2017 年度　博士学位請求論文

顧客の行動の多様性を用いた行動予測手法の提案
－多様性変数を活用した消費者の閲覧・購買の予測と精度の検証－

名古屋大学大学院経済学研究科

指導教員　根本　二郎 教授

氏　　名　新美　潤一郎

目次

表目次

図目次

第 1 章

はじめに

1.1 本研究の背景

1.1.1 電子商取引の興隆と消費者の多様化

現在では IT 技術の発達によりインターネット上の EC（電子商取引）サイトやスマートフォン上のモバイルアプリを利用したオンラインショッピングが活発化している。特に ICT 技術の向上（通信速度の向上によるインターネット上のコンテンツのリッチ化・スマートフォンをはじめとした高機能端末の普及等）に伴い EC はその市場規模を年々拡大させており，B2C EC（消費者向け EC）は 2022 年度には 26 兆円に達するという試算 [1] もある。2016 年の同市場規模はおよそ 15 兆円 [2] であり，今後も市場が継続的に成長していくことが示唆されている。

また先述のスマートフォンの普及により日常的なインターネット利用の習慣にも変化が起きている。実際に 10 代では週 1 回以上のインターネット閲覧時に使用する端末について，スマートフォンのみであると回答した消費者が 7 割（PC と併用している場合を含めると 9 割超）に達している [3] ことから，今後は商品購買にあたってもウェブサイトを閲覧せずにモバイルアプリのみで注文まで完結するなど購買行動にも変化が発生していくことが予想される。

こういった現状で，企業には各顧客のデモグラフィック情報としての属性（年齢，性別，職業，地域等）や環境（閲覧端末，決済方法等），サイコグラフィック情報（嗜好・消費意識等）など様々な条件に最適化したアプローチとしての One-to-One マーケティングが求められている。特に企業が自社で保有する顧客を適切に管理する顧客関係管理（CRM）の実施にあたって，顧客のロイヤルティ（自社への愛着心）の高さに応じてクーポン等を提供するロイヤルティマーケティングは広く実施されており，実際にポイントやマイレージといったプログラムによる発行額は 2020 年に 1 兆円相当を突破する [4] と見られており，消費者の企業に対するロイヤルティの高さの把握が CRM の観点から重要視されているといえる。

1.1.2 データ分析の需要増とモデルの複雑化

EC の活発化やスマートフォンの普及に伴い，消費者の生活に関する様々なデータが収集できるようになった。現在では社会的にビッグデータ活用の流れが加速しており，企業が蓄積する消費者のデータも日々莫大に増加しているといえる。しかしながら，一方でこのような莫大なデータ量の蓄積とともに，マーケティング分析の予測モデルの構築の際にも変数量の増加によるモデルの過剰な複雑化に繋がる恐れがある。これはモデリングの所要時間の増大や解析に必要な計算能力が必要になるなど，実務においても意思決定のコストを増大させてしまう。こういった点から，ビッグデータ時代のデータ分析と

して効率的な解析を実施するためにも予測モデルの構築にあたっても少ない変数で高い説明力が求められている。

また分析の需要増に伴いデータサイエンティスト（データ分析からビジネスにおける戦略立案までを一貫して行える人材）の育成が喫緊の課題となっており，文部科学省での「データ関連人材育成プログラム」の立ち上げやデータサイエンティスト協会の設立など，社会的にデータの分析を行える人材の需要が増している。

1.1.3　深層学習の発展と普及

データ分析の需要が高まる中で，近年では機械学習を用いた分析手法が AI（Artificial Intelligence, 人工知能）と呼ばれ大きく普及を進めており，中でも深層学習（ディープラーニング）をベースとした予測手法が次々に開発され実務においても積極的に活用されている。その多くは画像や音声データにおけるパターン認識への応用例が多いものの，中には社会科学系のデータとしての購買履歴データ等に応用することで消費者の購買行動を予測するようなものも存在している（第 2 章にて詳述）。特に深層学習はその独自の性質としての特徴抽出に着目されることが多い。この特徴抽出とは，与えられた説明変数を組み合わせることで，従属変数を最もよく説明できるよう自動的に説明変数と従属変数の間に潜在変数を生成する機能である。画像認識においては認識すべき画像の部分的なパーツが潜在変数として学習されることが一般的だが，深層学習自体がその構造として「ブラックボックス」と揶揄されるように，特徴選択がどの程度複雑な潜在変数を生成できるのかは特に社会科学系のデータにおいては未知数であり，通常の統計モデルと比較した網羅的な検証が必要である。

1.2　本研究の目的と構成

本研究ではこのような背景において，マーケティングの購買予測系の学術論文で過去に実施されてきた手法の限界について示した上で，消費者の行動を多面的に捉え，さらに行動予測の精度を向上できるような指標としての消費者の行動の多様性について提案するとともに，データ解析における深層学習の有用性とその限界について，統計モデリングと比較しながら議論する。

したがって本研究の構成としては，まず第 2 章で Web データを用いて顧客行動を予測しているマーケティング系の先行研究について俯瞰的なレビューを行うとともに，そこで用いられている主な手法における課題としての行動の幅広さが捉えられない点について述べる。そしてこの行動の幅広さを消費者の「行動の多様性」として指標化するため，多様性変数の定義と，これらが有効となりうる構造についての考察を行う。最後に多様性変数の有用性に関する予備的な解析として，近年着目されているマーケティング手法と比較した行動予測の精度について確認する。

次に第 3 章では，Web 行動に多様性指標を考慮することの有用性について網羅的に検証することを目的に，大手新聞社が運営するウェブのニュースサイトと大手 EC サイトについて競合関係にある各 2 社を取り上げて擬似的に自社と競合他社を設定，インターネット閲覧履歴データから収集した消費者の行動データを用いてウェブサイトへの訪問と購買について予測する。さらに同 EC サイトでの訪問と購買について同時に考慮した発展的なモデリングについても実施する。いずれの場合にも予測モデルへの多様性変数の投入の有無によるモデルの予測精度の変化を確認することで，多様性変数の有用性について議論する。

第 4 章では深層学習に関する網羅的な検証を目的に，一般的な統計モデルと深層学習の予測器を用いて消費者の訪問期間や購買回数について予測を行う。その際には第 3 章で用いられているような Web データを用いたものに加えて，小売企業から提供された実行動やモバイルアプリの使用履歴を組み合わ

せて解析を行うことで，消費者の購買方法の多様化に対応した多面的な解析例を提示する。そして第 3 章と同様に説明変数として多様性変数の投入の有無を変化させた複数の解析結果を比較することで統計モデルと深層学習での予測精度の差異について確認するとともに，深層学習内で多様性変数の有無を変化させることで特徴選択の精度の変化と多様性の有効性についての検証を行う。

最後に第 5 章では，ここまでの解析の結果から多様性変数の有用性やデータ解析における深層学習と統計モデリングの差異等についてまとめ，最後に今後の課題について示す。

注

[1] 野村総合研究所「2022 年度までの ICT・メディア市場の規模とトレンドを展望 ~AI や IoT を使いこなす「真の ICT 先進国」への道筋 ~」（2016 年 11 月 21 日発行）より。

[2] 経済産業省ニュースリリース「電子商取引に関する市場調査の結果を取りまとめました ~ 国内 BtoC-EC 市場が 15 兆円を突破。中国向け越境 EC 市場も 1 兆円を突破 ~」より。

[3]LINE 株式会社「〈調査報告〉インターネットの利用環境定点調査（2017 年上期）」より

[4] 野村総合研究所「ポイント・マイレージの年間発行額は 2022 年度に約 1 兆 1,000 億円に到達 ~ 国内 11 業界の年間最少発行額について、2014 年度の推計と 2022 年度までの予測を実施 ~」（2016 年 10 月 5 日発行）より。

第 2 章

先行研究と多様性の理論的背景の考察

2.1　はじめに

本章では，まず本研究に関連するマーケティング系の先行研究について俯瞰的なレビューを行う。関連する先行研究として，1) 一般的な顧客関係管理（CRM）の手法，2) ウェブサイトを対象とした消費者の行動予測を扱った先行研究，3) 近年活用されている機械学習を用いた先行研究についてのレビューを行うことで，本研究の位置付けについて確認する。

次に本研究で提案している消費者の行動の多様性の定義を行なう。ここでは多様性変数をマーケティング分析に用いる意義とその理論的背景についても併せて考察し，解析での有用性について議論する。そして近年マーケティング分野で注目されている手法である RFMC 分析についてその定義と現状の課題について指摘したのち，前節で考察した多様性変数が有用となりうる要因に従って RFMC を拡張することで，既存の RFMC での予測よりも高い精度で消費者の行動予測が可能である点についてシミュレーションを行なって確認する。最後に 1) RFMC，2) 筆者が拡張した RFMC，3) 本研究で提案している多様性変数の 3 つの変数群で作成した異なる 3 つのモデルで同一の従属変数について実データを用いた行動予測を実施し，多様性を用いた場合で最もモデルが改善することをもって実データにおける多様性変数の有用性を示す。

2.2　マーケティング系の先行研究

2.2.1　CRM の手法

Jacoby and Chestnut (1978) ではそもそも消費者が企業や企業の保有するブランドに対して抱きうる忠誠心としてのロイヤルティを行動的ロイヤルティ (behavioral loyalty) と態度的ロイヤルティ (attitudinal loyalty) に分類している。前者は再訪問や再購買といった自社の利益に直接的に貢献する実行動に基づくもの，後者は自社ブランドへの好印象など他の消費者へのレコメンデーションにも繋がりうる認知・態度とされ，消費者の行動や認知といった複数の面から自社ブランドへのロイヤルティを計測できる。これらのロイヤルティの計測にあたり，EC サイトでは会員登録や発送先のデータを用いることで同一顧客による継続的な購買を追跡しやすいという利点がある。一方近年では，ポイントプログラムの導入により実店舗でも ID-POS をはじめとして消費者を識別してその購買行動を継続的に捕捉できるよう取り組みを行なっている企業も多い。こういった消費者の購買行動の追跡が可能となることで消費者の行動的ロイヤルティを計測しやすくなったといえる。

このような各消費者についてのロイヤルティの把握にあたり広く一般的に用いられている代表的な指標には（１）顧客生涯価値，（２）RFM 分析，（３）シェア・オブ・ウォレットの 3 つが存在している。

（1）顧客生涯価値

企業における顧客のロイヤルティの把握には様々な手法が開発されているが，特に行動的ロイヤルティの把握のために実務においても積極的に用いられている代表的な手法として，顧客生涯価値（Customer Lifetime Value, CLV または LTV）がある (Blattberg et al., 2008)。ある顧客の顧客生涯価値は，年間の純利益 GC，顧客一人当たりの年間維持コスト M，将来期間の長さ n，年間の顧客保持率 r，割引率 d を用いて

$$CLV = GC \cdot \sum_{i=1}^{n} \frac{r^i}{(1+d)^i} - M \cdot \sum_{i=1}^{n} \frac{r^{i-1}}{(1+d)^{i-0.5}} \tag{2.1}$$

のように表すことができる。これは顧客が自社に将来 n 期間にわたってもたらすと予想される利益の総和から離脱率と維持コストを考慮した割引現在価値の算出であり，行動的ロイヤルティの推定の一手法であるといえる。また，このような CLV を用いて将来的な離反時期の予測 (Berger and Nasr, 1998) 等も行われる。

（2）RFM 分析・RFMC 分析

CRM の手法としては，CLV 以外にも RFM 分析が存在する。RFM 分析では，各顧客の過去一定期間における購買行動から集計した最終購買日（recency）・購買頻度（frequency）・支出額（monetary）を用いた直近期間の行動的ロイヤルティの高さの算出 (Gupta and Lehmann, 2006) 等が行われており，実務でも積極的に活用されている。

しかしながらこの RFM 情報を用いた顧客管理には，次に示すような問題が存在する。図 2.1 に示すような購買パターン I，II が存在した場合に，その購買発生のタイミングのばらつきの大きさからパターン I は不定期的にまとまった利用を行なっている一方でパターン II は明らかな定期利用であるなど，自社利用の動機が異なっている可能性がある。それでも RFM 分析の指標の上では期間中の購買回数や最終購買日が同一であることから，このように購買パターンに異質性が見られるような複数の顧客について，RFM の数値上では完全に同一の行動的ロイヤルティとみなされてしまう。これは消費者の購買行動の識別性が十分でないという点で問題である。

そこで RFM 分析において区別できないこうした購買パターンの識別性を高めた概念として，近年では RFM の各指標に加えて新たに C 指標を用いた RFMC 分析 (Zhang et al., 2014; Platzer and Reutterer, 2016) が提案されている。C 指標は Clumpiness の有無（購買や訪問といったイベントが期間全体のうちのある短い期間に集中して発生しているかどうか）を表す指標である。具体的には購買等のイベントを時系列に集計し，イベント発生がランダムに行われているかを検定することでその不均一性が特定の有意水準を超えた場合に Clumpiness の発生と判断する。具体的な算出方法については中山 (2016) が詳しい。

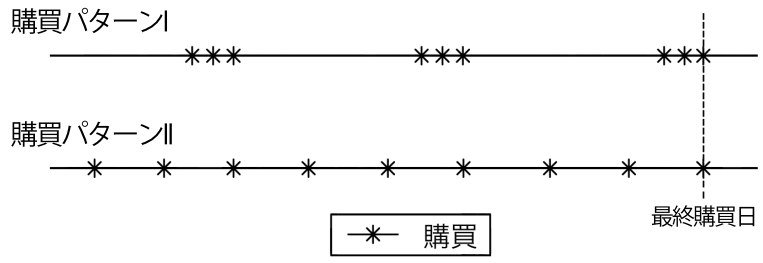

図 2.1　RFM 情報で区別できない購買パターンの例

Clumpiness は購買の不均一性の指標であり，自社購買における Clumpiness の発生は必然的にその前後に自社を利用していない空白期間が存在していることになる。これは企業が自社での購買データのみを考慮した場合には，顧客が単に当該カテゴリへの支出を行なっていないだけだと判断しかねないが，この空白期間に実際には競合他社を利用しており，データが取れていないだけで顧客は実際には継続的に利用を行なっているという可能性もある。すると Clumpiness の有無を競合購買の予測に考慮することで，消費者が競合他社にブランドスイッチしている可能性を適切に発見できることが考えられる。しかしながらこの RFMC は提案されて間もないことから活用例自体が希少なことから先行研究ではこのような点については言及されておらず，あくまでも顧客管理の指標としての購買パターンの異質性の考慮を目的に提案されている。先行研究としては近年台頭する複数の有料動画ストリーミングサービスの利用履歴のモデリングを行なった Zhang (2013) がある。

（3）シェア・オブ・ウォレット

一方近年では競合他社での購買を考慮した指標としてのシェア・オブ・ウォレット (SOW) の活用も盛んである。SOW は消費者ごとの特定の市場や商品カテゴリへの支出額全体に占める自社への支出額の割合として計算される。その定義から，算出することで各顧客について自社への行動的ロイヤルティが計算できる (Jones et al., 1995) ことに加えて，競合他社に支出している金額から潜在的な収益性 (potential profitability) も同時に把握できる (Glady and Croux, 2009) という点で有用とされる。Chen and Steckel (2012) では SOW を得ることにより顧客ごとの追加的な購買余地を把握することで，非ロイヤル顧客のみに対してクロスセリングやアップセリングを用いた効率的なプロモーションを実施できるとしている。

特定の市場や商品カテゴリにおける特定企業の SOW の推定は様々な市場を対象に実施されており，自社で得られない競合他社での購買情報は欠損データ予測の文脈として統計的データ融合の枠組みで議論されることが多い。統計的データ融合では調査対象者が異なるために通常では紐づけることの難しい複数のデータ（企業の購買履歴データと外部の調査データ等）を，2 つのデータに共通する変数としての共変量（年齢層や性別といった属性データであることが一般的）をキーとして統計的手法を用いて紐づけ，関係を明らかにする分野である。代表的なものとして因子モデルを欠損変数の予測に応用することでデータ融合を実施した Kamakura and Wedel (1997) があるが，具体的に特定業種でのデータを用いて SOW を予測したものとしては，銀行への支出に関する調査データを用いて複数銀行間での金融商品の SOW の推定を行った Du et al. (2007)，個人ごとの支出パターンの異質性や企業間・カテゴリ間での支出の関係の同時性を考慮しながら銀行内でのカテゴリごとの支出と銀行ごとの支出の同時モデリングを行うことで SOW の推定を行った Jang et al. (2016)，調査データ等の追加的な情報を用いずに自社への支出情報としての RFM とデモグラフィック情報のみで銀行の利用額としての預金残高の SOW を算出した Glady and Croux (2009)，複数のクレジットカード間での自社カードへの支出額に関する SOW を算出した先述の Chen and Steckel (2012)，小売企業を対象としたものとして Mägi (2003) などが存在する。また機械学習や深層学習を用いた先行研究として，欠損値の予測に用いることが可能なベイジアンネットワークで購買履歴の ID-POS データと調査パネルのデータ融合を実施した石垣他 (2011) や，深層学習の代表的な手法の一つである Deep Boltzmann Machine(Salakhutdinov and Hinton, 2009) を因子モデルの拡張として非線形の深層モデルに発展させてデータ融合に応用した新美・星野 (2017) などが存在する。

これらデータ融合系の先行研究を俯瞰した示唆は次のような 1) 消費者について得られている自社での支出額は競合での支出額とほとんど相関を持たないこと，2) ごく一部の消費者が競合利用の大部分を占めていること，の 2 点にまとめることができる。すなわち自社での購買額が小さいからといって競合

を積極的に利用しているわけではなく，そもそもカテゴリへの支出額自体が小さい消費者である場合も多いことから，SOW の低い顧客を適切に発見することが自社のシェアを改善するための有効な手立てとなる。

