〈博士学位請求論文〉

機械学習に基づくイーコマースにおける顧客流出予測に関する研究

Research on Consumer Churn Prediction in E-commerce Based on Machine Learning

カ　コウ　　ケン ジョウ

夏侯　賢城

(20DC02)

2023 年 1 月 10 日

大阪産業大学大学院　経営・流通学研究科

博士後期課程　経営・流通専攻

原田良雄教授研究室

〈博士学位請求論文〉

機械学習に基づくイーコマースにおける顧客流出予測に関する研究

Research on Consumer Churn Prediction in E-commerce Based on Machine Learning

カコウ　ケンジョウ
夏侯　賢城
学籍番号 20DC02
大阪産業大学大学院 経営・流通学研究科
博士後期課程 経営・流通専攻
原田良雄教授研究室

2023 年 1 月 10 日

〈論文の要旨〉

機械学習に基づくイーコマースにおける顧客流出予測に関する研究

Research on Consumer Churn Prediction in E-commerce Based on Machine Learning

氏　　名：　夏侯 賢城
指導教員：　原田 良雄

インターネットの普及とIT 技術の発達と共にオンライン購買が増え始めた。インターネット上で商品を売買するEC サイト市場の拡大を背景に、市場内での顧客獲得競争が起きている。顧客は企業の最も重要な資産の一つとして、企業の市場競争力と業績の向上に非常に重要な役割を果たしている。しかし、激しい市場競争の中で顧客は多くの製品やサービスプロバイダから容易に選択することができ、企業では顧客の流出が発生する可能性がある。一方、パレート法則によると、企業の80%の利益は20%の顧客から生み出す。これで、企業が市場の優位性を維持するために、既存の顧客資源をどのように流出を避けるか、そして、既存の顧客をどのように利用するはすでに企業が直面している重要な問題となっている。

顧客関係管理（Customer Relationship Management：CRM）とは情報技術の援用、すなわちデータを用いて顧客を識別し、顧客に対して顧客ごとに合ったダイレクト・メッセージを配信するなどで満足させ、顧客維持につなげていくものとされて、顧客と企業とのあらゆる接点や接触履歴を全て管理し、それをもとに顧客との関係性を深めて維持することで顧客満足度や顧客のロイヤルティを向上させ、収益の拡大を目指すマネジメント手法である。顧客ロイヤルティを高めるためには顧客データの分析が不可欠である。顧客関係管理に関する先行研究によると、顧客流出予測技術を用いて流出する可能性がある顧客を識別して、予測結果に基づいてマーケティング戦略を改善し、既存の顧客を維持することで業績損失を効果的に防止することができる。そのため、実務と理論の両側面から見ると、企業と顧客の関係性を構築するCRMシステムが重視しなければならない。

ほとんどのショッピングサイトでは、 Webサーバとデータベースサーバが連携して動作している。データベースサーバには、顧客情報、商品情報、在庫情報、販売情報などが保管され、Webサイトの訪問者が入力した情報が、リアルタイムにデータベースに書き込まれて更新される。一方、ビッグデータとクラウドコンピューティングなど情報技術の発達によって、消費者の購買データなどの消費者行動データは容易に収集、保存できるようになった。CRMは2000年頃から研究対象となったものの、データ

を活用するマーケティング手法である。しかし、EC企業にとってもCRMにおける顧客の購買行動データの分析と顧客流出予測に関する研究はあまり進んでいない。多様なデータに顧客流出予測手法が完全に対応されないため、各業界が多くのデータを保有しているにも関わらず適切な活用方法がまだ確立されていない。そのため、本文はCRMに関する先行研究文献の調査により、機械学習に基づくイーコマースにおける新たな顧客流出予測モデルを提案した。

近年では機械学習が実務でも浸透し始めたが、顧客流出予測に関する研究の多くは電信、銀行、小売などの業界に集中して、B2Cイーコマースの顧客流出予測に関する研究は少ない。さらに従来の研究はPOSデータに基づいて消費者セグメンテーションを行うこと多かったが、機械学習を用いてネットビジネスユーザーデータにおける顧客流出予測モデルの構築と消費者行動の予測が少ない。本研究の目的は機械学習AdaBoostを採用して「セグメンテーション・ファースト」（Segmentation-First）モデルの有効性を検討し、CRMの観点からB2Cイーコマースのために新しい顧客セグメンテーションと顧客流出予測の手法を提供し、人工知能技術を利用して顧客関係管理を行う。最終の目的は企業のために合理的で有効なマーケティング戦略を制定し、経営のコストを下げることである。

本研究では、イーコマースの発展とEC顧客購買行動の多様性及びビッグデータ情報技術の背景に基づいて、オンライン購買行動と購買時間の特徴を分析し、B2Cイーコマースのマーケティングの顧客セグメンテーションと顧客流出予測の学術論文におけるセグメンテーションと予測実施の手法を総括した上で、RF-PACVモデルと「セグメンテーション・ファースト」モデルを提案した。多種類の機械学習アルゴリズムを用いた顧客流出予測の精度を考察して、その後、各種の機械学習アルゴリズムの予測性能を比較し、B2Cイーコマースデータ分析における機械学習の有効性・実用性を議論する。

本論文は7つの章で構成されている。はじめに序論として本研究の背景と目的を説明し、イーコマース環境における消費者のオンラインショッピングの発展傾向とビッグデータ情報技術がイーコマース企業の管理に与える影響を分析した。イーコマース企業が競争しているため、オンラインショッピングの顧客はある会社から別の会社に転換しやすく、このような現状は企業の顧客の流出を招く可能性がある。そこで、本論文では顧客の流出予測の研究を展開する。

第Ⅰ章「顧客セグメンテーションとCRMに関する先行研究」では、従来の研究はRFMモデル分析に基づいて顧客セグメンテーションを行うことが多かったが、顧客の購買行動、購買意識と意思決定などの情報を把握することが難しい。このため、本章はKotler と Keller (2008) が提示した顧客セグメンテーションの基準に基づき、各種顧客セグメンテーション方法の特徴を詳しく分析し、B2Cイーコマース顧客セグメン

テーション方法を提示し顧客セグメンテーションの研究構想を明示する。一方、近年、企業と顧客との関係性の構築を重要視するマーケティングが注目されている。本章で、CRMの概念からCRM におけるデータと分析方法、およびCRM におけるデータの必要性とデータの位置づけを説き、また、CRMの先行研究について説明している。

第Ⅱ章「顧客流出予測に関する先行研究」では、まず機械学習の基本概念、機械学習におけるデータ活用の流れ、および基本手法とその応用を述べる。また、顧客流出は企業の経営業績に直接影響し、流出する可能性がある顧客をどのようにして残すかということに注目する。そのため、本章では先行研究の現状を分析し、電信、銀行、小売などの業界に顧客消費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の先行研究についてレビューを行うとともに、顧客流出予測の手法について述べる。

次に第Ⅲ章「顧客流出予測モデリングの基礎理論」では、予測モデルを構築するためには、多種類の機械学習アルゴリズムが必要になる。そこで、本章は既存の顧客流出予測アルゴリズムの理論とその方法を述べる。各種のアルゴリズムの手順をそれぞれ紹介する。本章で用いたアルゴリズムは、k平均法（k-means）、ランダムフォレスト（Random Forest：RF）、ロジスティック回帰（Logistic Regression：LR）、サポートベークタマシン（Support Vector Machine：SVM）、誤差逆伝播法によるニューラルネットワーク（Back Propagation Neural Network：BPNN）、アダブースト（AdaBoost）である。これらのアルゴリズムは第Ⅳ章の実証研究に用いられ、その後、これらの性能とその応用についても考察する

次に第Ⅳ章「顧客流出予測の実証研究」では、B2Cのオンライン購入行動と購入時間は他の購入方式と大きく異なるためである。本章は顧客流出予測に関する先行研究に基づき、実証研究を展開する。本研究はAlibaba Cloud Tianchiプラットフォームからデータベースからのデータセットをもとに原始データセットとして用いて、顧客流出予測モデル構築には、それぞれロジスティック回帰、サポートベクターマシン、誤差逆伝播法によるニューラルネットワークとアダブーストの4種類の予測アルゴリズムを用いた。そして、この4つのアルゴリズムの予測性能を比較する。これらの結果は、企業管理者または企業の顧客関係管理のための顧客保持ポリシーの策定に参考になる。

第Ⅴ章「考察」では、顧客セグメンテーションの結果に基づいて、顧客セグメンテーション後の顧客群の消費行動の特徴を検討する。また、4種類のアルゴリズムの予測モデルの結果を基に、予測モデルの有効性・有用性と理論価値を検討する。

最後に結論では、顧客セグメンテーションと各モデルの予測性能の結果について、顧客セグメンテーション方法と予測モデルがイーコマース企業のマーケティング戦略の制定に与える影響を議論する。実証研究の結果についての考察に基づき、本研究の学術的な貢献とB2Cイーコマース企業の顧客関係管理の現実的意義についてまとめ、今後の課題を述べる。

目　次

表 目 次

図目次

序論

1　研究の背景

IT 技術の発達によりインターネット上のEC（Electronic Commerce: イーコマース）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。ネットビジネスの規模が年々拡大し、「令和2年度　年次経済財政報告」によると、消費者との関連が深いBtoC及びCtoC市場も年々拡大しており、2019年における両市場を合わせて21.1兆円となっている[1]。ネットショッピングはすでに普及して、市場の変化に伴って企業競争も深刻化していく。インターネット上で商品を売買するイーコマースサイト市場の拡大を背景に、市場内での顧客獲得競争が起きている。顧客は企業の最も重要な資産のひとつであり、企業の市場競争力と業績の向上に非常に重要な役割を果たしている(Bi、2019)。市場競争の中で、顧客は多くの製品やサービスプロバイダから容易に選択することができる、そのため、企業では顧客の流出が発生する可能性がある(Maria　et al.、2017)。研究によると、新規顧客を開発するコストは、既存顧客を残すコストよりも高いことが多い(Roberts、2000)。企業と顧客が長期にわたって良好な関係を維持すれば、企業は既存の顧客からより多くの利益を獲得し、顧客保持率が5%増加するごとに、企業の純現在価値は25%-95%増加する(Reichheld et al.、1990)。顧客流出率が5%減少すると、企業の平均利益率は25%-85%増加する(Jones et al.、1998; Nie et al.、2011)。

顧客関係管理（Customer Relationship Management：CRM）[2]とは情報技術の援用、すなわちデータを用いて顧客を識別し、顧客に対して顧客ごとに合ったダイレクト・メッセージを配信するなどで満足させ、顧客維持につなげていくものとされて、顧客と企業とのあらゆる接点や接触履歴を全て管理し、それをもとに顧客との関係性を深めて維持することで顧客満足度や顧客のロイヤルティを向上させ、収益の拡大を目指すマネジメント手法である。この顧客ロイヤルティを高めるためには顧客データの分析が不可欠である。そのため、実務と理論の両側面から、企業と顧客との関係性の構築を重要視するCRMマーケティングが注目されている。また、パレート法則[3]によると、企業の80%の利益は20%の顧客から生み出す。市場の優位性を維持するために、既存の

---

[1] 令和2年度 年次経済財政報告(経済財政政策担当大臣報告)― コロナ危機：日本経済変革のラストチャンス― 第4章 デジタル化による消費の変化とIT 投資の課題. p.172。

[2] CRM という概念は、1990 年代半ばに IT 業界で認識され始め、主としてコンサルティング・ファームによって 90 年代後半にはその概念が普及され始めた。また、2000 年頃からは実務家間だけにとどまらずに研究者間でも本格的に研究対象となった。この CRM の背景には大きく 2 つの潮流があり、それがリレーションシップ・マーケティングとデータベース・マーケティングである。藪野祥太、「データ視点からの CRM（顧客関係管理）の再考」、経営研究 第 71 巻第 3 号、2020、p.87-107。

[3] パレートの法則は、イタリアの経済学者ヴィルフレド・パレートが提唱した法則で、80:20 の法則とも呼ばれる。

顧客資源をどのように利用し、既存の顧客の流出を避けるかはすでに企業が直面している重要な問題となっている(Gordini and Veglio、2017)。顧客関係管理に関する先行研究によると、顧客流出予測技術を用いて流出する可能性のある顧客を識別し、予測結果に基づいてマーケティング戦略を改善し、既存の顧客を維持することで業績損失を効果的に防止することができる。

ほとんどのショッピングサイトでは、Webサーバとデータベースサーバが連携して動作している。データベースサーバには、顧客情報、商品情報、在庫情報、販売情報などが保管され、Webサイトの訪問者が入力した情報が、リアルタイムにデータベースに書き込まれて更新される。一方、ビッグデータとクラウドコンピューティングなど情報技術の発達によって、消費者の購買データなどの消費者行動データは容易に収集、保存できるようになった。CRMは2000年頃から研究対象となっているが、データを活用するマーケティング手法である。しかし、EC企業にとってもCRMにおける顧客の購買行動データの分析と顧客流出予測に関するの研究はあまり進んでいない。多様なデータに顧客流出予測手法が完全に対応していないため、各業界が多くのデータを保有しているにも関わらず適切な活用方法がまた確立されていない。

一方、マーケティング･マネジメントにおいて、こういった多様化に対応するためには、消費者の行動や嗜好を理解し、究極的には個々の消費者のニーズや問題解決に適切な財やサービスを提供しなければならない。Kotler と Keller[4]は、このマーケティングの目的を達成するためには、マーケティング･マネジメントが発生すると述べている。彼らはマーケティング･マネジメントを「ターゲットとなる市場を選択し、優れた顧客価値を創造･提供･伝達することにより、顧客を獲得･維持･育成していく技術と科学」と定義している。STPおよび4P[5](Segmentation：セグメンテーション, Targeting：ターゲティング, Positioning：ポジショニング, Product：製品, Price：価格, Place：流通チャネル, Promotion：コミュニケーション)に代表されるマーケティング･プロセスは、マーケティング活動を円滑にかつ効率的に進めるためのフレームワークであり、特にセグメンテーションはその第一歩である。つまり、企業がマーケティングの内部統制要因である4Pの戦略を決定する前段階として、適切なセグメンテーションを通して、自社のターゲットを明らかにし、自社の戦略ポジショニングを決定することが重要となる。セグメンテーションでは、購買行動において似通っている顧客層の共通するニーズに着目し、市場を意味がある集団に分類する。その後、

[4] フィリップ・コトラー (Philip Kotler) 教授は、「マーケティングの神様」、「近代マーケティングの父」などと呼ばれるマーケティング界の第一人者である。彼の本は、世界中の大学や企業などマーケティングを学ぶ人たちに読まれ、大学教授、企業のエグゼクティブでも、フィリップ・コトラーの『マーケティング・マネジメント』がバイブルという人は多い。コトラー氏の STP 理論については、『Marketing Management, 15th ed.; Pearson Education Ltd.: Edinburgh, UK, 2016』などを参照。

[5] 4P は、統合型マーケティング、これは別名マーケティング・ミックスと呼ばれる。Product、Price、Place、Promotion の 4P を組み合わせて、顧客に製品やサービスといった[価値]を届けるマーケティングのことである。

ターゲットとする集団に対して経営資源を集中投下することで、効率的かつ効果的なマーケティング戦略を可能にする。

イーコマース消費者のショッピング行動データには顧客の問い合わせ、消費者の好み、顧客の購入頻度、購入日付、受注情報などがある。ネットビジネスにおいて顧客を理解するため、適切な顧客セグメンテーションが欠かせない。例えば、中村ら (2011) はPOSデータにおいて顧客セグメンテーションの諸手法と効果を説明した。彼らは、顧客セグメンテーションの手法を教師なしデータによるセグメンテーション手法と教師付データによるセグメンテーション手法という二つの手法に分けている。また、佐藤ら (2017) はロジスティック回帰分析によりゴルフのECサイトにおけるリピート顧客の特徴を分析した。しかし、従来の研究はPOSデータに基づいて消費者セグメンテーションを行う、機械学習を用いてネットビジネスユーザーデータにおける消費者行動の予測が少ない。

