データ解析コンペティション課題設定部門
—EC サイト顧客の顧客セグメントの予測—

松本　健，西郷　彰

平成 23 年度データ解析コンペティション「課題設定部門」は，ゴルフ用品の販売データが提供され，初期購入時点で将来において重要な顧客になるかどうかの判別をしたいという課題が課せられた．予測精度を決める要素として，どのモデルを採用するかが重要な要素になっているが，モデル以外の要素として，変数の加工や予測手順なども重要な要素である．われわれは，限られた時間のなかでこれらのバランスを考えながら分析を行った．以下で，どのような分析戦略のもと，予測を行ったかについて紹介する．

キーワード：予測精度，予測手順，決定木，RandomForests，説明変数作成

1. はじめに

平成 23 年度データ解析コンペティションにおいては，フリー部門・課題設定部門ともに（株）ゴルフダイジェスト・オンライン（以下，GDO）[1]からデータが提供された．課題設定部門は，顧客セグメントを予測するという課題が課され，精度によって評価される．精度を決める要素として，どのモデルを採用するかが重要な要素になっている．しかし，モデル以外の要素として，変数の加工や予測手順なども重要な要素である．われわれは，限られた時間のなかでこれらのバランスを考えながら分析を行った．以下で，どのような分析戦略のもと予測を行ったかについて紹介する．

GDO のビジネスモデルは，広告ビジネス，ゴルフ用品の販売ビジネス，ゴルフ場の予約ビジネスの 3 つを主に行っている．今回のデータ解析コンペの設定部門では，ゴルフ用品の販売データが提供され，初期購入時点で将来において重要な顧客になるかどうかの判別をしたいというのが，ひとつの課題になっている．具体的な提供データは，2010 年 7 月度に新規登録した会員のデータで，入会後 90 日間のデータを用いて 1 年後のセグメントを予測するというものである．

セグメントは，RFM（Recency: 直近購買日までの間隔，Frequency: 累積購買回数，Monetary Value: 累積購買金額）のうち F と M から構成され，Frequency で 3 セグメント，Monetary で 3 セグメント，合計 3×3 の 9 セグメントを予測するものとなっている．

2. モデルの基本戦略

このような FM で 9 セグメントを予測するうえで，われわれは大きく以下 5 つの方法を検討した．

1) カスタマが 1〜9 のどのセグメントに入るかを予測する（図 1）．
2) カスタマが Monetary の 3 セグメントに入るか，Frequency の 3 セグメントに入るかを別々に予測する（図 2）．
   - ステップ 1：Frequency を予測する．（3 カテゴリの予測）
   - ステップ 2：Monetary を予測する．（3 カテゴリの予測）
   - ステップ 3：予測された Frequency と Monetary をかけ合わせる．
3) セグメントを予測する方法（図 3）として，
   i. セグメントをカテゴリ変数とし，直接セグメ

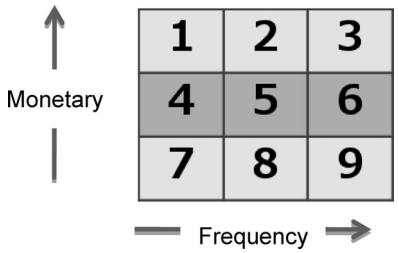

図 1　方法 1

まつもと　たけし
（株）リクルートライフスタイル
さいごう　あきら
（株）リクルートテクノロジーズ
〒 100–6640 東京都千代田区丸の内 1–9–2
グラントウキョウサウスタワー

