EC サイトにおける購買タイミングの
レコメンデーションのための変数選択法

生 田 目　　崇
鈴　木　元　也

Variable Selection for Recommendation of Purchase Timing for EC Site

Takashi Namatame
Motoya Suzuki

EC サイトにおける購買タイミングの レコメンデーションのための変数選択法

生 田 目　　崇
鈴　木　元　也

1. はじめに

現在、小売市場においても電子商取引（Electronic Commerce; EC）の市場規模はますます増大している。経済産業省によれば、2013 年の BtoC 電子商取引市場は 11.2 兆円とついに 10 兆円の大台に達し、前年比 17.4%増と近年の成長率をさらに加速させている[1]。

EC においてもっとも顕著なマーケティング活動は、レコメンデーションであろう。EC の雄であるアマゾンが急成長した理由の一つとして、顧客別のレコメンデーションをいち早く行ったことが挙げられる。EC は品揃えの観点からは理論的には無限の品揃えが可能であるが、その選択肢の幅の広さが逆に顧客の情報探索コストの増大につながり、客離れを引き起こしかねない[2]。そこで、アマゾンは、協調フィルタリングの技術を援用し、過去の購買・閲覧行動をもとに、訪問者の訪問時点におけるレコメンド商品の選定し表示することで訪問者に利便性を提供し、ロイヤルティの確保につなげた[7]。特に、書籍のようなカテゴリのようにジャンルが購買に大きな影響を及ぼす商品においては高い効果を上げたといわれている。

ただし、こうしたレコメンデーションは「どの商品」を推薦するかの仕組みであり、「いつ」という観点には立っていない。購買決定に関する研究は大きく分けると、購買生起、ブランド選択、購買量に大別されるが、定量分析の観点からはブランド選択および購買量に関する研究が広く行われている反面、特に個人を対象とした購買生起に関する研究は少ない。日常的に反復購買されるような消費財などの場合は、こうしたタイミングについてはさほど考慮する必要はないと考えられるが、耐久品や専門品に近い反復購買品といった、高額あるいは購買頻度が高くないカテゴリにおいては、どのようなタイミングでどのような商品をレコメンドするべきかといった、日用品とは異なるレコメンドが必要となる。

また、実際に EC においてレコメンデーションを行う場合は、リアルタイムにデータベースとレコメンドシステムを連携させて訪問者に対して適切なレコメンドを行わねばならず、

蓄積されるデータとシステムに合わせた方法が重要となる。

著者らは、これまでに EC サイトにおける購買タイミングの予兆を発見するモデルを提案したが、システム実装するためには上記のように蓄積された購買データとシステムに合わせた方式が必要であり、そのまま利用することは困難である。そこで本稿では、前稿同様ゴルフ用品を扱う EC サイトを取り上げ、既存の顧客行動分析をベースに実際にシステム化を意図したレコメンド方式の提案を行う。

2. 先行研究

本稿では、EC におけるレコメンデーションの最適タイミング策定を目的とする。著者らはすでにレコメンドタイミングに関する分析モデルを示しており、統計的モデルによって､EC サイト訪問者の閲覧行動から購買につながるタイミングを評価する分析を行った。本研究は既存の研究をベースとして、レコメンドシステムに実装することを念頭に置いた方式提案を行う。以下ではまず、著者らが提案した購買予兆発見モデル[3] についてまとめる。

著者らの従来のモデルは､矢野らの先行研究[4] をもとに､耐久消費財に近いレジャー（ゴルフ）用品を扱う EC サイトを対象とした購買予兆発見モデルである。モデルの検証にあたっては､株式会社ゴルフダイジェスト・オンライン社[1]（以下､GDO）に協力いただきデータを提供いただいている。利用した GDO のウェブサイトの購買・予約データ並びにログ・データの概要は表 1 のとおりである。

表 1　データ概要

項目	データ
期間	2012 年 1 月 1 日～6 月 30 日
会員数（人）	629,902
セッション回数（回）	23,824,781
セッション pv 数（ページ）	13,626,756,342
セッション時間（秒）	3,663,149
受注回数（回）	252,742
予約回数（回）	618,245

※pv は pageview を表す

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

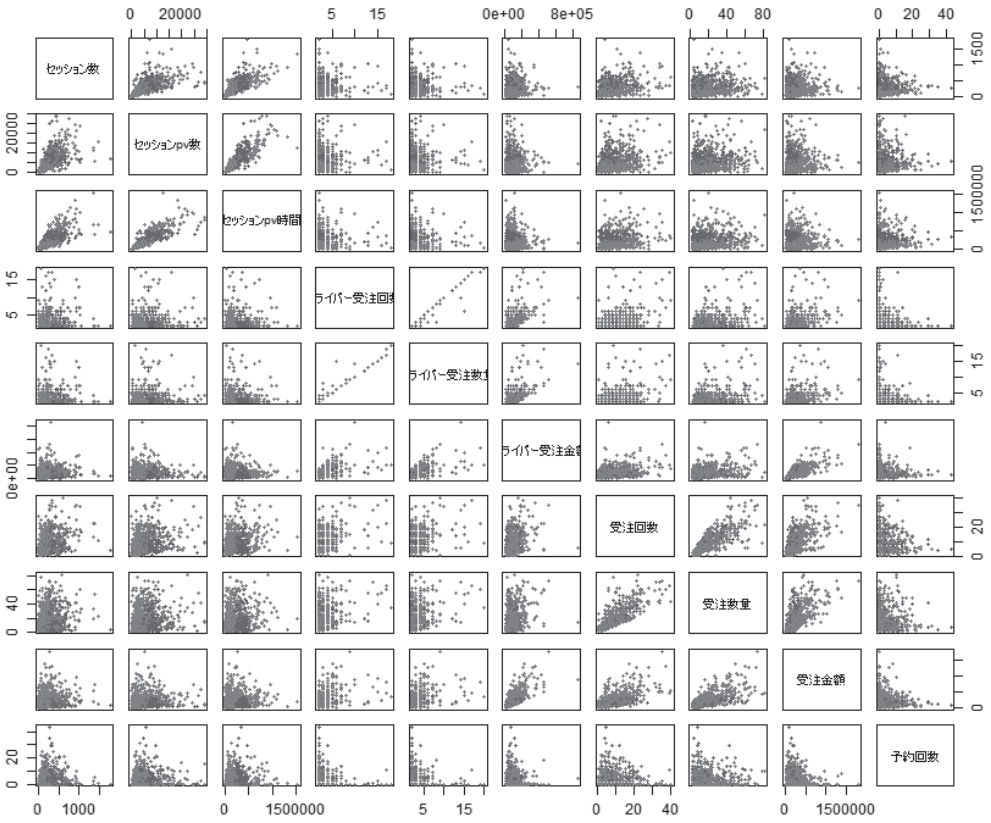

図１　登録会員の閲覧行動

ウェブ・アクセス・ログ・データはセッション番号に加え、セッション開始時間や、アクセス・ページといったアクセスに関するデータ、およびコンバージョンの有無などの情報が付与されている。また、受注データとしてはブランドやアイテムなどの名称、通常価格、割引後価格、受注数量、受注日、ポイントの付与や使用の情報、クーポン利用などの情報が含まれる。また、表１の他、プライバシーに関わる部分を削除した会員に関する一部のデータの提供を受けた。なお、データ全体から外れ値と考えられるデータについては分析から除外している。この結果、閲覧行動に関する散布図を図１に示す。

従来モデルでは､まずサイト訪問行動の多様性を考慮するために､モデル分析に先立ち、閲覧行動について非階層クラスタ分析の一つである k-means 法によって訪問者を複数のセグメントに分割している。たとえばドライバの場合、クラスタの統計量および解釈の容易さから５つのセグメントを抽出した。そして、次回購買に影響を与えると考えられる行動

変数を抽出し、購買直前の閲覧を購買準備期間としてこの期間に該当するかどうかを求める購買予兆発見モデルを、ロジスティック回帰モデルを用いて提案している。また、そのセグメントごとにロジスティック回帰分析モデルにおいて、統計的に有意となるようなものに変数を制限するように変数選択を行った。この結果、表 2 に示すようなパラメータが得られた。なお、予測の評価値について表の下部にまとめた。また、ドライバの他、パター、ボールについての分析を行っている。