2 研究の目的と意義

近年では機械学習が実務でも浸透し始めたが、顧客流出予測に関する研究の多くは電信、銀行、小売などの業界に集中して、B2Cイーコマースの顧客流出予測に関する研究は少ない。さらに従来の研究はPOSデータに基づいて消費者セグメンテーションを行うことが多かったが、機械学習を用いてネットビジネスユーザーデータにおける顧客流出予測モデルの構築と消費者行動の予測が少ない。本研究の目的はAdaBoostを採用して「セグメンテーション・ファースト」モデルの有効性を検討し、CRMの観点からB2Cイーコマースのために新しい顧客セグメンテーションと顧客流出予測の手法を提供し、人工知能技術を利用して顧客関係管理を行う。最終的目的は企業のために合理的で有効なマーケティング戦略を策定し、経営コストを下げることである。

3 研究の方法

本研究は基本的な文献調査を通して理論の枠組みを提示し、アリババグループの「TIANCHI天池」ビッグデータプラットフォーム[6]からユーザーデータを調べ、MySQL[7]をデータベースとしてデータを格納され、統計解析ソフトウェアSPSSを使ってデータを整理する。そして、k-meansアルゴリズムを採用して顧客をセグメントし、Random Forestアルゴリズムを利用して特徴を選択する。最後にPython[8]によりデータの抽

[6] ユーザーデータについては、Alibaba Tianchi Cloudプラットフォームが公開した科学研究とビッグデータコンテスト用のデータセットである。データサービスについては、Available online: https://tianchi.aliyun.com/datase を参照されたい。
[7] MySQLは、スウェーデンのMySQL AB社によって開発されているリレーショナルデータベース製品。MySQLについては、『MySQL徹底入門 第3版、日本MySQLユーザー会 著、2011、翔泳社』を参照。
[8] Python（パイソン）はインタープリタ型の高水準汎用プログラミング言語である。Pythonについては、『Python3 プログラミング徹底入門、マーク·サマーフィールド 著、長尾高弘 訳、2009』を参照。

出・可視化・分析を行って、機械学習のロジスティック回帰(Logistic Regression：LR)、サポートベクターマシン(Support Vector Machine：SVM)、誤差逆伝播法によるニューラルネットワーク(Back Propagation Neural Network：BPNN)とアダブースト(AdaBoost)アルゴリズムを用いて流出予測モデルを構築し、実証研究を行い、実験結果について考察を行った。

4　本論文の構成

本研究では、イーコマースの発展とイーコマース顧客購買行動の多様性及びビッグデータ情報技術の背景に基づいて、オンライン購買行動と購買時間の特徴を分析し、B2Cイーコマースのマーケティングの顧客セグメンテーションと顧客流出予測の学術論文におけるセグメンテーションと予測実施の手法を総括した上で、RF-PACVモデルと「セグメンテーション・ファースト」モデルを提案した。多種類の機械学習アルゴリズムを用いた顧客流出予測の精度を考察して、その後、ロジスティック回帰、サポートベークタマシン、BPニューラルネットワーク、AdaBoostの予測性能を比較し、B2Cイーコマースデータ分析における機械学習の有効性・有用性を議論する。

本論文は6つの章で構成されている。はじめに序論として本研究の背景と目的を説明し、イーコマース環境における消費者のオンラインショッピングの発展傾向とビッグデータ情報技術がイーコマース企業の管理に与える影響を分析した。イーコマース企業が競争しているため、オンラインショッピングの顧客はある会社から別の会社に転換しやすく、このような現状は企業の顧客の流出を招く可能性がある。そこで、本論文では顧客の流出予測の研究を展開する。

第Ⅰ章「顧客セグメンテーションに関する先行研究」では、従来の研究はRFMモデル分析に基づいて顧客セグメンテーションを行うことが多かったが、顧客の購買行動、購買意識と意思決定などの情報を把握することが難しい。このため、本章はKotlerとKeller (2008) が提示した顧客セグメンテーションの基準に基づき、各種顧客セグメンテーション方法の特徴を詳しく分析し、B2Cイーコマース顧客セグメンテーション方法を提示し顧客セグメンテーションの研究構想を明示する。一方、近年、企業と顧客との関係性の構築を重要視するマーケティングが注目されている。本章で、CRM の概念からCRMにおけるデータと分析方法、およびCRM におけるデータの必要性とデータの位置づけを説き、また、CRMの先行研究について説明している。

第Ⅱ章「顧客流出予測に関する先行研究」では、まず機械学習の基本概念、機械学習におけるデータ活用の流れ、および機械学習の基本手法とその応用を述べる。また、顧客流出は企業の経営業績に直接影響し、流出する可能性がある顧客をどのようにして残すかということに注目する。そのため、本章では先行研究の現状を分析し、電信、

銀行、小売などの業界に顧客消費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の先行研究についてレビューを行うとともに、顧客流出予測の手法について述べる。

次に第Ⅲ章「顧客流出予測モデリングの基礎理論」では、予測モデルを構築するためには、多種類の機械学習アルゴリズムが必要になる。そこで、本章は 既存の顧客流出予測アルゴリズムの理論とその方法を述べる。各種のアルゴリズムの手順をそれぞれ紹介する。本章で用いたアルゴリズムは、k-means、ランダムフォレスト、ロジスティック回帰、サポートベークタマシン、BPニューラルネットワーク、AdaBoostである。これらのアルゴリズムは第Ⅳ章の実証研究に用いられ、その後、これらの性能とその応用についても考察する。

次に第Ⅳ章「顧客流出予測の実証研究」では、B2Cのオンライン購入行為と購入時間は他の購入方式と大きく異なるためである。本章は顧客流出予測に関する先行研究に基づき、実証研究を展開する。本研究はAlibaba Cloud Tianchiプラットフォームのデータベースを原始データセットとして用いて、4種類の予測アルゴリズムを採用し、それぞれロジスティック回帰(Logistic Regression：LR)、サポートベクターマシン(Support Vector Machine：SVM)、誤差逆伝播法によるニューラルネットワーク(Back Propagation Neural Network：BPNN)とAdaBoostアルゴリズムの4種類の予測アルゴリズムを用いた。そして、この4つのアルゴリズムの予測性能を比較する。これらの結果は、企業管理者または企業の顧客関係管理のための顧客保持ポリシーの策定に参考になる。

第Ⅴ章考察では、顧客セグメンテーションの結果に基づいて、顧客セグメンテーション後の顧客群の消費行動の特徴を検討する。また、4種類のアルゴリズムの予測モデルの結果を基に、予測モデルの有効性・有用性と理論価値を検討する。

最後に結論では、顧客セグメンテーションと各モデルの予測性能の結果について、顧客セグメンテーション方法と予測モデルがイーコマース企業のマーケティング戦略の制定に与える影響を議論する。実証研究の結果についての考察に基づき、本研究の学術的な貢献とB2Cイーコマース企業の顧客関係管理の現実的意義についてまとめ、今後の課題を述べる。

なお、第Ⅳ章の「顧客セグメンテーションと顧客クラスタ」は、人工知能学会全国大会に投稿、掲載された論文(夏侯・原田、 2022)を加筆・修正し、第Ⅱ章と第Ⅳ章の内容は、American Journal of Industrial and Business Management、Journal of Theoretical and Applied Electronic Commerce Researchに投稿、掲載された論文(Xiahou and Harada、2022)を加筆・修正したものである。

図序-1　本研究の枠組み

序論

背　　景：　イーコマースの発展と顧客消費行動の多様性、機械学習技術の発展

問題意識：　激しい競争の背景の下で顧客の流出は避けられない

ビッグデータ時代の顧客流出予測研究は不足している

目　　的：　機械学習を用いた顧客関係管理、顧客と良好な関係を築くための手法

顧客セグメンテーションに関する先行研究

顧客セグメンテーションの基準と分析手法、顧客セグメンテーションの先行研究、顧客関係管理におけるデータと分析手法

顧客流出予測に関する先行研究

機械学習の概念、マーケティング分野における機械学習技術の応用、顧客流出予測の手法、顧客流出予測の先行研究

顧客流出予測モデリングの基礎理論

k-means、ランダムフォレスト

ロジスティック回帰、サポートベクターマシン、AdaBoost

誤差逆伝播法によるニューラルネットワーク

顧客流出予測の実証研究

顧客セグメンテーションと顧客クラスタ、ランダムフォレストと予測変数の選択

顧客流出予測モデルの評価

考察と結論

ジャーナルに掲載された本研究の結果、

本研究の新規性と有用性、学術的な貢献と現実の意義、今後の課題

出所：筆者作成。

Ⅰ　顧客セグメンテーションとCRMに関する先行研究

はじめに

本章では、マーケティングのSTP理論の視点から、顧客セグメンテーションに関する先行研究を説明する。関連する先行研究は主に以下の内容を含む。（1）顧客セグメンテーションの基準、（2）顧客セグメンテーションの分析手法、（3）ネットショッピング消費者を研究対象とする顧客セグメンテーションの先行研究。そして、CRMに関する先行研究を説明する。主にCRMを含む内容は、（1）CRM の概念、（2）CRMの先行研究。

限られた経営資源の中で、すべての顧客を満たすのは容易ではない。マーケティング資源を様々な顧客と顧客グループにどのように有効に利用するに対して、顧客セグメンテーションが極めて重要である。セグメンテーションとは、マーケティング環境分析と消費行動分析の結果を踏まえて、不特定多数の人々を同じニーズや性質を持つ固まりに分けること、他社に対する優位性を築くことを目指す。マーケティング·マネジメントにおいて、市場と消費者の多様化により企業が消費者の趣味と習慣や、消費行動の特徴などを理解し、究極的には個々の消費者のニーズや問題解決に適切なサービスを提供しなければならない。

マーケティングを考える前に、企業がまず明確にしておかなくてはならない重要なことは、誰を対象とするかが明確でなければ、理想的なマーケティング・ミックスは組めない。企業が新製品とサービスを市場に投入する際には、消費者の属性を趣味と習慣や、消費行動などでセグメンテーションを行って、購買行動において似通っている顧客層を区別する。企業はマーケット・セグメンテーションからコアターゲットを明確に定めることによって、市場で競争優位を得ることができるようになる。

1 顧客セグメンテーションについて

1-1　顧客セグメンテーションの基準

セグメンテーションを行う際、様々な切り口で分けていく必要があり、その切り口を変数と言う。顧客セグメンテーションについて様々な基準があるが、現在広く使われている基準はKotler and Keller（2008）によって提案された4つのタイプの変数である。Kotler and Kellerは、顧客の地理と人口統計学的属性、消費心理の変化と購買行動の変化に基づいて4つの変数タイプを顧客セグメンテーションの基準としてまとめた、すなわち、「地理的変化数」、「人口統計の変化数」、「心理的変化数」、「行動変化数」を用いている。表Ⅰ-1はセグメントの基準変数を示している。この4

つの変数タイプのうち、「地理的変数」と「人口統計的変数」のデータはすでに存在もしくは入手は容易なものが多い。観測可能性の観点から地理的変数、人口統計的変数は把握しやすく、現代のイーコマースでは、異なる地理地域で消費購買行動を行うことはよくあることであり、地理変数は顧客セグメンテーションにとって重要ではないかもしれない。心理的変数と行動変数はそれらに比べて観測が困難である。しかし、ビッグデータ技術の発展により現代のイーコマースでは、顧客の消費心理と購買行動のデータは獲得が容易になり、消費心理変数と購買行動変数はデータドリブンマーケティングに向けた価値を持つ可能性がある。本研究で利用するデータには、顧客セグメンテーションや流出予測に使用される購入希望（カート数）や、購入意図（フェイヴァリット数）などの心理的変数と行動変数が含まれている。

表Ⅰ-1　セグメントの基準変数

地理的変数	地域、人口規模、人口密度、気候帯
人口統計的変数	年齢、世帯規模、ライフサイクル、性別、所得、職業、教育水準、宗教、人種、国籍、社会階層
心理的変数	文化志向、スポーツ志向、アウトドア志向、神経質、社交的、権威主義的、野心的
行動変数	オケージョン：日常、特別
	ベネフィット：品質、サービス、経済性、迅速性
	ユーザーの状態：非ユーザー、元ユーザー、潜在的ユーザー、初回ユーザー、敵的ユーザー
	使用頻度：ライト、ミドル、ヘビー
	ロイヤルティ：なし、中程度、強い、絶対的
	購買準備段階：認知なし、認知、情報あり、関心あり、購入希望、購入意図
	製品に対する態度：非常に肯定的、肯定的、無関心、否定的、非常に否定的

出所：Kotler and Keller, 2008より筆者作成。

1-2 顧客セグメンテーションの分析手法

セグメンテーションはマーケティング分野においては重要な位置を占めているが、顧客セグメンテーションの分析手法については統一的な見解というのは存在せず、それぞれの主体や目的、顧客特性に合わせた手法が選択される。

従来のマーケティングに関する文献には顧客セグメンテーション手法の多くはRFM法を用いて顧客を分類しており、RFM法では、主な3つの変数は直近購入日（Recency：F）、累積購入回数（Frequency：F）、累積購入金額（Monetary Value：M）である。顧客の直近購入日や購買金額、購買回数によって顧客の購買行動を集計し、その値の高い順に並び替えをする、そして、上位を優良顧客と特定し、マーケティング･プログラムを効果的に実施し、これら優良顧客の離反防止を行おうとする手法である。ただし、どういった顧客が優れているかは、様々なセグメンテーションの手法がある。これで、いくつの代表的な手法を以下で説明する。

(1) クラスタ分析

クラスタ分析[9]は教師なしデータによる分類である。未知のデータから、それぞれの特徴を学習し、類似性をもとに、データのグループわけをよる。クラスタ分析は、類似度を距離で定量化し、その類似度をもとにグループ分けをする。グループごとのデータを集計することで、グループの典型的な振る舞いを推定することができ、顧客消費の観点から、クラスタ分析は、消費者を説明する変数の近接度をもとにグループにまとめていく手法である。主な適用場面は、セグメンテーション（消費者購買行動分析や競合分析、商品の分類など）が挙げられる。クラスタ分析は、階層型クラスタ分析と非階層型クラスタ分析に大別される。

A. 階層型クラスタ分析

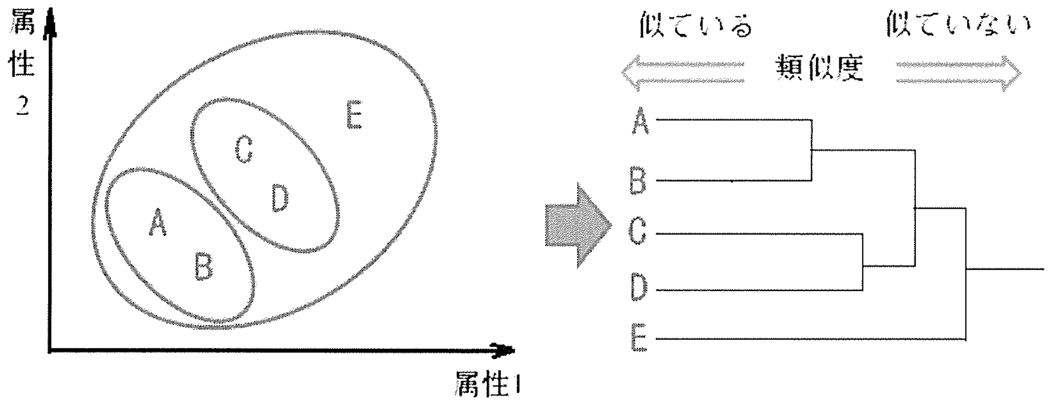
階層型クラスタ分析（Hierarchical clustering）は、クラスタ分けを最も似ている類似度が高いものからスタートし、徐々に似ていないものへと段階を踏んで行うものである。階層クラスタリングアルゴリズムは、データ間の距離に基づいて、階層アーキテクチャ方式を通じて、データを繰り返し集約し、階層を作成して所定のデータセットを分解する。個体間の類似度に基づいて、似ている個体から順次集めてクラスタを作っていく方法で、階層クラスタリングアルゴリズムは直観的なクラスタリングアルゴリズムであり、基本思想はデータ間の類似性を通じて、類似性が高い順に低い順に並べられた後に各ノードを再接続し、全体の過程はツリー構造を構築することであり、樹形図で示すことができる。階層型クラスタ分析のイメージ図が図Ⅰ-1である。