[1] http://www.golfdigest.co.jp/

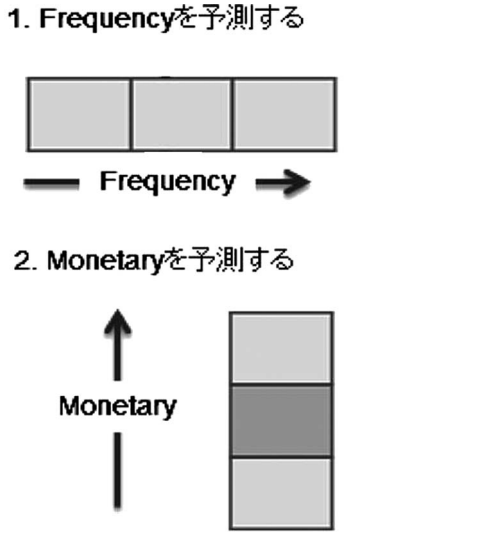

図 2　方法 2

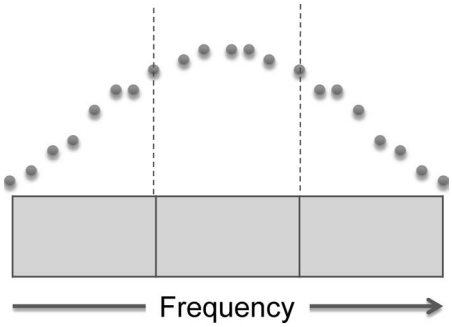

図 3　方法 3

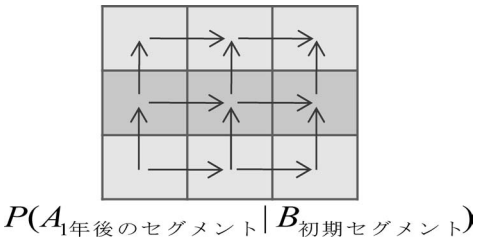

図 4　方法 4

ントを予測する.

ii. Monetary と Frequency の値（連続値）を予測し，次にセグメントに置換する.

4) 入会後 90 日間の時点でのセグメントから，1 年後にどのセグメントに遷移するかの予測を行う（図 4）.

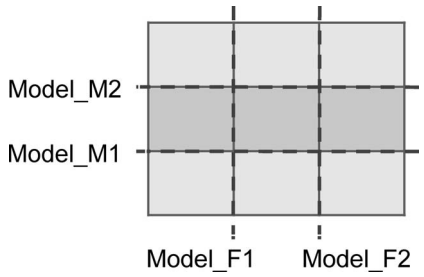

図 5　方法 5

5) 2 種類 (Frequency)×2 種類 (Monetary) の判別モデルの組み合わせを用い，セグメントを分割する（図 5）.

上記を初期モデル検討した結果，2) の Monetary3 セグメント × Frequency3 セグメントを予測し，9 セグメントを作成するものが一番精度よく，また，3) に関しては，i（カテゴリを予測）と ii（連続値を予測しセグメントに置換）を比較したところ，i の精度が高いことがわかった.

3.　手法の検討

3.1　モデル検討 1

今回は，われわれは独自の手法の開発にこだわりはなかった．その理由として，取り組める時間の制約を考えると，独自手法の開発を行うと，変数作成，頑健性，ポストフィルタリングといった部分の時間を取ることができなかったからである．今回の総合的な思想としては，バランスよく分析を行うことで総合的に精度を上げるということを考えた.

そのため，既存のモデルをベースに多くのモデルを試し（ステップ 1），その中で筋の良さそうなモデルのチューニングを行う（ステップ 2）という 2 段階の方針でモデルを検討することにした.

今回ステップ 1 でモデルを試したものは，以下のモデルである.