表 2　パラメータと評価

クラスタ	1	2	3	4	5
切片	-1.727	-2.303	-1.746	-1.787	-1.322
セッション回数	-0.007		0.003	-0.004	0.008
セッション pv	0.000	-0.000	0.000		0.000
購買回数		0.059	0.045	-0.042	-0.134
予約回数	-0.134			-0.072	-0.125
購買経過日数		-0.006	-0.016	-0.003	-0.005
予約経過日数				0.001	
正解率	43.9%	59.1%	50.3%	46.3%	63.6%
Precision	14.4%	9.8%	17.7%	13.3%	28.3%
Recall	72.4%	37.6%	64.3%	69.5%	42.2%
F 値	24.0%	15.6%	27.8%	22. 3%	33.9%

3. システム化に向けて

3.1　システム化における制約

前稿で提案したモデルは検証の結果、一定の有効性を示した。しかし、実際にシステム運用する上での問題も残されている。もっとも大きな問題点は、実システムにおいては、コマーシャルベースのレコメンデーション・システムを用いる場合、レコメンデーション対象の判定をするために、統計モデルを使うことができない点である。

もちろん、すべてのシステムを自社開発し、統計システムを開発したシステムの中に組

み込むことはができれば、こうした問題は発生しない。しかし、EC サイトの運営においては、ウェブサーバ、商品管理システム、購買管理システム、顧客管理システムなどを複雑に連携させている。その上で、レコメンデーションをするかどうかを判定してレコメンド対象を選定と実際のレコメンドを行わなければならない。

そのために、国内外のシステム企業が [8] のようなレコメンデーションのための専用システムを数多く開発している。こうしたシステムはレコメンデーションに加えて、AB テスト、ターゲティング、効果測定といったルールベースのウェブ・マーケティング管理などを行うことができる。顧客へのレコメンデーションは、観測される属性変数や行動変数に関して閾値を設定し、その閾値を超えるかどうかによってレコメンデーション対象かどうかを判別するといった方法が一般的に可能である。したがって、このようなシステムにおいて、前述のような統計モデルを直接利用することはできない。

3.2　システム化のためのルールベース・購買予兆発見方法の提案

前節でに述べたように、統計モデルは顧客の行動から実際の購買予兆を発見するために有効な知見が得られるにも関わらず、実システムで運用しようとしても、そのままでは実装できないという問題を含んでいる。

そこで、実システムで実現可能な形式で購買予兆発見を行うために、属性変数と行動変数を計測し、その閾値を求めることで購買のタイミングを計る方法を以下に提案する。

そのために、本稿では決定木分析[6] によるルール生成を行う。具体的には、前稿と同様、購買履歴と閲覧履歴から、購買より 1 週間前の期間のサイト訪問を購買予兆期間とし、その期間の閲覧を購買準備とみなし、それ以外の閲覧を購買予兆のないものとして、これらを教師データとする。そして、その教師データをうまく分割する変数とその閾値を、決定木分析を用いて求める。なお、決定木分析は前述した顧客セグメントの識別および購買予兆の判定のそれぞれについて行う。そして、得られた決定木の評価を行う。

本稿のアプローチを図 2 にまとめる。

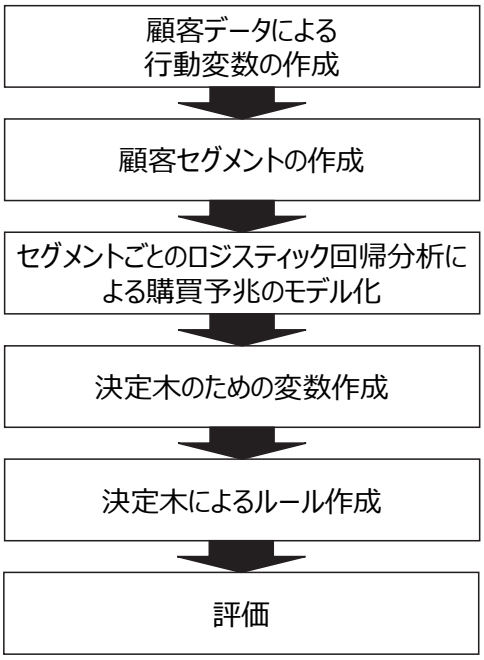

図 2　本稿のアプローチ

4. 実データによる分析と評価

4.1　変数の抽出

対象とするカテゴリの購買予兆を検知するためには、各訪問者の訪問機会ごとに購買に至りそうかどうかをそれぞれの閲覧状況から判定しなければならない。したがって、多くの場合は Cookie もしくは登録 ID で識別できる各訪問者それぞれのサイト訪問時点で、過去の訪問履歴から行動変数を作成しなければならない。実際のレコメンデーション・システムに実装するためにはそのシステムで求めることが可能な変数を用いなければならない。そのために、データ提供元企業とディスカッションをし、実システムにおける観測可能性、および顧客の訪問行動に関する特徴的な行動を考慮して、顧客セグメントの決定および購買予兆発見のそれぞれについて、以下に挙げる変数を用いることとした。

＜セグメント特定の決定木分析のための変数＞

- 期間内セッション回数
- 期間内セッション pv 数
- 期間内セッション pv 時間（秒）
- 期間内受注回数（対象カテゴリ）
- 期間内受注金額（対象カテゴリ）
- 期間内受注数量（対象カテゴリ）
- 期間内予約回数

＜購買予兆の決定木分析のための変数＞

- 前回購買からのセッション回数
- 前回購買からのセッション pv 数
- 前回購買からの購買回数（対象カテゴリ以外）
- 前回購買からの予約回数
- 前回購買からの購買経過日数（対象カテゴリ）
- 予約日からの予約経過日数

分析対象とする EC においては、ゴルフに関する様々なカテゴリの商品を販売していることに加えて、ゴルフ場の予約ができることが特徴であり、予約に関しても変数候補としている。これらを含めて計測可能な項目を変数として取り上げている。

4.2　ドライバ分析結果

前稿同様、ドライバの購買を対象として行う。また、他のカテゴリ（パター、ボール）についても分析したが、その結果は次節にまとめる。

分析に利用する決定木分析についてはいくつものアルゴリズムが提案されており、いずれを用いるかで分割結果は異なるが、事前の分析で二分木と多分木で結果に大きな差がないことを確認したため、本稿の分析に当たっては R の rpart ライブラリを用いた二分木を採用した。分割基準は Gini 係数を用い、過学習を防ぐための枝の剪定を行っている。