---

[9] クラスタ分析はデータ分析の方法である。階層分析と非階層分析の概念、原理、手法、手順及びクラスタ分析の注意点などについては、『入門 統計学 第2版 ―検定から多変量解析・実験計画法・ベイズ統計学まで―、栗原伸一 著、オーム社、2021』を参照。

代表的な方法として群平均法（Group average method）と最短距離法（Shortest distance method）がある。

図Ⅰ-1 階層型クラスタ分析のイメージ図



出所：筆者作成。

B. 非階層型クラスタ分析

非階層型クラスタ分析(Non-hierarchical clustering)は、階層型クラスタ分析のように総当りでの計算を行わず、データ計算を簡略化したものである。階層的クラスタに比べサンプルデータ計算量が少ないため、大規模サンプルデータに向いている。非階層型クラスタ分析は、いくつのクラスタに分類するかをデータ計算の前に決めなければならない。そのため、クラスタ数を変えながら何度か実際に計算してみるなど、適切な結果を得るために試行錯誤が必要となることがある。代表的な方法としてk-means法がある。k-means 法では、まず任意の個数のクラスタの中心を与える。次に、すべてのデータとクラスタの中心との距離を求めて、最も近いクラスタに分類し、そして、新たに形成されたクラスタの中心を求めて、最後に、データと新たなクラスタの中心との距離から分類し、クラスタの中心を求めて、クラスタの中心の位置が変化しなくなるまで繰り返す（Bradley and Fayyad、1998)。そのイメージ図は図Ⅰ-2のように示す。

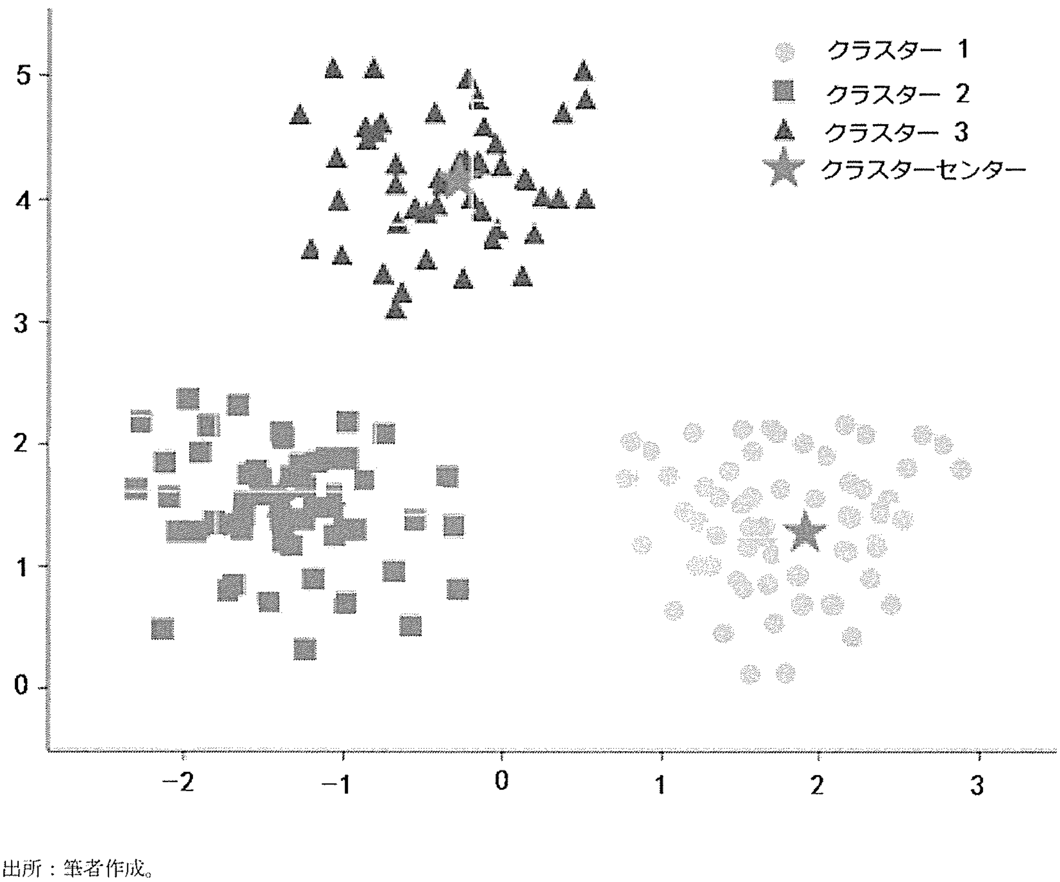
（2）判別分析

判別分析[10]は教師付データによる分類があり、すなわち自変数データもあれば、因変数データもある。既知のカテゴリのサンプルから提供された情報に基づいて、分類

[10] 「判別分析」は代表的な多変量データ分類手法である。決定木分析、ナイーブベイズ分析などについては、『Rで学ぶデータサイエンス データマイニングの基礎から深層学習まで、北 栄輔 著、オーム社、2018』を参照。

の規則性をまとめ、判別式と判別基準を確立することで、新しいサンプルがあれば、それに基づいてその所属カテゴリを判断することができる。判別分析の目的は、既知の分類のデータに対して数値指標からなる分類規則を確立し、そのような規則を未知の分類のサンプルに適用して分類することである。判別分析は通常はデータを2つに分ける必要がある。一部はトレーニングモデルデータであり、一部は検証モデルデータである。二群判別分析は回帰分析を用いて分析することが可能であり、多群判別分析の場合は正準相関分析により分析可能である（田中·垂水、1995）。代表的な方法として決定木（Decision Tree：DT）とナイーブベイズ（Naive Bayes：NB）がある。そのイメージ図は図Ⅰ-3のように示す。

図Ⅰ-2 K-meansクラスタリングのイメージ図



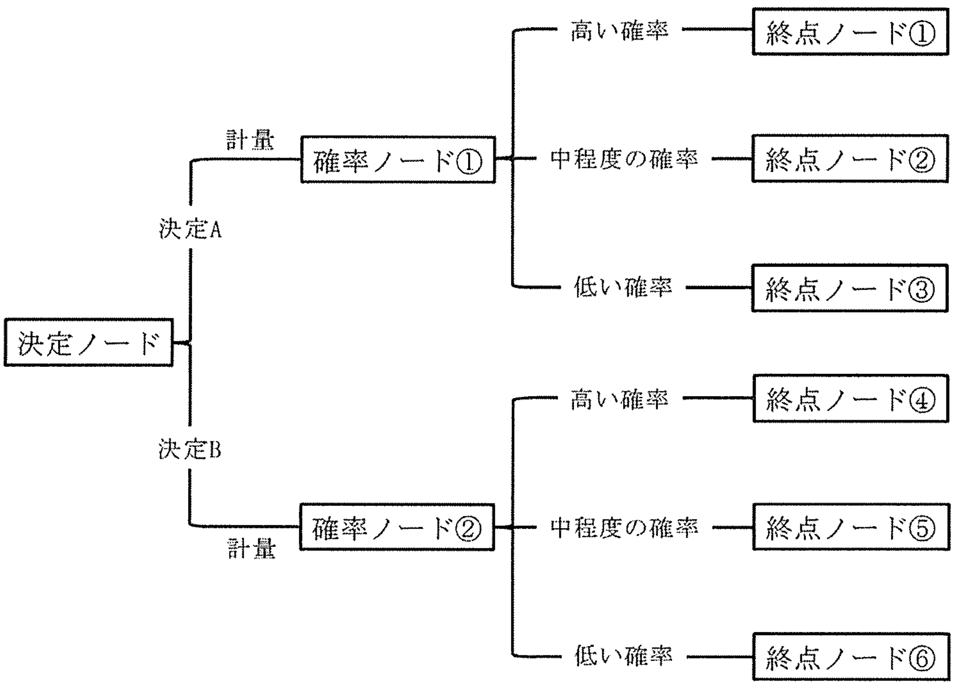
出所：筆者作成。

A. 決定木分析

決定木は、分類問題と回帰問題を解く教師あり学習のアルゴリズムの一つであり、分類の思想を利用し、データの特徴に基づいて樹モデルを構築し、データの分類を達成する。与えられたデータに対して、次々に条件を設けて、データを段階的に分類し

ていく手法である。与えられた「訓練データセット」に基づいて樹モデルを構築し、インスタンスを正確に分類できるようにする。本質的には、決定木は「訓練データセット」から分類規則のセットをまとめ、この特徴に基づいて訓練データを分割し、それぞれの「サブデータセット」に最も良い分類の過程を持つようにし、特徴選択、決定木の生成、決定木の剪定プロセスが含まれる。解析結果は、直感的に解釈しやすいのが特徴である。そのイメージ図は図Ｉ-3のように示す。

図Ｉ-3 決定木のイメージ図



出所：筆者作成。

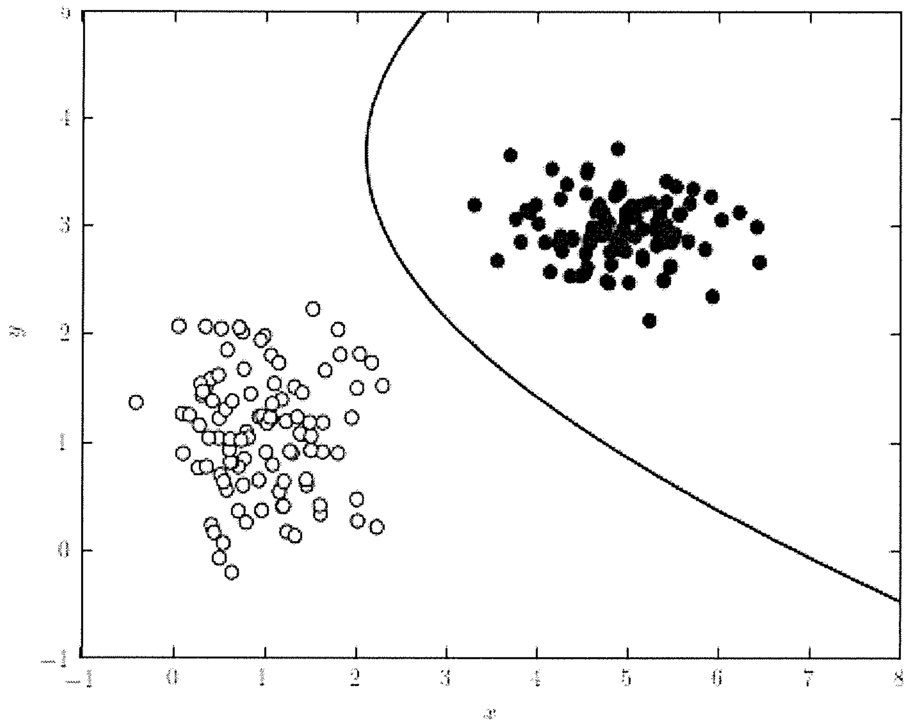
B. ナイーブベイズ分析

ナイーブベイズ(Naive Bayes)は教師付データによる一般的な分析手法であり、分類問題を解決することを目的としている。ナイーブベイズアルゴリズムには一定の仮定が必要であり、変数（特徴）間には相互に独立する必要があると仮定している。ナイーブベイズアルゴリズムの基本的な思想は、特徴確率を考慮することによって分類される。すなわち、与えられた分類されるサンプルに対して、このサンプルが現れる条件下で各カテゴリが現れる確率を求め、各確率に基づいて、サンプルがどのカテゴリに属することを分類される。そのイメージ図は図Ｉ-4のように示す。

現代のイーコマースでは、顧客の購買行動に影響を与える要素（変数）が多く、RFM法における3つの変数のほか、他の変数(商品の種類、消費者の好み、顧客の購入

頻度、購入日付、フェイヴァリット、クリークなど）は顧客セグメンテーションにおいて重要な要素になる可能性がある。本研究で利用する変数には、顧客セグメンテーションに使用される購入希望（カート数）や購入意図（フェイヴァリット数）などの変数が含まれている。

図Ⅰ-4 ナイーブベイズ分析のイメージ図



出所： www.astroml.org/book_figures/chapter9/fig_simple_naivebayes.html （検索日：2022年1月20日）。

1-3 顧客セグメンテーションの先行研究

顧客セグメンテーションの概念は1950年代に米国の学者ウェンデル・スミス（Wendel Smith、1956）が提案したものである。彼は顧客セグメンテーションは企業がマーケティング活動を展開する基礎であり、市場に同質性と資源の有限性があると考えている。そのため、企業は限られた顧客数と限られた資源の上で経営しなければならない。企業は特定のパターン下で顧客の属性、行動、需要、好み、価値などの多くの要素に基づいて顧客を分類して、さらに特色のある製品とサービスを提供する。そのため、企業は限られた顧客数と限られた資源の上で経営しなければならない。この分類方式は最も直観的で、必要なデータも最も入手しやすい。顧客関係管理の重要

な構成部分として、世界各国の学者が異なる角度から研究を行っている。消費者の顧客セグメンテーションに対して、現在は主に「人口統計に基づく」、「消費行動に基づく」、「顧客価値に基づく」などの3種類の顧客セグメンテーション方法がある。

(1) 人口統計に基づく顧客セグメンテーション

人口統計に基づく顧客セグメンテーションは、顧客の性別、年齢、職業、地域などの特徴に基づいて顧客を分類する。

Lazer (1963) は顧客の活動、興味と評価で顧客を分類することをいち早く提案した。ライフスタイルで顧客を識別し、どの顧客が「価値のあるユーザー」であるかを判断することを提案している。文献では顧客のライフスタイルのシステム性を強調しているが、顧客のライフスタイルに含まれる内容についての記述や詳細な論証はないため、この研究結果には一定の限界がある。Wells and Tigert (1971) は顧客のライフスタイルを研究し、AIOを用いて顧客のライフスタイルを記述することを提案している。AIOとは、活動 (Activity) 、趣味 (Interests) 、評価 (O pinion) でライフスタイルを表現することである。その後、Mitchell (1983) は社会階層、生活様式、個人の特徴に基づく顧客心理セグメンテーションモデルを提案し、人口統計の顧客セグメンテーション研究をさらに推進している。

劉ら (2006) は、人口統計と顧客のライフスタイルに基づいて、顧客セグメンテーションにおける人口統計変数が顧客セグメンテーションに与える影響を研究した。研究によると、ネット銀行の消費者にとって、地理変数は顧客の消費選好と購買行動との関連性が大きくないことが明らかになった。劉ら (2013) は顧客セグメンテーションモデルを構築する際に地理的要素をセグメンテーション変数とし、アパレル業界における異なる地域環境下の顧客の同じ製品に対する需要に差があることを研究している。向 (2015) は、主成分分析法を用いて人口統計に基づく住宅消費市場の顧客セグメンテーションを研究している。人口統計の変数は年齢、性別、収入、学歴、価格、環境などを含み、異なる顧客の人口因子と好みに対して顧客セグメンテーションを行う。

これまでの人口統計に基づく顧客セグメンテーション文献を分析することにより、人口統計セグメンテーション変数の選択プロセスにおいて、業界要素を考慮しなければならないことが分かったが、地理要素がインターネット業界の顧客セグメンテーションに与える影響はほとんどない。