2.2.2　インターネット閲覧履歴データの活用

次に Web を対象とした消費者の行動予測を扱った文献についてまとめる。そもそも Srivastava et al. (2000) ではウェブ上で収集できるデータを i) サーバで収集できる Clickstream Data などの web usage data，ii) 年齢性別といったデモグラフィック情報や嗜好などの user profile data，iii) EC サイトにおける商品情報など実際のテキストや画像データなどの contents data，iv) HTML 等を用いて表されるウェブサイトの構造を扱った structure data に分類している。中でも web usage data として得られる代表的なデータが，ウェブサイトの閲覧履歴を時系列で収録した Clickstream Data である。ウェブ系の論文について網羅的にレビューした Bucklin et al. (2002) によれば，Clickstream Data はユーザの閲覧したウェブページについてその URL・アクセス日時・滞在秒数・リファラ情報等の情報を逐次サーバで収集したものであり，各ユーザについてウェブサイト横断的に閲覧情報を収集した User-Centric Clickstream Data と，特定のウェブサイトについてそこにアクセスした全ユーザのサイト内での閲覧を収集した Site-Centric Clickstream Data に分類でき，本研究のようにオンラインストアでの購買をはじめとしたインターネット上の消費者の行動予測を扱ったマーケティング研究の多くで活用されている。企業が保有する EC サイト等のサーバから自社データとして得られるのは Site-Centric なデータのみであり，自社サイトにアクセスした顧客について，自社サイト外でのウェブ閲覧に関する情報を得ることはできない。そこでインターネットの利用に関する人口統計に基づいて収集された User-Centric なデータを特定のウェブサイトへのアクセスに限定することにより，企業が得られうる擬似的な Site-Centric Clickstream Data とすることで様々な研究に応用されている。

本研究と同様に Clickstream Data をマーケティング系に応用する研究としては，主に 1) 消費者の購買などウェブ上でのコンバージョン [5]，2) 特定ウェブサイト内での回遊行動，3) ウェブサイト内もしくは検索ポータルでの検索行動，4) オンライン広告のコンバージョンや広告効果の推定，の 4 分野でのモデリングが一般的である。まず 1) のコンバージョン予測の代表的なものとして，過去の購買行動からウェブサイトへの訪問パターンや購買パターンを発見することで将来的な購買を予測した Park and Park (2016) や，自社 EC への次期アクセスの際の購買有無の予測にあたって Clickstream Data や購買履歴データなどを組み合わせてロジスティック回帰分析で予測した Van den Poel and Buckinx (2005) など，2) については自社ウェブサイト内でのページ遷移の経路の考慮が購買有無の予測においても有用であることを示した Montgomery et al. (2004)，消費者の EC サイトへの購買にあたって各ユーザのアクセスの目的を購買や情報収集など 4 つに分類して購買行動の予測に考慮した Moe and Fader (2004) などが挙げられる。3) については Clickstream Data を消費者の情報探索としての EC サイト内での検索・閲覧行動の把握に応用した Moe (2003) や米 Amazon でのカメラの検索行動から購買をモデル化した Kim et al. (2010) のように単一 EC サイト内での検索行動を扱ったものが存在する一方で，消費者の検索ポータルでの検索行動を depth（検索した結果として訪問するサイト数）・dynamics（depth の時系列での変化）・activity（結果として購買する商品カテゴリ数）として指標化し，検索後にアクセスする EC サイトでの購買行動との関係を明らかにした Johnson et al. (2004) のように，検索ポータルを対象として複数の EC サイトへの流入や購買を扱っているものも存在する。4) のオンライン広告を扱ったものとして，ウェブ広告への接触から購買に至るまでの期間の推定を行なった Manchanda et al. (2006) や，複数のオンライン広告出稿先での消費者の異なる広告接触傾向と広告の形式によるクリック率の変化の関係を明らかに

した Nottorf (2014) 等が存在する。

本研究では特に web usage data としての Clickstream Data から得られる消費者のウェブ閲覧の特徴をアクセス・パターン情報 [6] と呼ぶ。web usage data からのアクセス・パターン情報の抽出と顧客に関する知見の創出は Büchner and Mulvenna (1998) をはじめとして古くから行われており，従来の実店舗では得ることの難しかった消費者の店舗内での動線や，その際にどういった商品が閲覧・検討されたのかなど，購買までの過程の行動を詳細に把握することが可能な点で実務・アカデミック共に広く活用されている。しかしながら先行研究におけるアクセス・パターン情報の活用には次に指摘する 2 点の特徴として，1) アクセス・パターン情報の抽出にあたっては多くの場合に量的な大きさ・平均値・最頻値が主に利用される点， 2) 企業にとって貴重な自社外での消費者の行動が得られるリファラ情報の活用が限定的である点，が挙げられる。

まず 1) については先の Srivastava et al. (2000) でも言及されているように，web usage data からマーケティングに有用なパターンを発見するには，アクセスされたページ数や滞在時間等の集計に対しての記述統計量の算出，最もよく閲覧されたページなどのカテゴリカルな情報としての最頻値変数 [7] の計測などが行われる。実際に多くのマーケティング系の先行研究において，競合購買や自社の将来購買の予測の際には自社購買の RFM 情報や，ウェブサイトの閲覧にあたって用いられたリファラの数や購買された商品カテゴリの多さなどの量的な情報が用いられており，そこでの特定の要素への依存度といった情報は考慮されていない。

次に 2) については，そもそも企業の戦略策定にあたって自社で保有する顧客の自社外での行動を得るためには，顧客へのアンケートの実施や外部企業からの調査データの購入等を行うことが必要であり，追加的な情報収集コストがかかることが一般的である。さらにそういったデータを外部から入手した場合でも，企業が自社で保有する顧客データとは収録されている対象者が異なることや，仮に一部が重複していてもユーザ ID 等が異なることからそのまま紐づけて解析を行うことが困難であるなどの問題が存在する。そういった状況の中で，ウェブサービスを運営する企業において自社顧客のサイト外での行動を得られる数少ない情報源がリファラである。Site-Centric Clickstream Data しか得られない企業においても自社サイトに流入する際のリファラは収集可能であり，流入経路として各顧客がどういったウェブサイトを閲覧しているのかを部分的にでも把握可能である。実際にこのリファラ情報を活用した先行研究としては，サイト内での閲覧ページ数の推定にあたって直前に閲覧したページ情報を考慮するためにリファラを活用した里村 (2007) や，Web サイト閲覧における各検索ポータルや自社関連サイトからの流入の識別のためにリファラを活用した勝又 (2010)，ウェブサイトの閲覧の有無と閲覧時間の同時分析にあたり流入経路の効果を推定するためにリファラを活用した猪狩・星野 (2014) 等，いずれも特定の流入経路を用いた場合に閲覧・購買行動に及ぼす効果の大きさの推定に用いるにとどまっている。しかしながらリファラ情報は企業において数少ない顧客の自社外の行動データを得られる貴重なソースである以上，この情報の活用により行動予測の精度を高めることは可能であると考えられる。

2.2.3 機械学習・深層学習のマーケティングへの応用とその課題

こういったマーケティング系の行動予測には，統計モデルを用いたものに加え近年では機械学習を用いたものも増加している。中でも深層学習は特徴選択による潜在変数の自動生成と非線形の活性化関数を複数重ねた深層構造が特徴的であり，分布を仮定しないノンパラメトリックな推定器として注目を集めている (岡谷, 2015)。代表的な手法としては Feed-Forward Neural Network（FFNN）と呼ばれる最も基礎的な深層ニューラルネットワークが存在している他，比較的初期に確立されたのが Convolutional Neural Network(LeCun et al., 1989) である。現在でも特に画像認識の分野においてこの CNN は高い性

能を示しており，先述の伊藤他 (2016) を始めとして多くのビジネス事例でも実際に用いられている。

このような分析モデルはマーケティング分野においても顧客行動の予測に応用され始めている。例として Vieira (2015) は web usage data としての Clickstream Data から消費者の EC サイトでの行動データを取り出し，あるセッションでの購買有無の予測とその際に購買される商品の予測をロジスティック回帰分析，機械学習の一手法である Random Forest(Breiman, 2001; Liaw and Wiener, 2002)，深層学習における離散選択の代表的なモデルである Deep Belief Network(Bengio et al., 2007) を用いて予測しその精度を比較している。Vieira (2015) をはじめとした情報学・工学分野においては，EC サイト内の商品の詳細な文章などの web contents data の情報を解析に加味することを目的として Word2Vec(Mikolov et al., 2013) が盛んに活用されている。Word2Vec とは文章のテキストコーパスを投入することで単語の関係を特徴量ベクトルに変換できるアルゴリズムであり，web contents data に適用することで商品情報データをベクトルに変換し，特徴量として投入することが可能である。

また，機械学習モデルを用いて企業のトランザクションデータ等を学習させ，特定の状況下においての顧客行動を予測した既存研究には，コンビニエンスストア（CVS）の POS（Point-of-Sales）データを用いて顧客エージェントを発生させ，欠品の状況下での購買シミュレーションを行った松村他 (2016) や，ID-POS のトランザクションデータと顧客の生活調査データから商品カテゴリと顧客のライフスタイルに関する潜在クラス変数を生成し，ベイジアンネットワークを用いて関係をモデル化することにより，季節や顧客のライフスタイル，時間帯等の様々な条件下でどのような潜在クラスの商品カテゴリが購買されるのかを明らかにした石垣他 (2011) がある。

また CRM における顧客の保持／離脱を予測した研究でも，同様に機械学習の手法が積極的に活用されている。例として Support Vector Machine (Boser et al., 1992) を用いて予測した研究に Zhao et al. (2005) が，またニューラルネットワークを用いて予測した研究には Sharma et al. (2013) などがある。

Deep Learning に限らない機械学習全般のマーケティングへの活用事例では，複数の消費者や商品同士の類似性を計算することで，購買される確率の高い商品を表示し顧客に推薦する協調フィルタリング (Goldberg et al., 1992) を用いた EC サイトでのレコメンデーション等の研究が行われており，既に Web 上で広く活用されている。このような協調フィルタリングを用いたレコメンデーションに関する既存研究やアルゴリズムについては，Park et al. (2012) において詳細にレビューされている。

このようにマーケティング分野においても少しずつ深層学習が活用され始めているが，特に実務におけるマーケティングデータの解析では顧客データが完全に得られることがほぼないという点で画像認識等とはそもそも異なる点が多い。深層学習では完全データを使用することが前提となっていたが，近年ではマーケティングデータのような欠測データも扱える手法として，Lopes and Ribeiro (2012) により一般的なニューラルネットワークを欠損値を扱えるよう拡張した Neural Selective Input Model (NSIM) が提案されるなど，徐々にマーケティングモデルに適した手法も登場している。

こういった深層学習では推定器の柔軟さから比較的予測力が高まる傾向にある。しかし一方で非線形の活性化関数と何層にもわたる潜在変数を持った複雑な構造であることから，特に本研究のような社会科学データへの応用においては実際の変数間の関係や構造を捉えているとは考えにくい。そのため一般の統計モデルが説明変数と目的変数の構造的な関係を把握することを目的に使用される一方で，深層学習は予測精度が高まることを目的に活用されることが一般的である。

表 2.1 先行研究で区別できないリファラの多様性の例

顧客	流入総数	リファラ元	流入回数	依存度	リファラ数
顧客 A	300	Google	300	高	1
顧客 B	300	Google	100	平等	3
		Facebook	100	平等	
		Twitter	100	平等	
顧客 C	300	Google	280	高	3
		Facebook	10	低	
		Twitter	10	低	

2.3 行動の多様性の定義と理論的背景

2.3.1 多様性変数の目的

ここで本研究で提案している多様性変数について述べる。まず前節で網羅的に調査したように，マーケティングとして消費者の行動予測を扱っている文献には様々な行動を対象としたものが存在するが，本研究で提案する多様性変数の目的は，主に先行研究の手法では識別できない消費者の行動の差異を捉えることにある。例として先行研究において活用が限定的であったリファラを取り上げて表 2.1 に架空のウェブサイトへの消費者ごとの流入回数とその経路についてまとめる。先行研究の特徴として挙げた量的変数や最頻値変数のみを用いて行動パターンを識別する場合には，顧客 A と B，C の差異についてはリファラ数のカウントで捉えることが可能であるが，一方でリファラ数が同一でありながら各リファラから均一な頻度で流入している顧客 B と Google からの流入に強く依存している顧客 C の行動の差異は識別することはできない。今回は例として web usage data における代表的な情報としてのリファラを取り上げたが，実際には購買された商品カテゴリやアクセスされた時間帯など，様々な情報についてその依存度の指標化を行うことで，消費者の購買行動を先行研究と比較して詳細に把握することが可能である。したがって本研究では，これら消費者の行動における特定の要素への依存度を「行動の多様性」として指標化し行動予測に考慮することを考える。

2.3.2 多様性変数の定義

ここで本研究で提案する多様性変数の指標化の方法について述べる。例としてリファラの多様性でいえば，リファラの要素全体に占める特定要素への依存の度合いを数値化することが必要となる。 そこで本研究では，依存度についてハフィンダール・ハーシュマン指数（HHI）を，要素数の多さについて累積パーセントを用いた数値化を試みている。

偏りの大きさの指標化

HHI は市場の寡占度を測定することを目的に用いられる経済指標であり，ある市場に参入している全企業について各企業の市場シェアの二乗和として算出される (Davis and Garcés, 2009)。従って特定の一社が完全に市場を独占している場合に最大値 10000 を取り，多くの企業が参入し尚且つ各企業のシェアの格差が小さい場合に最小値を取る。 この指標を顧客ごとの各リファラからの流入の割合に適用す

ることで，顧客の特定のリファラへの依存の大きさを数値化することが可能である。例として，ある顧客 j について顧客ごとの自社サイトへの流入時のリファラ i（但し $i = 1, 2, \cdots, m_j$）のシェア s_{ij} から，顧客 j のリファラの偏り $RefHHI_j$ は

$$RefHHI_j = \sum_{i=1}^{m_j} s_{ij}^2 \tag{2.2}$$

と表すことができる。但し本研究では HHI を 10000 で除し，0 から 1 までの値をとる百分率としている。この定義から，ある顧客について自社への流入経路が特定のリファラに依存しているほど HHI は上昇する。これは曜日や時間帯などでも同様の傾向を表すことから，自社顧客がウェブサイトにアクセスする際の様々な傾向について HHI を用いることで把握できる。

相対的な要素数の多さの指標化

次に，他の顧客と比較した相対的な要素数の多さについても指標化が必要である。特にリファラについては極端に多くのリファラを経由している顧客も存在していることから，そのような異常値の影響を除き正規分布に近づけることを目的としている。実際の指標化には累積パーセントを用いており，これはリファラを例にとると，ユーザ別に調査期間全体にわたった自社サイト流入時のユニークなリファラの数を計測し，これを値の小さいユーザから昇順に並べた上で，そのユーザが下位何パーセントに位置しているかを示したものである。従って最もリファラの数の少なかった顧客が 0% を，最も多かった顧客が 100% の値を取る。実際にモデルに説明変数として投入する際には，正規累積関数の逆関数を用いて算出している。

ジニ係数との比較検討

本研究において HHI を多様性の依存度の測定に使用したのは，HHI が市場の独占度合いの評価に使用されているという点で，それぞれの要素の格差を測定することに長けていると考えたためである。しかしながら格差を評価するような指標としては，HHI 以外にジニ係数という指標も広く使われている。ジニ係数とは，国内の所得分配の不平等度を計測することでその格差の程度を評価するために使用される指標であり，ローレンツ曲線を用いて算出することができる (Weymark, 1981)。リファラを例にジニ係数による不均一性を測定すると，ある顧客 j の m_j 個のリファラそれぞれからの自社サイトへの流入回数 t_{ij} から，

$$RefGini_j = \frac{\sum_{k=1}^{m_j} \sum_{l=1}^{m_j} |t_{kj} - t_{lj}|}{2\bar{t}_j m_j^2} \tag{2.3}$$

と表すことができる。HHI とジニ係数では表 2.2 に示すような 2 つの場合において表現に差異が現れる。本研究では消費者の行動の差異について詳細に指標化することを目的に多様性変数を提案しているため，行動の差異が具体的に数値差に現れる HHI を採用している。

2.3.3　多様性変数を用いた行動予測の有効性の考察

このように消費者の行動の依存度について指標化することは消費者の行動の差異を捉えるという点で競合購買や自社での将来の行動を予測するにあたって有用であると考えられる。しかしながら前節において網羅的に調査したように，消費者の行動予測を扱ったマーケティングの先行研究には本論文で提案するような消費者の行動の多様性と行動予測の関係に着目した指標を提案しているような事例は存在していない。したがって多様性変数の有用性に関する構造は明らかになっていないことから，本節では一

表 2.2　リファラの多様性における HHI とジニ係数での表現の差

顧客 A		
リファラ	流入回数	シェア（%）
Google	100	20%
Yahoo! Japan	100	20%
Facebook	100	20%
twitter	100	20%
アメーバブログ	100	20%
	ジニ係数:	0
	HHI:	2000

顧客 B		
リファラ	流入回数	シェア（%）
Google	100	33%
Yahoo! Japan	100	33%
Facebook	100	33%
	ジニ係数:	0
	HHI:	3267