まず、クラスタ特定のための決定木分析の結果とクラスタ正答率をそれぞれ図 3 と表 3 に示した。

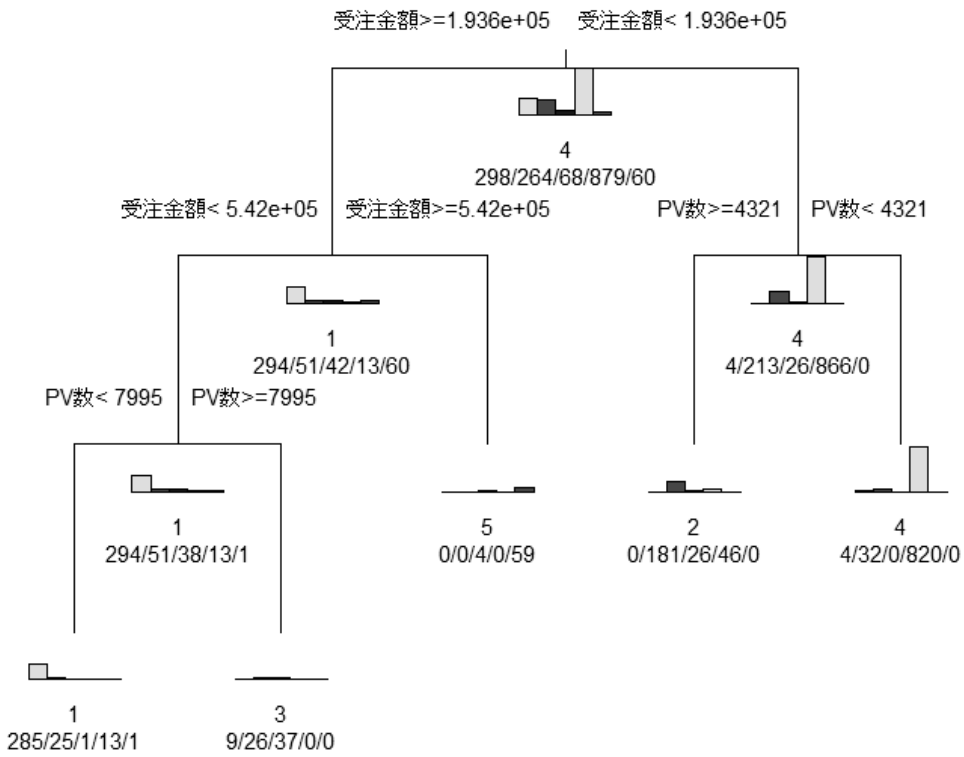

図 3　顧客セグメント識別のための決定木分析（ドライバ）

表 3　顧客セグメント正答率

セグメント	1	2	3	4	5
正答率	95.6%	68.6%	54.4%	93.3%	98.3%

表 3 は、決定木の各終端ノードに判別された顧客について、そのノードに属する顧客のセグメントのうち最も多いセグメントをそのノードのセグメントとしたときの正答率を示している。セグメント 2 とセグメント 3 の正答率が比較的低いが、図 3 のルートノードが示す[2] ように、セグメント 2 とセグメント 3 の構成比率はもともと高くなく、また前稿で考察したようにこのセグメントはいずれも購買回数が多くないセグメントである。したがって、ランダムな閲覧といった、購買以外の訪問が想定されるセグメントであるため、訪問時の行動にばらつきが大きいなどの理由であまりうまく判別できなかったと考えられ

[2] ルートノードにある各セグメントの構成比率のグラフは、左からセグメント 1、セグメント 2、…の順である。

る。しかし全体的には十分な精度ということができる。

次に、判別された各セグメントについて、購買予兆発見の決定木分析を行った。ただし、購買予兆でないとされる閲覧データが予兆とされるデータより絶対的に多い。このまま決定木分析を行うと、すべてのノード購買予兆でないと判定されてしまうため、購買予兆でないデータについてはランダムサンプリングをして、ルートノードにおける構成比率を 1:1 にしている。セグメントのサイズの大きいセグメント 4 に関する決定木分析の結果と各セグメントにおける精度について図 4、表 4 にそれぞれまとめる。なお、他のセグメントの決定木分析の結果は付録にまとめる。

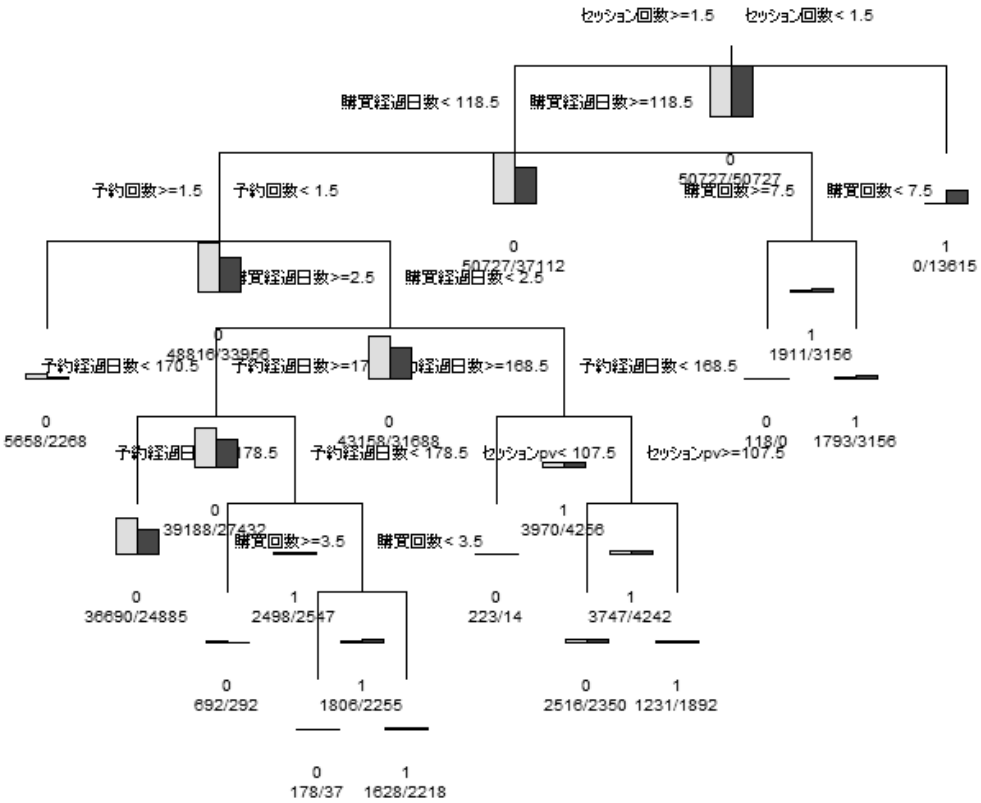

図 4　購買予兆発見のための決定木分析の結果（ドライバ、セグメント 4）

表 4　購買予兆発見のための決定木分析の正答率

セグメント	1	2	3	4	5
正答率	61.3%	60.5%	59.3%	62.4%	66.5%

表 3 を示したように、顧客セグメントほどの識別率は得られなかった。しかし、全体に 10%程度の正答率の向上が得られた。ただし、顧客セグメント分割に比べて精度は低く、購買行動に至る閲覧行動の多様性は捉えきれていない部分があるという結果となった。

4.3　その他のカテゴリの分析結果

前節では、ドライバに関する分析を行ったが、パターとボールを対象にした結果を以下の表 5 にまとめる。決定木分析の結果については付録にまとめる。表 5 に示すように、パターとボールで顧客セグメント数が異なるが、これはそれぞれの商品カテゴリについて k-means 法によるクラスタ分析を行ったが、統計的にもっとも有意なセグメント数（本稿では Calinski and Harabasz [5] の指標を用いた）を採用したためである。

表 5　パターとボールの分析結果概略

	セグメント	1	2	3	4	5	6	7	8
パター	サイズ	8	404	160	111	172	34	69	
	セグメント正答率	0.0%	92.1%	90.6%	93.7%	69.2%	0.0%	97.1%	
	購買予兆正答率	-	69.4%	65.9%	62.6%	64.1%	-	72.3%	
ボール	サイズ	771	264	48	213	1793	4962	33	979
	セグメント正答率	75.9%	0.0%	0.0%	0.0%	76.2%	91.4%	0.0%	87.0%
	購買予兆正答率	58.2%	-	-	-	62.2%	72.5%	-	63.2%

セグメントの特徴については、はっきりと特徴が表れた受注金額とセッション pv 時間について以下のようにまとめられる。

パターのセグメントは、セグメント 1 は受注金額が大変高く、セグメント 2 は受注金額、セッション pv 時間とも少ない、セグメント 3 はセッション pv 時間はセグメント 2 とあまり変わらないが、受注金額が多い、セグメント 4 と 5 は受注金額はセグメント 1 と変わら

ないが、セッション pv 時間が多いセグメントであり、セグメント 4 のほうがセグメント 5 よりも多い。セグメント 6 はセッション PV 時間が大変多いセグメントであり、セグメント 7 はセグメント 1 に次いで受注金額が多いセグメントである。

ボールについては､セグメント 3 と 7 がそれぞれセッション pv 時間および受注金額いずれも多いセグメントであり、セグメント 6 がいずれも最小のセグメントである。その他のセグメントについては、セグメント 4、1、5 は受注金額が少ないセグメントで、この順にセッション pv 時間が長い。またセグメント 2､8 はセッション pv 時間は比較的短いもののその割に受注金額が多く、この順で受注金額が多い。

なお、決定木分析によるセグメントの特定については、いくつかのセグメントの正答率が 0%となっているが、これは、終端ノードに属するサンプルのうち最大サイズのセグメントをそのノードのセグメントしたため、もともとのサンプルサイズの小さいセグメントについてはどのノードでも埋もれてしまったためである。ただし、全体の正答率は 80%を超えており、サンプル全体については比較的高い割合でセグメントを同定できているといえる。

また、購買予兆については、おおむね 2/3 以上の正答率となっており、前節で述べたドライバの場合と同様の精度は得られている。

5. まとめと今後の課題

本稿では、EC サイトの訪問者に対する購買タイミングに関するレコメンデーションのための分析方法を提案した。特に、実際のシステムへの実装を念頭に置いて、レコメンデーションすべきかの判定をするための変数選択法について言及した。分析結果から、本稿の決定木分析による判定は既存の統計的分析による判定に精度では及ばない面があるものの、一定の効果があることが確認できた。