(2) 消費行動に基づく顧客セグメンテーション

21世紀初頭に研究者は「顧客消費行動に基づく」顧客セグメンテーションについて研究を始めている。「人口統計に基づく」セグメンテーションの研究結果には大きな

違いがあり、これらの結果はすべての顧客の消費嗜好を説明することができないため、Hughes（1994）は初めてRFM分析に基づく顧客セグメンテーション手法を提案した。RFM分析は顧客の消費行動に基づくセグメンテーション方法であり、Hughesは顧客の消費データベースには「Recency：直近購買日」、「Frequency：累積購買回数」と「Monetary Value：累積購買金額」である。顧客が直近購買日に消費する時間が近づくほど、顧客を引きつけて再び消費する可能性が高くなるため、直近購買日に購買する時間が短いほど良い。「累積購買回数」と「累積購買金額」の間の多重共線性[11]という問題を解決するために、Marcus（1998）は従来の「累積購買金額」の代わりに「平均購買金額」を用いるとともに、「平均購買金額」と「累積購買回数」を利用して顧客価値の混同行列を構築し、顧客グループが多すぎる問題を解決する。その後、Cheng（2009）は顧客関係管理において「粗い集合」を理論として新しい顧客セグメンテーションプログラミングを行った。まず、RFMモデルに基づいて顧客データを分類し、このプログラムを演算した後、K-meansクラスタリング手法を用いてRFMモデルのデータ変数のクラスタリングを行い、最後に様々な顧客カテゴリを得て、各顧客の特徴を分析した。実証研究ではある電子業界のデータを用いて性能と効果検証を行った。その結果、顧客セグメンテーションの正確度が高いことが分かっている。顧客セグメンテーション研究の発展に伴い、Hsuら（2012）は顧客の消費商品が顧客分類に影響を与え、階層クラスタリング技術を用いて顧客セグメンテーションを提案した。その結果、階層クラスタリング方法は従来の顧客セグメンテーション方法よりも良い分類結果を持つことが分かっている。曽ら（2013）はRFMモデルを基礎として、R、F、Mの3つの変数を10の変数にセグメンテーションし、顧客の消費データに基づく多変数モデルを提案し、それから因子分析方法を用いて変数を3つの因子を抽出し、最後に区分に基づくクラスタリング技術によって顧客を6つの大分類に分けるとともに、クラスタリング結果とRFMモデルのクラスタリング結果を分析した。その結果、改善された多変数RFMモデルのクラスタリング効果が高いことが明らかになった。蔡ら（2015）も伝統的なRFMモデル変数の改善を行い、この研究では、商品の種類も顧客のセグメンテーションに大きな影響を与えると考えている。この研究のセグメンテーション手法は、セグメンテーション変数を3種類から4種類に拡大し、10つの変数である。顧客セグメンテーションは小売業界の運用から航空業や電力消費などの業界に徐々に広がっており、楊ら（2015）はK-meansクラスタリングアルゴリズムを用いて民間航空の顧客セグメンテーションモデルを構築し、民間航空の顧客を3つに分類し、各種類の顧客群の特徴を識別することで、民間航空市場における企業の競争力を高めるため

[11] 多重共線性(multicollinearity)：Goldberger(1968)は，多重共線性を「説明変量のうちのいくつかが相互に関連しており，そのために単独の影響を分離したり，効果を評価したりすることが不可能ではないにしても困難な状態」と定義している。
(Goldberger, A.S.,1968,Topics in regression analysis, Macmillan)

のマーケティング戦略を提案している。盧ら（2016）は電力会社の消費者の特徴に基づいて、K-meansを用いて電力会社の顧客を4つに分類し、各顧客グループの異なるニーズに応じて個性的な付加価値サービスを提供する上で顧客セグメンテーションに適するサービスシステムを開発した。趙ら（2013）はK-meansアルゴリズムを用いて商業銀行のファンド財テク顧客取引行動データを分類した。徐ら（2012）伝統的な小売業界の顧客が細分化したRFMモデルに、総利益属性を導入した。RFPモデルを作成し、K-meansを用いて、あるイーコマース企業の顧客に対してクラスタリング分析を行っている。Britoら（2015）は実例を応用してK-Medoidsを用いてイーコマースの顧客選好をよりよく理解し、企業が顧客のニーズによりよく応え、競争力を高めることができることを証明した。

(3) 顧客の価値に基づく顧客セグメンテーション

顧客の人口統計情報と消費購買行動情報を利用することで顧客を効率的に分類することができるが、人口統計に基づく顧客セグメンテーションは顧客の属性を反映するしかなく、企業に対する顧客の価値貢献を知ることも、顧客の消費行動を知ることもできない。これで、研究者は顧客の価値に基づいて様々な顧客セグメンテーション手法を提案している。Verhoef and Donkers（2001）は顧客の人口統計情報と消費取引データを利用して顧客の潜在価値を評価し、顧客セグメンテーションの際に企業が高潜在価値の顧客層に注目すべきだと考え、顧客の現在価値と潜在価値を採用して保険会社の顧客に対して顧客セグメンテーションを行った。しかし、顧客消費の多様化によりこの方法は顧客関係の長期性と安定性を考えていない。高い潜在価値を持つ顧客は必ずしも高い忠誠心ではないため、Hwang ら（2004）は顧客の現在価値、潜在的価値、顧客ロイヤルティの 3 つの要素を同時に考慮する顧客セグメンテーション手法を提案し、この三つの要素を基に顧客生涯価値を計算し、顧客生涯価値に基づいて顧客細セグメンテーションを行った。研究結果によると、この方法はより良い分類結果を持っていることが明らかになった。Kim ら（2006）は顧客の価値と顧客生涯価値を組み合わせた顧客セグメンテーションモデルを提案し、顧客の歴史的価値、将来価値、流出率を構築した LTV(Life Time Value)モデルを構築し、このモデルを利用して顧客セグメンテーションを行った。

顧客生涯価値の視点から、顧客セグメンテーションが明らかな利点があるため、研究者は様々な方法で顧客の潜在価値を評価し、顧客セグメンテーションを行う。綦（2013）は既存の文献による顧客セグメンテーションの研究を行っている。多くの文献は単一の視点から顧客セグメンテーションを行っており、構築されたモデルも顧客消費記録に基づいてセグメンテーションを行っただけで、顧客の将来価値を予測していない。そのため、綦らはこれに基づいて、多次元、動態性、予測性の3つの視点か

ら顧客セグメンテーションを行うことを提案している。趙と斉（2014）は団体購入サイトの顧客価値の分析に基づいて、顧客の「ウェブレビュー行動」を顧客価値の変数とし、古典的なRFMモデルを基礎として「オンラインレビュー行動:Publishing online reviews」に基づくRFMPモデルを構築し、2つのモデルを統合して顧客の最終価値を計算し、企業により正確なマーケティング意思決定と管理提案を提供している。劉ら（2015）は伝統的なRFMモデルの限界を分析し、この文献は3つの視点からRFMモデルの限界を述べ、第1はモデルの応用シーンであり、異なる業界の顧客セグメンテーションモデルの変数が異なるべきであるためである。次に、RFMモデルの変数はライフサイクルという要素を考えていない。最後にRFMモデルの応用には具体的な業務目標と計画を結合する必要がある。李ら（2015）は航空会社の顧客消費の現状を分析し、顧客価値に基づく顧客セグメンテーション手法を提案している。この文献では、複数の変数を選択して顧客の現在価値と潜在価値を評価し、その後、顧客の現在価値と潜在価値の2つの次元を通じて顧客の価値混同行列を構築し、航空会社の顧客を高現在価値顧客、高価値顧客、高潜在価値顧客、低価値顧客の4つの顧客群に分類し、最後に各顧客層に対して適合なマーケティング戦略を提案している。

以上より、顧客の消費行動分類に対して、主にRFMモデルを基礎として、異なる業界自身の特徴と結びつけたモデル変数に対して改善を行い、主にK-meansクラスタ方法を用いて顧客に対してセグメンテーションを行っている。顧客価値の分類については、顧客価値に基づくセグメンテーションについて顧客の潜在価値を測定し、より正確に顧客セグメンテーションを行うために、研究者は各種セグメンテーションモデルを構築するだけでなく、各業界のデータを結合し、各種技術を利用して実証研究を行っている。一部の文献は顧客の現在価値と潜在価値を変数として顧客セグメンテーションを行っている。一方、一部の文献は顧客価値とロイヤルティを変数として顧客セグメンテーションが行われ、少数の文献はRFMモデルのR、F、Mの3つの変数の改善を行い、顧客を高、中、低の3つの異なる価値の顧客層に分類している。

現在、伝統的な業界（電気通信、保険、金融業界など）の顧客セグメンテーションに関する研究文献は比較的多いが、B2Cイーコマースに対する顧客セグメンテーション研究は比較的少なく、本研究では従来の研究者の顧客セグメンテーションに関する方法を参考にして、主にイーコマースショッピングサイトで商品購買行為を展開する顧客に対してセグメンテーション研究を行い、その後、セグメンテーション結果に基づいて異なる顧客群に対して流出予測の研究を行う。顧客セグメンテーションの実証研究段階では、本研究ではRFMモデルに基づいて時間次元、消費嗜好、購買意欲などの変数を細分化モデルの変数として提案し、K-meansクラスタリング方法を用いて改善されたRFMモデルのデータ変数をクラスタリング研究し、最後に各種顧客カテゴリを得て、各顧客の特徴を分析した。

2　CRMに関する先行研究

2-1　CRM の概念と特点

(1) CRM の概念

CRM は、1999 年に Gartner Group によって最初に提案された。学者や企業が異なれば、CRM の概念についての定義も異なる。 Gartner Group は、CRM がマーケティング戦略であると考えている。顧客の分類に従って企業リソースを効果的に編成し、顧客センターのビジネス行動を促進し、顧客センターのビジネスプロセスを実施し、これを企業の収益性、利益、および顧客満足度を向上させる手段として使用する。IBM は、CRM が企業製品のパフォーマンスを改善し、顧客サービスを強化し、顧客提供価値と顧客満足度を改善し、顧客と長期的で安定した信頼できる関係を確立することにより、新規顧客を引き付け、既存顧客を維持し、企業のビジネスパフォーマンスと競争上の優位性を向上させると考えている。SAP 会社は、CRM が顧客データに対する管理方法であり、顧客データベースが企業の重要なデータセンターであり、マーケティングの過程で企業と顧客の間のさまざまな相互作用、およびさまざまな関連アクティビティのステータスを記録し、各種データモデルを提供し、後続の分析と意思決定をサポートすると考えている。

CRM の概念（薮野祥太、2020）によると、CRM はコンピュータやネットワークなどの情報技術を利用し、すなわち大量の顧客データを用いて顧客を識別し、顧客一人ひとりに適したダイレクトメッセージを顧客に送信するなどして満足させ、ダイレクトメッセージを利用して外部からメッセージの内容を覗くことができない形式で連絡できる。最終的な形式は、IT とビッグデータ技術を利用したデータ分析に基づいて、顧客ごとに適切な形式で CRM を実施し、企業と顧客の関係を構築および維持し、顧客ごとの収益（顧客生涯価値）を最大化することである。この目的を達成するために、企業は個々の顧客情報データを収集、分析し、最も効果的な顧客セグメンテーションを行う必要がある。

(2) CRM の特点

① CRM は管理理念である

CRM は「データベースマーケティング」、「リレーショナルマーケティング」、「マンツーマンマーケティング」などの最新の管理思想の統合である。顧客との個性的なコミュニケーションを通じて顧客のニーズを把握し、その上で顧客に個性的な製品とサービスを提供し、企業が顧客に提供する提供価値を絶えず増やし、顧客の満足度とロイヤルティを高める。

② CRM は管理メカニズムである

CRM は、企業のマーケティング、販売、サービス、テクニカルサポートなど、顧客に関連する分野に応用できる。サービスの質を高めるとともに、情報共有と業務プロセスの最適化により経営コストを効果的に削減する。

③ CRM は管理ソフトウェアと技術である

CRM はインターネットとイーコマース、マルチメディア技術、データベース、データマイニング、エキスパートシステム、人工知能などの先進的な IT 技術を集積し、製品販売、顧客サービス、意思決定サポートなどに自動化されたソリューションを提供する。

2-2 CRM の活用と研究のポイント

CRM活用は、まず顧客の購買履歴の情報のセグメンテーションをはかる。利益貢献度が高い顧客の特性や購買商品特性などのデータを集めて、分析を行う。さらに、その他の顧客の属性（年齢、地域、性別、好みや行動心理・スタイル）を分析することで、これらの顧客と商品の関係性を発見ている。CRM活用においては顧客データ情報チャネルをどのように活用し、データ情報を一元化するかが大切である。情報管理センターは、パソコンおよび情報管理システムを活用して、顧客情報を自動的にデジタル化できる。その顧客との会話情報、営業と顧客の接点、POSなど購買時点の情報、広告等の反応情報をデータベースに入れ、必要に応じてデータを引き出して一元的に分析している。

CRMは個性化する消費者ニーズを満たし、顧客ロイヤルティを高め、販売コストを下げ、利益収入を増やし、市場を拡大し、企業の収益力と市場競争力を全面的に向上させる目的である。CRM研究のポイントは主に（1）どのように企業の経営効率を高めるのか、（2）既存の顧客を維持し、新規顧客を拡張するのか、（3）どのように企業の市場空間を拡大し続けるのかという点にある。

2-3 CRMの先行研究

20 世紀 90 年代にはインターネットやイーコマース技術の向上に伴い、CRM も急速に発展した。2000 年頃から、CRM は企業管理者や学者の研究対象になりつつある。蘇（2010）は、CRM の主な研究内容が、顧客関係の構築、顧客関係の維持、顧客関係の回復を含むと考えている。

顧客関係の構築について、Ramendra ら（2016）は、企業が持続可能な競争優位性を獲得するには、顧客ニーズと価値を理解しなければならず、顧客関係ごとの価値を理解することで、企業は顧客を「関係の組み合わせ」に細分化して企業の顧客関係の

収益率を高めることができ、企業は顧客関係の構築に重点を置き、顧客価値に応じて市場資源を合理的に分配し、顧客のポートフォリオ管理を実現すると考えている。Baydar ら（2002）は、激しい市場競争の中で小売業は企業と顧客の関係を効果的に把握しなければならず、CRM を実施する際には顧客を分類研究するだけでなく、顧客個体と製品の間の需要関係を発見し、顧客個体に対する需要管理を実現しなければならないと提案した。伍（2017）は、CRM の重点は企業と顧客の関係をうまく処理することであり、顧客の満足度とロイヤルティに注目し、顧客関係の構築、顧客の保持と維持、顧客の流出などの面で有効な管理を行う必要があると考えている。Yuen（2014）は、製品品質、サービス品質と顧客ロイヤルティの関係について研究を行い、マンマシンインタラクション、製品信頼性、問題解決方案が顧客満足度と顧客ロイヤルティに与える影響を分析し、そして、企業が CRM を実施することは顧客満足度とロイヤルティを高めるだけでなく、製品の販売量を効果的に高めることができると指摘した。

顧客関係の維持において、沢登（2000）は顧客情報収集、顧客分類、顧客コミュニケーション管理などの e-CRM の経営理念を提出し、イーコマース企業に対してシステム化された管理体系を構築し、21 項目の顧客関係の取得と維持の方式を提案した。徐（2011）は CRM を研究し、e-CRM には 4 つの特徴があり、それぞれ企業内部で顧客資源を共有でき、ネットワークの相互作用性が強く、情報伝播がタイムリーで、顧客データベースに記録された顧客情報がマーケティングに基礎データを提供でき、そして e-CRM システムを利用して顧客を細分化でき、「データ」を通じて顧客特徴を分析でき、それによって顧客満足度とロイヤルティを効果的に高めることができると指摘した。Verma ら（2013）は、携帯電話の付加価値サービスの応用に伴い、携帯電話端末の CRM が企業管理の重要な内容になると考えている。Soltani and Navimipour (2016) は、データマーケティングライブラリ技術、インターネット技術、オブジェクト指向技術、デジタルマーケティングなどの関連技術成果を統合した CRM を包括的なシステムとして研究している。業務のデジタル化の観点から見ると、CRM は企業が適切な製品マーケティング方案を提出し、顧客サービスと顧客管理のために有効なサポートを提供するのに役立つ。

小括

本章では、顧客セグメンテーション基準と分析方法について説明し、顧客セグメンテーションの先行研究についてレビューした。従来のSTP研究に関する文献は主に顧客セグメンテーションの基準に基づいて顧客セグメンテーション方法を研究し、セグメンテーション法を選択する際に、主にR、F、Mを変数としてどの顧客が優良顧客であるかを区別し、その後、顧客のR、F、M変数に基づいて、自社製品に適した消費者

層を探し出し、自社製品の独自性と差別化を明確にする。コトラーセグメンテーション基準を出発点とし、市場、顧客及びその需要がすでに存在する場合、消費者の購買意欲をセグメンテーション要素に組み入れ、顧客の購買意欲を分析することは、今後のSTPの研究重点内容のひとつである。一方、顧客ごとに消費時刻が異なるため、顧客の商品消費時間を考慮し、各時刻の消費需要を分析することもSTP研究のポイントのひとつである 。

また、本章ではCRMの目的と特徴、CRMの活用などについて説明し、CRMの先行研究について概説した。従来のCRM文献の中で、多くの研究はアンケート調査を採用して顧客データを収集し、顧客満足度と忠誠度に対して研究を行ったが、データベースを利用した顧客データ情報を利用してCRMを研究することは少なく、特に機械学習を利用して顧客の流出を研究する文献は極めて少ないため、機械学習を利用してCRMを研究することは今後の研究の重点である。