般消費者の Web 利用について調査している文献から多様性変数が有用となる理論的背景について考察する。

まず考えられる理由として，多様性変数を用いることで 1）ウェブ利用における消費者の異質性を部分的に把握できるため，2）ウェブ利用の経験や習熟度を反映しているため，の 2 点が挙げられる。1）に関しては，Dembczynski et al. (2008) においてそもそも Web 上での消費者の閲覧行動には大きな個人差が存在する可能性が指摘されている。星野 (2009) では，マーケティングの実証研究で用いられている web usage data や購買履歴データはいずれも結果のみを記録した行動データであり，その行動がどういった嗜好や属性に起因して発生しているかは不明であることから，Clickstream Data に関しても個人の異質性が発生しているとしている。新美・星野 (2015) では多様性変数を用いない従来的な変数では Clickstream Data 上に捉えられない Web 閲覧パターンの差異が存在することについて指摘しているが，アクセス・パターン情報の多様性を用いることでこういった個人の Web 利用の異質性を部分的に捉えていることが予測精度の改善に繋がっていると考えられる。

2）については Johnson et al. (2004) において，書籍や CD といったコモディティ類似型商品（commodity-like products）では Web 利用の習熟度が高く EC サイトの利用経験が多い消費者ほど購買に利用するウェブサイト数が減少し，特定サイトに集中することが示されている。Web 利用の習熟度が高まることで習慣的に特定のサイトで購買を済ませやすいという傾向から，よく訪れるウェブサイトはブラウザのお気に入りに登録して簡単にアクセスできるようにするといった行動に起因して Web 行動の多様性に変化（特にリファラの多様性の減少や特定時間帯への集中等）が発生していると考えられる。Web の習熟度が効果を持つ理由として，商品購買の意思決定のために必要な情報の探索コストが異なってくる点がある。というのも実店舗で競合する複数の店舗に赴いて商品を比較検討する必要があるのと同様に，オンラインショッピングにおいても複数店舗を Web 上で閲覧し商品の品質や価格について限られた情報と選択肢の中で最適な意思決定を行わなければならない。そこで Web への習熟度の低い消費者では意思決定のための情報探索コストが高まることから様々なウェブサイトを経由して自社サイトに訪問することが考えられる。これは経済学における今井他 (2007) 等から考えても自然な構造である。

2.3.4　実行動の多様性が有用な背景

次に実行動における多様性の有効性についてその理論的背景を考察する。理由としては 1）消費者の異質性の一部を捉えられるため，2）デモグラフィック情報の一部を捉えられるための 2 点が考えられる。1）については，特に実店舗での購買においては行動の結果としての購買履歴データしか得られないことが一般的であり，その購買プロセスとしての比較検討の過程を得ることは困難であった。しかしながら本研究の第 4 章で用いている小売ブランドでは，仮に実店舗で購買を行わずとも訪問した際にモバイルアプリから「チェックイン」を行うことでポイントが得られるインセンティブを与えたり，店内に置かれた商品についてモバイルアプリから詳細な情報を得られたりといった形で，実購買を伴わない訪問についても情報を得られる仕組みが整備されている。こういった購買プロセス中の消費者ごとに異なる行動パターンを多様性で捉えることで，消費者の異質性を部分的に捉えている可能性が考えられる。2）については，消費者の職業などの属性や店舗の扱う商品カテゴリによっては，平日の日中に実店舗へ訪問し購買することは難しいなど，年齢層や職業といったデモグラフィック情報とアクセスできる曜日や時間帯の多様性が関係している可能性がある。

2.3.5　多様性変数と RFMC の関係

競合他社での行動を考慮した解析を行うにあたって，Park and Fader (2004) では競合するウェブサイトでは訪問期間が互いに影響し合う点について指摘しており，これは複数サイトにおける価格や商品の比較検討が目的とされている。**??**節でも述べたように，ある企業での利用が空白になっている期間には競合する他企業の利用を行なっている可能性があることから，自社の利用間隔の不均一性を示すことのできる RFMC では競合他社での利用についても予測が可能であると考えられている。ここでは多様性変数が有効となる主要因が消費者の異質性の判別能力の高さであることを確認するために，近年着目されている RFMC における Clumpiness の有無を表す C 指標の課題について議論した上で，C 指標における消費者の異質性の判別能力を高めるよう拡張することで予測精度が向上する可能性について検討する。

そもそも RFMC の C 指標は，ある調査期間における潜在的な購買機会（日次や週次で基準化された調査期間）において，ある顧客 i の実購買の時期から算出される購買間隔のばらつきの指標 H_i と，ランダムに同様の回数の購買が行われた場合のシミュレーションから計算できる一般的な購買間隔のばらつきの上側 $\alpha\%$ での臨界値 H_0 を比較した場合に，当該顧客の C 指標は

$$C_i = \begin{cases} 1 & (H_i \geq H_0) \\ 0 & (otherwise) \end{cases} \tag{2.4}$$

と定義される。これは消費者 i の購買行動が $\alpha\%$ の統計的有意差をもって非ランダムであることを示していることから，短期間のまとまったイベント発生（Clumpiness）と結論づけられる。

??節で述べたように，先行研究においては確かにこの手法で競合の利用を高い精度で発見できるという示唆が得られている。しかしながらこの C 指標を用いてもなお購買パターンの異質性は考慮しきれていない。C 指標の定義について中山 (2016) では「(イベントが）均等な間隔に従わない度合い」とされているものの，式 (1) からもわかるように実際には C 指標は離散値であり，仮に期間中に複数の Clumpiness が存在したり極端に偏った時期に購買が行われたりする場合（いずれも Hi の値が大きく上昇）にも通常の Clumpiness の発生と同様に $C_i = 1$ と判定されてしまう。購買の間隔の不均一性が強まるほど自社での購買回数に対しての自社を利用していない継続的な空白期間が長くなることから，購買

の不均一性についても他の RFM の 3 指標と同様に連続値で比較できることが望ましい。しかし中山 (2016) でも指摘されている通り，C 指標の算出のための H_i を複数のサンプル同士で比較する場合は，サンプル同士で潜在的な購買機会と実際の購買回数が同一である必要がある。特定の期間における複数の顧客に対して解析を行う場合には潜在的な購買機会が異なることは考えにくいが，実際の購買回数についてはサンプル間で同一になることは少なく，同一条件のみでの比較では解析に用いるにあたって不便である。そこで検定時に算出した臨界値 H_0 に対する H_i の比率として新たな指標である C 比率を

$$C_i^* = \frac{H_i}{H_0} \tag{2.5}$$

と定義する。C 比率では同一条件下の「不均一さの基準」としての臨界値と実際のばらつきの大きさの比をとっているため，平均的な基準から比較してどれほど購買が不均一かについてその度合いを連続値で定義することが可能である。この場合には C 指標と比較して明らかに購買パターンの異質性が詳細に把握できることから，多様性変数を用いる場合と同様に通常の RFMC に比べて購買予測の精度が向上すると考えられる。そこで本研究では通常の RFMC や提案する多様性変数に加えて，この C 比率を用いたモデルについても比較を行う。

2.4　解析 1：多様性変数の RFMC との比較

2.4.1　シミュレーションによる C 指標と C 比率の比較

ここで擬似的に作成した購買データを用いたシミュレーションと，実データを用いた解析を行う。本研究ではアクセス履歴データを用いることを前提としているため，RFM 分析における支出額としての M 指標は考慮しない。

多様性変数の有用性について実データで確認する前に，ここで C 指標と C 比率の有効性について簡易的に作成した消費者の疑似購買データを用いて，シミュレーションを行う。シミュレーションで多様性変数を用いないのは，購買行動と紐づいた疑似的な web usage data の作成が困難であるためである。

ここでは擬似データを作成し，それを RFMC 情報のみで予測することを考える。その際に説明変数を変更した 2 つのモデルを作成しており，モデル 1（通常の RFMC 情報を用いた場合）とモデル 2（RFMC で C 指標の代わりに C 比率を用いた場合）それぞれでの情報量基準を比較することでその妥当性について検証する。

まず市場として EC サイト A 社と競合 B 社を仮定し，潜在的な購買機会 N におけるある消費者 i の市場全体での購買回数（トータルウォレット）を非負の整数 Y_i で定義した。Y_i はポアソン回帰関数を用いた 3 因子 f_{ij} $(j = 1, 2, 3, i = 1, ..., 3000)$ に従う線形モデルとして，

$$P(Y_{1i} = k \mid \beta) = \lambda^k \frac{e^{-\lambda}}{k!} \tag{2.6}$$

$$\lambda = \beta_0 + \sum_{j=1}^{3} \beta_j f_{ij} + \epsilon_i \tag{2.7}$$

で作成した。消費者 i の t 回目の購買（ただし $t = 1, ..., Y_i$）の発生時期 T_{it}（ただし $T_{it} \leqq N$）を期間 N に対する一様分布で設定することで，期間中のどのタイミングでも同様の確率で購買が発生することを仮定している。また購買 t での利用企業 q_{it} については，購買に A 社を利用した場合に 1，競合の場合に 0 をとる離散値として定義し，その確率にはベルヌーイ分布を用いて，

$$P(q_{it} = 1) = \begin{cases} 0.5 & (t = 1) \\ 0.4 + 0.2q_{it-1} & (otherwise) \end{cases} \tag{2.8}$$

とした。これは初回購買 ($t = 1$) は完全にランダムに決定されるものの，2 回目以降の購買では前回利用した企業を再度利用する確率が微増するモデルであり，ブランドスイッチにより購買が連続しやすいことから Clumpiness が発生しやすくなることを目的としている。

作成したシミュレーションデータから各消費者についての A 社での R 指標，F 指標，C 指標，C 比率を算出し，競合 B 社での購買回数についてポアソン重回帰分析を行なった結果を表 2.3 に示す。ここでは作成した 2 つのモデルの適合度について情報量規準 AIC・BIC で比較した結果を用いており，競合での購買回数の予測では C 比率を用いたモデル 2 の場合でモデルの適合度が向上したことがわかる。

表 2.3　解析 1-1: RFMC の拡張の有無による情報量基準の比較

	モデル 1	モデル 2
AIC	9298.15	9297.89
BIC	9327.55	9327.29

2.4.2　実データ解析による C 指標，C 比率，多様性変数の比較

次に実データを用いた解析を行なう。使用したデータは株式会社ビデオリサーチインタラクティブより提供いただいた web usage data としての Web Report で，これはインターネット利用の人口統計に基づいて収集された大規模なウェブ閲覧履歴のパネルデータである。データ内から 2012 年 1 月の 1 ヶ月間の通信販売の大手 2 社（A 社・R 社）へのアクセスデータのみを抽出したところ，期間中 1 度でも A 社もしくは R 社にアクセスしたユーザは 5677 人であった。企業が自社データを用いて競合他社での購買状況を予測することを想定し，自社企業を A 社，競合を R 社として，A 社で 1 度でも購買している顧客に対して R 社での購買を予測するモデルを作成したところ，対象者は 450 人となった。

解析にあたっては該当期間の A 社での閲覧履歴から RFMC 情報，C 比率，多様性変数を計測し，シミュレーションと同様に説明変数のみを差し替えた複数のモデルを作成し，各モデルを用いて競合 R 社での同じ期間の購買回数についてポアソン重回帰分析を用いて予測する。その結果から情報量基準を用いて変数の妥当性について議論する。A 社・R 社での購買情報やその他の準備した変数について，モデル 1（一般的な R，F，C 指標を使用），モデル 2（一般的な R，F 指標と C 比率を使用），モデル 3（多様性変数を使用）の 3 モデルを作成した結果の記述統計量について表 2.4 に示す。（ただし A 社顧客には R 社では購買していないサンプルも存在しているため，R 社に関する変数の一部の統計量は記載していない。）

C 指標が $\{0, 1\}$ の離散値である一方で，C 比率では最大値が 2.97 となっており，基準値と比較しておよそ 3 倍程度の不均一性が存在していることがわかる。C 比率が 1 以上のサンプルについては C 指標では全て 1 と判定されてしまうことから，C 指標を用いることで従来の RFMC と比較して高い精度で競合購買を予測できると考えられる。さらに多様性変数については，RFMC のような購買に関わる情報を一切用いていないものの，アクセス・パターン情報として消費者のウェブ閲覧パターンの異質性について詳細に把握できることからさらに高い精度で顧客行動の予測が可能であることが予想される。

実データ解析の結果について表 2.5 に示す。として，まずは C 比率を用いて拡張した RFMC を用いた場合には，シミュレーションと同様に通常の RFMC に比べて情報量基準を改善させることができた。また特筆すべき点として，R，F，C 指標という自社での購買に関する情報を一切説明変数に用いずに Web の閲覧行動としての流入経路や閲覧の時間帯・曜日の多様さについてのみ考慮したモデル 3 において最良の結果が得られている。

表 2.4 記述統計量

変数名	投入モデル	平均値	標準偏差	最小値	最大値
A 社閲覧ページ数	-	122.44	162.62	4.00	1412.00
R 社閲覧ページ数	-	284.46	775.40	0.00	8013.00
A 社購買回数（F 指標）	1, 2	1.44	0.90	1.00	8.00
R 社購買回数	目的変数	0.20	0.56	0.00	4.00
A 社最終購買（R 指標）	1, 2	14.50	8.70	1.00	30.00
R 社最終購買	-	11.11	8.40	-	29.00
C 指標	1	0.56	0.50	0.00	1.00
C 比率	2	1.23	0.69	0.40	2.97
リファラ HHI	3	0.66	0.25	0.10	1.00
時間帯 HHI	3	0.27	0.22	0.05	1.00
曜日 HHI	3	0.36	0.23	0.15	1.00

表 2.5 解析 1-2: 購買予測における RFMC と多様性変数での情報量基準の比較

	モデル 1	モデル 2	モデル 3
AIC	466.68	465.10	458.78
BIC	486.86	485.29	478.97

2.4.3 結果の考察

まず近年着目されている RFMC については，消費者の行動の多様性の概念を応用して連続値に拡張することでシミュレーション，実データ解析ともにその予測精度を高めることができた。これは多様性変数の有用性でも触れた消費者の行動の異質性を詳細に考慮できたことが要因であると考えられる。しかしながら，この拡張した RFMC と比較しても，なお多様性変数のみを用いた場合の方が競合他社での購買の予測精度を高めることができた。RFMC と比較した多様性変数の有用性に関しては，RFMC に比べても詳細に消費者の異質性を考慮できる点に加えて，RFMC 情報があくまで購買履歴という売上の結果データを集計したものに過ぎない一方で，多様性変数はその多くがアクセス履歴データに由来するものであり，購買に至る前に商品を検討する過程を考慮できる点に起因すると考えられる。

2.5 まとめ

本章では本研究全体で用いている自社顧客の行動の多様性について，先行研究と比較した際のその理論的背景の考察を行った。そこでは多様性変数の有用性の構造としての消費者の異質性について詳細に考慮できる点や，Web 利用の習熟度との関係，デモグラフィック情報と相関している可能性等について検討した。また近年積極的に用いられている RFMC の C 指標に着目しその課題である離散変数としての定義について，多様性変数に倣って消費者の異質性を判別できるよう連続変数として拡張することで，シミュレーション・実データ解析ともに通常の RFMC と比較してモデルの適合度が改善されたことから，消費者の異質性の考慮により競合他社の購買を高い精度で予測できる点についても示すことができた。それに加えて実データ解析では RFMC と比較して多様性変数が高い説明力を持ったことから，

学問研究に限らず実務的にも幅広く活用できる指標であるといえる。今後の課題として，まず本研究ではデータの都合上アクセス履歴データのみを用いての解析を想定しているため，RFMC 分析において支出額を加味することができなかった。これに関しては，Clickstream Data と ID-POS が共通の ID で紐付けされているようなデータセットを用いることで考慮できると考えられる。また RFMC における C 指標の問題は今回扱ったもの以外にもまだ存在している。Clumpiness の発生有無の検定においては，潜在的な購買機会と実際の購買回数が一致した場合など，極めて高頻度に購買が行われている場合には，購買の回数が均一にも関わらず Clumpiness の発生確率が大きく上昇してしまう。これは基準化されている潜在的な購買機会の設定について，その基準化の方法の恣意性に加えて Clumpiness の算出方法の問題でもあると考えられる。これに関しては調査期間内での購買タイミングの偏りについて，多様性変数の算出のために用いている HHI などの指標を応用するなどすることで柔軟な連続値として定義できる可能性がある。

注

[5] コンバージョンとは EC サイトにおける購買，ニュースサイトで表示されている広告リンクのクリック，不動産仲介サイトにおける資料請求など，Web サイトで行う行動のうち企業利益に結びつくもの。

[6] ウェブ行動のデータマイニングやパターン認識などを扱った情報系の文献において，ウェブの利用履歴から得られるユーザごとのウェブ閲覧時の特徴は access patterns(Cooley et al., 1997) などと呼称されており，本論文では筆者訳のアクセス・パターン情報を用いる。

[7] 本研究ではウェブサイトに最もアクセスした時間帯や最も利用された流入経路などのカテゴリカル変数を最頻値変数と呼称する。

第 3 章

Web 行動の多様性の検討

3.1　はじめに

本章では消費者の Web 上での行動に焦点を当てた解析を行っていく。そもそもウェブサービスにおいては実店舗での小売等とは異なり，必ずしもサービス利用者と企業の間で直接的に金銭のやりとりが発生するものばかりではない。例として不動産の仲介業等においては資料請求がコンバージョンとなりうるし，ウェブサイトへの訪問・閲覧により表示される広告を主な収益源としたウェブサイトも多数存在している。広告収入を主要モデルとした代表的なものとして，大手新聞社等が運営するニュースサイトでは新聞の購読者向けに電子版の全コンテンツを公開する一方で，非購読者に対しても部分的にコンテンツを公開するとともに Web ページ上に表示される広告で収益を得ることが一般的である。