- IBM SPSS Modeler
  - ▶ CHAID
  - ▶ C&R Tree
  - ▶ C 5.0
  - ▶ ニューラルネットワーク
  - ▶ ロジスティック回帰
  - ▶ Support Vector Machine (SVM)
  - ▶ Naïve Bayes
- KXEN
  - ▶ K2R

- Visual Mining Studio
  - ▶ 決定木
  - ▶ SVM
- R
  - ▶ RandomForests

3.2 モデル検討 1

分析を進めるにあたって，既存のモデルを試しつつ，オープンソースに実装されている新しい手法へのチャレンジも行った．具体的には統計解析言語 R に実装されている RandomForests について取り組んだ．RandomForests は 2001 年に Breiman により提案された比較的新しい手法で近年実務ビジネスのデータ解析・機械学習においても使われるようになってきた [1]．変数をランダムにサンプリングしつつ，Bagging を行うので高次元のデータに向いている．今回のデータから作成した変数は後述するが，約 500 であり，RandomForests は有効な手法と考えられた．図 6 はテストデータにおける学習回数と誤り率で，今回のモデル作成においては，学習回数を 1,000 回とした．

4. モデル精度の向上

4.1 変数作成

使用する変数として，性別，年齢，居住地，ハンディキャップといったデモグラフィック属性と，サイトの訪問回数，流入経路といった行動履歴の 2 種類の変数を用意した．特に今回の目的変数は，Frequency と Monetary といった行動履歴が目的変数となっているため，どのような行動履歴の変数を作成するかが重要になると考えた．変数を作成する過程として，やみくもに変数を作成するのではなく，重要となるであろう項目について仮説を置き，その仮説を検証しながら変数を作成していくことを考えた．

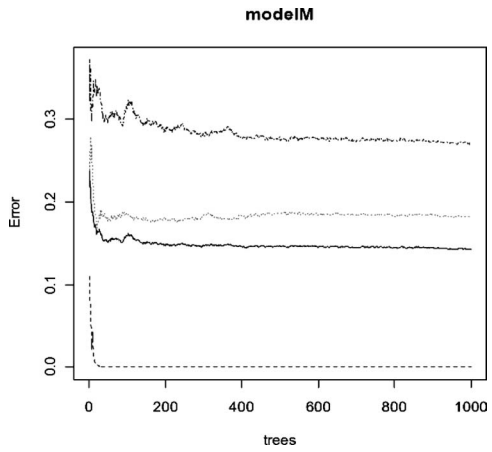

図 6　RandomForests の収束状況

具体的には次の変数を作成した．ログデータから作成した変数は 305，予約データから作成した変数は 80，受注データから作成した変数は 130，会員データから作成した変数は 8，合計 523 変数を作成した．

- 訪問回数
- 来訪時間と日時のセグメント
- 流入経路
- 入口／出口
- 各ページの PV 数
- ゴルフ場の予約日数
- ゴルフ場のお気に入り度
- ゴルフ用品の購入日数
- ゴルフ用品の購入金額
- ゴルフ用品の回数

以下にいくつかの仮説とその検証についてまとめる．

訪問回数に関する仮説
平均的に来訪している，登録後だけ来訪する，直近の来訪が増えてきたといった来訪パターンの変化が考えられ，これらの差異が 1 年間の Frequency と Monetary に影響を及ぼすのではないか？

変数を作成するにあたり，仮説としてカスタマの来訪間隔は週単位で考慮したほうがよいのではと考えた．そこで，来訪日から次の来訪日までの間隔をプロットした（図 7）．

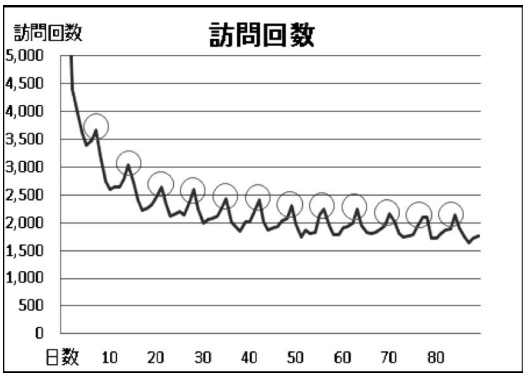

図 7　来店間隔ごとの訪問回数

図 7 をみると 7 の倍数ごとに訪問回数が多くなっていることがわかる．そこで，1 週間ごとに訪問回数のピークがあると考え，1 週間単位で各変数を作成することにした．