今後の課題としては、提案したモデルを実際のレコメンデーション・システムに実装した上で効果測定をすることが挙げられる。その実施には企業の協力が必要であるが、残念ながら現状では実施に至っていない。本稿では、商品カテゴリを限定した分析モデルを示したが、購買は他カテゴリについても行われるため、カテゴリをまたいだクロスセルに関する言及も必要となる。また、購買予兆に関しては、ある程度の精度を示したものの精度の改善が求められる。以上より、現状のシステムに実装することを考えると、本稿提案が

実現可能性の点では現実的なモデルであるとはいえるものの精度の向上を目指したモデル化、システム化の改善についての余地は残されている。

謝辞

本研究にあたり、株式会社ゴルフダイジェスト・オンライン　マーケティング部の福永和洋様、光山勝之様よりディスカッションを通じて有益なコメントを多くいただきました。ここに謝意を表します。

参考文献

[1] 経済産業省、『平成 25 年度我が国情報経済社会における基盤整備（電子商取引に関する市場調査)』、2014 年。
http://www.meti.go.jp/press/2014/08/20140826001/20140826001-4.pdf (2014 年 11 月 10 日アクセス)

[2] 野村総合研究所、松下東子、濱谷健史、日戸浩之、『なぜ、日本人はモノを買わないのか?』、東洋経済新報社、2013 年。

[3] 生田目崇、鈴木元也、「EC サイトにおけるサイト閲覧行動と購買行動」『刑事英情報学会論文誌』第 22 巻、第 4 号、273-278 ページ。

[4] 矢野順子、佐治美歩、中川慶一郎、高橋彰子、山中啓之、生田目崇、「クレジット・カード利用顧客のデフォルト予兆発見分析」『オペレーションズ・リサーチ第 51 巻, 2 号, 2006 年, 104-110 ページ.

[5] Calinski, R.B. and J. Harabasz, “A Dendrite Method for Clustering Analysis,” *Communications in Statistics*, Vol.3, 1974, pp.1-27.

[6] Dahan, H., S. Cohen, L. Rokach and O. Maimon, *Proactive Data Mining with Decision Tree*, Springer, 2014.

[7] Sarwar, B., G. Karypis, J. Konstan, and J. Riedl “Item Based Collaborative Filtering Recommendation Algorithms,” *Proceedings of the 10th International Conference on World Wide Web*, 2001, 285-295.

[8] R Toaster ウェブサイト、http://www.rtoaster.com/ (2014 年 11 月 10 日アクセス)