第 2 章で行動の多様性を考慮する意義について考察し予備的な解析を行ったが，あくまで短期間の購買データを用いた簡易的な解析であり，より長い期間のデータと複数の対象企業で同様の傾向が確認されるなどの包括的な検証が求められる。そこで本章では，Web 上で収集した消費者の行動の多様性の有効性について広く検証することを目的に，実際に消費者のウェブサイトへの訪問・購買に関する行動の予測モデルの構築を実施する。具体的には複数のニュースサイトと EC サイトを対象に，擬似的に設定した自社・競合他社について実際の web usage data から収集した消費者の行動を用いた行動予測を行う。その際には Web 上における行動の多様性の有効性について検証することを目的に，web usage data から消費者の行動の多様性を算出して解析に考慮する。

このような競合他社での行動予測は，企業が本来得ることのできない情報であるという点で顧客データにおける欠損データの補完という文脈で理解することもできる。本章で行う全ての解析は，実務への応用可能性の観点から，解析に用いる説明変数は全て企業が自社の Web サーバから収集できるもののみを用いている。

3.2　分析方法

3.2.1　使用データと分析対象

本研究では株式会社ビデオリサーチインタラクティブのサービスである Web Report から得られた非集計レベルのアクセスログのパネルデータを分析に使用しているが，これは先述の User-Centric Data に当たる。このデータでは，ウェブ利用に関して日本の人口分布に基づき，調査に参加しているユーザ（およそ 1 万 3 千人）が自宅のパソコンで閲覧したウェブサイトについて，アクセスした日時，アクセスした URL，そのウェブページへの滞在秒数，リファラ情報などが収録されている。本研究ではこのデータを加工してアクセスした URL を特定のウェブサイトに限定することにより，擬似的に企業がそ

のウェブサイトについて得られる Site-Centric Clickstream Data として使用している。

本研究では，ウェブ閲覧行動関連の分析として新聞社の運営するニュースサイトを，そしてウェブ上での購買行動関連の分析として EC サイトを分析対象として扱っている。また，本論文においては，データ使用の都合として分析対象のサイト名を，新聞社大手 2 社の運営するニュースサイトについて N 社，M 社，同様に EC サイト大手 2 社について A 社，R 社としている。その上で，ウェブサイトで得られるユーザ別アクセス・パターン情報の多様性が, 自社・競合他社での行動予測を行うにあたって活用できる可能性に着目して分析を行っている。

3.2.2 使用モデル

予測モデルの作成にあたっては，目的変数である訪問・購買の回数の予測を行うため，解析 1 と同様にポアソン分布を用いた GLM での重回帰分析を実施した。分析を行う際には得られているデータ全てを用いてモデリングした結果としての全量解析に加えて，新規データを用いた場合の汎化的な予測精度としてクロスバリデーション（結果を示した表では CV と表記）も実施しており，この場合には得られたデータから無作為に抽出した 7 割の自社顧客のデータからモデリングを行い，作成したモデルを残り 3 割の顧客に適用し予測値を算出した。予測精度の比較には，作成したモデルから算出した従属変数の予測値と真値での相関係数を用いた。多様性変数の有効性の判断にあたっては，多様性変数を用いたモデルの CV での相関係数が最良の値になることをもって多様性が有効であると判断する。

本研究の各分析において使用された変数は表 3.1 の通りである。年代・性別・居住地域等の，企業の持つアクセスログやユーザ登録情報などによって得られる可能性のある情報としての基本変数，同様にアクセス・パターン情報から測定可能な多様性変数として「リファラの多様性」「自社にアクセスした時間帯／曜日の多様性」「自社で閲覧した店舗数の多様性」「自社サイトでアクセスしたドメインの多様性」等を用意した。ここで店舗数の多様性とは，本研究において自社として取り上げた A 社がサイト内に多数の EC 業者を持つ電子モールであることから，A 社サイト内で閲覧した店舗の数について測定した多様性変数である。またドメインの多様性とは，EC サイト以外にも多様なサービスを提供する A 社ではサービスごとに異なったドメイン名が設定されていることから，ユーザがアクセスしたドメインの多様性を測定することで自社で提供されているサービスをどれほど多様に利用しているかを測定した多様性変数である。

リファラの多様性に関しては，その流入元の URL のドメインの種類を測定することにより，いかに多様なウェブサイトから自社サイトに流入しているかを測定したものである。この情報は，アクセス・パターンとして自社で得られる情報の中で唯一「どれほど多様なウェブサイトを閲覧しているか」を測定できる変数であるため，その多様性の定義により自社顧客に関してより詳細な情報を得られる可能性がある。

その他の変数として，自社サイトでアクセスしたドメインの多様性などを用意した。ウェブサービスにおいては EC サイト以外にもブログや検索ポータルをはじめとした様々なサービスを提供している場合も多く，今回自社サイトとして取り上げた A 社では，サービスごとに異なったドメイン名が設定されている。例として，オークションのサービスを提供する auction.***.jp，ブログのサービスを提供する blog.***.jp など。ただし***は A 社のウェブサイトのドメイン名を表す。従って，ユーザがアクセスしたドメインの多様性を測定することによって，自社で提供されているサービスをどれほど多様に利用しているかが判明することから，自社へのロイヤルティの高さを測定できると考えられる。

また，最頻値変数として顧客ごとに自社への流入の際に最もよく使われているリファラやアクセス時間帯などについての情報を取得し，ダミー変数化した。この情報を使用したのは，「特定のブログペー

表 3.1 用意した変数一覧

	#	変数名	説明
	1	ユーザ ID	ユーザごとに共通化された番号（匿名化済）
基	2	調査期間	ユーザごとのアクセス履歴取得期間
本	3	年代・性別	ユーザの年代・性別のダミー変数
変	4	居住地域	ユーザが居住している地域のダミー変数
数	5	購買回数	A 社での累積購買回数
	6	購買間隔	A 社での直近の購買間隔
	7	お気に入り流入率	A 社への流入時にお気に入りを使用した割合
	8	平均閲覧秒数	自社でのページ単位の平均滞在秒数
	9	アクセス間隔	自社へアクセスする平均間隔
多	10	検索頻度 **	自社での検索行動の多さ
様	11	検索カテゴリ数 **	検索時に使用したカテゴリの多さ
性	12	リファラ HHI *	自社でのリファラの多様性
変	13	リファラ数 **	自社でのリファラの多さ
数	14	ドメイン HHI *	自社でのドメインごとのアクセス多様性
	15	ドメイン数 **	自社で訪問したドメインの多さ
	16	閲覧店舗数 **	自社で閲覧した店舗の多さ
	17	閲覧店舗 HHI *	自社で閲覧した店舗の多様性
	18	アクセス曜日 HHI *	自社にアクセスした曜日の多様性
	19	アクセス時間帯 HHI *	自社にアクセスする時間帯の多様性
最	20	最頻リファラ	最もよく使用されているリファラ
頻	21	最頻時間帯	最もよく自社にアクセスしている時間帯
値	22	最頻曜日	最もよく自社にアクセスしている曜日
	23	最頻検索エンジン	自社流入時に最もよく使用されている検索エンジン

*: 多様性の指標として HHI を使用

**: 多様性の指標として累積パーセントを使用

ジから頻繁に流入している顧客は購買頻度が上昇する」など，アクセス・パターンの多様性ではないものの分析の推定結果を改善する可能性があると考えられるためである。これらの変数はいずれも，「自社で得られる情報であること」「ユーザのウェブ使用の多様性を測定できること」を前提とし，それぞれの分析において使用を検討している。

また，推定の際には個人ごとの調査期間が異なる点を考慮するため，「個人ごとの調査期間」も含めている。

3.3　解析 2：ニュースサイトにおける閲覧頻度の予測

3.3.1　自社と競合他社における閲覧頻度の関係

オンラインにおいては，必ずしも企業は顧客行動の購買のみをコンバージョンとしているわけではない。例として新聞社のウェブサイトにおいては，無料で提供している記事に対するオンライン広告での収益や，最終的な有料記事の購読への誘導など，無料の自社ページへの訪問だけでも重要な顧客行動となりうる場合がある。本節では，新聞社のウェブサイトを対象として，一方のサイトで得られるアクセス・パターン情報から，競合他社である他方のウェブサイトの閲覧頻度の推定を行う。

推定においては，自社サイトへのアクセス頻度などの基本的な情報に加えて，アクセス・パターン情報の多様性を加えることにより予測精度が向上することを示す。分析対象には日本の新聞社大手 2 社を取り上げ，自社サイト C 社から，そこで得られたアクセス・パターン情報の多様性を用いて競合他社 D 社での閲覧回数を目的変数とした回帰分析を行う。分析対象者の抽出の条件は C 社でアクセスログを得た際の調査期間が 30 日以上であるユーザで，分析対象者は 1943 人である。調査期間を 30 日以上としたのは，アクセス・パターン情報の算出において自社サイト内での行動についての十分な情報量を確保することを目的としている。

表 3.2　閲覧頻度と多様性の相関について

	自社閲覧（実測値）	競合閲覧（実測値）	リファラ多様性（HHI）	曜日多様性（HHI）	時間帯多様性（HHI）
自社閲覧（実測値）	1.000	0.209	-0.073	-0.271	-0.271
競合閲覧（実測値）	0.21	1.000	-0.136	-0.183	-0.205

全ての相関係数について有意水準 1% を満たす。

表 3.3 および図 3.1，図 3.2 は，ニュースサイトでの自社顧客における自社および競合他社のウェブサイトについて測定した累積閲覧頻度（対数化）と，自社で測定した多様性変数（対数化）の相関係数と散布図を示したものである。これらからも閲覧頻度とそれぞれの多様性についての相関係数は決して高いとは言えず，従って閲覧頻度等の高い顧客だからといって必ずしも各多様性変数が上昇，下落といった特徴的な変化をするとは言えないことがわかる。すなわち，従来では閲覧時間の長さや訪問頻度の高さをはじめとして，自社サイト上での行動に関する量的変数の大小によってのみ比較できていた顧客を，その多様性を測定することにより分散的な指標で比較することが可能となるのである。これにより，ユーザの特定のリファラや時間帯，曜日などへの依存度合いの把握が可能となることから，閲覧や購買といった顧客ごとの行動の特性の数値的比較がより容易となることが考えられる。

3.3.2　解析 2-1：自社サイトの閲覧頻度の推定

自社顧客の閲覧行動についての推定に関して，最初に自社顧客の閲覧頻度の推定結果を示したのが表 3.3 である。推定結果では，推定の際に表 4.1 における基本変数に加えて最頻値変数と多様性変数を追加した 3 つのモデルについて比較している。顧客の自社での情報の推定については多様性を用いずとも比較的高い推定結果を得られているが，多様性変数の追加でより結果が改善していることがわかる。この結果から，自社顧客に関する情報を活用することにより，競合他社における閲覧頻度など，自社の持

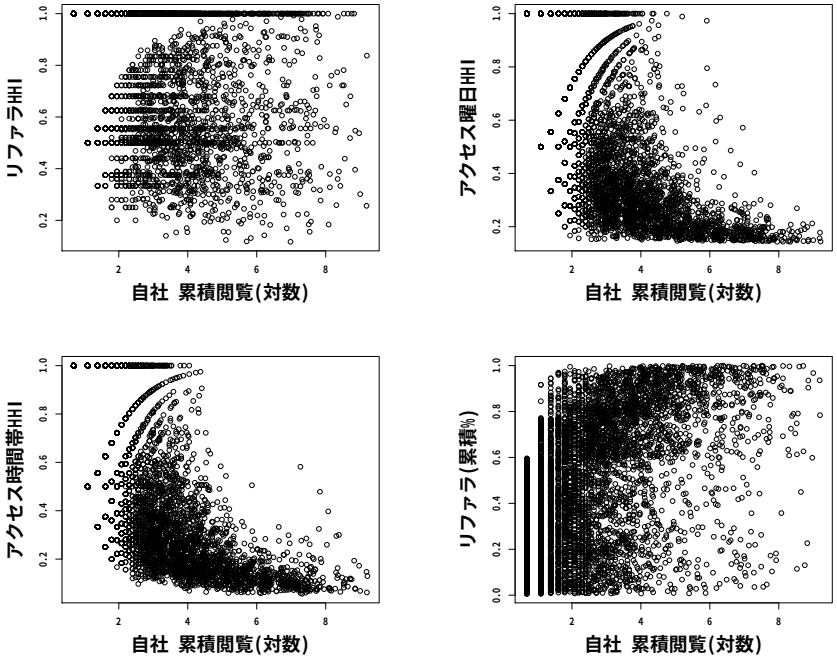

図 3.1　自社における閲覧頻度と多様性変数の関係

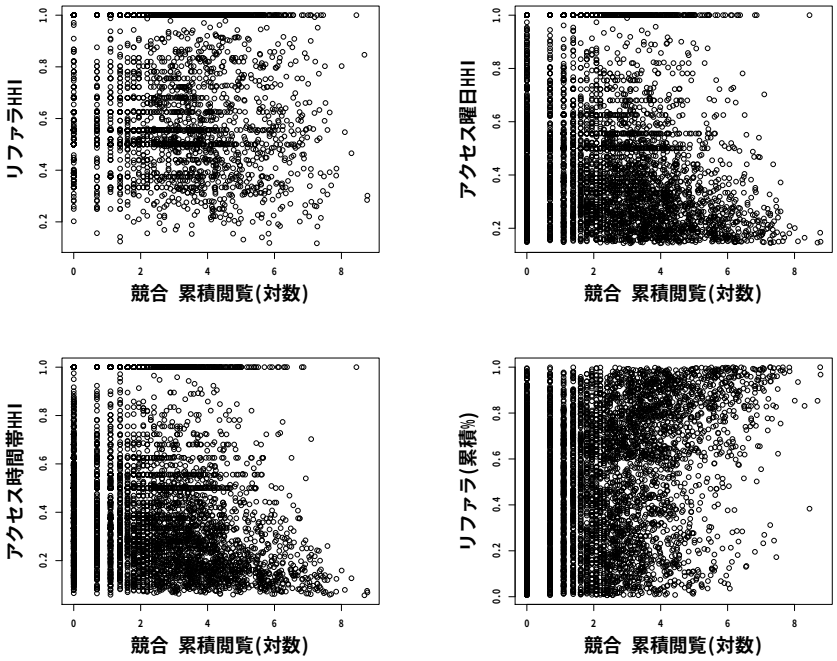

図 3.2　競合他社における閲覧頻度と多様性変数の関係

ちうる情報のみではわからない情報を推定できる可能性が考えられる。

表 3.3　解析 2-1: 自社閲覧行動 予測精度の比較（相関係数）

	推定値 （多様性 + 最頻値 + 基本変数）	推定値 （最頻値 + 基本変数）	推定値 （基本変数のみ）
自社閲覧実測値（全量）	0.850	0.806	0.778
自社閲覧実測値（CV）	0.529	0.433	0.429

全ての相関係数について有意水準 1% を満たす。

3.3.3　解析 2-2：競合他社サイトの閲覧頻度の推定

そこで，自社サイトと同様に競合他社サイトにおいても自社顧客の閲覧頻度に関する分析を行った。モデリングの際には，使用する説明変数はすべて自社サイトにて測定している。その結果を示したのが表 3.4 である。こちらも分析 1 と同様に基本変数に加えて最頻値変数と多様性変数を加えた結果について示している。多様性変数の投入により推定値の相関係数が改善されており，自社顧客について競合他社での閲覧回数を推定する上で，ユーザのアクセス・パターンの多様性が有効であることが示された。

表 3.4　解析 2-2: 競合閲覧行動 予測精度の比較（相関係数）

	自社閲覧頻度 （実測値）	推定値 （多様性 + 最頻値 + 基本変数）	推定値 （最頻値 + 基本変数）	推定値 （基本変数のみ）
競合他社閲覧（全量）	0.209	0.542	0.394	0.383
競合他社閲覧（CV）	-	0.286	0.224	0.202

全ての相関係数について有意水準 1% を満たす。

3.4　解析 3：EC サイトにおける購買頻度の予測

3.4.1　EC サイトでの多様性変数と購買の関係

続いて，EC サイトにおける購買回数の推定に関する分析を行う。分析に先立って，EC サイトでの多様性変数と購買回数の関係について示す。表 3.5 では，使用データにおけるユーザごとの調査期間から，ユーザ全体での自社での 1 ヶ月あたりの平均購買回数を算出した。その上で自社，競合他社それぞれについて 1 群（平均購買回数以上の顧客群）と 2 群（平均購買回数未満の顧客群）を設定し，自社 EC サイトで測定したアクセス・パターン情報の多様性の平均値を群間で比較している。t 検定の実施により各群についていずれも有意差が観測されたため，各群によって多様性に傾向がある可能性が考えられる。自社，競合それぞれで各群のとった値を比較すると，時間帯や曜日，閲覧店舗数の多様性においては概ね同じ傾向にあることが分かる。しかしながら，リファラ HHI に限っては，自社では 1 群にてより高い値をとっているが，一方で競合では 2 群がより高い値をとっている。リファラ HHI が高いほど「自社サイトへの流入経路が特定のサイトに依存している」ことを鑑みると，自社サイトでの購買頻度の高い顧客については流入経路が限定されていることから，予め自社で購買することを決めた上で，

使い慣れている経路を通じて自社へ流入している可能性が考えられる。一方で，競合他社でよく購買を行っている顧客については，自社サイトへの流入の際には多様なウェブサイトから流入している場合が多い。理由のひとつとして，様々なウェブサイトのブラウジングを行う中で，商品を紹介しているブログなどの多様なサイトから自社サイトへ誘導された可能性が考えられるだろう。

表 3.5　購買回数による多様性平均値の差

		リファラ	時間帯	曜日	閲覧店舗数
	人数	HHI	HHI	HHI	HHI
A 社 1 群	2402	0.745	0.146	0.224	0.125
A 社 2 群	7578	0.587	0.387	0.468	0.424
R 社 1 群	2304	0.599	0.190	0.267	0.198
R 社 2 群	7676	0.625	0.376	0.458	0.398