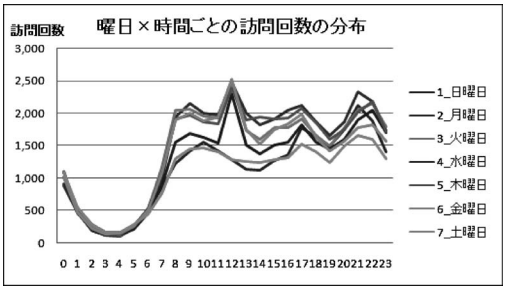

図 8　曜日×時間ごとの訪問回数の分布

来訪時間と日時のセグメントに関する仮説
サイトに夜だけ訪問したり，週末だけ訪問したりするカスタマが多く，訪問日時の差異が 1 年間の Frequency と Monetary に影響を及ぼすのではないか？

図 8 に示すように曜日×時間ごとの訪問回数の分布を作成した．曜日によって特に日中の訪問回数にばらつきがあることがわかる．また，夕方 17 時と夜 22 時頃にもピークが発生していた．さらに，平日については昼に最大のピークがあるが，週末はこのピークが観測されない．これらの結果から，曜日については各曜日に分けた 7 セグメント，時間については「朝：4 時 ～ 10 時」，「昼：11 時～14 時」，「夕方：15 時～19 時」，「夜：21 時～3 時」の 4 セグメントに分け，これを掛け合わせて $7 \times 4 = 28$ セグメントを作成した．

4.2　モデル作成時の留意点

今回作成した変数の数は約 500 に上る．今回の学習用データに対して予測精度は高まるが，モデルが過学習している可能性が高い．そのために，データを学習用，テスト用，検証用の 3 つに分けることにした．

それぞれのデータの役割として，学習用のデータを使いモデルを作成した．そして，モデルのパラメータを変えたり，変数を増減したりしながら，最良のモデルを作成した．その際，テスト用データを使いモデルの調整を行った．しかし，複数のモデルを 1 つのテストデータで調整してしまうと，今度はそのテストデータに対しての過学習の可能性が出てくる．そのため，最終的なモデルの結果を判断するために検証用データを用い，各モデルの精度を比較した．

