〔論　　文〕

樹木構造接近法による与信リスク因子の探索*

衛　藤　俊　寿**

Searching for Credit Risk Factors by the Tree-Structured Approaches*

Toshihisa ETO**

**School of Engineering, Nippon Bunri University

Abstract

In this paper, the data mining methods are applied to analyze the credit risk management of a financial business, especially for consumers. To evaluate the present credit scoring model, we search for the credit risk factors affecting credit contract defaults. In this case, two tree-structured approaches, namely, the CART and CHAID methods are applied. Through the process of data analysis, the risk factors related to the credit scoring model and their specific profiles are discussed.

キーワード：データマイニング，樹木構造接近法，与信リスク因子，消費者金融
Keywords：data mining, tree-structured approaches, credit risk factors, consumer finance

１．はじめに

近年，企業ではデータウェアハウスの構築が完了し，蓄積されたデータを活用するための新しいデータ解析の概念が利用されている。その考え方や方法は，社内の各部門で断片的に蓄積・廃棄されていたデータを集め，顧客データ解析を行ったうえで，顧客重視の事業戦略を立て事業構造転換を図るCRM（Customer Relationship Management）の分野でとくに注目されている。

データマイニング（Data Mining）と呼ばれるこの概念は，近年に注目されているデータサイエンスの概念に包含され，収集・蓄積された膨大な情報（データ）に対して，データ解析の目的を設定し，必要なデータを抽出・加工し，マイニング手法を駆使することによって意味のある結果（知見）を導き出す一連の過程（プロセス）のことである。その適切な使い方は，データマイニングの考え方や方法を業務プロセスに結びつけ，有益で実行可能な情報を作り出すことである[1]。

本稿では，金融業，とくに消費者金融における与信リスク管理へのデータマイニングの適用を試みた。消費者に無担保で融資を行う消費者金融では，入会時の限られ

*2022年６月13日受理
**日本文理大学工学部 教授

た情報（申込用紙に記述された性別，年齢，月収といった情報）によって融資の可否と融資限度額を決定しなければならない。このときの融資には貸倒れ，すなわち契約の償却というリスクを伴い，金融業者にとってこのリスクをうまく管理し制御することが重要な関心事である。これを「初期与信リスク管理」と呼ぶ（因みに，融資途上において融資限度額を変更し危険顧客を抽出することを「途上与信リスク管理」という）。ここでは，初期与信リスク管理に焦点をあて，初期与信時の融資契約が償却するか否かに影響を及ぼす因子をデータマイニングによって抽出した。

本稿では，まず，消費者金融業における与信リスク管理の現状での問題点と解決へのアプローチについて述べる。次に，データマイニングの概要について触れ，その手順とここで使用した方法を紹介する。さらに，実際のデータセットへの適用事例について述べ，解析結果を考察する。

2．与信リスク管理の問題点と解決へのアプローチ

従来，消費者金融業では，初期与信時の融資の可否と融資限度額の決定に「伝統的・経験的」に作成された貸付採点表が利用されてきた。貸付採点表とは，初期与信時に顧客への点数づけ（スコアリング）を支援するための点数表のことである。金融業者は貸付採点表から顧客情報に対応した点数を導きだし，その合計点数に基づいて融資の可否と融資限度額を決定する（例えば，L件数（他の金融業者から融資を受けている多重債務件数）が０件の顧客には５点，１件には４点，２件には３点…，年収が100万円未満の顧客は１点，100万円以上300万円未満には２点…など）。しかし，効率的で比較的安全な融資を実施するという観点から，この伝統的・経験的な貸付採点表の信頼性がどの程度のものであるかを科学的な見地から評価することが必要となっている。

そこで，償却の有無（すなわち償却する契約と償却しない契約）に真に影響を及ぼしている要因を特定することによって，貸付採点表の項目が妥当であるか否かを判断することとした。さらに，要因探索の過程でその影響の度合を考察することができれば，顧客への点数づけ（スコアリング）の妥当性を検討できる。