付録 A　ドライバの他のセグメントの決定木分析の結果

以下に、本論中で示さなかったドライバのセグメント 4 以外の決定木分析の結果をまとめる。

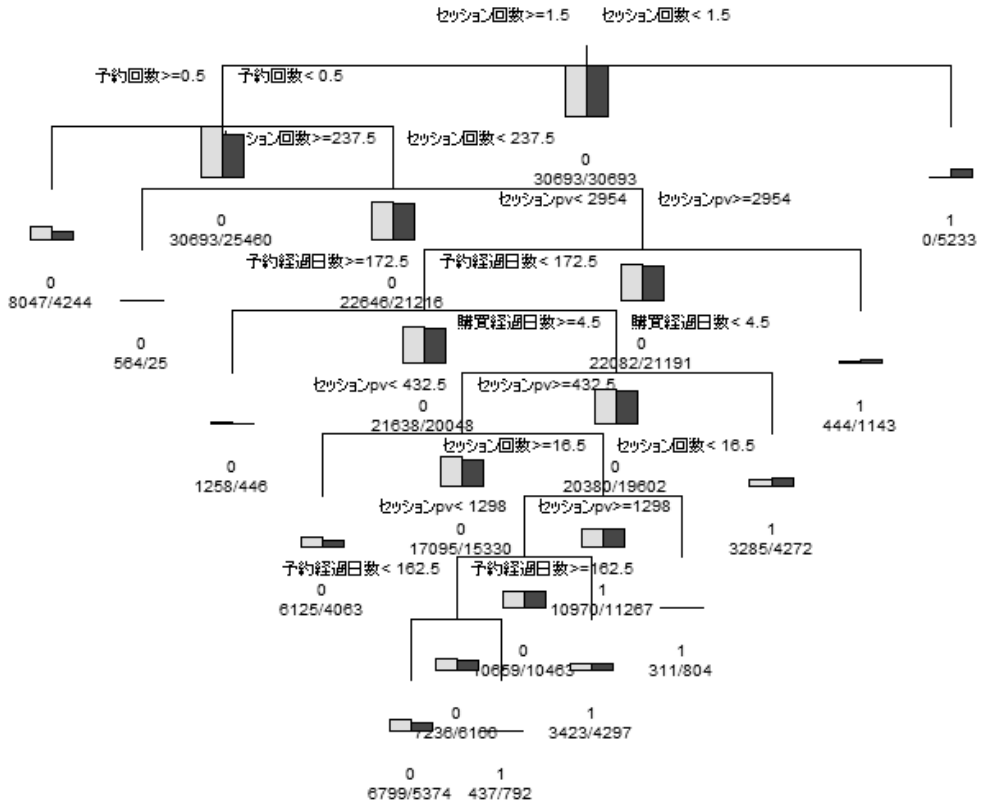

図 5　ドライバ　セグメント 1

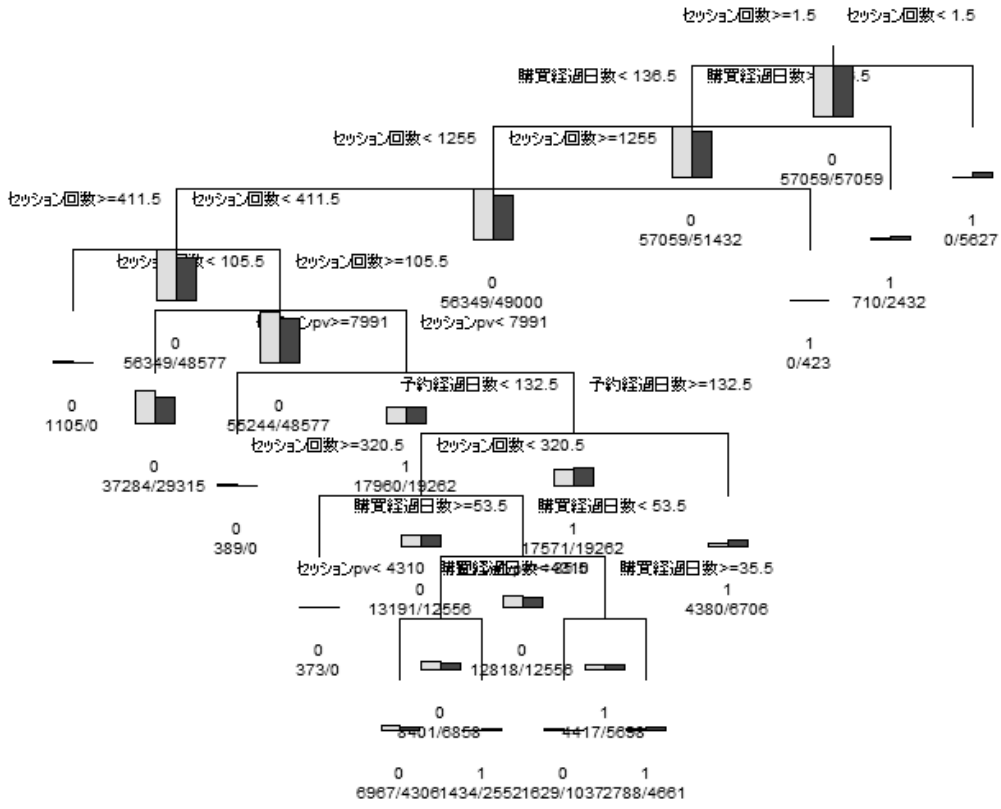

図 6　ドライバ　セグメント 2

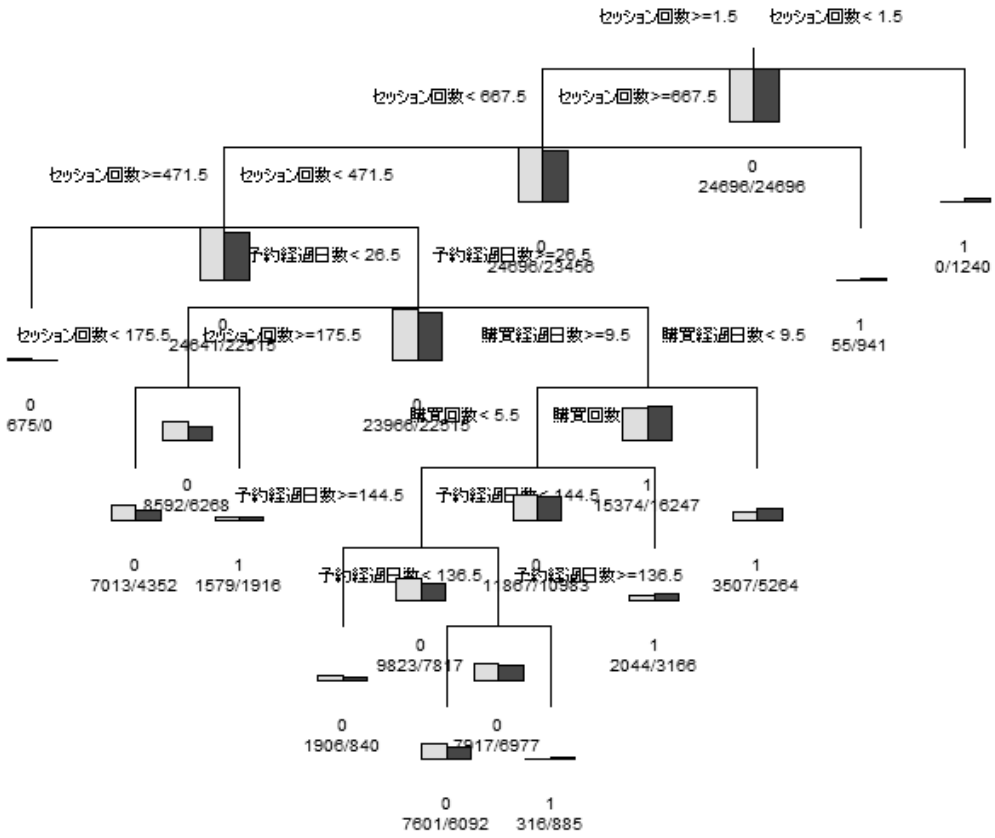

図 7　ドライバ　セグメント 3

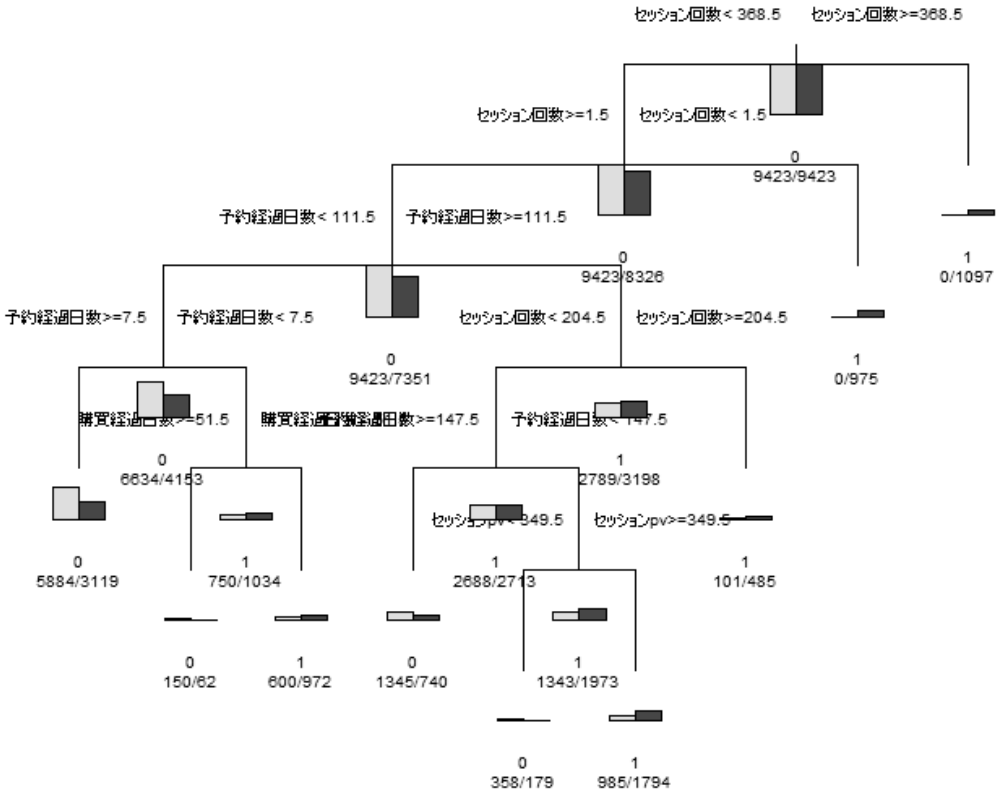

図8　ドライバ　セグメント5

付録 B　パターの決定木分析の結果

パター購買のセグメント特定のための決定木分析の結果は図 9 の通りである。図 10～14 は特定された各セグメントにおける予兆発見のための決定木分析の結果である。