別途，モデルの安定性や頑健性を強めていくために，次の 2 つのことを行った．

1 精度が大きく落ちない程度にバランスよく，変数を削除した．
2 Bagging や RandomForests などの集団学習アプローチをとった．

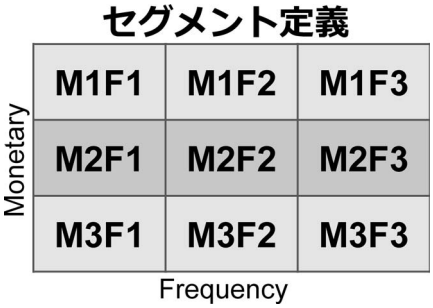

図 9　セグメントの定義

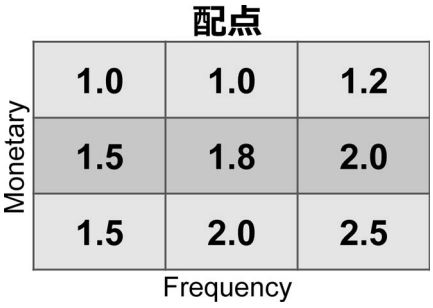

図 10　セグメントごとの配点

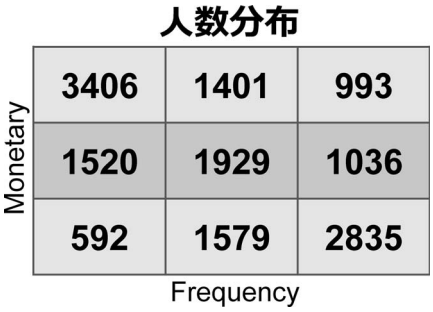

図 11　セグメントごとの人数分布

4.3　1 票の重さと利得

各セグメントを図 9 のように名づける．今回のコンペティションにおいては，各セグメントにおける正答の配点が異なっている．具体的には図 10 に示したように，FM が高くなるにつれ配点が高くなっている．そのため，直感的に配点の高いセグメントに予測割り当てをすると総スコアが高くなると考えられる．一方で，図 11 に示すように各セグメントの人数分布に偏りがあるため，その出現率は異なっている，このため，配点の低いセグメントにおいても，1 人あたりのスコアの持ち分が変わってくる．つまり，どこのセグメントにユーザーが布置するのかを割り当てる際の 1 票の重さが異なっていることがわかる．このように各セグメントにおける，1 人あたりの配点比率を $score_j$ とし

セグメント j の $Score$

Monetary \ Frequency			
	0.294	0.714	1.208
	0.987	0.933	1.931
	2.534	1.267	0.882

図 12　セグメントごとのスコア

ユーザ i の $prob_{ij}$

Monetary \ Frequency			
	0.20	0.15	0.10
	0.15	0.10	0.07
	0.10	0.07	0.06

図 13　あるユーザーのセグメントごとの存在確率

ユーザーの期待利得

$$g_{ij} = prob_{ij} \times score_j$$

Monetary \ Frequency			
	0.059	0.107	0.121
	0.148	0.093	0.135
	0.253	0.089	0.053

図 14　あるユーザーの期待利得

て，そのセグメントの正答 1 人の利得として定義した（図 12）．

また，一方でモデル作成によって与えられるユーザー i の各セグメント j における存在確率を $prob_{ij}$ として定義し（図 13），ここに前述の正答 1 人の利得 $score_j$ の積を求めることにより，ユーザーがセグメントに割り当てることによる期待利得 g_{ij} が求まる（図 14）．今回はこの期待利得 g_{ij} が最も高いセグメントを予測結果として割り当てた．

例えばあるユーザー i における存在確率 $prob_{ij}$ は下記のようになっており，M1F1 が 0.20 と最も高いが，期待利得 g_{ij} は下記のようになっており，最も高いセグメントは M3F1 となる．当該ユーザーにおいて予測セグメントは M3F1 となり，コンペティションの回答提出においては，ユーザー 13,563 人に関して，同様の方法で予測セグメントを割り当て，提出を行った．

4.4　Post Scoring

モデル作成を進めていくなかで，正答率の偏りが顕著に見られた．例えば，RandomForests については，単体モデルで予測を行うと M2F3 のセグメントが極端に低い正答率となってしまうことがわかった．

そのため，低い正答率のセグメントだけを別の予測モデルで補う，PostScoring という手法を用いて全体精度の向上を目指した．

例えば，RandomForests に関しては，M2F3 のセグメントの予測を CHAID で補うことによって，ほかのセグメントの予測精度がほとんど変わらないまま，正答率を 36.1%から 48.8%に向上させることができた（図 15）．

RandomForests　→9.95

Monetary \ Frequency	1	2	3
1	96.1%	78.9%	81.1%
2	85.0%	71.2%	36.1%
3	69.2%	55.5%	72.0%

RandomForests
×CHAID_M2F3　→10.14

Monetary \ Frequency	1	2	3
1	96.1%	78.9%	81.1%
2	85.0%	71.2%	48.8%
3	69.2%	55.5%	69.5%

図 15　PostScoring 前（上）と PostScoring 後（下）

決定木やニューラルネット，SVM などの先に挙げたさまざまなモデルを試しながら，最もよい組み合わせを探索しつつ，最終モデルを検討した．その際に以下に示すようないくつかの知見が得られた．