3.4.2　EC サイト間での多様性変数の関係

アクセス・パターンの多様性を用いない場合には，競合他社での顧客行動を推定するにあたって使用できる変数には，EC サイトであれば累積の購買回数や購買期間，購買した商品のカテゴリについての量的変数など自社の購買履歴を用いた変数や，検索ポータルからの流入率といった既存研究で使用されている変数，そしてユーザ登録に関連したデモグラフィック変数程度に限られることが予想される。しかしながら，仮に同様のデモグラフィック変数を持った顧客であっても，ウェブの使用頻度やその使用目的の幅広さによって購買行動が異なることも考えられる。すると，シェア・オブ・ウォレットを推定するような目的の場合にはこれらの変数が必ずしも強い説明力を持つとは言えず，企業が競合他社の情報推定をするにあたって有効な変数は限られてくるであろう。そこで，自社サイトにおいて測定した多様性変数が，競合他社で測定される多様性変数と近い性質を持っていた場合，競合他社における顧客情報の推定においても強い説明力を示すことが考えられる。表 3.6 に A 社，B 社のそれぞれのウェブサイト内において測定されたアクセス・パターンの多様性の相関係数を示した。この表より，それぞれの多様性は異なったウェブサイト間でも少なくとも弱い相関を，時間帯関係の多様性については比較的強い相関を持っていることがわかる。使用できる説明変数が少ない企業活動において，競合他社での自社顧客の行動と同じ傾向を持つ変数が得られることは，分析における活用可能性も大きいといえるだろう。

表 3.6　EC サイト間での多様性変数同士の相関

		R 社			
		リファラ（HHI）	リファラ（累積 %）	時間帯（HHI）	曜日（HHI）
A 社	リファラ（HHI）	0.290	-0.086	0.266	0.263
	リファラ（累積 %）	-0.04	0.304	-0.066	-0.070
	時間帯（HHI）	0.276	-0.087	0.558 *	0.547 *
	曜日（HHI）	0.268	-0.089	0.546 *	0.566 *

全ての相関係数について有意水準 1% を満たす。

* : 相関係数の絶対値が 0.5 以上

3.4.3　解析 3-1：自社での累積購買回数

ここで，多様性の有効性の検証として多様性を用いた自社の EC サイトにおける購買頻度の推定を行なう。また自社の購買に関しては顧客ごとの将来購買の予測も行う。表 3.7 に，自社での累積購買回数の推定結果を示した。推定結果より自社購買の場合では多様性変数を用いていなくとも推定結果の相関係数が比較的高い値を取っている。しかしながら，モデルに多様性変数を投入することにより，推定結果をより改善できたと言える。

表 3.7　解析 3-1: 自社累積購買行動 予測精度の比較（相関係数）

	推定値 （多様性 + 最頻値 + 基本変数）	推定値 （最頻値 + 基本変数）	推定値 （基本変数のみ）
自社購買（実測値） 全量	0.813	0.756	0.713
自社購買（実測値） CV	0.505	0.495	0.494

全ての相関係数について有意水準 1% を満たす。

3.4.4　解析 3-2：自社での将来的な購買回数

続いて，自社での将来的な購買回数の推定を行った。具体的な推定方法としては，Web Report でのそれぞれのユーザの調査期間を最新の 30 日間（期間 I）とそれ以前の期間（期間 II）に分割し，期間 II で測定した変数を使用してモデリングを行い，期間 I で実際に購買を行った回数をポアソン分布による重回帰分析を用いて推定する。分析対象者は，A 社で得られる調査期間が 60 日以上のユーザであり，少なくとも一度 A 社で購買を行っているユーザ 4146 人である。表 3.8 に，期間 I での購買回数の実測値に対する期間 II での推定結果の相関係数を示した。まず，多様性を用いた場合での推定値は，期間 II の実測値や多様性を用いずに行った推定のいずれよりも相関係数が改善する結果となった。この結果より，自社顧客の将来的な購買に関しての推定精度を改善するにあたって多様性変数が有効であることが示された。ただし多様性を用いない推定からの改善幅が小さい点については，期間 I と期間 II での購買回数の相関の高さからもわかるように，既に強い説明力を有していることからさらなる改善が難しいことが原因であると考えられる。

表 3.8　解析 3-2: 自社将来購買行動 予測精度の比較（相関係数）

	実測値 （期間 II）	推定値 （多様性 + 最頻値 + 基本変数）	推定値 （最頻値 + 基本変数）	推定値 （基本変数のみ）
実測値（期間 I） 全量	0.538	0.576	0.547	0.543
実測値（期間 I） CV	-	0.397	0.234	0.224

全ての相関係数について有意水準 1% を満たす。

3.4.5　解析 3-3：競合他社での累積購買回数

続いての分析では，シェア・オブ・ウォレットの推定として，顧客ごとの競合他社における累積の購買回数を推定する。元のデータセットより自社顧客の競合他社である B 社での累積購買回数を計測し，

目的変数として回帰分析を行った。

モデリングの際のパラメータ推定値に関しては，多様性変数に関して有意に説明力を持ったものとしてリファラの多様性，ドメインの多様性，曜日／時間帯の多様性などがあった。そこで，競合他社の分析に関しても，自社顧客と同じように「リファラに偏りがなく種類が多いほど（様々なウェブサイトから自社に流入している顧客ほど）自社での購買回数が増加する」，「アクセスする曜日／時間帯に偏りがないほど（様々な曜日／時間帯に自社サイトにアクセスしている顧客ほど）自社での購買回数が増加する」というように解釈できる。多様性変数に関しては，自社／競合他社の購買ともに，パラメータの正負の符号の傾向は一致しているが，本分析においてはドメインの多様性が有意に説明力を持っており，「自社の提供する様々なサービスを偏りなく使用している顧客ほど，競合他社での購買回数が減少する」と解釈できる。自社のサービスの幅広い利用は必ずしも自社購買の増加にはつながらないものの，競合他社の利用頻度が減少するという点においては自社のシェアオブウォレット向上の要因とも考えられるのではないだろうか。

表 3.9 に，競合他社での累積購買回数の実測値に対する，自社での累積購買回数の実測値，多様性を用いない場合の推定値，そして多様性を用いた場合の推定値の相関係数を示した。

表 3.9　解析 3-3: 競合累積購買行動 予測精度の比較（相関係数）

		自社購買（実測値）	推定値（多様性 + 最頻値 + 基本変数）	推定値（最頻値 + 基本変数）	推定値（基本変数のみ）
全量	競合他社購買（実測値）	0.159	0.333	0.297	0.246
	4 回以上購買 正判別率	-	42.15%	40.82%	36.08%
CV	競合他社購買（実測値）	-	0.286	0.264	0.228
	4 回以上購買 正判別率	-	35.56%	34.89%	32.17%

全ての相関係数について有意水準 1% を満たす

本分析では相関係数にみる推定値の改善幅は比較的小さいという結果に留まっているが，本分析を行う目的はシェア・オブ・ウォレット自体の推定に加えて，自社でのシェア・オブ・ウォレットの小さな顧客を見つけることである。推定された競合他社での累積購買回数に従って，期間中に 4 回以上競合他社にて購買を行った顧客を抽出した場合，その正判別率は 42.15% となった。本モデルを使用せずに自社顧客からランダムに 4 回以上購買者を抽出する場合，その確率は 13.4% となるため，ランダムサンプリングに比べても有効性を持つことがわかる。

また，同様にモデルを評価するために，図 3.3 に，推定値に従ったゲインズチャート [8] を作成した。ゲインズチャートとは，作成したモデルによる顧客の抽出精度をランダムサンプリング時の結果と比較することでそのモデルの予測力を示すことのできる図表である。モデルの推定結果に従って自社顧客を競合他社での購買回数が多い順に並べ，その上位 X% を抽出した際に，実際に 4 回以上購買している顧客全体のうち Y% が含まれるというこの X, Y の関係をグラフ化した。一般に，自社顧客全体からランダムに X% の顧客を抽出した場合には，母集団のうち 4 回以上購買している顧客についても同様に X% が抽出されるはずである。これがランダムサンプリング時の性能として 45 度線に示されており，従って 45 度線と比較して上方向の膨らみが大きいほど推定精度が高いことを表している。また，図表における予測精度の上限とは，仮にモデルが非常に高い予測力を持っていた場合に，推定値に従って抽出した上位 X% の消費者のうち全員が 4 回以上購買を行った顧客であったとしても，抽出した人数より多くの 4 回以上購買者が存在した場合にはその全てを抽出することができないという点で，この上限以上の性能を取ることはできないことを示している。

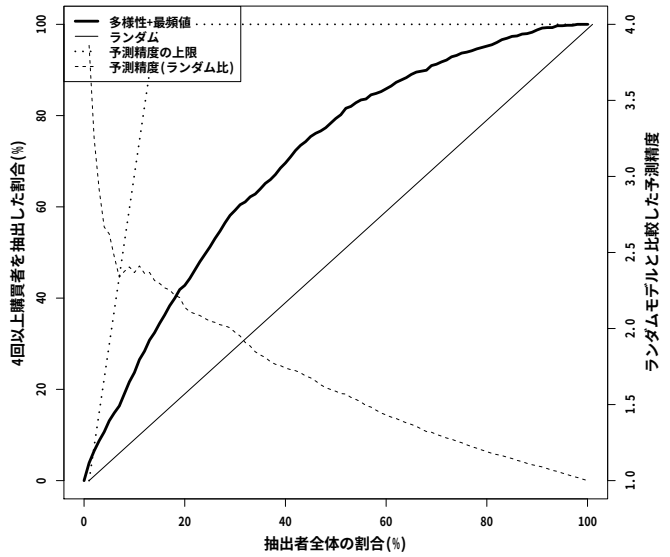

図 3.3　解析 3-3 の結果に従ったゲインズチャート

このゲインズチャートより，本モデルに従って競合他社にて頻繁に購買を行っていると推定された顧客の上位 20% を抽出した場合，その中には 4 回以上購買者全体のうち 42.78% の顧客が含まれていることがわかる。この推定精度は 4 回以上購買した顧客をランダムに抽出した場合の精度 20% の 2 倍以上であり，当モデルを使用した分析の有効性が示されていると言える。

3.5　解析 4：繰り返しのある継続時間と購買有無の同時モデリング

3.5.1　解析の目的

ここまで，本論文ではユーザ別アクセス・パターン情報の多様性を用いた場合の推定精度の改善について分析を行ってきた。多様性変数で Web 上での個人の異質性を捉えることによりその予測精度が向上することを示し，多様性変数の有効性について検証した。しかしながら，これまでの分析には次に挙げるような問題が考えられる。まず一つ目に「個人の異質性」を完全には考慮できていないことである。モデルにおいて個人の異質性を捉えることのみを目的にするのであれば，多様性変数を用いずともモデルに変量効果として投入することで個人の異質性は考慮することが可能であり，多様性変数の必要性はなくなってしまう。しかしながら第 2 章で示しているように，Web usage data から収集した多様性変数を用いることは個人の異質性以外の効果も考えられることから，本節では多様性変数以外の手法を用いて個人の異質性を考慮しているモデルにおいて，さらに多様性変数を投入することでモデルの改善が見られるのかについて検証する。

また二つ目の問題として，購買と閲覧を同時に考慮できていないという点がある。これまでの解析では自社・競合における累積・将来期間での閲覧・購買の回数について個別に予測するモデルを作成している。しかしながら，実際には次回のある 1 回の訪問の時期の予測と，その際の購買の有無について繰

り返し予測を行う方がより実務的貢献が高いことが考えられる。そこで本章で実施する解析では，各消費者について繰り返しのある訪問（各 1 回の訪問について繰り返し予測）とその際の購買有無についてのジョイントモデルを作成して予測する。

3.5.2　使用モデル

実際の同時分布モデルは次のように表現できる。d_{ij} を個人 i が第 j 回目に特定サイトへ訪問した時から第 $j+1$ 回目に訪問するまでの時間間隔とする。そして $j+1$ 回目でそのサイトからの購入を行った場合に 1，購入していない場合には 0 となる 2 値の離散変数を y_{ij} とする。また，個人 i が解析の対象となっている時間幅においてそのサイトを訪問した回数を J_i とする（従って $j = 1, \cdots, J_i$ となる）。

ここで個人 i の共変量（以降個人共変量と呼ぶ）の値を $\boldsymbol{x}_i$，個人 i の j 回目から $j+1$ 回目の訪問の間に特定の値を持つ時変共変量を $\boldsymbol{w}_{ij}$ とし，継続時間として d_{ij} には比例ハザードモデルと加速故障時間モデル両方の性質を満たす唯一のモデルであるという利点を有すること，およびマーケティングモデルとしても先行研究で利用されることが多いため，実務的な有用性という観点からワイブルモデルを利用する。

$$f\left(d_{ij}|\boldsymbol{w}_{ij}\right) = \alpha d_{ij}^{\alpha-1}\lambda_{ij}\exp\left(-\lambda_{ij}d_{ij}^{\alpha}\right) \tag{3.1}$$

さらに

$$log\left(\lambda_{ij}\right) = \boldsymbol{w}_{ij}^{t}\boldsymbol{\beta}_i \tag{3.2}$$

とし，

$$\boldsymbol{\beta}_i = \boldsymbol{B}\boldsymbol{w}_i + \epsilon_i, \quad \epsilon_i \sim N\left(0, \boldsymbol{\Phi}\right) \tag{3.3}$$

とする。但し $\boldsymbol{\Phi}$ は対角行列である。ここで訪問回数が十分複数回観測されるように時間幅を十分にとることで，打ち切りの問題は考えないとする。

さらに $j+1$ 回目のサイト訪問での購入を表す変数 y_{ij} はロジスティック回帰モデルに従い，共変量に依存する

$$logit[p\left(y_{ij}^{B} = 1|\boldsymbol{w}_{ij}\right)] = \boldsymbol{w}_{ij}^{t}\boldsymbol{\gamma}_i \tag{3.4}$$

但し

$$\boldsymbol{\gamma}_i = \boldsymbol{C}\boldsymbol{w}_i + \eta_i, \quad \eta_i \sim N\left(0, \boldsymbol{\Psi}\right) \tag{3.5}$$

とする。但し $\boldsymbol{\Psi}$ は対角行列である。

このとき，繰り返しのある継続時間と購買有無の同時分布は

$$p(y_{i1}^{*}, \cdots, y_{iJ_i}^{*}|\boldsymbol{w}_{i1}, \cdots, \boldsymbol{w}_{iJ_i}, \boldsymbol{x}_i) = \int\Big(\prod_{j=1}^{J_i} p(d_{ij}|\boldsymbol{\beta}_i, \boldsymbol{w}_{ij}, \boldsymbol{x}_i)p(y_{ij}|\boldsymbol{\gamma}_i, \boldsymbol{w}_{ij}, \boldsymbol{x}_i)\Big)p(\boldsymbol{\beta}_i.\boldsymbol{\gamma}_i|\boldsymbol{x}_i)d\boldsymbol{\beta}_i d\boldsymbol{\gamma}_i \tag{3.6}$$

と表現される（但し $y_{ij}^{*} = (d_{ij}, y_{ij})^{t}$ とする）。

上記のモデルは一般化混合線形モデル (McCulloch and Neuhaus, 2001) であり，変量効果についての多重積分を含み解析が難しいが，本研究では積分段階的ラプラス近似法（integrated nested Laplace approximation, INLA）を利用する。INLA は Rue et al. (2009); Martino and Rue (2009) によって，正規分布に従う変量効果モデル（正確にはより一般的な latent Gaussian model）でのラプラス近似の新しい方法として提案され，また Fong et al. (2010) は一般化混合線形モデルへの INLA の一般化を行い，近年様々な応用研究に利用されている推定法である。

このような INLA を用いた同時モデリングとしては，Guo and Carlin (2004) において Gaussian model に基づく個人の縦断的データと Weibull 分布に基づく生存時間データに関するジョイントモデルを作成することで情報量規準の改善に成功している。これは縦断的データの結果が個人の生存時間データと相関を持っているために，同時モデリングを行うことでよりモデルの説明力があがっていることに起因している。また，星野 (2013) においては，消費者の購買行動をはじめとして個人の行動を反映した結果としてのこのようなログデータには大きな個人差が存在する可能性について指摘している。そして顧客ごとの訪問タイミングに関する継続時間とその訪問ごとの購買の有無に関する離散選択の同時分析について，変量効果の投入により個人差を考慮した上でのモデル作成を行っている。

3.5.3　解析結果

以下の表 3.10 で，モデリング結果の比較として，モデルに投入する説明変数として多様性変数の有無と変量効果の有無を変化させた 4 つのモデルでの情報量規準の値を比較している。結果として変量効果があり，なおかつ多様性変数も含んだモデルで情報量基準が最善の値を取っていることがわかる。また表 3.11 には変量効果を持つモデルでのパラメータ推定値の一覧を示した。閲覧，推定のいずれのモデルにおいても，HHI や累積パーセントといった多様性変数のうちの多くが有意に説明力を持つ結果となっている一方で，訪問ごとに異なる経由リファラに関するカテゴリカル変数などは十分な説明力を持たない結果となった。これは，特定のリファラからの流入が購買を促進するといった可能性よりも，多様性変数による顧客行動の質的な測定がより強く購買を説明できているということが考えられる。

表 3.10　解析 4: 訪問と購買の同時モデリングでの報量基準の比較

	変量効果なし		変量効果あり	
	多様性なし	多様性あり	多様性なし	多様性あり
AIC	58961.61	56675.76	56650.87	56612.64
DIC	57567.79	55198.60	54523.10	54414.68

3.6　まとめ

本章では Web 上の多様性を用いての自社・競合他社での訪問・閲覧について予測モデルの作成を行った。結果としていずれも多様性変数を用いた場合に予測精度が改善しており，多様性変数の有効性が示されたといえる。また個人の異質性を考慮したモデルにおいても多様性変数を用いることでモデルの情報量規準に改善が見られたこと k ら，個人の異質性以外にも Web への習熟度等の別の効果も内包していると考えられる。

