変数候補があるとき，探索的に変数選択を行うということであれば，かなり「予測モデリング文化」に近づいてくるが，評者などはそれも含めて通常の統計的モデリングではないかと感じる．

では，予測精度をひたすら追求してブラックボックスに突き進まない本書のスタンスをどう解釈すべきだろうか．恐らくそれは，アクチュアリーないしクォンツアナリストとして，モデルの解釈性・可読性は実務上不可欠であるという経験から来るものだろう．モデルの予測性能が劣化したときに，「何が効かなくなって当たらなくなったか」を（上司に）説明できないことは実務上致命的であり，それゆえ生成モデリング的な「変数の顔が見えている」モデルもキープしておかなければならないのである．

さて，本書における「リスク管理」に関する内容は，個人向け自動車保険での保険金請求に関するデータを扱った第 9 章にほぼ集中している．このデータセットは，保険金請求の有無という切り口で考えれば分類問題であり，事故件数に注目すれば回帰問題（ポアソン回帰）として取り扱える，非常に興味深いデータである．

評者には，分類ないし判別問題こそリスク解析の第一歩ではないかと思われるのだが，判別・分類のための Hello World とも言うべきロジスティック回帰（これも GLM の範疇）は，9.3 節に至ってようやく登場し，併せて ROC，AUC といった基本的な指標も紹介される．9.5 節のポアソン回帰における，観察期間を表す変数の取り込み方についての注釈は，リスク管理のためのモデリングにおける著者らの経験を感じる部分である．

キーワードの最後は「R」であるが，R によるサンプルビニェットは随所に織り込まれており，R に対する習熟度の高低に拘わらず円滑に読み進めることができるだ

ろう．R をインストールした経験もない読者は付録 A から，多少はさわったことがあるが基本操作に関して自信がないという読者は付録 B から先に目を通しておくとよいだろう．本編では，第 4 章，第 5 章が R に関して記述のまとまっている部分である．以後は，R の関数として基本的なものでも，必要が生じた段階で紹介されている印象である．

ここまでは表題から拾ったキーワードを軸に内容を紹介してきたが，以下にもう少し本書の特色を拾い上げておきたい．

第 2 章において予測モデリングの基本用語や基本概念を紹介した上で，第 3 章では予測モデリングの手順を紹介している．その際，モデリングの手前で行う探索的データ解析が 3.6 節で，モデルの選択・評価に関する手法が 3.8 節で丁寧に述べられている．

データの前処理から探索的データ解析を行う部分は，第 6 章でボストン住宅価格データを例に，コードとともに具体的な分析が例示されている．特に 6.7 節において，ランダムフォレストのようなブラックボックス型予測モデルで特徴量と予測値との関係を可視化する有効なツールである，Partial Dependence Plot（PDP）と Individual Conditional Expectation（ICE）に紙幅を割いて丁寧に説明しているところは，他の日本語テキストにはない特長と思われる．

モデルの選択・評価に関しては，3.8 節で概念的な整理を行い，実例は第 8 章で展開されている．扱うデータはボストン住宅価格データで，分析結果はさほど印象を強く残すものではないが，モデルの予測能力の比較検証手順が丁寧に示されていて，すぐれたチュートリアルである．

評者の見るところ，著者らが一番に推す手法は正則化 GLM なのだろうという印象を受けたが，その一方で，恐らく他の手法との記述バランスを考えた上で，本編では敢えて突出しないように記述を収めた部分もあったかもしれない．それを補う形で，補論にパッケージ aglm が紹介されている．

まとめ：本書は予測モデリングと生成モデリングの両方を尊重しながら，分析手続きとしては予測モデリングの手順に沿って，現代的な統計モデリング，機械学習の手法を実例とコード付きで示した良書である．必ずしも self-contained なテキストではないが，読者が自身の関心に従って次の文献・書物を探すには，良い出発点を与えると思われる．特に，実務上リスク管理のためのデータ分析に携わる人々には，本書のスタンスは共感を持って迎えられるであろう．