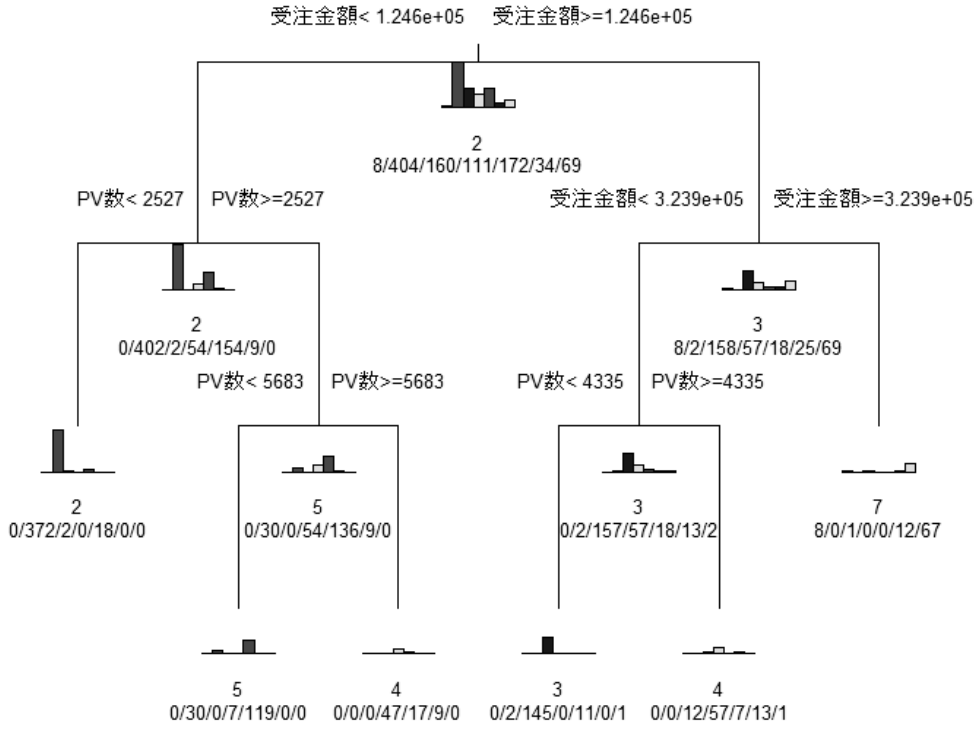

図 9　パターの顧客セグメント識別のための決定木分析

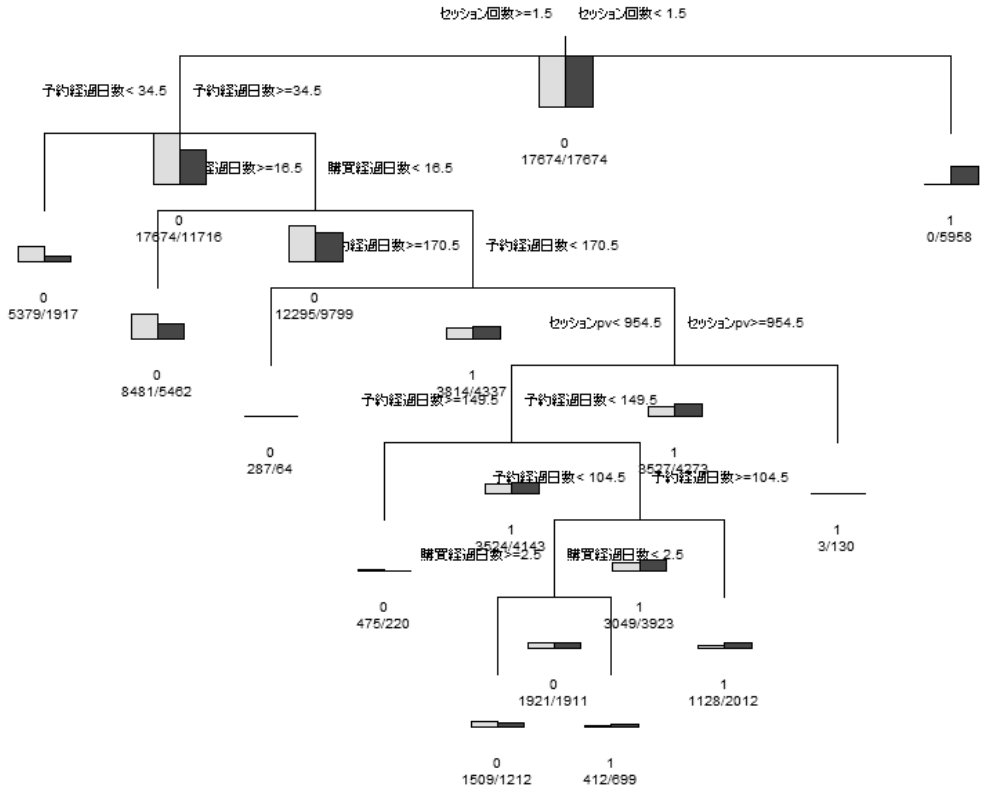

図 10　パター　セグメント 2

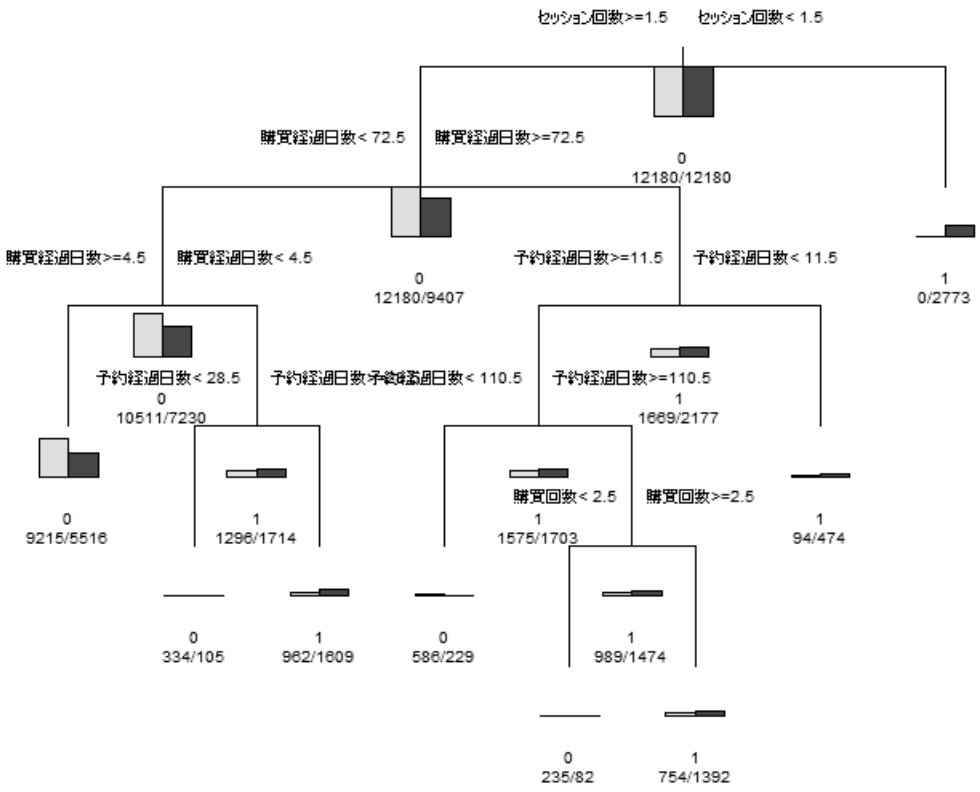

図 11　パター　セグメント 3

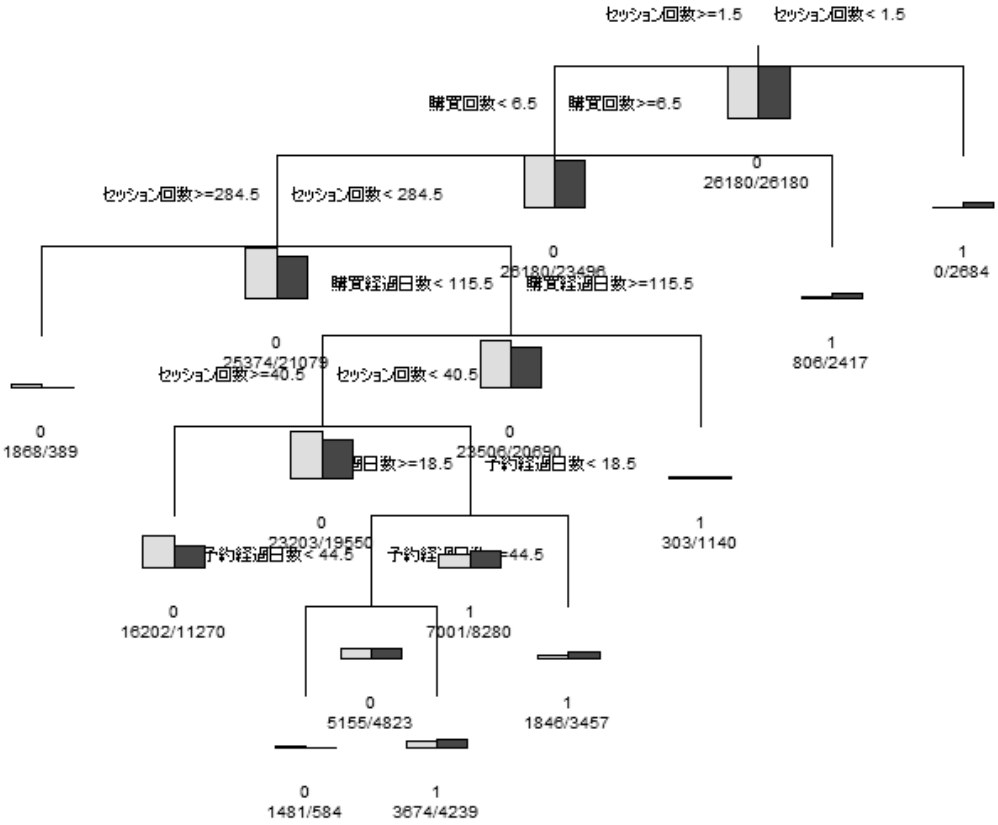