ここまでの分析と検証より，自社顧客の競合他社での行動を予測する際に，アクセス・パターン情報の多様性を計測することによって，推定の精度の向上をはかることや分析における従来以上に柔軟な説明変数を用いた予測ができた。実際の企業活動においては顧客の情報を推定するにも活用できる変数が少ないことも考えられる。特に時間的な余裕やインターネットの活用具合などに関しては，正確なデータを得るためには顧客へのアンケートを実施するなど追加的なコストが必要となることも考えられる。しかしながら，このような情報に関して今回定義した多様性変数を代替的に用いることで推定が可能であるという結果が得られたことにより，企業のマーケティングをはじめとした顧客管理を行う上でより幅広い推定を行ない，効率的に戦略を策定することが可能になるであろう。

表 3.11　INLA ジョイントモデル パラメータ推定値

	購買（binomial）		訪問間隔（weibull）	
変数名	事後平均	事後標準偏差	事後平均	事後標準偏差
Intercept	-0.069	0.119	0.11	0.095
リファラ HHI	0.351	0.120 **	-0.069	0.119
サイト内ドメイン HHI	0.039	0.014 **	-0.109	0.013 **
訪問店舗 HHI	0.000	31.623	0.000	31.623
アクセス曜日 HHI	0.032	0.011 **	-0.03	0.011 **
アクセス時間帯 HHI	0.047	0.016 **	-0.137	0.015 **
リファラ（累積 %）	0.051	0.017 **	-0.212	0.015 **
サイト内ドメイン（累積 %）	0.491	0.167 **	1.353	0.159 **
訪問店舗（累積 %）	0.486	0.166 **	0.809	0.163 **
検索頻度（累積 %）	0.491	0.168 **	1.144	0.162 **
平均訪問間隔	0.496	0.170 **	0.712	0.156 **
リファラ　infoseek	0.000	0.000	-0.001	0.0001 **
リファラ　Yahoo!	0.171	0.147	0.468	0.0311 **
リファラ　Google	0.152	0.144	-0.296	0.03 **
リファラ　お気に入り	0.169	0.201	-0.302	0.0397 **
リファラ　FC2	0.144	0.125	-0.035	0.027
リファラ　pixiv	0.188	0.448	-0.213	0.0841 *
リファラ　livedoor	0.199	2.386	-1.409	0.4398 **
リファラ　EC ナビ	0.183	0.654	-0.185	0.122
リファラ　価格.com	0.186	0.367	0.527	0.0699 **
リファラ　live メール	0.193	0.516	-0.181	0.097
前回訪問で購入（0/1）	0.152	0.633	0.656	0.1187 **
前回と同一リファラ（0/1）	0.351	0.120 **	-0.069	0.119

∗: 有意水準 5% 以下を満たす

∗∗: 有意水準 1% 以下を満たす

注

[8] アメリカ南メソジスト大学の Thomas B. Fomby 氏によるゲインズチャートの解説を参考に著者が作成。
（http://faculty.smu.edu/tfomby/eco5385/lecture/Lift\%20Charts.pdf）

第 4 章

深層学習を用いた Web と実行動の多様性の検討

4.1 はじめに

4.1.1 研究の背景

昨今，実店舗を持つ小売企業では，Amazon や楽天といった通信販売サイトの興隆に対抗して顧客を実店舗と Web 上で相互に送客しあう Online to Offline / Offline to Online (O2O) の試みや，さらにはオムニチャネル [9] による実店舗，Web の垣根を超えたマーケティング戦略を実施しており，その戦略策定はますます高度で複雑なものとなっている。企業が顧客管理に用いることのできる情報は，従来では店舗への訪問の際の購買履歴をポイントカードなどを用いて継続的に収集することで，顧客の購買間隔や商品カテゴリ等の情報を得るに限られており，店舗訪問時の購買関連の情報以外を得ることは困難であった。しかしながら，このようなオムニチャネルの進行にともなって企業が得られる顧客データは非常に多面的になっている。実店舗・Web サイト・モバイルアプリなどの様々なチャネルから消費者にアプローチするオムニチャネルでは，多くの場合自社顧客を共通の ID 等で紐付けていることから，店舗外での日常生活まで含めて幅広い情報を得られるようになった。

企業が顧客管理を行うにあたって，従来からマーケティング分析の手法として一般的であるセグメンテーションなどを用いて顧客を分類することはもちろん有用である。しかしながら，趣味嗜好が多様化した現代社会において，購買回数等の量的変数やデモグラフィック情報のみに頼っての画一的な予測は難しいことが予想される。様々なデータを得られるようになった一方で，多量かつ多様なデータを未だ顧客の分析に活用しきれていないのが現状である。企業が新たに得られるようになった顧客の店舗訪問時以外の日常生活における Web 閲覧やモバイルアプリの使用データなど，今後の顧客分析にあたっては幅広い顧客行動を考慮することが重要になると考えられる。

また，現在様々な分野で機械学習，殊に深層学習を用いた予測モデルの構築や解析例が増えている。一般的には画像や音声といったパターン認識分野に強いといわれる深層学習であるが，近年ではビジネスへの応用も増加している。例として NTT コミュニケーションズにおいては，深層学習を含む AI 関連技術の活用による電話応対サービスや，得られた家計簿データを分析してアドバイスを行う機能，さらには画像認識により消費者が所有する画像を自動分類するサービスなどが提供されている (伊藤他, 2016)。 マーケティング分野においても今後益々の活用が期待される深層学習について，より精度の高い分析が求められる。

4.1.2　特徴選択の精度の問題

現在の深層学習においては柔軟かつ高い精度での予測が可能とされている一方で，現在の深層学習全般でひとつの課題となるのは特徴選択である。特徴選択では組み合わせ爆発や計算時間の増大を避けるために，主に予測すべき目的変数に対して説明力のない／小さい特徴量 [10] をモデルから取り除いたり，与えられた特徴量を組み合わせることで従属変数の説明に最も適した潜在的な特徴量をモデルの中間層（隠れ層）に自動的に生成する。しかしながらどれほど複雑な変数が内部で生成されているのかに関しては，画像認識などの分野とは異なり社会科学分野，とりわけマーケティング分析においては議論されていない。これらの構造が複雑に多層にも及び非線形の回帰関数を挟んでいる点こそ，深層学習がブラックボックス的解析と呼ばれる所以でもある。

4.1.3　本研究の目的と構成

本研究では，小売企業が自社顧客の将来的な行動を予測するにあたって，顧客行動の「多様性」を考慮した場合にその予測精度が向上する可能性について論じている。マーケティングにおける顧客関係管理の諸研究では，既存顧客を維持することは新規顧客の獲得に比べて低コストであり，新規顧客獲得の戦略策定やコストを考慮すれば既存顧客の維持は企業にとって重要な課題である (Blattberg et al., 2008)。しかしながら，そのためには顧客行動をより詳細に把握し，分析に用いることが必須となる。本研究では，従来では得ることの難しかった顧客の購買などといった実店舗内での行動に加えて，実店舗に訪問していない場合でも顧客の店舗外での自社ブランドに対する情報収集の過程としてのスマートフォンアプリでの閲覧履歴や，オンラインストアでの購買情報など，幅広い日常的な行動の多様性を加味した解析を行うことで，行動予測の精度の改善につながる可能性について検証する。

本論文の構成は，本章以降 2 節において顧客行動のモデリングや深層学習を用いて解析を行っている既存研究に関するレビューを，3 節において本研究で着目する顧客行動を多様性変数という形で定義する。そして 4 節，5 節で実際に行った解析例を示し，最後に本研究の考察と今後の課題を示す。

4.2　解析 5：EC サイトにおける将来の訪問・購買の予測

4.2.1　解析の目的

まず解析 5 においては，EC サイト上での Web 閲覧データから抽出した購買頻度データを用いて予測を行う。本解析では国内の大手 EC サイトを対象として，サイト内で購買を行ったユーザを企業の自社顧客として扱い，彼らの将来の購買についての予測を行う。これは多様性変数が顧客の将来購買の予測にあたって有用であることを示すことを目的としており，実際の解析にあたっては第 3 章までと同様に投入する変数の組み合わせを変えた複数のモデルを作成し，結果を比較している。またその解析については深層学習とパラメトリックな統計モデルのそれぞれで実施し，その予測精度についても比較する。

また多様性変数の有用性の検証と同時に，深層学習においての特徴選択の精度の検証も実施する。第 3 章で網羅的に検証したように一般的な Web 行動の予測モデルにおいては多様性変数の投入により予測精度は改善すると考えられる。深層学習といえども多様性変数の投入によって予測精度が向上するのは同様であると考えられるが，一方で深層学習での解析では特徴選択により最適な潜在変数が生成され変数選択が不要であると言われる。つまり仮に深層学習での予測にあたり多様性変数のような複雑な変数を投入して予測精度が向上しないようであれば，同一もしくは類似した変数が内生的に考慮されてい

ると考えられる。

深層学習を扱った一般的な情報学分野における解析と異なり，マーケティングの研究分野においては消費者の行動における購買の量や次の訪問・購買までの時間・頻度といった量的な情報に関心を持って研究が行われることが一般的であるため，本研究では購買有無や訪問有無ではなく，これらを量的変数としたまま解析を行った。

4.2.2 使用データ

この解析で用いたデータは，解析 2，3 と同様のアクセスログである。このデータは実購買等のトランザクションデータは持たないものの，日本のインターネット利用の人口統計に基づき代表性を担保して抽出された 13000 人程度のユーザが PC で閲覧した Web サイトについて，アクセスされた日時・URL・ページ単位での滞在時間（秒数）・リファラ情報などがサイト横断的に収録された大規模な User-Centric Clickstream Data である。本解析においてはこのデータを特定の Web サイトに限定することにより，企業が得られる擬似的な Site-Centric Clickstream Data として扱う。

実際の解析では今回使用したパネルデータに参加している対象者のうち，1 年以上の期間にわたって対象の EC サイトにアクセスしていること，直近の購買間隔を算出するために累積で 2 回以上購買していることを満たす 4146 人である。購買の有無は EC サイトで実際に購買した後に表示されるサンキューページの URL へのアクセスをもって判断しており，誤操作や買い忘れを考慮するため 1 日に複数回発生した購買については同一の購買と判定している。

4.2.3 使用モデル

本解析では，深層ニューラルネットワーク系の代表的な手法として Feed-Forward Neural Network (FFNN) を用いており，入力層から 3 層の隠れ層と各層に 10 個の素子 [11] を経由して出力層へ向かう構成とした。学習方法にはベイズ正則化を用いており，これは Moré (1978) による Levenberg-Marquardt 法 [12] を元に，過学習を防ぐために損失関数に正則項を設定した手法である。利点として情報量基準を用いて汎化誤差の上昇前にモデルの学習を停止することができるため，過剰な精度の低下を防ぐことにより通常の誤差逆伝播モデルより頑健であるとされる (Burden and Winkler, 2009)。

本モデルで用いた m 個の入力 x_i とそれに対応する重み付け w_{ji}，素子ごとのバイアス b_j を持つある素子 h_j の出力は，

$$h_j = \Phi(b_j + \sum_{i=1}^{m} w_{ji} x_i) \tag{4.1}$$

$$\Phi(z) = \frac{2}{1 + e^{-2z}} - 1 \tag{4.2}$$

と表せる。ただし関数 Φ は今回活性化関数 [13] として用いている Hyperbolic tangent sigmoid function である。それぞれの解析では，2 つの目的変数「将来 1 ヶ月間の購買回数」，「将来の訪問期間」についての予測を行う。使用データから顧客ごとに最新 12 ヶ月間の Clickstream Data を取得し，その期間を最初の 11 ヶ月 (t_1) とそれ以降の最新 1 ヶ月間 (t_2) に分割した。顧客の将来の行動を予測するため，t_1 を既知の期間，そして t_2 を将来の未知の期間としている。まず「将来 1 ヶ月間の購買回数の予測」では，自社における顧客の将来の購買回数の予測を行うため，既知の学習データである 11 ヶ月間の購買データやその他の行動データから学習を行った上で，将来 1 ヶ月に何回の購買が行われたかを予測している。次に「将来の訪問期間の予測」では，既知の期間に観測された最後の訪問日時から数えて何日目に再度訪問したかを予測した。未知のデータとして扱っている将来 1 ヶ月間の間に複数回の訪問があった

場合には，既知期間から最も直近の 1 訪問のみを対象とした。

比較に用いるパラメトリックな統計モデルでの予測には，GLM による重回帰分析での最尤推定を行う。解析では目的変数として「累積購買回数」「将来の購買期間」を採用しているが，それぞれの予測にあたり一般的に利用されるモデルとしてポアソン分布による回帰モデルとワイブル分布による回帰モデルを用いた。予測精度の算出の際には，汎化誤差で評価する深層学習との公平性を期すため，統計モデルにおいても元データのうち 75 % のデータからモデリングし，作成したモデルに対して残り 25 % のテストデータを適用する形でクロスバリデーションを実施し，これを汎化誤差として比較している。

また多様性変数の有効性の評価にあたっては，これまでと同様に，説明変数として用意した基本変数，最頻値変数，多様性変数の 3 つの変数群の組み合わせにより複数のモデルを作成している。訪問時期と累積購買回数ともに予測値と実測値での平均平方二乗誤差（RMSE）を算出しモデル間で比較した際に，多様性変数が説明変数として用いられているモデルで最も RMSE が改善していることをもって多様性変数が有効であると判断する。

4.2.4　使用した変数

ここで本解析で用いた変数の一覧を表 4.1 に示す（但し解析 5 と解析 6 で用いた変数をともに示している）。基本変数，最頻値変数，多様性変数の分類については，前章までに示した解析で用いているものと基本的には同様である。

まず基本変数には過去の購買・Web サイトへのアクセス・モバイルアプリの使用の頻度情報などを始めとして，基本的な顧客情報を表しうる量的変数が主となっている。次に最頻値変数では，最もよく購買された商品カテゴリや最もよく来店した自社店舗など，顧客の嗜好を表す変数が中心となっている。最頻値変数は全て深層学習モデルに投入しやすいようダミー変数としている。また多様性変数には，どれほど幅広い商品カテゴリを購買したか，どれほど特定のカテゴリに依存しているか，といった質的な情報を多様性変数として集めている。

本来であれば最頻値変数の中に「t_1 期間に自社で最も購買された商品カテゴリ」や「t_1 期間に最も来店した自社店舗」についても投入すべきであったが，本研究で対象にしている企業の展開する店舗数が非常に多く，次元数が極めて増加するため今回は解析から除いた。

また一部の変数には解析 5，6 の一方でしか用いていないものがあるが，これらは各解析で異なるデータを使用していることから，データ上の都合として測定が不可能であったことを示す。例として純粋な Clickstream Data のみで解析を行っている解析 5 では価格に関する情報は得られず，解析 6 では Clickstream Data の都合からリファラ情報を得ることができなかった。

4.2.5　解析結果

将来の購買回数と訪問期間について，深層学習と GLM を用いて予測した結果をそれぞれ表 4.2，4.3 に示す。まず RMSE の値が最も改善したのは，深層学習を用いた際に基本変数に加えて多様性変数のみを説明変数として加えたものであった。解析結果からわかる特徴について述べると，深層学習と GLM 間での予測結果の比較では，いずれの場合においても常に深層学習の方が予測精度が改善している。また深層学習については購買回数の予測に比べて訪問期間を扱った場合には予測精度の改善幅が小さくなっている。これらの特徴については解析 6 での結果も加味してまとめにて考察する。

表 4.1 用意した変数一覧

変数名	変数使用 解析 5	変数使用 解析 6	説明
目的変数			
Buy_{t_2}	○	○	目的変数 1: t_2 期間の自社での累積購買回数
$Duration_{t_2}$	○	○	目的変数 2: t_2 期間中の最新の訪問間隔（日数・対数化）
基本変数			
Buy_{t_1}	○	○	t_1 期間の自社での累積購買回数
$Duration_{t_1}$	○	○	t_1 期間の自社での平均購買間隔（日数・対数化）
$Access_{t_1}$	○	○	t_1 期間中に自社へアクセスした平均間隔（日数・対数化）
$PriceDifference_{t_1}$	-	○	t_1 期間中に購買された商品価格の最大値と最小値の差
$BuyQuantity_{t_1}$	-	○	t_1 期間中の 1 取引での平均購買品目数
$CheckinRange_{t_1}$	-	○	t_1 期間中のチェックイン店舗の範囲
$AppBrowseProduct_{t_1}$	-	○	t_1 期間中にモバイルアプリで閲覧した商品数
$Action_{t_1}$			t_1 期間中の次のそれぞれの行動の回数
	○	-	- 1: サイト内での商品検索
	○	○	- 2: 自社 Web サイトへの流入
	-	○	- 3:モバイルアプリの起動
	-	○	- 4: 店舗への訪問
	-	○	- 5: アプリでの商品詳細の閲覧
最頻値変数			
$MostRef_{t_1Web}$	○	-	t_1 期間に最も使用した自社サイトへの流入経路
$MostTime_{t_1Web}$	○	○	t_1 期間に自社 Web サイトへ最もアクセスした時間帯
$MostWeekday_{t_1Web}$	○	○	t_1 期間に自社 Web サイトへ最もアクセスした曜日
$MostTime_{t_1App}$	-	○	t_1 期間に自社アプリを最も使用した時間帯
$MostWeekday_{t_1App}$	-	○	t_1 期間に自社アプリを最も使用した曜日
$MostTime_{t_1Visit}$	-	○	t_1 期間に自社店舗へ最も来店した時間帯
$MostWeekday_{t_1Visit}$	-	○	t_1 期間に自社店舗へ最も来店した曜日
多様性変数（HHI ／累積パーセント）			
$Referer_{t_1}$	○	-	t_1 期間の自社 Web サイトへの流入経路の多様性
$AccessTime_{t_1Web}$	○	○	t_1 期間の自社 Web サイトへアクセスした時間帯の多様性
$AccessWeekday_{t_1Web}$	○	○	t_1 期間の自社 Web サイトへアクセスした曜日の多様性
$ViewCategory_{t_1}$	○	○	t_1 期間に自社で閲覧した商品カテゴリの多様性
$PurchaseCategory_{t_1}$	-	○	t_1 期間に自社で購買された商品カテゴリの多様性
$VisitStore_{t_1}$	-	○	t_1 期間に来店した自社店舗の多様性
$AccessTime_{t_1App}$	-	○	t_1 期間の自社アプリを使用した時間帯の多様性
$AccessWeekday_{t_1App}$	-	○	t_1 期間の自社アプリを使用した曜日の多様性
$AccessTime_{t_1Visit}$	-	○	t_1 期間の自社店舗へ来店した時間帯の多様性
$AccessWeekday_{t_1Visit}$	-	○	t_1 期間の自社店舗へ来店した曜日の多様性