Ⅱ　顧客流出予測に関する先行研究

はじめに

本研究は機械学習技術を利用してイーコマース環境下の顧客流出予測問題を研究するものであり、本研究の目的を実現するためには、機械学習の基本概念と手法を理解しなければならない。本章では、まず機械学習の概念、プロセスと手法、マーケティングにおける機械学習技術の応用を紹介する。そして、本研究に関連する顧客流出予測に関する先行研究のレビューを行う。関連する先行研究は主に以下の内容を含む。（1）顧客流出と顧客流出管理の定義、（2）顧客流出予測の手法、（3）顧客流出予測の先行研究。

市場競争と消費者行動の多様性に適応するために、多くの企業のマーケティング戦略は過去の製品宣伝を主とすることから、保有顧客の維持と顧客の流出へ関心が移っている。このように、流出する可能性がある顧客を引き留めることはすでに企業の顧客関係管理の重要な内容となっている。しかし、大量の広告の投入や製品の最適化など従来のマーケティング戦略は、消費者行動の多様性を理解することが難しい、企業がコストを上がるリスクに直面しやすい。一方、ビッグデータ時代に伴って、顧客の購買データを収集することがやすくなっている。顧客の購買データを基に流出可能な顧客をセグメンテーションして、顧客流出予測モデルを構築する方法は企業にとって顧客消費行動を効率的に理解し、市場需要の変化に対応するなどの新しい経路を提示することができる。

1　機械学習について

コンピューターの性能が大きく向上したことにより、機械であるコンピューターが「学ぶ」ことができる。それが人工知能（AI）の中心技術、機械学習である。「AI」とは、人間の思考プロセスと同じような形で動作するプログラム、あるいは人間が知的と感じる情報処理・技術といった広い概念で理解されている[12]。

① 人工知能[13]
- 人間の思考プロセスと同じような形で動作するプログラム全般
- あるいは、人間が知的と感じる情報処理・技術全般

② 機械学習
- AI のうち、人間の「学習」に相当する仕組みをコンピューター等で実現するもの

---

[12] 総務省令和元年版情報通信白書第 1 部第 1 章第 3 節 ICT の新たな潮流（2 AI に関する動向），p. 82

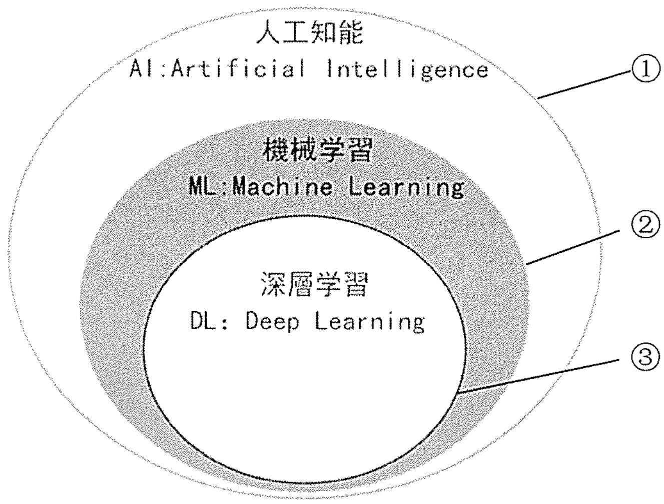
[13] 人工知能の概要と分野内の諸領域、人工知能研究の歴史及び機械学習の原理、機械学習の方法、知識表現と推論、ニューラルネットワーク、深層学習などの内容については、『基礎から学ぶ　人工知能の教科書、小高知宏 著、オーム社、2019』を参照。

- 入力されたデータからパターン/ルールを発見し、新たなデータに当てはめることで、その新たなデータに関する識別や予測等が可能

③ 深層学習

- 機械学習のうち、多数の層から成るニューラルネットワークを用いるもの
- パターン/ルールを発見する上で何に着目するか(「特徴量」)を自ら抽出することが可能

図Ⅱ-1　人工知能・機械学習・深層学習の関係



出所：総務省令和元年版情報通信白書第1部第1章第3節ICTの新たな潮流。

人工知能に関わる分析技術として機械学習は大量の多様なデータを処理することができる。マーケティング分野では、機械学習というデータ分析手法が注目を集めている。人工知能と機械学習は混同して語られることもあるが、人工知能の方がより広義な概念であることに対して、機械学習は人工知能よりも狭義の概念である。「人工知能」、「機械学習」、「深層学習：ディープラーニング」、その関係は図Ⅱ-1のとおりである。

1-1　機械学習の概念

機械学習（マシーンラーニング：ML）とは、人間の学習に相当する仕組みをコンピューター等で実現するものであり、一定の計算方法（アルゴリズム）に基づき、入力されたデータからコンピューターがパターンやルールを発見し、そのパターンやルールを新たなデータに当てはめることで、その新たなデータに関する識別や予測等を

可能とする手法である[14]。

機械学習を理解するためには、統計学との違いを理解することが大切であり、「機械学習」は機械が自動的に学習するものであるのに対し、「統計学」はデータのルールやパターンを統計的に判断する。統計学も機械学習も、データから、ルールやパターンを見つけ出し、モデルを構築することが同じである。統計学の場合はデータの「説明」を目的としており、機械学習の場合は「予測」を目的としている。統計学においても回帰モデルなどを使うことで、予測に活用することもできる。しかし、統計学の主目的はデータの背景にあるルールをより正しく説明できているかどうかを重視し、機械学習の主目的は、より正しく予測できているかどうかを重視しているのである。統計学では、ある程度、直感的に理解できる説明変数で構成されたモデルが多くなるが、機械学習の場合は、直感的には理解できない説明変数も考慮されるため、より精度が高まる可能性がである。

機械学習の研究初期には「学習する」点に注目されたが、現在では「学習に基づいて予測・判断する」点に注目されることが多くなっている。機械学習はデータから規則性と判断基準を学習し、それに基づいて予測と判断を行うことができる。機械学習においては、データセットが最も重要である。データセットとは、人工知能が学習を行う際に必要な学習データのことで、データとラベルをセットにしたものである。機械学習には「学習」と「推論」の2つのプロセスがあり、この2つのプロセスでそれぞれ異なるデータを用いることとなる。学習のプロセスとは、入力されたデータを分析することにより、コンピューターがデータの識別等を行うためのパターンを確立するプロセスである。推論のプロセスとは、学習のプロセスを経て出来上がった学習済みモデルにデータを入力し、確立されたパターンに従い、実際にそのデータの識別等を行うプロセスである。

1-2　機械学習におけるデータ活用のプロセス

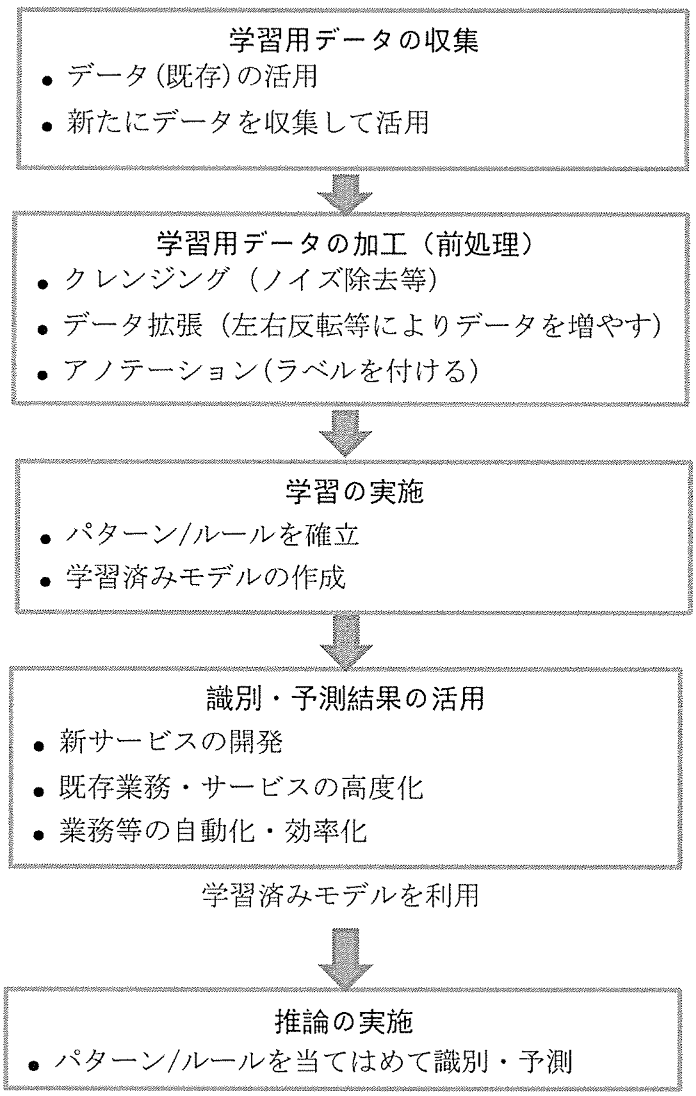
CRISP-DM[15]フレームワークでは、データ活用のプロセスを（1）ビジネスの理解、（2）データの理解、（3）データの準備、（4）モデル作成、（5）評価、（6）展開と6つのフェーズに分けて定義する。CSRISP-DMは、汎用性の高いよくできたプロセスであるが、本研究の分析プロセスとして活用するにあたっていくつかの改良を加えている。機械学習におけるデータ活用のプロセスは、まず、学習データからコンピューターが法則性を学習し、モデルが生成される。生成されたモデルに対して、テストデ

[14] 総務省令和元年版情報通信白書第1部第1章第3節 ICT の新たな潮流（2 AI に関する動向），p.83。

[15] CRISP-DM とは、Cross-Industry Standard Process for Data Mining の略で、1996 年には、SPSS、ダイムラー・クライスラーと NCR 社は、データマイニング方法とプロセスの基準を確立するための興味グループを共同で設立した。そして 1999 年に正式に CRISP-DM プロセスを抽出した。

ータを入力すると、コンピューターが学習データから得られた法則性に従って、教師あり学習であれば予測、教師なし学習であれば識別を行うといった流れとなる。機械学習モデルの具体的な学習プロセスを示したのが図Ⅱ-2である。本研究の顧客流出予測実験における学習データはトレーニングデータセットとテストデータセットに区分される。トレーニング段階では、アルゴリズムは新しいモデルを生成したり、事前トレーニングモデルを特定の応用に再調整したりして、モデルがそのパラメータを学習するのに役立ちする。テスト段階では、学習で得られたパラメータに基づいて新しいデータを推定し、決定する。機械学習の技術応用の視点から機械学習の技術を使ってできることは２つ、「顧客識別：グループ分け」と「顧客流出予測」である。

図Ⅱ-2　機械学習におけるデータ活用のプロセス



出所：筆者作成。

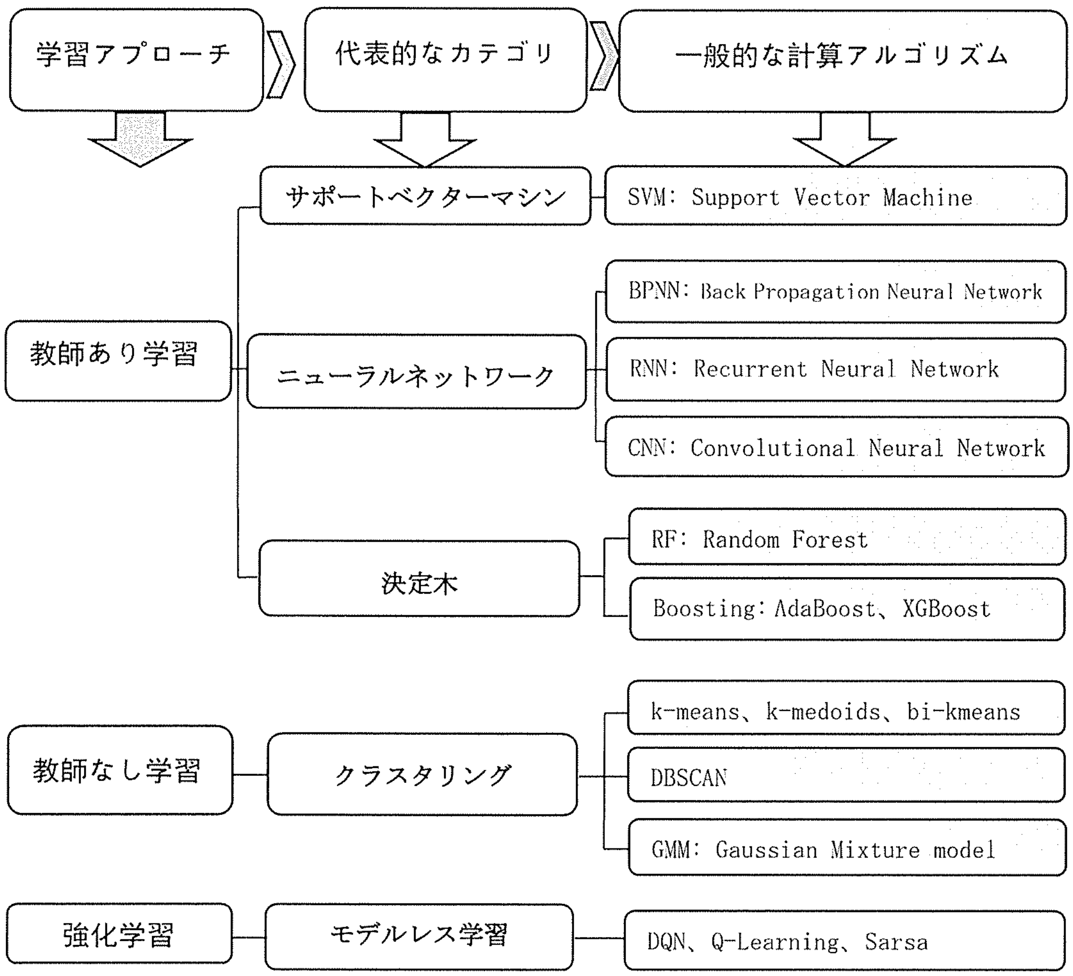
1-3　機械学習の手法

機械学習の手法は主に「教師あり学習」、「教師なし学習」、「強化学習」の3つが存在する。望ましい結果を導き出すために、機械学習を用いる目的に応じて、最適な手法を選択できるようにする。機械学習の手法分類を示したのが図Ⅱ-3である。

(1) 教師あり学習

教師あり学習とは、あらかじめ定めてある正解のラベルに基づき、機械が出力する学習方法である。教師あり学習では具体的な正解のデータに基づいて明確な結果を導き出せるため、基本的な機械学習の手法として用いられている。教師あり学習のモデルはトレーニングプロセスで予測を出しながら構築され、予測に間違いがある都度に訂正される。

図Ⅱ-3 機械学習の手法分類



出所：筆者作成。

(2) 教師なし学習

教師なし学習とは、正解のデータを与えない状態で学習させる方法である。教師なし学習には、データの構造や分布を明らかにするという目的がある。入力の学習データはラベルが付けられておらず、既知の結果も有しない。教師なし学習のモデルは入力データを統計的に解析し、データに存在する構造を抽出して構築される。コンピュータによってデータのパターンの抽出や識別するため、データの数が多い場合や正解のデータがない場合に活用される。

(3) 強化学習

強化学習とは、これまで出力したデータを価値付けすることによって、その価値を最大化するために学習する方法である。「教師なし学習」と同様で正解のデータが与えられないため、最適な出力だった場合に、コンピュータに報酬を与えることで、コンピュータに継続して最適な出力ができるように仕向ける。

従って、前述の3つの手法に参考する上に、ユーザーデータの縦断的な時間性と多変数の特徴に基づいて、本研究は顧客流出予測実験における顧客セグメンテーション(識別)は教師なし学習の手法が用いられ、顧客流出予測は教師あり学習の手法が用いらる。