図 12　パター　セグメント 4

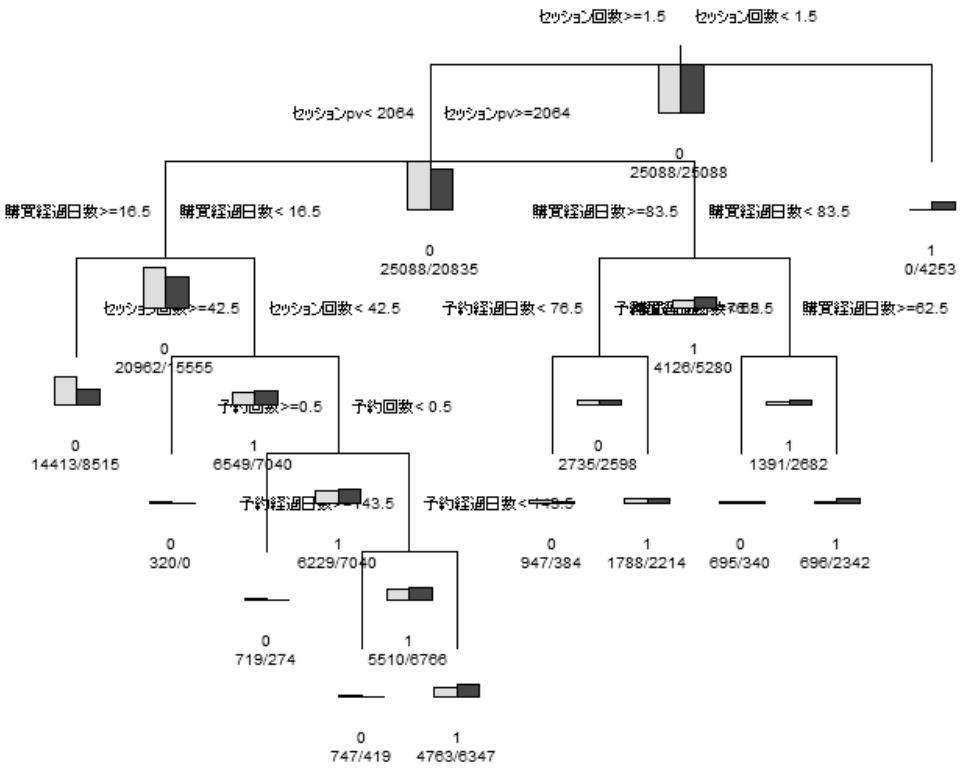

図 13　パター　セグメント 5

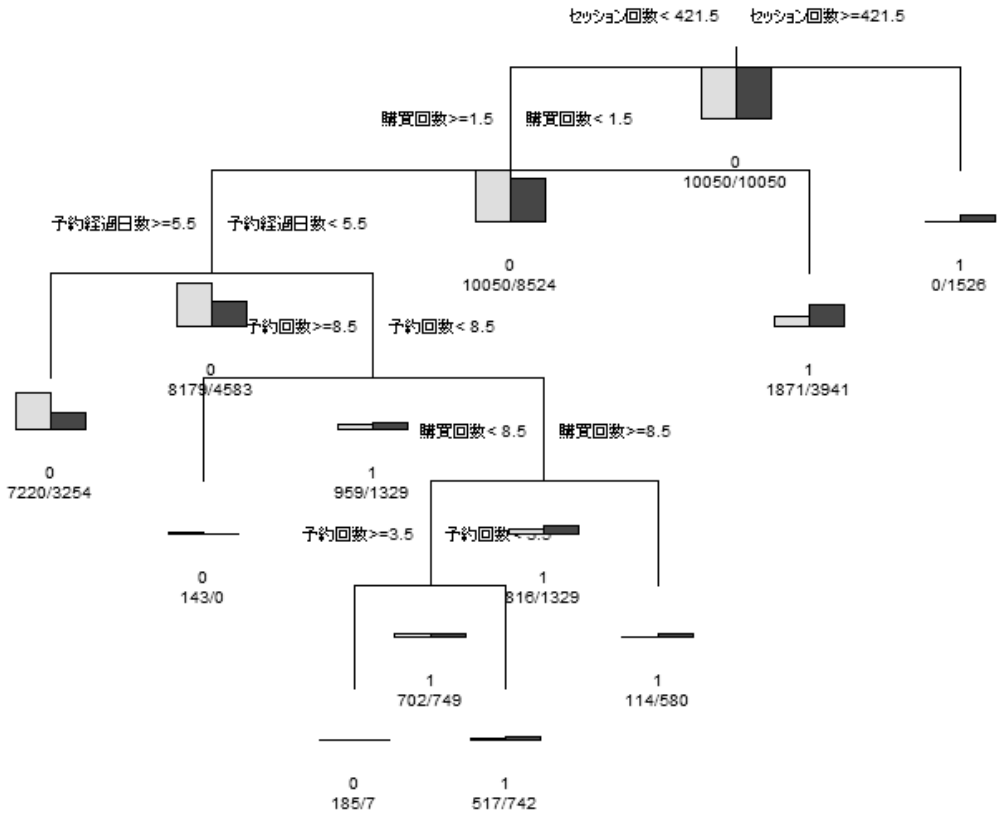

図 14　パター　セグメント 7

付録 C　ボールの決定木分析の結果

ボール購買のセグメント特定のための決定木分析の結果は図 15 の通りである。図 16～19 は特定された各セグメントにおける予兆発見のための決定木分析の結果である。