表 4.2 解析 5-1: 将来の購買回数 予測精度の比較（RMSE）

予測方法	基本	基本 + 最頻	基本 + 多様性	基本 + 最頻 + 多様性
深層学習	1.57	1.57	1.38	1.84
GLM	2.36	3.07	2.42	3.12

表 4.3 解析 5-2: 将来の訪問期間 予測精度の比較（RMSE）

予測方法	基本	基本 + 最頻	基本 + 多様性	基本 + 最頻 + 多様性
深層学習	1.16	1.17	1.14	1.17
GLM	2.19	2.19	2.19	2.19

4.3 解析 6：小売企業における将来の訪問・購買の予測

4.3.1 解析の目的

次に解析 6 では，プライベートブランドを扱う小売企業の収集した複数のデータを用いて同様の分析を行う。顧客のアクセスログや EC サイトでのトランザクションデータなどの Web 情報は一般的には消費者のオンライン上での行動の予測に活用されることが多い。しかしこのような顧客行動を継続的

に取得した場合，アクセスしやすい時間帯やよく閲覧している商品カテゴリ等，顧客の行動や趣向に関して店頭の POS データだけでは得ることの難しい多様な情報を得られることが予想され，これは実店舗での顧客行動においても広く有用である可能性が考えられる。そこで本解析では，アクセス履歴等の Web データから顧客ごとの行動パターンを発見し分析に考慮することが，Web 以外の実店舗での購買や訪問間隔等の行動を予測するにあたっても広く有用であることを示す。また解析 5 と同様に，参考として GLM を用いた予測も行うことで結果を比較している。

4.3.2　使用データ

本解析では，自社プライベートブランド商品を販売する小売企業から提供された各種データを使用した。このデータには顧客のデモグラフィック情報（年齢，性別，居住都道府県等），自社 EC ストアと自社店舗での購買履歴データ（受注 ID，受注日時，販売商品の JAN コード，商品ジャンル，価格，割引の有無，返品の有無等），アプリ内での行動履歴（アプリ起動，チェックイン店舗コード，閲覧商品等），自社 Web ページ内の閲覧履歴（閲覧 URL，欲しいなどのコメント等）の情報が含まれている。そしてこれらデータは，顧客がユーザ登録を行っている場合には，全て共通のユーザ ID によって紐付いており横断的に参照することができる。従って，多くの顧客について Web での商品情報の閲覧や，購買の有無に関わらない実店舗への訪問，さらには最終的な購買の有無など，オンライン・オフラインを問わず行動を時系列に従って継続的に追跡することが可能となっている。購買を伴わない店舗への訪問は，モバイルアプリによる店舗への「チェックイン機能」から判別が可能である。顧客はチェックインごとに特典ポイントを得られることから，顧客は来店した際に積極的にチェックインを行うインセンティブを持つ。

実際の解析では，この使用データからいくつかの条件に従ってユーザを抽出した。まず，分析対象者がユーザ登録を行っていること。これは将来的な訪問や購買の有無／回数を予測するにあたって，ユーザ ID の登録により継続的に購買情報が得られている必要があるためである。また，多面的な分析を行うため，Web アクセスやスマートフォンアプリとの連携が取れている顧客に絞って分析を行う。さらに，将来の行動の予測を行うにあたって十分に顧客行動の分析が行えるよう，Web 閲覧やアプリの使用を含めて少なくとも 1 年間以上継続的にデータを得られた顧客のみを扱う。また，得られた Web 閲覧データは，自社サイトに限定された Site-Centric Clickstream Data である。調査期間が 1 年以上でありながらアクセス回数が極端に少ない顧客については分析が困難であると判断し，対象から除外した。結果として分析対象の顧客は 56710 人となった。

使用した変数は 4.1 に示しているが，設定した説明変数として $CheckinRange_{t_1}$ では顧客がチェックインした全店舗の地理的な幅広さを計測している。例として顧客 k が訪問した店舗 i（$i \in N_k$）の緯度，経度がそれぞれ y_i，x_i の時，顧客ごとの訪問店舗の重心 $G_k = (\overline{x}, \overline{y})$ であり，重心と各店舗の距離 $length_i$ は

$$length_i = \sqrt{(\overline{x} - x_i)^2 + (\overline{y} - y_i)^2} \tag{4.3}$$

である。顧客の行動範囲としての訪問店舗の範囲がほど重心からの店舗までの距離の最大値が増加するため，顧客ごとに期間中に訪問した全店舗の $length$ を算出し，その最大値を $CheckinRange_{t_1}$ として設定することにより顧客の地理的な行動範囲の大きさを数値化している。

4.3.3　解析結果

表 4.4，4.5 に，解析 6 における「将来 1 ヶ月間の購買回数の予測」,「将来の訪問期間の予測」それぞれの予測精度の一覧を示した。結果としては原則的に解析 5 と同じ傾向にあり，深層学習で基本変数

表 4.4 解析 6-1: 将来の購買回数 予測精度の比較（RMSE）

予測方法	基本	基本 + 最頻	基本 + 多様性	基本 + 最頻 + 多様性
深層学習	3.10	3.78	2.84	4.54
GLM	4.19	4.20	4.10	4.32

表 4.5 解析 6-2: 将来の訪問期間 予測精度の比較（RMSE）

予測方法	基本	基本 + 最頻	基本 + 多様性	基本 + 最頻 + 多様性
深層学習	1.03	1.08	1.01	1.05
GLM	4.21	4.23	4.09	4.86

に加えて多様性変数のみを加えたモデルで RMSE が最も改善する結果となった。解析 5 と同様の点としては，深層学習について購買回数の予測に比べて訪問期間を扱った場合に予測精度の改善幅が小さくなっている点があるが，一方で異なる点として GLM で多様性変数を用いた場合に購買・訪問ともに予測精度が改善している。

4.4 まとめ

4.4.1 多様性変数の有用性

まずは 2 つの解析の結果から，非線形の予測手法としての深層学習と多様性変数の有用性について考察する。得られた解析の結果から，解析に使用するデータ，モデル，解析内容による予測精度の変化については次の 2 点にまとめることができる。特徴 1 として本章で行なったすべての解析において常に深層学習で基本変数と多様性変数を用いた場合に最も RMSE が改善した点，そして特徴 2 としていずれの解析でも深層学習では購買回数に比べて訪問間隔の予測時に多様性変数の改善幅が小さい点である。

まず特徴 1 に関してであるが，すべての解析について深層学習と GLM の間での予測精度の変化と，各同一モデル内で変数を変えた際の予測精度の変化について俯瞰すると，全体的に深層学習を用いた場合に RMSE は大きく改善する傾向にあり，その中でも基本変数と多様性のみを用いた場合が常に最良となっている。この結果から，深層学習で捉えられている現象について，より単純なモデルとしての GLM では捉えきれていない点があるといえる。また GLM では解析 5-1 など一部の結果において基本変数と多様性を用いたモデルよりもその他のモデルの方が予測精度が向上している場合があるが，いずれにしても深層学習の場合と比べて大幅に予測精度は悪化しており，あくまで現象を説明できていないモデル内での挙動であるといえる。

次に特徴 2 に関しては，企業にとっては訪問間隔よりも購買量がより重要であるが，購買と訪問という行動自体の差異に着目すると，Web サイトや実店舗といった形態にかかわらず購買は単なる訪問と比較すると意思決定と費用負担が発生するという点で異なっている。実店舗や Web サイトへの訪問では自社以外を含めて複数企業に様々に訪問し得るが，商品購買についてはある商品カテゴリについて特定の店舗で購買した場合に他の店舗でも購買するということは少ないと考えられる。

深層学習の結果からは，コストを伴わない訪問の間隔は最頻値変数という単なる個人の Web 等での活動量の大きさで説明しやすいが，購買回数のように意思決定と費用を伴う行為の予測には多様性変数を用いることで予測精度が高まるということは注目に値する。これは今後さらに発展的な研究を実施する意義があるが，多様性変数を用いることで消費者がより多様な情報源から探索しようとしているかど

うかについての情報など，自社以外での行動要素が多様性変数に含まれているということを示唆している。

さてここまでで各解析結果から多様性変数を用いたモデルの挙動について考察を行った。結果として将来の行動を予測するにあたって既に得られている顧客行動から多様性変数を算出することで予測精度の向上をはかることができた。一般的には顧客の行動についての情報を得るためには，アンケートによる情報収集をはじめとして追加的なコストや顧客への負担のかかるものも多い。そういった中で本研究ではモバイルアプリやオンラインでの閲覧履歴，実購買のデータなど，いずれも企業が通常の企業活動を行う上で得られるであろう顧客行動の実データのみから計測した多様性変数を解析に用いている。そのような点からも，本研究で提案する多様性変数は企業の実務においても十分に活用できるものであると考えられる。

4.4.2　多様性変数の比較検討

前章までに顧客の行動の幅広さを行動の多様性として解析に用いることを提案し本章でもそれらを用いて解析を実施したが，このような顧客行動の多様性を測定する指標としては第 2 章でも検討したように HHI の他にも所得の不平等を計測し数値化するジニ係数が存在している。そこで参考までに解析 5，6 の深層学習での予測について HHI とジニ係数で比較検討した結果についても検証するため，表 4.6 にそれぞれの指標を説明変数として用いた場合の RMSE を示した。いずれの解析についても HHI を用いた場合にジニ係数よりも RMSE が改善しており，本解析における多様性変数の測定には HHI が適した指標であることがわかる。

表 4.6　深層学習におけるジニ係数と HHI での RMSE の変化

使用変数	解析 5-1	解析 5-2	解析 6-1	解析 6-2
基本変数 + HHI	1.38	1.14	2.84	1.01
基本変数 + ジニ係数	1.57	1.18	2.92	2.50

4.4.3　深層学習を用いた予測モデルの構築について

本研究では深層学習での予測を行うにあたって多様性変数を用いることに加えて，線形モデルとしての GLM との予測精度の比較を行った。解析ではいずれも基本変数に加えて最頻値変数を用いた場合には常に予測精度が悪化している。これは従来マーケティングを行うにあたって有用であると考えられてきたような変数が，深層学習のにおいてはむしろノイズになってしまう可能性を示唆している。この点については次のような 2 つの要因が考えられる。

まず最初に「最もよく利用する店舗」や「最もよく購買する商品カテゴリ」などをダミー変数として投入する場合，展開している店舗数や取り扱う商品カテゴリの数によっては入力の次元数が非常に大きくなってしまう。次元数が過剰に大きくなる場合にはそれに呼応して十分なサンプル数が必要となるのはもちろんのこと，最適化に必要な計算量や所要時間についても大きく増加してしまう。つまりモデルが複雑化することから多様性変数のみで行なっていたような 3 層 10 素子の構造では現象が捉えきれなかった可能性がある。

またもうひとつの要因として，特徴選択によって隠れ層の内部で最頻値変数と類似した情報を持つ潜在変数が基本変数と多様性変数から生成された可能性もある。これに関しては多様性変数の投入によりそもそも予測精度が向上したことから，多様性変数のような複雑な変数を自動生成することは難しい

が，最頻値変数で説明できるような情報に関しては考慮ができていたとも考えられる。

いずれにせよ多様性変数を用いることで最頻値変数などの次元が過剰に大きくなる変数を用いずとも予測が可能であることは，ビッグデータ時代として予測モデルが莫大な規模になりがちな現代において少ない変数で効率的な予測が可能となることから，実務でのマーケティング分析などの産業応用にあたっても効率的である。

本研究における深層学習での予測には隠れ層を 3 層持つ FFNN を用いているが，この構成の設定にあたっては予測精度と解析の所要時間に関して予め網羅的な解析を行っている。参考までに表 4.7 に隠れ層の数と学習の所要時間，そして予測精度について示す。隠れ層を増やした場合には構造が過剰に複雑になってしまい予測精度が低下する，あるいは解析時間が莫大に増加する一方で予測精度には大きな変化が見られないなどの結果となっており，本研究では 3 層の FFNN を用いることが最適であると判断した。

表 4.7　解析 6-1 での隠れ層の数と予測精度，解析所要時間の関係

隠れ層	1 層	3 層	5 層	10 層
RMSE	17.35	2.84	3.41	2.85
所要時間	3 分	8 分	15 分	40 分

4.4.4　限界と今後の課題

今回は使用しているデータ上の都合から，解析 6 の小売店での購買データについては，Clickstream Data の分析にあたって通常であれば得られるはずのリファラ情報を分析に考慮することができなかった。EC サイトにおいてリファラ情報は顧客の自社外での行動に関して企業が得られる数少ない情報である。実際に企業で行われる分析ではサーバから得ることが可能であるため，リファラの多様性を分析に加えることで更に予測精度が向上する可能性がある。

また，今回分析対象とした企業ではモバイルアプリへのプロモーション情報の通知なども行っている。さらに Web 上では季節に応じたキャンペーンや特集ページなど，様々な企画が行われている。本研究では分析モデル上の限界からこのような情報については加味しなかった。しかしながらこのようなマーケティング情報の受信により来店までの期間や購買頻度が変化する可能性も考えられるため，今後はさらに幅広い情報を用いて分析を行う必要がある。またマーケティング情報としては，閲覧した商品の詳細情報から word2vec を用いて商品の情報を特徴量ベクトルとして加味することで，カテゴリよりもさらに詳細な情報を分析に取り入れることも可能であると考えられる。

注

[9] 顧客の認知から購買までのプロセスに，実店舗，モバイル，Web サイト等の複数チャネルで複合的に対応すること。

[10] 画像認識においてはある物体を認識するために必要な特徴 (色や形，大きさなど) を指し，ベクトルで記述される。より一般的にはある性質／概念を特徴づける変数。

[11] ニューラルネットワーク内の構成要素であり，ニューラルネットワークはこの演算素子を層状にした階層構造が互いに結合した構成となっている (萩原, 2006)。

[12] 最適化のための手法である最急降下法と Gauss-Newton 法を組み合わせた，非線形型関数の最小化の標準的な手法。

[13] 前層からの出力を重み付きの線形和として入力することで当該素子の出力を生成する関数。例としてステップ関数を用いると出力が $\{0, 1\}$ の二値になるなど，活性化関数を帰ることにより様々な状態を表現することが可能。

第 5 章

まとめ

5.1　各章で実施した解析の俯瞰

本研究では全体を通していくつもの解析を実施した。まずはすべての解析の目的と結果について表 5.1 に示す。

表 5.1　各解析の目的と結果

章	解析	目的
第 2 章	解析 1	- 識別性の向上による RFMC の拡張 - 拡張した RFMC と多様性変数の比較検討
第 3 章	解析 2	- ニュースサイトへの訪問における多様性の有効性の検証
	解析 3	- EC サイトでの購買における多様性変数の有効性の検証
	解析 4	- 個人の異質性を考慮した場合の多様性変数の有効性の検証 - 訪問と購買の同時モデリング - INLA による大規模データの効率的な解析例の提示
第 4 章	解析 5・6	- 深層学習での多様性変数の有効性の検証 - 特徴選択の精度の検証 - 深層学習と GLM の比較
	解析 6	- 実データにおける多様性の有効性の検証 - 大規模データの同時モデリング

まず解析 1 では，近年大きく注目されている CRM 手法である RFMC について，新たに定義された C 指標の課題について示し，その識別性を高める形で拡張することで購買予測に活用できることを示した。そして拡張した RFMC と多様性変数で購買予測を行うことにより，購買履歴情報を用いていない多様性変数のみのモデルで高い精度での解析を実施できることを示した。これは主に消費者の購買パターンの異質性を考慮できることに加えて，購買プロセスを考慮することで高い説明力が得られたためと考えられる。解析 2・3 では個別の行動についてより詳細にモデルを作成して比較を実施しているが，解析 2 ではウェブサービスにおける訪問行動について，解析 3 では EC サイトにおける購買行動について，いずれにおいても従来積極的に用いられている量的な変数に加えて多様性変数を用いることでよりよい結果が得られており，消費者が購買に至るまでの過程における異質性を考慮することで行動予測に貢献できることを示した。解析 4 については，INLA を用いた効率的な解析で，通常では実施の難しい

訪問と購買の同時モデリングを実務にも容易に応用可能な形で実施した。理論的背景でも述べた通り多様性変数では消費者の異質性を捉えることがひとつの目的とされているため消費者の異質性を考慮した場合に予測精度の向上は見られないことも考えられるが，結果として消費者の異質性を変量効果として投入した場合にもなお多様性変数を用いた場合でモデルの情報量基準が改善していることから，多様性変数の有用性については消費者の異質性の考慮以外にも存在していることがわかる。