1) モデルごとに得意・不得意がある
それぞれのセグメントを判定するモデルのアルゴリズムに起因すると思われる．低い確率のセグメントを充てられるかどうかという点で主に違いがみられた．
2) RandomForests はベースモデルで精度がよい
変数の多さや集団学習のアプローチのためと考察

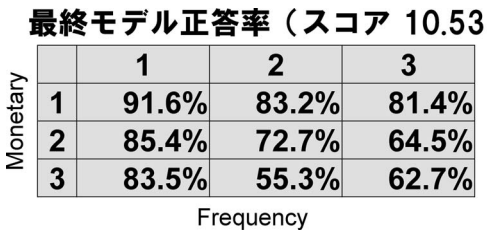

図 16　最終モデルの正答率

した．比較的，頑健性の高いモデルが 500 変数で 3 カテゴリ判別という条件にフィットしたものと考察された．

3) ニューラルネットはベースモデルでは精度が良くない
多くの変数があるなかで，変数の取捨選択や，中間層の検討を行いきれなかったことが原因と考えられた．

4) Tree 系は PostScoring において精度向上に寄与
細かく切り分けた 1 セグメントを判別するのに適していることが考察された．

4.5　最終モデル

前述の方法でモデル作成と予測を行い，検証データセットにおける最終正答率とスコアを算出した（図 16）．モデルによる予測セグメントを提出した．最終モデルは，株式会社数理システム VMS: 決定木 (decision tree) を IBM SPSS Modeler: CHAID で PostScoring したものであった．予測スコアは 10.53 であった（図 16）．また，次点のモデルは RandomForest に IBM SPSS Modeler: C5.0 で PostScoring を追加したモデルで，同様に検証データで評価したところ 10.49 と僅差であった．

5.　まとめ

5.1　予測精度向上について

今回のコンペティションの取り組みにおいて気づかされラーニングになったことをまとめると「バランスよく検討する」ということであった．予測精度を上げるために利用できる手法はさまざまで，例えば，どのモデルを使うのか，モデルの変数は何にするのか，頑健性（ロバスト性）や安定性はどう担保するのか，また評価指標は何にするのか，などを慎重に検討することが重要と考えられる．

例えば，評価指標においては，配点が高い F3M3 を当てることが必ずしも重要ではない．バランスよく当てることが大切であり，前述にも述べたが，特に 1 人正解することによって得られる利得が高いセグメントが重要である．実際のビジネスにおいても，購入金額が大きいロイヤル顧客を当てることが必ずしも重要でないシーンがあり，それと同じことが考察できる．

また，実務での分析業務でも同じことが言えるが，今回はコンペティションの期間が限られていた．つまり限られた時間制約の中でどこが重要なのか見定め，どこかにフォーカスしてしまいやすいが，そうではなく，バランスよく分析・実装を進めるアプローチが大切であると考察できる．

5.2　さいごに

実務に関しての適応をテーマとした，大変意味のあるコンペティション課題であった．今回提供されたようなネットビジネスにおける分析業務においては，大量データを加工する SQL や，マーケティングや消費者行動を読み取った変数作成，またロバストなモデルを作成するための統計解析・機械学習の知識などが，幅広く必要である．かつ，実際にサイト上の結果に反映するために，実務では，マーケティング施策の企画業務や，システム開発業務なども必要である．

昨今，この領域は，周辺領域を取り込み，データサイエンティストなる業務として認知され始めている．欧米のネット企業においてその重要性が取りざたされていること [2] がきっかけなのだが，すでにノースウェスタン大学などの名門大学といわれる教育機関でもその認識が広がり，専門講義が開催されている．現在，情報処理・統計解析・機械学習・最適化などの研究に携わっている読者（特に学生）の方々においては，長期的な視野でぜひデータサイエンス業務を目指していただきたい．

参考文献

[1] L. Breiman, “Random Forests,” *Machine Learning*, **45**, 5–23, 2001.
[2] 城田真琴，『ビッグデータの衝撃』，東洋経済新報社，2012.