1-4　マーケティング分野における機械学習技術の応用

人工知能技術はすでにマーケティング·マネジメント業界で変革を起こし、急速な進歩を推進し続けている。

米国企業の60%以上が人工知能技術を既存のマーケティング案に応用しており、その中で最も一般的な方法は機械学習技術で市場をセグメンテーションし、顧客のニーズを予測し、広告の投入を最適化し、顧客に製品を推薦し、ユーザーの気分や状況によって推薦システムを開発することである。イーコマースマーケティングの分野では、消費者の取引データが急速に増加しているだけでなくデータのタイプも多様化している。消費者データはユーザーと商品ID、商品種類、時刻印など取引記録、クリックストリームと消費者行動意図など多くの形式を含み、データのタイプも従来の構造型データから非構造型データに広がっている。データ環境と技術のアップグレードは機械学習の深い発展を推進している。競争環境が激化していく中でイーコマース企業の顧客流出の現象が深刻で、そのため、顧客流出予測も研究の焦点となっている。機械学習技術を用いてこれまでのデータベースや審査機能よりも精度の高いデータ分析が可能となり、既存顧客の流出を防止する。表Ⅱ-1はこの5年間のマーケティング分野における機械学習技術の応用例を示している。

表Ⅱ-1 マーケティング分野における機械学習技術の応用例

応用分野	目的	アルゴリズム	使用するデータ	出所
商品購買予測	ウェブ検索ログを用いてカメラの購入予測	サポートベクターマシン、 ロジスティック回帰	ECサイトデータ	中田・村本ら(2021)
	店舗経営、在庫管理	ランダムフォレスト	POSデータ	安田·吉野·松山(2019)
	ECマーケティング等の戦略策定	一般の統計モデル ワイブルモデル	Web Reportからのパネルデータ	新美(2017)
	顧客の定着予測、顧客再購買の特有の要因解明	ランダムフォレスト	ECサイトデータ	北澤・生田目ら(2020)
顧客行動分析	Web ページ閲覧履歴を用いた情報収集行動	k-means法	Web ページデータ	佐藤·高久(2020)
	優良顧客の購買行動の理解	ロジスティック回帰	ポータルサイトデータ	中里見·生田目ら(2020)
	商品購買行動の傾向や遷移	自己組織化マップ	POSデータ	川名·諏訪·関(2018)
	小売店の売上の向上	マルチエージェントシミュレーションシステム	POSデータ	中村・酒井ら(2020)
	販売促進施策	EM アルゴリズム、 マルコフモデル	ECサイトデータ	松寄·三川ら(2017)
市場セグメンテーション	実店舗マーケティングモデル	トピックモデル、 主成分分析、因子分析	POSデータ	李·照井(2018)

出所：筆者作成。

2　顧客流出予測について

市場競争と消費者行動の多様性に適応するために、多くの企業のマーケティング戦略は過去の製品宣伝を主とすることから、保有顧客の維持と顧客の流出の関心に移っている。これにより、流出する可能性がある顧客を引き留めることはすでに企業の顧客関係管理の重要な内容となっている。しかし、大量の広告の投入や製品の最適化など従来のマーケティング戦略は、消費者行動の多様性を理解することが難しい。企業はコストの上昇リスクに直面しやすい。一方、ビッグデータ時代に伴って、顧客の購買データの収集かたやすくなっている。顧客の購買データを基に流出可能な顧客をセグメンテーションして、顧客流出予測モデルを構築する方法は企業にとって顧客消費行動を効率的に理解し、市場需要の変化に対応するなど新しい経路を提示することができる。

2-1　顧客流出と顧客流出管理の定義

顧客流出とは、顧客がある企業の製品やサービスの使用を放棄することであり、別の競合企業の製品またはサービスを使用することである(Amin et al.、2017)。Jahromiら（2017）は、あるイーコマースサイトで半年以内に何の消費行動もない顧客を流出顧客と見なしている。すなわち、顧客の消費頻度で流出を定義している。消費頻度に基づく定義方法とは異なり、Migueisら（2012）は消費金額を用いて流出を定義し、すなわち、顧客のt番目の期間における購入金額がt-1番目の期間における消費総額の40%未満である場合、流出と定義する。電気通信業界の顧客に対して、Huangら（2013）とCoussementら(2017)は電気通信顧客が番号を抹消すれば、顧客流出と定義する。張ら（2014）は電信顧客の停止、強制閉鎖、ID抹消などの状態を流出と確定した。金融業界の顧客に対して、Larivièreら（2004）は顧客が金融口座を閉鎖したことを流出と見なしている。顧客流出に関する研究分野では、流出顧客に関する定義は業界によって異なる。また、顧客を失う意思から見ると、顧客を失うことは2つに分けることができる。1つは自発的な流出であり、もう1つは非自発的な流出である。非自発的な流出とは、顧客がサービスを乱用したり、サービスに料金を払ったりしないなどの原因で企業に取り消された顧客を指し、非自発的に流出した顧客は企業管理者に識別されやすい。自発的に顧客を失うとは、顧客が企業の製品やサービスとの関係を自発的に終了し、別の企業の製品やサービス(Hadden et al.、2007)を選択することである。そのため、自発的に顧客を流出する原因は比較的複雑で、流出に影響する要素が多く、自発的に顧客を流出することは企業流出管理の重点的な関心の対象であり、研究者の重点的な研究の課題でもある。

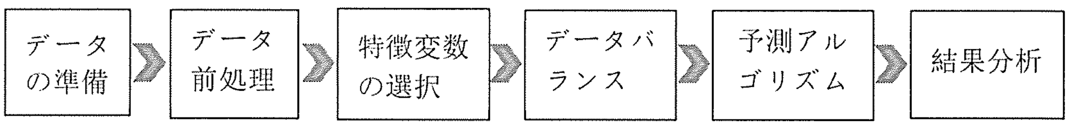
顧客流出管理とは、顧客消費の履歴情報により、顧客の将来の可能な消費行動を

予測することである。顧客の流出確率値を計算することにより、高流出確率の顧客を企業が顧客保留のマーケティング戦略と経営管理活動を展開する主な対象とする(Coussement et al.、2017)。Dattaら（2000）は2000年に顧客流出管理の枠組みを提案し、この枠組みの中で、顧客流出管理の内容は消費データの選択、消費行動の理解、消費特徴の選択、予測モデルの構築と検証を含む。Limaら（2011）は顧客流出管理プロセスを6つのステップに分解した：業務理解、データ理解、データ前処理、予測モデリング、モデル評価、管理実施。これらの先行研究では、顧客流出管理に異なる流出予測モデルと顧客セグメンテーション手法を採用し、異なる研究結果も得た。全体的に言えば、学者たちは顧客流出管理の内包に対して一致した基本的な観点を形成した。すなわち、顧客流出管理は複雑なプロセスであり、顧客が流出するかどうかは、多くの要素の影響を受けている。顧客流出管理の焦点は、流出する可能性のある顧客を正確に識別し、判断し、識別結果に基づいて顧客保留措置を実施することである。

2-2　顧客流出予測の手法

セグメンテーションの視点から見ると、顧客流出は二分類手法で分析できる。すなわち、顧客を「流出」と「非流出」という二つのクラスタに分けている。データマイニング技術を用いて流出予測のフローには、データの準備、データ前処理、特徴変数の選択、予測モデル、結果分析など複数のステップが含まれている。流出予測のフローは図Ⅱ-4 に示す。その中には、従来の研究は特徴変数の選択、データバランスと予測アルゴリズムという3つのステップについて注目しているため、以下では特徴変数の選択、データバランス、予測アルゴリズムの3つのステップの手法について説明する。

図Ⅱ-4　顧客流出予測のフロー



出所：筆者作成。

(1) 特徴変数選択の手法

本研究で取り扱うデータの中にはショッピング時間、商品ページのクリック、商品のコレクションといった高次元データが存在するが、これらのデータの特徴を把握することは困難である。用いられるデータは、高次元データであり、顧客データセットには無関係な変数と余分な変数が多く含まれており、データマイニングで「データセット」を訓練する際に、多すぎる特徴変数は多重共線形、オーバーフィッティング、

過パラメータ化などの問題をもたらすため、予測モデルを構築する前に特徴変数の選択を行う必要があり、そこで、実際にはこの特徴変数選択のプロセスは次元削減のプロセスである。次元削減とは、高次元のデータをできる限り重要な情報を保持したまま低次元データに変換する手法のことである。データ次元削減の方法は線形次元削減と非線形次元削減に分けることができ、非線形次元削減は核関数に基づく方法と特徴値に基づく方法に分けることができる。線形次元削減方法には、主成分分析（PCA)、独立成分分析（ICA)、線形決定分析（LDA)、局所特徴分析（LFA）などがある。核関数に基づく非線形次元削減方法には、核関数に基づく主成分分析（KPCA)、核関数に基づく独立成分分析（KICA)、核関数に基づく決定分析（KDA）などがある。固有値に基づく非線形次元削減方法にはISOMAPとLLEがある。 次元削減の手法は色々とあるが、代表的な手法である「主成分分析:PCA」、「t-SNE」、「ランダムフォレスト:RF」、「UMAP」の4つについて説明する。

A. 主成分分析

主成分分析(Principal Component Analysis : PCA)は、データの分散共分散行列もしくは相関行列に基づき、複数の変数を幾つかの合成変数に縮約することによって、次元の圧縮やデータの可視化、特徴分析を可能にする分析手法である。主成分分析とは、多くの変数の情報をできるだけ損なわずに、少数の変数に縮小させることを目的とした解析手法のことである。

主成分分析の基本ステップは、① データの標準化、中心化、分散の基準化、②分散共分散行列の計算、③分散共分散行列を固有値固有ベクトル分解、④固有値の大きい方からいくつかの固有値固有ベクトルを取ってくる、⑤主成分にデータを射影して視覚化および回帰などの処理を続行。主成分分析のメリットは、データ数（変数の数）を少なくして、その後の可視化や回帰分析を容易にすることができる。デメリットは、元の情報量よりは情報量が落ちる。分析結果は研究者の判断に委ねられる。

B. t-SNE

t-SNE[16]はMaaten and Hinton（2008）によって提案された新しい次元削減アルゴリズムである。t-SNE（T-distributed Stochastic Neighbor Embedding）は柔軟性を持つアルゴリズムであり、データの局所的な特徴を保持できる。基本的な要求は元の距離が近いデータであり、次元を下げた後も距離が近い。元の距離が遠いデータは、次元が削減されてからも距離が遠い。「距離の遠近関係」を確率分布に変換し、各確率分布は「サンプル間距離の遠近」の関係に対応している。次元削減前後のデータはそ

[16] t-SNEが次元圧縮のアルゴリズムであり、特に高次元データの可視化に適すると評価されている。t-SNEはHintonとRoweis (2002) によって提案されたSNE(Stochastic Neighbor Embedding)を改良した方法である。

れぞれ一つの確率分布に対応している。基本的なステップは、まず同じ数の低次元データをランダムに生成し、損失関数を計算し、「勾配降下法」でこれらのデータを更新し、最終的に要求を満たす低次元データを得ることである。

C. ランダムフォレスト

ランダムフォレスト(Random Forest:RF)は決定木に基づく次元削減の手法であり、次元削減の基本原理は特徴変数の重要度に基づいて判断される。変数の重要度は、決定木の「feature importances」によって記述され、この指標は、ある変数の各本数における混雑度の低下の累積平均値と標準差を計算することによって表される。ランダムフォレストを利用するには、PCAのように変数を再構築する必要はない。純度が下がるほど、変数が重要になる。

D. UMAP

UMAP[17] (Uniform Manifold Approximation and Projection : UMAP) は、2018年に提案され、元の特徴空間上で近い点が圧縮後にも近くなるように圧縮される新しい次元削減の手法である。元のデータの特徴を最大限に残すことができるとともに、特徴次元数を大幅に下げることができる。

基本思想は「マニホールド近似」と「投影技術」を利用して、次元削減の目的を達成することである。基本ステップは、まず高次元空間における点間の距離を計算し、それらを低次元空間に投影し、この低次元空間における点間の距離を計算する。そして、[勾配降下法]はこれらの距離間の違いを最小化する。

UMAPのメリットは、t-SNE次元削減手法に対して、UMAPが大規模な情報の損失を避けることができる。UMAPのデメリットは、計算時間が遅く、大きなデータセットを有効に表すことができないことである。

以上より、高次元データに対する次元削減手法についていくつか説明した。消費者行動データを用いてデータの質や量、分析の目的などによってどの手法が最適かというのは異なるので難しい部分もあるが、実際に使用する場合に適切な次元削減の手法を選択する必要がある。

本研究では、次元削減手法の選択は主に4つの要素から考えられ、すなわち、データ量、次元数、データ訓練処理時間、アルゴリズムの計算複雑性と誤差である。ランダムフォレストアルゴリズムは上記の4つの要求を満たすことができ、他のアルゴリズムに比べて大きなメリットがあり、ランダムフォレストの訓練処理時間は短く、高い次元のデータ（つまり、多くの特徴のデータ）を処理することができる。また、デ

---

[17] UMAPはMcInnes and Healy et al. (2018) によって提案された代数トポロジーとリーマン幾何学の理論が元になっている新しい次元削減アルゴリズムである。

ータ訓練プロセスで、特徴変数間の違いを検出することができ、データ訓練が完了した後、ランダムフォレストは特徴変数の重要度を与えることができる。

(2) データバランスの手法

機械学習において分類問題などを扱う場合、実環境において収集したデータの分布が偏りがよく発生して、特に異常時のデータは平常時のデータと比較してサンプル数が少なくなる。このようなデータの分布の偏りは、データの不均衡問題をよく招いて、予測の確率を低くにする。このような場合には、件数が多い方のデータを削減する「アンダーサンプリング」や件数が少ない方のデータを増幅する「オーバーサンプリング」を行い、データのバランスを均一に近づけてから学習を行う、データバーランスという手法がある。本研究における予測は、流出顧客の割合が高くて、顧客クラスタ（流出顧客クラスタと非流出顧客クラスタ）についてデータの不均衡問題が顕著であるため、データバランスの手法を利用して、不均衡データを処理する必要がある。データの視点から、データ不均衡を解決する手法は、少数クラスのデータをオーバーサンプリングしたり、多数クラスのデータをアンダーサンプリングしたり、デフォルトの0.5閾値を調整したりして、既存のサンプルを用いて新しいサンプルを生成したりする。次に、よく使われるデータ不均衡の処理手法を2つ説明する。

A. SMOTEアルゴリズム

SMOTE（Synthetic Minority Over-Sampling Technique）は、少数クラスサンプルを分析し、少数クラスサンプルに基づいて新しいサンプルを人工的に合成し、データセットに追加する。基本的な手順は以下の通りである。

① 最近接アルゴリズムを用いてサンプリングし、少数類サンプル毎のK個の近接サンプルを計算する。

② K個の近傍サンプルからランダム線形補間を行うためにN個のサンプルをランダムに選択する。

③ 新しい少数クラスサンプルを構築する。

④ 新しいサンプルと元のデータを組み合わせて、新しいデータセットを構成する。

B. ADASYNアルゴリズム

ADASYN（Adaptive Synthetic Sampling）は、少数クラスデータサンプルの分布に応じて適応的に生成される少数クラスデータサンプルである。基本的な手順は以下の通りである。