さらに解析 5・6 では予測に深層学習を用いたが，いずれの場合にも基本変数に加えて多様性変数のみを加えた場合に予測精度が最良となった。解析 5・6 では多様性変数の有効性の検証に加えて，特徴選択の精度の検証も目的としていた。つまり深層学習について一般的に言われている，潜在変数として自動的に複雑な変数を生成できることから説明変数には単純な生データのみを投入すればよいという考え方について，実際には多様性変数の投入で汎化誤差が低下したことから多様性変数のような複雑な変数の生成は困難であるということがわかる。ビッグデータ活用の時代の流れの中で様々な変数を投入してブラックボックス化しがちな深層学習では，カテゴリカル変数である最頻値変数を説明変数として投入すると次元が極端に大きくなりモデルが複雑化してしまう。そういった中で従来よく用いられてきた最頻値変数を用いずに次元の少ない多様性変数でより良い精度を出せるという点で有用性は高いと考えられる。また解析 6 の結果より実行動での多様性を用いても予測精度が改善されたことから，Web データに限らず企業がマーケティング活動の中で得られる様々な顧客データから多様性変数を算出することで，消費者の行動を多面的に捉えることが可能である。

本研究で扱った多様性変数はいずれも企業が自社データから収集できる情報を用いたものである。なおかつこれまで一般的に用いられて来た最頻値変数のようなカテゴリカルな情報では次元数が莫大な数になってしまう大規模な解析においても，多様性変数のみを用いることで次元数を抑えながら効率的に予測モデルを作成可能である点を示した。本研究のようなマーケティング分野の学術研究においては実際の産業応用の可能性という観点からの貢献が重視されるが，本研究で提案する多様性変数や解析例に関しては，実務でのマーケティング分析の実施にあたっても解析時間の短縮や効率的なモデリングの実施につながると言える。

5.2 本研究の貢献

5.2.1 自社顧客の行動の多様性を用いる意義

本研究では消費者の行動を予測するにあたり，先行研究で広く用いられている説明変数での予測手法では捉えきれていなかった消費者の行動の異質性を捉えることを目的に，新たに消費者の行動の多様性を定義して考慮することにより，全ての解析で一貫して予測精度が改善する結果となった。特に，自社で保有する顧客といえども競合他社をどれほど利用しているかなどの自社外での情報を得ることは難しい。本研究では第 3，4 章ともに Web usage data を用いた解析にはリファラ情報の多様性を用いている。リファラは自社外での行動に関して得られる希少な情報であることから，その多様性を定義することで消費者がどれほど幅広い Web サイトを利用しているかといった情報を考慮できるようになったことから，競合他社での購買予測について高い精度での予測を実現した。このような自社外での行動の幅広さについて部分的にも把握できる情報はリファラ以外にも存在していると考えられることから，企業が保有する顧客行動に関するデータから幅広く多様性変数を定義して用いることにより，今後も予測精度を改善できると予想される。

5.2.2 深層学習に多様性を用いる意義

また本研究では深層学習の有用性に関しても検証を行った。あらゆる解析にあたり適用することで高い精度での予測が可能であるように喧伝されがちな深層学習であるが，そもそもビッグデータの活用と称して多様で大規模な種々のデータを投入することで解析を行うためには高い計算能力が必要となる。しかしながら迅速な PDCA に基づく意思決定にシームレスにデータ解析を組み込むためには，分析モデルが過剰に複雑になり解析の所要時間が伸びることは防がなければならず，そういった点からも効率的な解析が必要である。本研究では深層学習における特徴選択の精度の検証と称して本研究で提案している多様性変数を用いた少ない次元数での高精度な解析手法について示したが，これはモデルが膨大な規模になりがちなビッグデータ時代に用いるべきデータ解析としての目的に合致しており，今後の産業応用においても十分な可能性を持っているといえる。

5.3 今後の課題

まず本研究では，データ上の都合から企業の自社での購買履歴と競合他社での購買情報を同時に考慮した解析を行うことはできなかった。この点に関しては，ある消費者について 1 企業での購買履歴データと競合他社での購買履歴が紐づいた状態でのデータセットの入手がそもそも非常に困難であることから，今後もし入手できれば SOW の予測などの研究にあたっても大きなインプリケーションを持った研究を実施できることが予想される。またアクセス履歴データ以外の企業データを用いた場合には，アプリデータの活用幅が限定的であったことと，さらに近年の情報学・工学分野では位置情報データを活用した解析も盛んに実施されている。このように本研究で扱ったデータ以外にも様々な消費者の行動データが存在していることから，将来的にはこれらを組み合わせることでより消費者の生活を特定のチャネルに限定されず多面的に捉えた解析が可能となると予想される。

また web contents data としては商品情報以外にも商品を購買した消費者が自身で投稿する商品のレビュー文等も考えられる。本研究ではモデルが過剰に複雑になることを考慮して用いなかったが，インフルエンサー [14] 等の評価が重要視される現状から，他の消費者のレビューが購買行動に与える影響も今後は加味していきたい。

また本研究における訪問と購買の同時モデリングには GLMM としての INLA を用いたが，深層学習においても近年では単一の予測器で複数の従属変数を扱うような多変量の予測手法が用いられ始めていることから，深層学習の活用という観点でも今後より幅広い手法を用いた検証が必要である。

注

[14] 他の消費者に対して強い影響力を持つ消費者であり，一般消費者からその購買行動を参考にされることが多い。

謝辞

本研究の実施にあたり，指導教員の根本二郎教授ならびに前期課程時代の指導教員であります慶應義塾大学の星野崇宏教授には多大なご指導を賜るとともに，副指導教員の宮崎正也准教授，セミナー担当教員として山田基成教授からも大変有用な助言を頂戴しました。また各ゼミナールでの諸先輩方や後輩のみなさんにも様々な観点から勉強の機会をいただきました。さらに解析にあたっては株式会社ビデオリサーチインタラクティブ様をはじめとした各社様より大変貴重なデータをご提供いただきました。ここに御礼を申し上げます。

本研究は科学研究費補助金基盤 (B)26285151 および名古屋大学が実施する博士課程教育リーディングプログラム「PhD プロフェッショナル登龍門」の助成を受けています。

参考文献

Bengio, Yoshua, Pascal Lamblin, Dan Popovici, Hugo Larochelle et al. (2007) "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, Vol. 19, p. 153.

Berger, Paul D and Nada I Nasr (1998) "Customer lifetime value: Marketing models and applications," *Journal of interactive marketing*, Vol. 12, No. 1, pp. 17–30.

Blattberg, Robert C, Byung Do Kim, and A. Neslin Scott (2008) *Database Marketing: Analyzing and Managing Customers*: Springer.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992) "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, ACM.

Breiman, Leo (2001) "Random forests," *Machine learning*, Vol. 45, No. 1, pp. 5–32.

Büchner, Alex G and Maurice D Mulvenna (1998) "Discovering internet marketing intelligence through online analytical web usage mining," *ACM Sigmod Record*, Vol. 27, No. 4, pp. 54–61.

Bucklin, Randolph E, James M Lattin, Asim Ansari, Sunil Gupta, David Bell, Eloise Coupey, John DC Little, Carl Mela, Alan Montgomery, and Joel Steckel (2002) "Choice and the Internet: From clickstream to research stream," *Marketing Letters*, Vol. 13, No. 3, pp. 245–258.

Burden, Frank and Dave Winkler (2009) "Bayesian regularization of neural networks," *Artificial Neural Networks: Methods and Applications*, pp. 23–42.

Chen, Yuxin and Joel H Steckel (2012) "Modeling credit card share of wallet: Solving the incomplete information problem," *Journal of Marketing Research*, Vol. 49, No. 5, pp. 655–669.

Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava (1997) "Web mining: Information and pattern discovery on the world wide web," in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*, pp. 558–567, IEEE.

Davis, Peter and Eliana Garcés (2009) *Quantitative techniques for competition and antitrust analysis*: Princeton University Press.

Dembczynski, Krzysztof, Wojciech Kotlowski, and Dawid Weiss (2008) "Predicting ads click-through rate with decision rules," in *Workshop on targeting and ranking in online advertising*, Vol. 2008.

Du, Rex Yuxing, Wagner A Kamakura, and Carl F Mela (2007) "Size and share of customer wallet," *Journal of Marketing*, Vol. 71, No. 2, pp. 94–113.

Fong, Youyi, Håvard Rue, and Jon Wakefield (2010) "Bayesian inference for generalized linear mixed models," *Biostatistics*, Vol. 11, No. 3, pp. 397–412.

Glady, Nicolas and Christophe Croux (2009) "Predicting customer wallet without survey data," *Journal of Service Research*, Vol. 11, No. 3, pp. 219–231.

Goldberg, David, David Nichols, Brian M. Oki, and Douglas Terry (1992) "Using Collaborative Filtering to Weave an Information Tapestry," *Commun. ACM*, Vol. 35, No. 12, pp. 61–70, December, URL: `http://`

doi.acm.org/10.1145/138859.138867, DOI: http://dx.doi.org/10.1145/138859.138867.

Guo, Xu and Bradley P Carlin (2004) "Separate and joint modeling of longitudinal and event time data using standard computer packages," *The american statistician*, Vol. 58, No. 1, pp. 16–24.

Gupta, Sunil and Donald R Lehmann (2006) "Customer lifetime value and firm valuation," *Journal of Relationship Marketing*, Vol. 5, No. 2-3, pp. 87–110.

萩原克幸 (2006) 「ニューラルネットワークの基礎と理論的に重要な課題 (ニューラルネットワークの応用と今後の発展)」,『プラズマ・核融合学会誌』, 第 82 巻, 第 5 号, 282–286 頁.

石垣司・竹中毅・本村陽一 (2011) 「日常購買行動に関する大規模データの融合による顧客行動予測システム実サービス支援のためのカテゴリマイニング技術」,『人工知能学会論文誌』, 第 26 巻, 第 6 号, 670–681 頁.

伊藤浩二・西土祥子・山崎智章 (2016) 「NTT コミュニケーションズにおける AI を活用した事業変革への挑戦 (特集 NTT グループにおける AI の取り組み)」,『NTT 技術ジャーナル』, 第 28 巻, 第 2 号, 26-29 頁, URL : http://ci.nii.ac.jp/naid/40020732970/.

Jacoby, Jacob and Robert W Chestnut (1978) *Brand loyalty: Measurement and management*: John Wiley & Sons Incorporated.

Jang, Sungha, Ashutosh Prasad, and Brian T Ratchford (2016) "Consumer spending patterns across firms and categories: Application to the size-and share-of-wallet," *International Journal of Research in Marketing*, Vol. 33, No. 1, pp. 123–139.

Johnson, Eric J, Wendy W Moe, Peter S Fader, Steven Bellman, and Gerald L Lohse (2004) "On the depth and dynamics of online search behavior," *Management science*, Vol. 50, No. 3, pp. 299–308.

Jones, Thomas O, W Earl Sasser et al. (1995) "Why satisfied customers defect," *Harvard business review*, Vol. 73, No. 6, p. 88.

Kamakura, Wagner A and Michel Wedel (1997) "Statistical data fusion for cross-tabulation," *Journal of Marketing Research*, pp. 485–498.

Kim, Jun B, Paulo Albuquerque, and Bart J Bronnenberg (2010) "Online demand under limited consumer search," *Marketing science*, Vol. 29, No. 6, pp. 1001–1023.

LeCun, Yann, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel (1989) "Backpropagation applied to handwritten zip code recognition," *Neural computation*, Vol. 1, No. 4, pp. 541–551.

Liaw, Andy and Matthew Wiener (2002) "Classification and regression by randomForest," *R news*, Vol. 2, No. 3, pp. 18–22.

Lopes, Noel and Bernardete Ribeiro (2012) "Handling missing values via a neural selective input model," *Neural Network World*, Vol. 22, No. 4, p. 357.

Mägi, Anne W (2003) "Share of wallet in retailing: the effects of customer satisfaction, loyalty cards and shopper characteristics," *Journal of Retailing*, Vol. 79, No. 2, pp. 97–106.

Manchanda, Puneet, Jean-Pierre Dubé, Khim Yong Goh, and Pradeep K Chintagunta (2006) "The effect of banner advertising on internet purchasing," *Journal of Marketing Research*, Vol. 43, No. 1, pp. 98–108.

Martino, Sara and Håvard Rue (2009) "Implementing approximate Bayesian inference using Integrated Nested Laplace Approximation: A manual for the inla program," *Department of Mathematical Sciences, NTNU, Norway*.

松村直樹・和泉潔・山田健太 (2016) 「POS データに基づく欠品時の顧客行動を考慮した小売店舗の購買シミュレーション」,『人工知能学会論文誌』, 第 0 号.

McCulloch, Charles E and John M Neuhaus (2001) *Generalized linear mixed models*: Wiley Online Library.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013) "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119.

Moe, Wendy W (2003) "Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream," *Journal of consumer psychology*, Vol. 13, No. 1, pp. 29–39.

Montgomery, Alan L, Shibo Li, Kannan Srinivasan, and John C Liechty (2004) "Modeling online browsing and path analysis using clickstream data," *Marketing science*, Vol. 23, No. 4, pp. 579–595.

Moré, Jorge J (1978) "The Levenberg-Marquardt algorithm: implementation and theory," in *Numerical analysis*: Springer, pp. 105–116.

新美潤一郎・星野崇宏 (2015)「ユーザ別アクセス・パターン情報の多様性を用いた顧客行動の予測とモデリング」,『応用統計学』, 第 44 巻, 第 3 号.

Nottorf, Florian (2014) "Modeling the clickstream across multiple online advertising channels using a binary logit with Bayesian mixture of normals," *Electronic Commerce Research and Applications*, Vol. 13, No. 1, pp. 45–55.

Park, Chang Hee and Young-Hoon Park (2016) "Investigating purchase conversion by uncovering online visit patterns," *Marketing Science*, Vol. 35, No. 6, pp. 894–914.

Park, Deuk Hee, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim (2012) "A literature review and classification of recommender systems research," *Expert Systems with Applications*, Vol. 39, No. 11, pp. 10059–10072.

Park, Young-Hoon and Peter S Fader (2004) "Modeling browsing behavior at multiple websites," *Marketing Science*, Vol. 23, No. 3, pp. 280–303.

Platzer, Michael and Thomas Reutterer (2016) "Ticking away the moments: Timing regularity helps to better predict customer activity," *Marketing Science*, Vol. 35, No. 5, pp. 779–799.

Van den Poel, Dirk and Wouter Buckinx (2005) "Predicting online-purchasing behaviour," *European Journal of Operational Research*, Vol. 166, No. 2, pp. 557–575.

Rue, Havard, Sara Martino, Finn Lindgren, D Simpson, A Riebler, and ET Krainski (2009) "INLA: functions which allow to perform a full Bayesian analysis of structured additive models using Integrated Nested Laplace Approximation," *R package version 0.0*.

Salakhutdinov, Ruslan and Geoffrey Hinton (2009) "Deep boltzmann machines," in *Artificial Intelligence and Statistics*, pp. 448–455.

Sharma, Anuj, Dr Panigrahi, and Prabin Kumar (2013) "A neural network based approach for predicting customer churn in cellular network services," *arXiv preprint arXiv:1309.3945*.

Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan (2000) "Web usage mining: Discovery and applications of usage patterns from web data," *Acm Sigkdd Explorations Newsletter*, Vol. 1, No. 2, pp. 12–23.

Vieira, Armando (2015) "Predicting online user behaviour using deep learning algorithms," *CoRR*, Vol. abs/1511.06247, URL: `http://arxiv.org/abs/1511.06247`.

Weymark, John A (1981) "Generalized Gini inequality indices," *Mathematical Social Sciences*, Vol. 1, No. 4, pp. 409–430.

Zhang, Yao (2013) "New clumpiness measures and their application in customer evaluation," *Journal of*

Applied Statistics, Vol. 40, No. 11, pp. 2533-2548.

Zhang, Yao, Eric T Bradlow, and Dylan S Small (2014) "Predicting customer value using clumpiness: From RFM to RFMC," *Marketing Science*, Vol. 34, No. 2, pp. 195–208.

Zhao, Yu, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren (2005) "Customer churn prediction using improved one-class support vector machine," in *Advanced data mining and applications*: Springer, pp. 300–306.

岡谷貴之 (2015)『深層学習』，講談社.

今井亮一・工藤教孝・佐々木勝・清水崇 (2007)「サーチ理論－分権的取引の経済学」.

勝又壮太郎 (2010)「ウェブコンテンツ評価モデルの比較と活用-滞在時間とページビュー双方の観点から-」,『マーケティングサイエンス』，第 18 巻，第 1 号，1–27 頁.

新美潤一郎・星野崇宏 (2017)「Deep Boltzmann Machine を用いたデータ融合手法の提案」,『2017 年度人工知能学会全国大会（第 31 回）』.

星野崇宏 (2009)「調査観察データの統計科学: 因果推論・選択バイアス・データ融合」.

——— (2013)「継続時間と離散選択の同時分析のための変量効果モデルとその選択バイアス補正—Web ログデータからの潜在顧客への広告販促戦略立案—」,『日本統計学会誌』, 第 43 巻, 第 1 号, 41-58 頁.

中山雄司 (2016)「顧客関係管理研究の新動向: 来店/購買間隔の不均一性を測るクランピネス指標」,『甲南経営研究』，第 57 巻，第 2 号，161–181 頁.

猪狩良介・星野崇宏 (2014)「階層ベイズ動的サンプル・セレクションモデルによる Web サイトへの誘導とサイト閲覧行動の同時分析」,『日本統計学会誌』，第 43 巻，第 2 号，157–183 頁.

里村卓也 (2007)『EC サイトの閲覧・購買行動のモデル分析』，千倉書房.