実際のデータセットでは，要因の候補として，契約情報データ上の顧客背景因子（性別，年齢，住所など），顧客資金因子（住居，会社情報，収入など），顧客取引因子（カード発行回数，利用目的，CL 変更記録など）が観測されていた。ここで，償却契約とは，①データ上に事故コードが入っている契約または31日以上延滞中の契約，または②事故を起こした契約であり，それ以外の契約のことを正常契約とした。また，事故とは，①償却（債権放棄）したもの，または②例外処置（利息免減などによる債権減少）をした契約のことである。

3．データマイニングの手順と方法

品質改善という目的に対して，実験を実施する際の一連のサイクルは「計画（plan）」「実行（do）」「検討（study）」「行動（act）」というプロセスの繰り返しである[2]。このサイクルと科学的に同等と考えられるプロセスは「デザイン（design）」「実施（execute）」「解析（analysis）」「予測（predict）」となるであろう[3]。この科学的サイクルを本マイニングにあてはめると図１のような一連のプロセスとなる。なお，ここで実施したデータマイニングでは，図１に示すマイニングプロセスのうちデータ解析過程のすべてとフィードバック過程の一部を実施するにとどまったことに留意されたい。

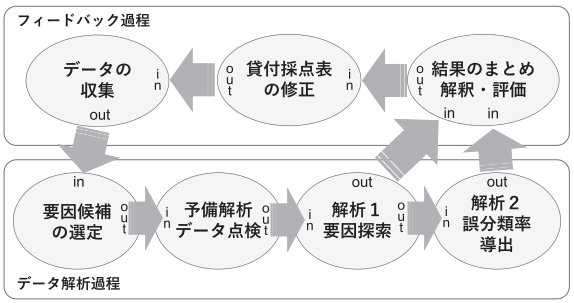

図１　本マイニングのプロセス

データ解析過程では，まず，以下の手順で予備解析を実施した。第一に，連続データおよびカテゴリーデータの分布を検討した。前者については，要約統計量を吟味し，分布を視覚的に確認するためにヒストグラムおよびBoxplot を作成した。その結果，異常値が発見されれば，これらの異常値は欠測値として扱うことにした。後者については，度数表およびヒストグラムを作成した。その結果，数個の異常値が発見されれば，それらはその異常値を含まない最大のカテゴリー区分のデータとして包含することにした。第二に，要因候補変数間の関連を検討した。連続データについては相関解析を実施し，変数間の関連を視覚的に観察するために散布図を作画した。カテゴリーデータについては m×n 度数表（m および n は変数それぞれのカテゴリー区分数）を作成し

独立性のカイ二乗検定を実施した。

次に，要因探索過程については，償却の有無に影響を及ぼす因子を探索するという目的を考慮して樹木構造接近法を採用した。樹木構造接近法は，分類や予測に広く使われているデータ解析法であり，誰にでもわかりやすい“ルール”によって作成される点に特徴がある（詳細については［1］を参照されたい）。

結果が樹木図で表現される樹木構造接近法には数種類のアルゴリズムがあり，それぞれ解析の目的やデータの属性によって使い分けられる（データの属性によるアルゴリズムの体系的整理については[4]を参照されたい）。ここでは，要因候補のデータ属性に従って，CART（連続量データに対して適用される）と CHAID（カテゴリーデータに対して適用される）を採用した。

CART（Classification And Regression Tree）[5]は分類回帰樹木と呼ばれる２分割成長アルゴリズムである。CART は，データを２つのサブセットに分割し，各サブセット内の均質性が分割前のサブセット内の均質性より増すように分割される（具体的には，分割前のデータ群内の平方和と２分割されたサブセット内の平方和の合計との差が最大となる分割が採用される）。そして，所与の等質性基準あるいは分割停止基準が満たされるまでこの分割を繰り返す。

一方，CHAID（Chi−squred Automatic Interaction Detection）[6]はデータのセグメンテーションに非常に効果を発揮するデータ解析法である。CHAID は，まずデータ内で均質と判断されたカテゴリー内データを結合し，異質と判断されたデータと区別する。次に最も均質と判断されたサブセットと異質と判断されたサブセットに分割される。この分割過程を所与の分割停止基準が満たされるまで繰り返す。CHAID では，分割の基準として統計的仮説検定の p 値を利用して潜在的説明変数のカテゴリー対のデータをすべて評価する。