① 不均衡度を計算する。

② 合成が必要なサンプル数を計算する。

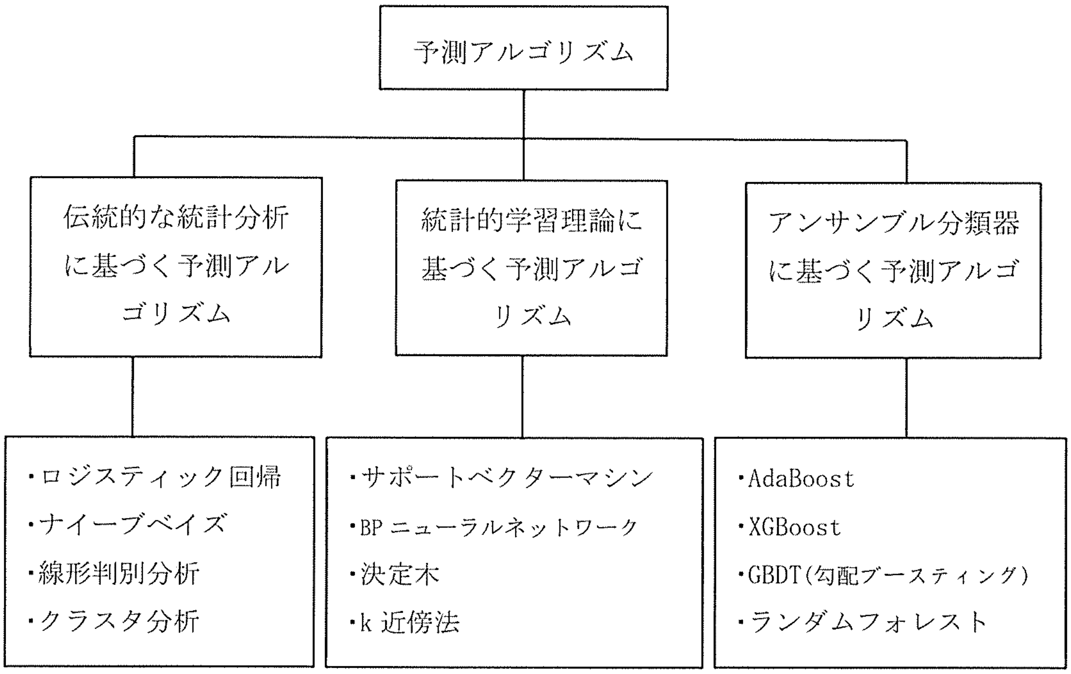
③ 少数クラスサンプルに属する各サンプルに対してユークリッド距離を用いて k 個の近隣サンプルを計算する。

④ 各少数クラスサンプルの周囲の多数クラスサンプルを計算する。
⑤ 少数クラスサンプルごとに合成サンプルの数を計算する。
⑥ 合成すべき少数クラスサンプルの周囲の k 個の近隣サンプルの中から、1 個の少数クラスサンプルを選択し。
⑦ サンプルの必要性が満たされるまで合成を繰り返す。

(3) 予測アルゴリズム

予測アルゴリズムには、大まかに3つのタイプがある。それぞれ、「伝統的な統計分析に基づく予測アルゴリズム」、「統計学習理論に基づく予測アルゴリズム」、「アンサンブル分類器に基づく予測アルゴリズム」である。代表的な流出予測アルゴリズムは図Ⅱ-5 に示す通りである。

図Ⅱ-5 代表的な流出予測アルゴリズム



出所：筆者作成

A. 伝統的な統計分析に基づく予測アルゴリズム

伝統的な統計分析に基づく予測アルゴリズムには代表的なロジスティック回帰、ナイーブベイズ、線形判別分析、クラスタ分析がある。

① ジスティック回帰

ロジスティック回帰（Logistic Regression：LR）は、既知の引数を使用してイベ

ントの発生確率を予測する離散型の引数の値であり、ロジスティック回帰予測は確率値であり、出力値は0から1の間であるべきロジスティック関数（Logistic Function）をフィッティングすることによってイベントの発生確率を予測する。ロジスティック回帰は形式が簡単で、説明性が非常に良いというメリットがある。訓練速度が速い、リソースの消費量が少なく、特にメモリが少ない、最後の予測結果を簡単に得ることができる。ロジスティック回帰のデメリットは精度が高くないこと、データの不均衡の問題を処理するのは難しい、他の方法を導入しない場合は、線形分割可能なデータしか処理できない。

② ナイーブベイズ

ナイーブベイズ（Naive Bayes：NB）は訓練データを用いて P（XY）と P（Y）の推定を学習し、連合確率分布を得る：P(X, Y)=P(Y)P(X|Y)確率推定方法は極大尤度推定またはベイズ推定であることができる。ベイズのメリットは、安定した予測効率があること、欠落したデータに敏感ではなく、アルゴリズムは比較的簡単である。デメリットは、入力データの形式に敏感であり、データセットの属性に関連性がある場合、予測の効果が非常に悪いことである。

③ 線形判別分析

線形判別分析（Linear Discriminant Analysis：LDA）は、データセットを一直線上に投影し、同類サンプルの投影点をできるだけ近づけ、異なる種類のサンプルの投影点をできるだけ遠ざけ、サンプル訓練が完了した後、新しいサンプルをこの直線上に投影し、投影点の位置に基づいて新しいサンプル点のカテゴリを決定する。予測結果は、クラスごとの判別値を算出し、クラスを判別値が最も大きいクラスと予測することにより得られる。線形判別分析のメリットは、ディメンションを下げる過程でカテゴリの事前知識経験を使用できることである。デメリットは、K 個のカテゴリがある場合、LDA は K-1 次元の数までしか下がらないことである。

④ クラスタ分析

クラスタ分析（Cluster Analysis：CA）は既知のデータに基づいて、各観察変数間の相互関係の統計量（例えば、距離、相関係数など）を計算し、ある基準（最短距離法、最長距離法、中間距離法、重心法など）に基づいて、同じクラス内の差を小さくし、クラスとクラス間の差を大きくする。クラスタ解析のメリットは、結果が直感的で結果形式が簡潔であることである。デメリットは、サンプル量が多い場合、予測結果が不正確であることである。

B. 統計的学習理論に基づく予測アルゴリズム

統計学習理論に基づく予測アルゴリズムには代表的なサポートベクターマシン、誤差逆伝播法によるニューラルネットワーク、決定木、k近傍法がある。

①サポートベクターマシン

サポートベクターマシン(Support Vector Machine : SVM)は、ニューロンのモデルとして単純な線形しきい素子を用いて、2 クラスのパターン識別器を構成し、空間を超平面で分割することにより 2つの分類からなるデータを分類し、 2つの分類が正例と負例からなるものとすると、 学習データにおける正例と負例の間隔が 大きいものほどデータで誤った分類をする可能性が低いと考えられる。

②誤差逆伝播法によるニューラルネットワーク

誤差逆伝播法によるニューラルネットワーク(Back Propagation Neural Network : BPNN)は、ネットワークに入力データを与えて、入力層から出力層に向かって順方向の計算を行う。この際、各ニューロンにおいて計算された活性化関数の微分値は次の逆方向計算のために保存しておく。ネットワークの出力データと教師データの差から構成された目的関数に基づいて、出力層から入カ層の方向に向かって目的関数の重み微分値を順次計算する。メリットは、非線形マッピング能力と柔軟性の高いネットワーク構造である。デメリットは学習速度が遅く、局所極小値に陥りやすい、ネットワーク階数とニューロン個数の選択が指導できないことである。

③ 決定木

決定木 (Decision Tree : DT) は、与えられたデータに対して次々に条件を設けて、データを段階的に分類していく手法である。解析結果の特徴は容易かつ直感的に説明することができる。決定木の基本的なプロセスはデータ情報利得を最大となるようにデータを複数のカテゴリに分割し、次に、それぞれのカテゴリに対して情報利得が最大となるようにデータを分割する。そのプロセスでこのような作業を適切な回数で繰り返すことで、決定木が作られる。決定木を作成するときに決定木枝の分割条件が数値の大小比較となるので、データの特徴量を標準化する必要はない。特徴量を用いて標準化しなくても結果には何の影響もない。

④ k近傍法

k近傍法(K-nearest Neighbor : K-NN)は、他の機械学習アルゴリズムとは異なり、教師データからモデルのパラメーターを推測するというステップが存在しない。入力された教師データを記憶するだけである。k近傍法で作られたモデルは、距離に基づいてデータを分類しているため、適切な距離を選ぶことが非常に重要となってくる。距離は特徴量に応じて決められ、K値についてデータセットの特徴にあった数値を選ぶ必要がある。

C. アンサンブル分類器に基づく予測アルゴリズム

アンサンブル分類器に基づく予測アルゴリズムには代表的なアダブースト(AdaBoost)、極端勾配ブースト (XGBoost) 、勾配ブースティング(GBDT)、ランダム

フォレスト（Random Forest：RF）がある。

① アダブースト

アダブースト（AdaBoost）は、精度の高くない識別器（弱識別器と呼ばれる）をいくつか組み合わせることで全体として一つの精度の高い識別器（強識別器と呼ばれる）を生成するアンサンブル学習のアルゴリズムである。AdaBoostアルゴリズムは、逐次的に指標を選択するが、ある時点で選ばれた指標によって誤識別される学習サンプルの重みを次の繰り返しステップでは増加させる。

② 極端勾配ブースト

極端勾配ブースト（XGBoost）は、Boosting ベースのリフティングツリーモデルであり、k モデルの加算モデルであり、多くの弱い分類器を統合して強い分類器を形成する。XGBoost はリフティングツリーモデルのため、多くのツリーモデルを統合して強い分類器を形成し、使用するツリーモデルは CART 回帰ツリーモデルである。

③ 勾配ブースティング決定木

勾配ブースティング決定木(Gradient Boosting Decision Tree：GBDT)は、ブースティングアルゴリズムの一種であり、学習器にあまり高性能なものを使わずに、予測値の誤差を新しく作った弱学習器がどんどん引き継いでいきながら誤差を小さくしていく方法である。複数の弱学習器を統合して全体の学習器を構成する手法である。弱学習器には回帰木を用いることが多く、本研究では弱学習器として回帰木を用いる。

④ ランダムフォレスト

ランダムフォレスト（Random Forest：RF）は、与えられたデータセットから 組のブートストラップサンプルを作成する。各々のブートストラップサンプルデータを用いて未剪定の最大の決定・回帰木を作成する。分岐のノードはランダムサンプリングされた変数の中の最善のものを用いる。全ての結果を統合・組み合わせ、新しい予測器を構築する。

2-3 顧客流出予測の先行研究

顧客流出予測技術を用いて流出する可能性がある顧客を識別し、予測結果に基づいてマーケティング戦略を改善し、既存顧客を維持することで業績減損を効果的に防止することができる。先行文献の収集、分析と調査を通じて、顧客流出予測に関する先行研究レビューを行う。

(1) 伝統的な統計分析に基づく顧客流出予測

伝統的な統計分析に基づく予測方法は主にロジスティック回帰（LR）、ナイーブベイズ（NB）、線形判別分析（LDA）、クラスタ分析（CA）などがあり、例えば、Renjith ら（2015）はロジスティック回帰を用いたイーコマース顧客の流出を予測し、機械学

習方法を用いた個性的な顧客保存戦略を提案した。Caignyaら（2018）は、電気通信業界の顧客流出を予測するために、ロジスティック回帰と決定木を結合した。Jahromiら（2014）はLogitモデルを用いたオーストラリアのB 2 Bイーコマースプラットフォームの顧客流出問題を研究し、決定木モデルとBoostingモデルを比較した。その結果、Logitモデルは顧客流出を効果的に予測できるが、予測精度は他の予測モデルに及ばないことが分かった。Nieら（2011）はある銀行のクレジットカード顧客データに対してLogitモデルを構築し、その潜在的な流出顧客を識別し、決定ツリーモデルの予測効果を比較した結果、Logitモデルは比較的に良い予測効果があることを示した。Pmarら（2011）は、ある通信会社の顧客がナイーブベイズを用いた顧客の流出状況を予測した結果、顧客の平均通話時間が顧客の流出と強い相関性があることを示した。上述のこれらの研究は伝統的な統計分析方法を用いた予測を行い、構築された予測モデルは強い解釈性があるが、これらの方法はビッグデータと多次元性変数データを処理する際に限界があり、予測効果は明らかではない。

(2) 統計学習理論に基づく顧客流出予測

統計学習理論に基づく予測方法は主にサポートベクターマシン（SVM）、人工ニューラルネットワーク（ANN）、決定木（DT）などがあり、例えば、Farquadら（2014）は銀行クレジットカード顧客の流出予測を行い、サポートベクターマシンから規則を抽出する混合方法を提案した。Gordiniら（2017）はB2Bイーコマースの顧客に対して流出予測を行った結果、SVMはノイズ、アンバランス、非線形のB2Bイーコマースデータを処理する際に良い予測効果があることを示した。Tianら（2007）は電信顧客の流出状況を予測し、2層ニューラルネットワーク技術を用いて生データから適切な変数を抽出し、人工ニューラルネットワークに基づく流出予測モデルを提案し、その結果、この方法の予測効果は決定木とナイーブベイズの予測効果より良いことを表明した。Yuら（2018）は電信顧客の流出状況を予測研究し、粒子群最適化（PSO）を用いて予測モデルを訓練し、粒子群最適化に基づくBPニューラルネットワークを提案し、その結果、この方法は流出予測の正確性を高めたことを表明した。Neslinら（2006）は、決定木が顧客流出予測の実際の工程に広く応用されており、決定木アルゴリズムを流出予測の基礎モデルとして応用できると考えている。Zhangら（2015）はC 5.0決定木を用いた電気通信企業の郵便SMSサービスに流出予測を行った結果、C 5.0決定木予測モデルの精度が高いことを示した。

(3) アンサンブル分類器に基づく顧客流出予測

アンサンブル分類器に基づく予測方法は、複数の弱分類器を統合して強分類器を構成する方法である。基本モデルが異なることと統合規則が異なることから、決定木

統合方法（Abbasimehr et al.、2014）、サポートベクターマシン統合方法（Vafeiadis et al.、2015）、ニューラルネットワークの統合方法（Gordini and Veglio、2013）、線形判別法の統合方法（Xie and Li、2008）がある。一般的なアンサンブル分類器の方法としては、AdaBoost、XGBoost、Random Forestなどがある。例えば、Wuら（2016）はイーコマース顧客の流出予測問題を研究した。サンプリング比によって正負類データをバランスさせ、データセットの規模を低減するとともに、同時にAdaBoostアルゴリズムを結合し、分類器の分類精度を高めた結果、AdaBoostは良好な予測効果を持っていることが分かった。時間的特徴を持つ電気通信業界の顧客データセットについて、Jiら（2021）はXGBoostに基づく混合特徴選択アルゴリズムを提案した。このアルゴリズムは2つの視点から特徴選択を行い、予測顧客の流出に最も重要な特徴を選別することができ、冗長な特徴を除去することができ、実験結果はこの方法が良好な予測性能を持っていることを示した。Ahmedら（2019）は電気通信業界の顧客流出予測を研究し、組合せヒューリスティックアルゴリズムに基づく予測モデルを提案した。Yingら（2010）は銀行顧客の二分類問題に対して研究を行い、集積したLDAとBoostingを用いた顧客流出を予測し、良好な予測効果が得られた。Zhangら（2014）はCARTとBoosting統合モデルを用いた電気通信顧客の流出問題を予測した結果、この方法は高い予測精度を持っていることが分かった。

(4) 電信など業界の顧客流出の応用シーン

業界の応用から見ると、先行研究の文献は主に以下の3つの業界にある。

A. 電信業界の顧客流出の応用シーン

Coussementら（2017）はヨーロッパ電信会社の顧客流出問題を研究し、元のデータは30,104の顧客を含み、データ変数は956個を含み、データ変数は人口統計学的特徴、通話行動、運営者との相互作用、コース購読数に関連している。流出顧客は総顧客の4.52%を占め、この研究はLogitモデルを採用して顧客流出予測を行っている。Huangら（2013）は104,199個の電気通信顧客データセットを分析し、データ変数は121個であり、データ変数は顧客の人口統計学情報、アカウント情報、通話情報に関連し、そのうち流出顧客の割合は5.8%を占め、K-means混合モデルに基づいて流出予測を行っている。張ら（2014）は184,761社の電信会社の顧客データを研究し、データには15変数が含まれており、そのうち流出顧客の割合は7.3%を占め、CARTとBoostingアルゴリズムの統合モデルを通じて顧客流出予測を行った。Masandら（1999）はGTE社の20の最大の携帯電話通信市場の顧客を対象とし、単回帰（Simple Linear Regression: SLR）、最近傍分類器（Nearest Neighbor Classifiier: GNC）、決定木（Decision Tree: DT）、人工ニューラルネットワーク（Artificial Neural Networks: ANN）を用いた

顧客流出予測を行っている。丁ら（2015）は、ある電信会社の2013年9月から2014年2月までの7,913件の顧客データを収集し、その中で、流出顧客が3.3%を占め、改善されたランダムフォレスト(Random Forest: RF)アルゴリズムを通じて顧客流出を予測している。羅ら（2018）はBPニューラルネットワーク、RBFニューラルネットワーク、Elmanのネットワーク（Elman Neural Network：ENN）、一般化回帰ニューラルネットワーク(Generalized Regression Neural Network: GRNN)の4つのサブ分類器を線形集積し、人工蜂コロニー（Artificial Bee Colony Algorithm：ABA）アルゴリズムを用いて線形組合せの重みを最適化し、ある電気通信企業の20,000の顧客データの流出予測分析を行っている。