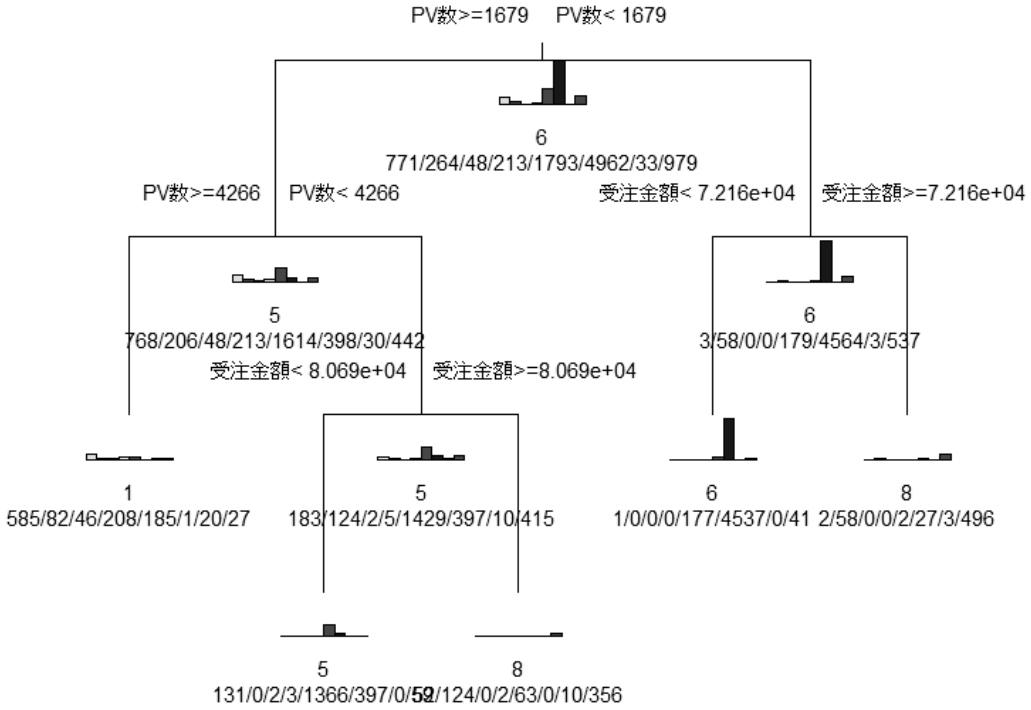

図 15　ボールの顧客セグメント識別のための決定木分析

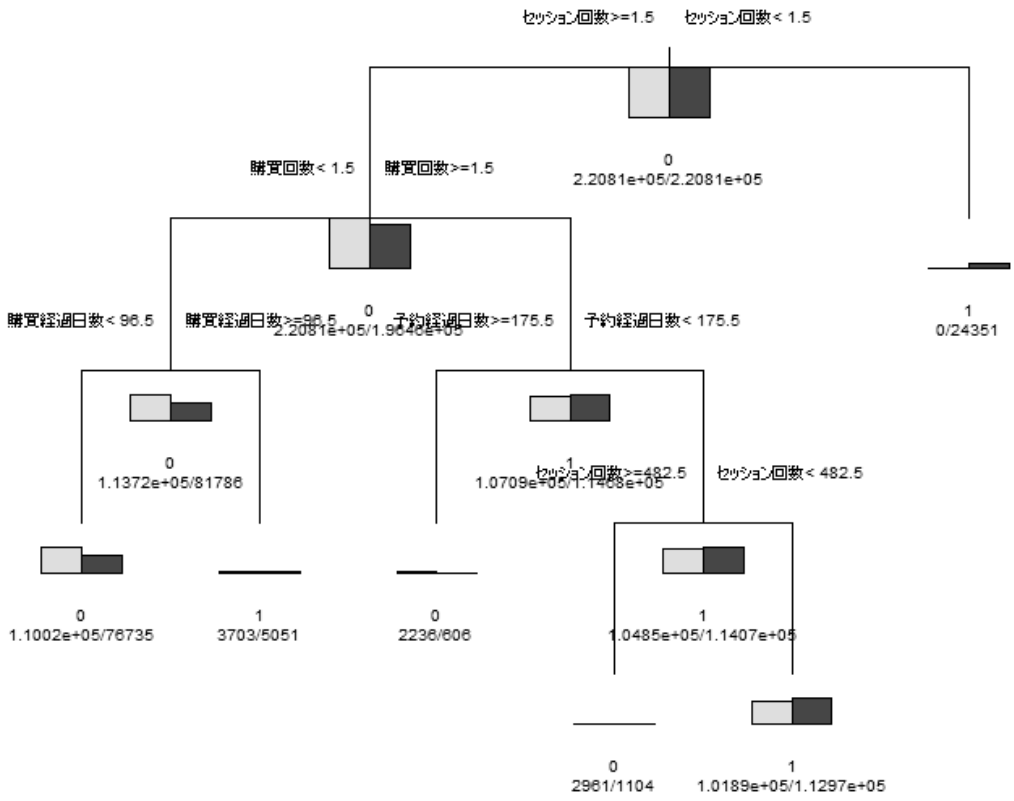

図 16　ボール　セグメント 1

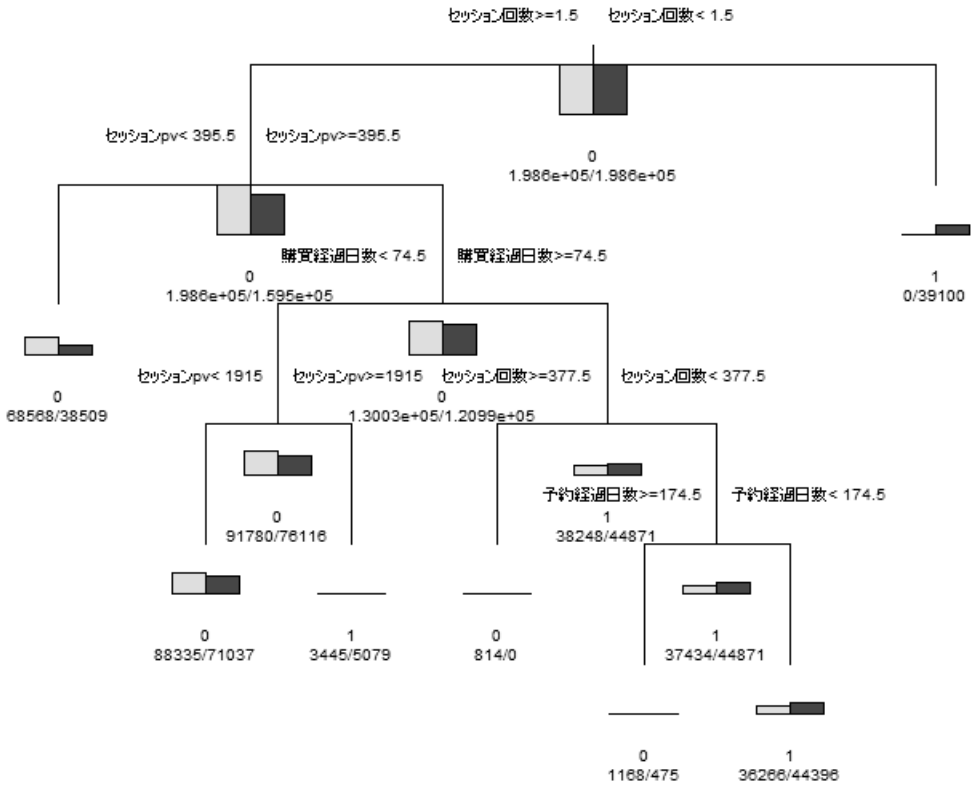

図 17　ボール　セグメント 5

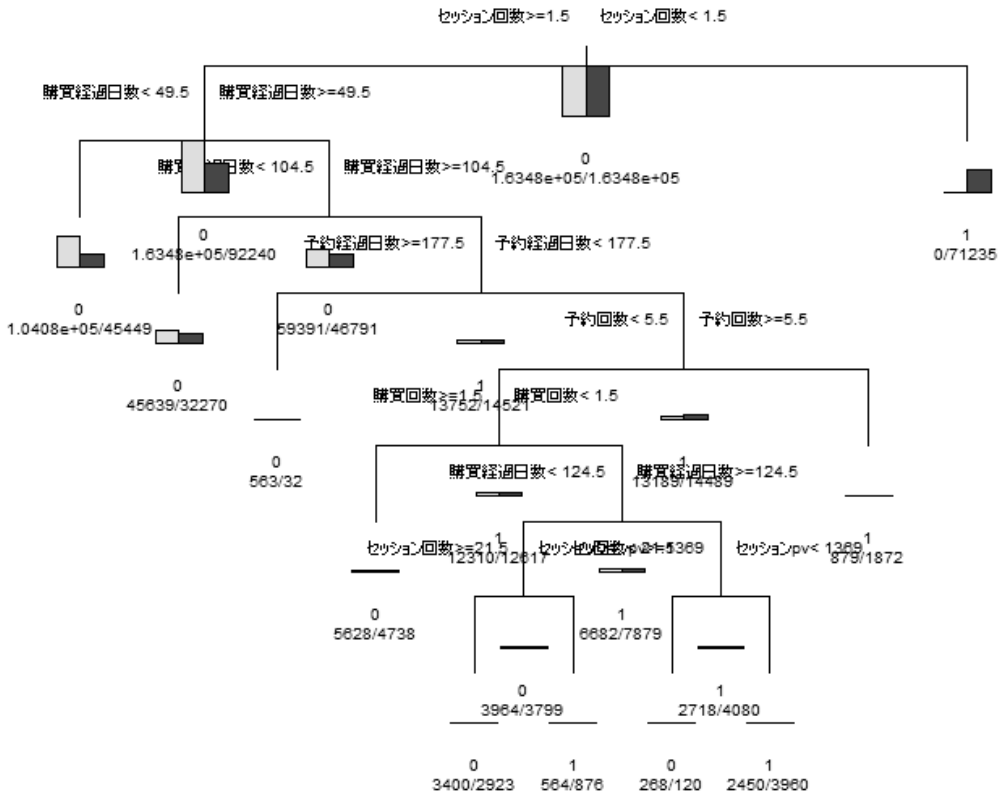

図 18　ボール　セグメント 6

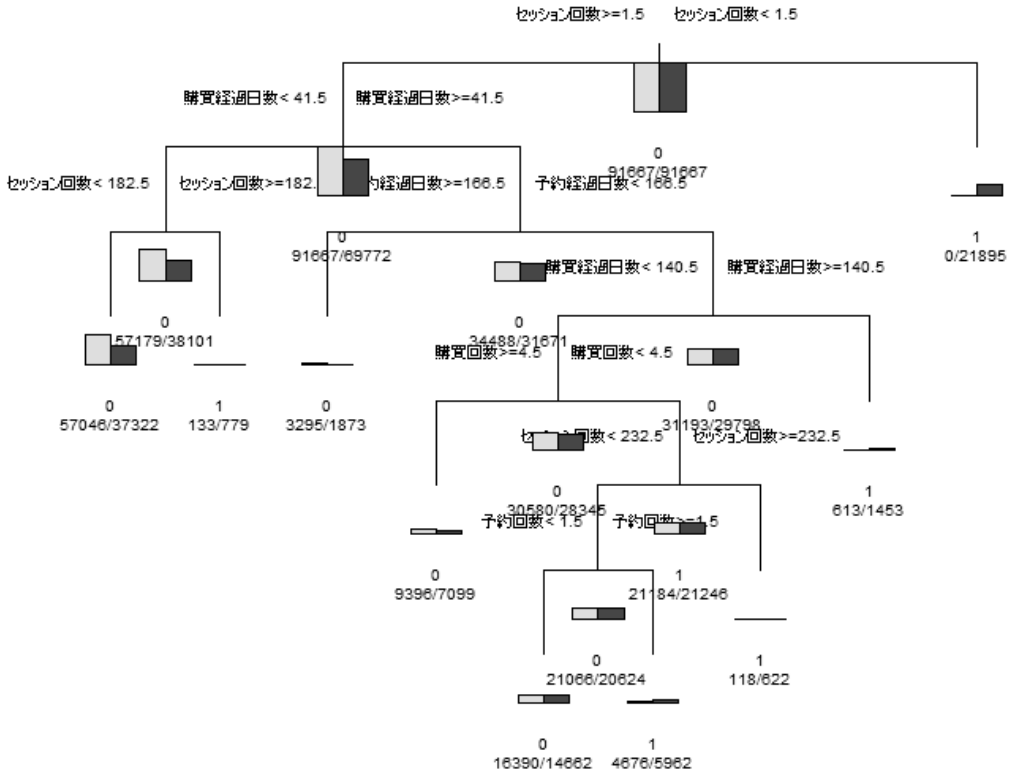

図 19　ボール　セグメント 8