最後に，樹木構造接近法により抽出された影響因子の妥当性を評価した。ここでは，判別解析を実施し，その判別性能を評価するために，交差確認法（Closs−Validation）によりその誤分類率をもとめた。

この誤分類率が現状の誤分類率（35％）より向上することが確認できれば本マイニングによる業務改善の効果があったと判定できる。

４．与信リスクデータセットへの適用と結果

４．１．データセットの内容

消費者金融で実施する消費者ローンに契約した顧客を対象に，初期与信時における契約情報が収集されていた。ここでは，初期与信リスクに関与すると考えられる53個の因子候補および当該契約の償却の有無に関するデータが観測されていた。因子候補はそれぞれ，属性因子（背景因子，職業因子，住居因子），借入因子，貸付因子，償却因子と考えることができた。なお，初期与信時には，償却因子および貸付因子のデータは発生しないことから，初期与信リスクの解析ではこれらの因子は採用しなかった。データは顧客の既存システムのデータベースに格納されていたが，本解析のため CSV テキストファイル形式に抽出された。このとき，顧客既存システムが最新版となった1997年以降のデータがその信頼性の観点から解析対象データとして抽出された。その結果，解析対象データは12,461件であった。表１および表２に観測された項目とそのカテゴリー区分を示す（償却の有無については，１：償却なし，２：償却あり）。

４．２．解析の内容と結果

影響因子候補から償却の有無に影響を及ぼす因子を探索するために，償却の有無を応答とする樹木構造接近法を作成した。ここでは，説明変数を連続値（あるいは順序のあるカテゴリー値）とカテゴリー値に分けて，それぞれ CART および CHAID で解析した。

CART による解析では，説明変数として，性別，年齢，既婚区分，同居人数，子供人数，勤続年数，給与支給日，役職，申込限度額，申込借入総額，従業員数，賞与，月収，住宅費月払い，住宅費ボーナス払い，住居年数，L 件数，総額残高金額を採用した。なお，解析では樹木を４段階まで成長させた。CART 解析の結果を図２に示す。

この結果，償却の有無へは，L 件数，申込限度額，申込借入総額がこの順に影響を及ぼしていることが示唆される。とくに，L 件数が５件以下で申込限度額が5,000円以下の契約は償却になる傾向があり（応答の平均値：2.0，件数：123件），逆に，L 件数が３件以下で申込限度額が5,000円より大きい契約は償却とならない傾向のある（応答の平均値：1.07，件数：7,445件）ことが示唆される。

また，L 件数が５件以下の契約では申込限度額が影響を及ぼしているのに対して，L 件数が６件以上の契約では申込借入総額が償却の有無に影響を及ぼしている傾向がある。

一方，CHAID による解析では，説明変数として，同居人数，職種，役職，会社ランク，月収，住宅形態，住宅費区分，住居年数，持参資料，利用目的，L 件数，年

表1 観測された項目のカテゴリー区分：属性因子

因子候補	カテゴリー	因子候補	カテゴリー	因子候補	カテゴリー
(1)性別	男（8314 件） 女（4147 件）	(7)役職	係長以上（752 件） 課長以上（540 件） 部長以上（988 件） その他（1964 件）	(13)住宅形態	持家（4009 件） 社宅（653 件） 賃貸（6169 件） 同居（1335 件） 独身寮（220 件） その他（42 件）
(2)既婚区分	未婚（6371 件） 既婚（6090 件）	(8)従業員数	0 人（7691 件） 1〜99 人（3322 件） 100〜999 人（1041 件） 1000〜9999 人（335 件） 10000 人以上（72 件）	(14)会社ランク	官公庁（329 件） 都市上場（569 件） 地方上場（356 件） その他（9273 件）
(3)同居人数	0 人（3275 件） 1 人（2007 件） 2〜3 人（4611 件） 4〜5 人（2237 件） 6 人以上（331 件）	(9)賞与	0 ヶ月（4769 件） 1〜3 ヶ月（2987 件） 4〜6 ヶ月（2377 件） 7〜9 ヶ月（314 件）	(15)住宅費区分	ローン（2175 件） 賃貸（6738 件） なし（3425 件）
(4)子供人数	0 人（7491 件） 1 人（1904 件） 2 人（2156 件） 3 人以上（910 件）	(10)月収	10 万未満(718 件) 10〜19 万(2432 件) 19〜27 万(3143 件) 27〜35 万(2352 件) 35 万以上(3816 件)	(16)住宅費月払い	0 万（11849 件） 1〜4 万（2342 件） 5〜9 万（1538 件） 10〜19 万（471 件） 20 万以上（47 件）
(5)年齢	20〜25 歳（1779 件） 26〜34 歳（3398 件） 35〜55 歳（5873 件） 56〜60 歳（745 件） 60 歳以上（666 件）	(11)勤続年数	0 年（310 件） 1〜2 年（2332 件） 3〜6 年（3998 年） 7〜9 年（1800 件） 10〜14 年（1522 年） 15〜24 年（1239 年） 25 年以上（787 件）	(17)住宅費ボーナス払い	0 万（11849 件） 1〜19 万（267 件） 20〜39 万（292 件） 40 万以上（53 件）
(6)職種	事務管理職(1460 件) 労務（3877 件） 専門技術（1193 件） 販売・営業（3006 件） 経営（361 件） 自由業（189 件） その他（1295 件） 雇員（1964 件）	(12)給与支給日	上旬（1862 件） 中旬（1732 件） 下旬（7672 件）	(18)住居年数	1 年未満（1624 件） 1〜2 年（2332 件） 3〜4 年（1718 件） 5〜9 年（2863 件） 10 年以上(3924 件)

表2 観測された項目のカテゴリー区分：借入因子

因子候補	カテゴリー	因子候補	カテゴリー	因子候補	カテゴリー
(19)申込借入総額	借入なし（4512 件） 1〜100 万未満（4361 件） 100〜200 万未満（2606 件） 200〜300 万未満（807 件） 300 万以上（175 件）	(21)利用目的	生活費（1816 件） 出産・教育（531 件） 慶弔費（577 件） 旅行・レジャー（2979 件） 車・車検（1519 件） ショッピング（1727 件） その他（3312 件）	(23)L 件数	0 件（1934 件） 1 件（1844 件） 2 件（1895 件） 3 件（1850 件） 4 件（1977 件） 5 件（1718 件） 6 件以上（1243 件）
(20)申込限度額	0〜10 万未満（179 件） 10〜50 万（2516 件） 50〜100 万未満（9727 件） 100 万以上（38 件）	(22)持参資料	保険証（7256 件） 免許証（4632 件） その他（573 件）	(24)総額残高金額	残金なし（53 件） 1〜50 万未満（1940 件） 50〜100 万未満（2006 件） 100〜500 万未満（2325 件） 500 万以上（6137 件）

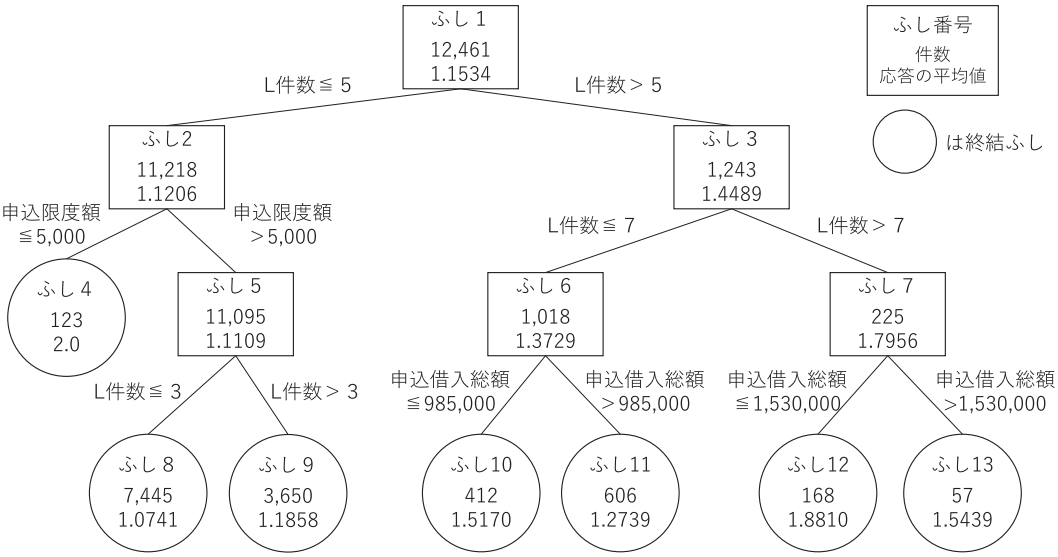

図２　連続データに対する CART 図

齢，勤続年数を採用した。ここで，連続データについては，表１および表２に示すカテゴリー区分でカテゴリー化した。また，CHAID のカイ二乗検定の有意水準は 0.05とし，樹木の分割深度は４段階とした。CHAID 解析の結果，償却の有無へは，月収，Ｌ件数，住居年数，住宅形態，勤続年数が影響を及ぼしていることが示唆された。

とくに，月収が10万～19万円未満でＬ件数が６件以上である契約は償却となる傾向のあることが示唆される（応答の平均値：1.64，件数：144件）。逆に，月収が10～27万円未満でＬ件数が１件以下，住宅形態が独身寮・社宅・同居である契約は償却とならないことが示唆される（応答の平均値：1.0024，件数：421件）。また，月収なし，10万～27万円未満，27万円以上の各層では償却の有無に影響を及ぼす要因が異なっていた（月収なし：住居年数・Ｌ件数，10万～27万円未満：Ｌ件数・勤続年数・住宅形態，27万円以上：住宅形態・Ｌ件数）。

以上の解析結果から，初期与信リスクに影響を及ぼす因子として，Ｌ件数，申込限度額，申込借入総額，住居年数，勤続年数，月収，住宅形態が重要であり，とくに，Ｌ件数および申込限度額は契約の償却の有無に大きな影響を与えていることが示唆される。

４．３．評価と考察

償却の有無に影響を及ぼす因子と判断された因子（Ｌ件数，申込限度額，申込借入総額，月収，住居年数，勤続年数，住居形態）がどの程度に実際の償却の有無を正しく判断しているか否かを評価するために，影響因子を説明変数とする判別解析を実施した。判別解析の誤分類情報を表３に示す。

表３　判別解析における交差確認法の誤分類情報

応答/判別結果	償却	正常	合計	誤分類率
償却	287 件 (16.69%)	1483 件 (83.31%)	1780 件 (100.0%)	83.31%
正常	141 件 (1.38%)	10049 件 (98.62%)	10190 件 (100.0%)	1.38%
合計	438 件 (3.66%)	11532 件 (96.34%)	11970 件 (100.0%)	13.57%

判別解析および交差確認法の結果から，償却の有無に影響を及ぼすと思われる上記の因子を説明変数とする場合，正常契約を「償却」または償却契約を「正常」と誤って判断する率は低かった（誤分類率：13.57%）。また，実際には正常である契約を「正常」と正しく判断する傾向が強く（分類率：98.62%），逆に「償却」と誤って判断する傾向は弱かった（誤分類率：1.38%）。一方，実際には償却となる契約を「償却」と正しく判断する傾向は弱く（分類率：16.69%），逆に「正常」と判断する傾向が強かった（誤分類率：83.31%）。ただし，初期与信においては，なるべく多くの顧客と契約し契約件数を拡

表４　償却の有無に関する契約プロフィール

償却傾向	プロフィールの識別	応答の平均値	影響因子		
			L件数	申込限度額	申込借入総額
大↑|↓小	a	2.0	5件以下	5,000円以下	―
	b	1.88	7件より大	―	1,530,000円以下
	c	1.54	7件より大	―	1,530,000円より大
	d	1.52	5件より大7件以下	―	985,000円以下
	e	1.27	5件より大7件以下	―	985,000円より大
	f	1.19	3件より大5件以下	5,000円より大	―
	g	1.07	3件以下	5,000円より大	―

大することに留意すると，実際には償却である契約を「償却」と判断することより，実際に正常な契約を「正常」と判断することがより重要であろう。

従来の貸付採点表での契約における正常／償却の誤分類率が35％（すなわち分類率は65％）であったことを考慮すると，本マイニングで抽出された因子に重点を置いて貸付採点表を再検討することによって誤分類率を13.57％（すなわち分類率86.43％）まで向上させることができる。すなわち，樹木構造で表された顧客プロフィールに留意して貸付採点表の採点項目に，L件数，申込限度額，申込借入総額，月収，住居年数，勤続年数，住居形態を取り込むことによって，業務を効率的・経済的に運営することが可能となる。

また，図２に示すCARTの結果から，償却の有無に影響を与える契約（顧客）のプロフィールを整理した（表４を参照）。このプロフィール表から，貸付採点表での点数づけの方針を提案することができる。すなわち，表４で上部の契約プロフィールほど償却する傾向が強く，下部の契約プロフィールほど償却しにくい傾向があるため，これを参考に点数づけを行えばよい。

例えば，L件数が５件以下で申込限度額が5,000円以下の契約は償却する傾向が非常に高いため点数づけを小さくする。逆に，L件数が３件以下で申込限度額が5,000円より大きい契約は償却しにくいので高い点数づけをおこなってもよい。

５．結びにかえて

本稿では，金融業，とくに消費者金融におけるデータマイニングのプロセスと契約の償却にデータマイニング手法の一つである樹木構造接近法を適用した。そこでは，CARTおよびCHAIDによって抽出された影響因子を採用することでより良好な業務運営が可能なことが提示された。

今後には，本稿で非定型的に実施されたデータマイニングの一連のプロセスをどのようにシステム化し，定型的に業務へ結びつけることができるか否かに留意しなくてはならない。

また，本マイニングの結果を参考に新しい貸付採点表を作成し，そこで新たな契約データを収集することが必要である（図１のフィードバック過程を参照）。さらに，そのデータをマイニングにかけることよって今回の結果を評価し，解析結果を業務にフィードバックしていくことが重要であろう。このように，図１のマイニングプロセスを繰り返し適用していくことで，より精度の高い信用リスク管理を実現することができる。

一般に，樹木構造接近法に基づく方法では，説明変数（ここでは要因候補となった項目）と応答（ここでは償却の有無）の絡み具合を定量的に評価することが困難であり，説明変数の分割に基づく標本の分離に伴い過大あてはめになりがちである。したがって，実地の適用では諸種の妥当性確認手段を組み入れた結果を提示し，経験則やデータの背後にある知識を整備して解釈に活かすことが肝要であろう。

参考文献

［１］Berry, M. J. A & Linoff, G.（1997）. Data Mining Techniques : For Marketing, Sales, and Customer Support. John Wiley & Sons, Inc.［江原淳・佐藤栄作共訳（1999）．データマイニング手法．海文堂］

［２］Deming, W. E.（1986）. One of the Crisis. Cambridge : Massachusetts Institute of Technol-

ogyPress.
［3］Nelder, J. A.（1999）. Statistics for the Millennium : From statistics to statistical science. The Statistician, 48（2）, 257–269.
［4］Eto, T.. Survival CART and its applications. Department of Informatics and Mathematical Science, Graduate School of Engineering Science, Osaka University.
［5］Breiman, L. , Friedman, J. H., Olshen, R. A. & Stone, C. J.（1984）. Regression trees. Classification and Regession Trees. Ch. 8, 216–265, Wadworth.
［6］Kass, G. V.（1980）. An exploratory technique for investigating large quantities of categorical data. Appl. Statist. , 29, 119–127.