B. 金融業界の顧客流出の応用シーン

Nieら（2011）はある銀行の5,456枚のクレジットカードデータセットを分析している。データ変数は135個をであり、データ変数は所有者の情報、クレジットカード情報、取引情報、異常使用情報に関連し、その中で、非流出顧客は8.1%、流出顧客は91.1%を占め、K-means混合型を用いて顧客流出予測を行っている。Lariviら(2004)はベルギーの金融サービス機関の顧客流出問題を研究した。応ら(2007)はある銀行の12万件の個人信用顧客の取引データを研究し、データ変数は16個を含み、改善されたサポートベクトルアルゴリズムを用いて信用顧客の流出予測問題を研究した。

C. イーコマース業界の顧客流出の応用シーン

Gordini ら（2016）は、イタリアのある会社が 2013 年 1 月から 2014 年 1 月までの間に 80,000 件の顧客データを分析し、その中で、流出顧客が 10%を占め、データ変数は登録情報、取引情報、アクセス情報に関連し、採用した予測アルゴリズムは改善されたサポートベクターマシンである。Yu ら（2011)は中国のあるイーコマースサイト 50,000 人の顧客の登録情報、取引情報、サイトログ情報のモデリングを分析し、改善されたサポートベクトル機を通じて顧客流出予測を行っている。Jahromi ら（2014）は Logit、決定木、Boosting などの複数のモデルを用いて、オーストラリアのある B2B イーコマース企業の顧客流出問題を研究している。

以上より、以前の文献では、研究者は各種予測方法を用いて電気通信業界、銀行業界とB2Bイーコマース企業の契約型取引[18]の顧客の流出問題に対して深い研究を行い、そして各種方法の優位性を検討し、契約型取引の顧客の流出予測の研究に大きな貢献をした。文献からは、一部の研究の実験結果に不一致が生じることがあることがわかる。また、顧客流出予測に関する研究は主に電信と金融業界に集中し、B2Cイーコマ

[18] 契約型取引とは、通信、金融、定期購読の新聞や雑誌、フィットネスクラブなどに典型的にみられるように、初期契約時に一定期間でのサービスを利用することを前提に交わされる契約を指す。『小野譲司、(2008) 契約型サービスにおける顧客関係、マーケティングジャーナル、 Vol.28 No.2、p15-27。』

ース業界に注目する研究はわずかである。B2Cイーコマース企業の顧客流出問題が非常に際立っているため、このような顧客のショッピング行為は多次元性、ショッピングの意欲と傾向が個性化されているため、顧客データの特徴に基づいて、B2C環境下の流出予測モデルを開発し、B2C環境の様々な特徴変数（商品の種類、商品コレクション、商品のショッピングカートへの加入、商品の好み、ショッピング時間）を十分に考慮する必要がある。本研究はB2Cイーコマース企業の非契約型取引[19]の顧客の流出問題について研究を展開する試みである。

小括

本章では、機械学習の概念、手法及び機械学習におけるデータ活用のプロセスを紹介し、この5年間のマーケティング分野における機械学習技術の応用例を列挙した。本研究の顧客流出予測における変数は時間変数、購買意図変数、商品種類変数などを含み、17の変数があり、高次元データに属する。変数が多いと流出予測の精度に影響するため、高次元データを削減する必要がある。そこで、主成分分析（PCA）、ランダムフォレスト（RF）など4種類の変数特徴選択手法の基本原理とそのメリットとデメリットを紹介し、本研究の実証研究部分では特徴変数の重要度の判断手法としてランダムフォレストを選択して変数特徴選別を行う。また、流出と非流出の種別データの差が大きい、すなわち、データサンプルが不均衡であるため、よく使われる2つのデータ不均衡処理手法を紹介した。関連する流出予測の先行研究として、伝統的な統計分析に基づく顧客流出予測の先行研究、統計学習理論に基づく顧客流出予測の先行研究とアンサンブル分類器に基づく顧客流出予測の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。また、電気通信業界、金融業界、イーコマース業界の顧客流出予測の応用シーンをレビューした。

先行文献の収集、分析と調査を通じて、顧客流出予測に関する先行研究は、B2Cイーコマース業界に注目する研究はわずかである。

[19] 非契約型取引とは、製品やサービスに限らず、都度払いで売買が行われる単発的取引である。非契約型の取引では、購買と使用の意思決定がその都度、繰り返し行われる。そして、企業にとっての顧客関係の管理とは、契約型取引では初期契約で獲得した顧客の離脱をいかに少なくするかであり、非契約型取引では顧客の反復購買をいかに増やすかに焦点が向けられる。『小野譲司、(2008) 契約型サービスにおける顧客関係、マーケティングジャーナル、Vol.28 No.2、p15-27。』

Ⅲ　顧客流出予測モデリングの基礎理論

はじめに

本研究では、顧客セグメンテーションと流出予測の 2 つの重要な段階がある。どのクラスタリングアルゴリズムとどの予測アルゴリズムを採用するかは決定しなければならない重要なことである。クラスタリングアルゴリズムの選択は、データのタイプとクラスタリングの目的に依存する。この 5 年間の顧客セグメンテーションに関する文献では、イーコマース業界(Gordini and Veglio：2017, Wu et al.、2020, Wu et al.、2021)、小売業界 (Li et al.、2021, Christy et al.、2021)、金融業界 (Abbasimehr and Bahrini、2021)、電気通信業界 (Coussement et al.、2017, Alboukaey et al.、2020, Zhou et al.、2020) の顧客セグメンテーション方法は大規模なデータを処理でき、計算が簡単で演算効率が高いという特徴があり、k-means の優位性が明らかで、他の業界(Li et al.、2020)でも広く使用されている。予測アルゴリズムでは、ほとんどの文献は、統合データセットを使用して複数のモデルの予測効果を比較することにより、最適アルゴリズムを決定している。流出予測の文献(Gordini and Veglio、2017)では、LR、SVM、Neural Network、Decision Trees などの複数の予測アルゴリズムの比較により、SVM モデルが良好な予測性能を持ち、訓練速度が速いことが明らかになった。本研究も各種アルゴリズムの比較の方法を用いて、予測アルゴリズムを選択する。そこで、本章では、使用する顧客細分化アルゴリズムと特徴変数選択アルゴリズム、および 4 種類の予測アルゴリズムの原理とそのフローについて説明する。

1　k-means について

元田ら(2007)を参考にして、一般的な k-means 法について述べる。

D次元ユークリッド空間上のデータ集合$\{x_1, x_2, \cdots, x_N\}$を$K$個のクラスタに分割(分類)することを考える。このとき、各データ x が K 個のクラスタのいずれに属するのかを示す、2 値変数 rnk∈{0, 1} (k=1, …, K) を定める。データ点xがクラスタkに割り当られている場合にrnk = 1 となり j≠ k の場合にr_{nj} = 0 となる。データ集合を K 個のクラスタに分割するということは、以下の目的関数 J を最小にすることで実現される。

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \left\| x_n - \mu_k \right\|^2 \tag{3-1}$$

k-means 法において最適なクラスタ数(分類数)を決定する手法は確立されていな

い。このような問題への簡単な対処法としてエルボー図を用いる方法がある。エルボー図とは、クラスタ数Kと残差平方和$SSE(K)$(Sum of Squared Errors of Prediction)の関係を示したグラフである。

文献（Birodkar et al.、2019, Onoda et al.、2011）によると、k-means法及びその手順を以下に示す。

① 任意の初期重心点をデータセットからランダムに集めた少量のサンプルデータから決定する。

② 各サンプルから最も近い距離の重心点を計算によって求め、クラスタを構成する。

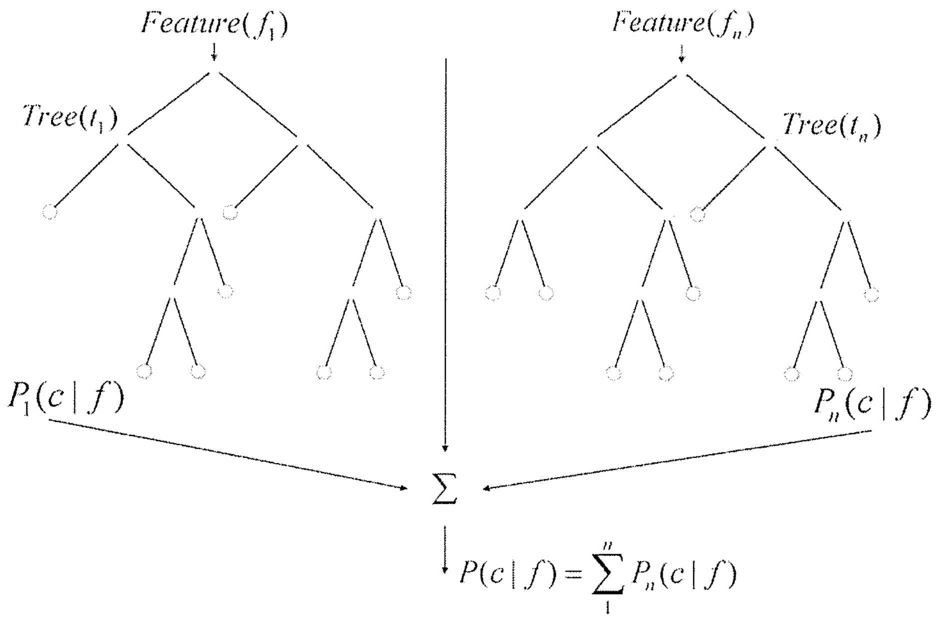
③ ステップ２ で求めたクラスタごとに重心を求め、ステップ２ を再度実行する。

ステップ2〜3 を決められた回数繰り返し実行し、クラスタに大きな変化がなくなるまで計算する。

2　ランダムフォレストについて

ランダムフォレスト（Random Forest : RF）は決定木による複数の学習器を統合させる手法である。関数RandomForestは決定木CARTによって未剪定の木を大量に生成し、多数決をとる方法で判別モデルを構築するアルゴリズムである。決定木より性能の良い識別や予測が可能になり、選択肢を増やしてもエラーが生じにくくなる。文献（金明哲、2009）に基づいて、ランダムフォレスト法を以下に示す。

図Ⅲ-1 ランダムフォレストのアルゴリズムで生成した木



出所：Xuら（2019), Research on a Mixed Gas Classification Algorithm Based on Extreme Random Tree, Appl. Sci. 2019, 9, 1728; p.7 Figure 4により。

$$P\left(c|f\right)=\sum_{1}^{n}P_n\left(c|f\right) \tag{3-2}$$

ランダムフォレスト法の手順を以下に示す。

① データセットからnセットのブートストラップ・サンプル$t1$, $t\ 2$, …, tnを作成する。ただし、構築したモデルを評価するために約1/3のデータを取り除いておいてサンプリングする。取り除いたデータをOOB (Out-Of-Bag) データと呼ぶ。

② t_k (k=1, 2, …, n) におけるM個の変数の中からm個の変数をランダムサンプリングする。m はMより小さい値であり、$m=\sqrt{M}$ が多用されている。

③ ブートストラップ・サンプルt_kのm個の変数を用いて未剪定の最大の決定木T_k を生成する。

④ n個のブートストラップ・サンプルt_k の決定木T_k について、OOBデータを用いてテストを行い、推測誤差を求める（OOB推測誤差と呼ぶ）。その結果に基づいて、新たに分類器を構築する。回帰の問題では平均、分類の問題では多数決をとる。

3　ロジスティック回帰について

ロジスティック回帰 (Logistic Regression : LR) は、一つのカテゴリ変数（二値変数）の成功確率を、複数の説明変数によって説明し、予測する多変量解析の一つである。ロジスティック回帰は古典的な分類方法 (Lee et al.、2006, Minka et al.、2004) であり、分類と予測の問題を解決するための一般的なアルゴリズムの1つであり、既存のカテゴリラベルのデータセットによって未知のカテゴリのデータセットが属するカテゴリの確率を予測し、条件確率分布の形式でP(Y|X)と表すことができる。二分類問題を例にとると、Xはn次元ベクトルであり、Yは0または1の値をとると、予測結果は以下のようになる。

$$P\left(Y=1|X\right)=\frac{exp(wx+b)}{1+exp(wx+b)} \tag{3-3}$$

$$P\left(Y=0|X\right)=\frac{1}{1+exp(wx+b)} \tag{3-4}$$

ここで、線形回帰モデルの実数値はSigmoid関数を介して[0, 1]間の値に変換される。すなわち、あるサンプルXが式(3-3)または(3-4)の計算を経た結果、そのサンプル点Xがあるカテゴリに属する確率の大きさとなる。閾値を設定することで、カテゴリの区

分ができる。一般にSigmoid関数から算出される値が0.5以上の場合はカテゴリ1、値が0.5 以下の場合はカテゴリ0となる。ロジスティック回帰のメリットは、連続性とカテゴリ性の変数に適した結果の解釈が容易であることである。

ロジスティック回帰法の手順を以下に示す。

① 入力引数の特徴。

② 引数の線形結合yを定義する。

③ 線形回帰結果yをsigmoid関数にマッピングし、0～1の範囲の関数確率値を生成する。

④ 確率値に基づいて閾値（通常0.5）を定義する、分類結果の正負を判定する。

4　サポートベクターマシンについて

サポートベクターマシン（Support Vector Machine：SVM)は、数理計画法による判別分析である（Vapnik、2000, Scholkopf、2002)。この分類器は、二次最適化問題を解くことにより、データセット間に最適な分離平面を構築する。非線形分類問題にも核関数の変換により適用することができる。SVM はラグランジュ乗数法によって元の問題をより解きやすい対偶問題に変換し、

$$L(w,\ b,\ \alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}^{m}a_i\left(1-y_i\left(W^T x_i+b\right)\right) \tag{3-5}$$

それぞれ w と b に対して偏導を求め、0 にする、以下の式を得ることができ、

$$W=\sum_{i=1}^{m}a_i x_i y_i=0 \tag{3-6}$$

$$\sum_{i=1}^{m}a_i y_i=0 \tag{3-7}$$

式（3-6)、(3-7）を式（3-5）に代入し、最終的に元の問題を次の目標関数（3-8)に変換して解くことができる。

$$max\sum_{i=1}^{m}a_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}a_i a_j y_i y_j x_i x_j$$

$$s.t.\quad \sum_{i=1}^{m}\alpha_i y_i=0,\quad \alpha_i\gg 0,\quad i=1,2,\ldots,\quad m \tag{3-8}$$

最大の特徴としては、観測データの計量空間で直接判別のための関数を求めるのではなく、特徴空間という元の値から見ると非線形の空間上にデータを射影して判別が行われる。SVM は核関数による非線形決定境界の良好なシミュレーションだけでなく、

「オーバーフィット」も良好に制御することができる。SVMの利点は、[汎化能力]の向上と[高次元データ]処理の問題点である。

サポートベクターマシン法の手順を以下に示す。

① データを入力し、正規化処理を行う。

② データセットを構築し、モデル訓練を行う。

③ モデルパラメータの表示。

④ モデル予測。

⑤ モデルの可視化。

5 誤差逆伝播法によるニューラルネットワークについて

誤差逆伝播法（Back Propagation Neural Network：BPNN）によるニューラルネットワーク（BPNN）で採用されている伝達関数は非線形変換関数であり、Sigmoid関数を用いて実現されている（Ji et al.、2021）。Sigmoid関数は

$$f(x)=\frac{1}{1+e^{-x}} \tag{3-9}$$

関数の定義域は実数全体であり、値域は[0, 1]の間である。xの値が十分に大きい場合は、0または1の2種類の問題とみなすことができ、0.5より大きい場合は1種類の問題とみなし、逆に0.5より小さい場合は0種類の問題とみなすことができる。

入力層・隠れ層・出力層で構成される三層のパーセプトロンを例に、ネットワーク出力と所望出力が異なる場合、出力誤差Eが存在し、以下のように定義

$$E=\frac{1}{2}(D-O)^2=\frac{1}{2}\sum_{k=1}^{l}(d_k\text{-}o_k)^2 \tag{3-10}$$

以上の誤差定義式を隠れ層に展開すると

$$E=\frac{1}{2}\sum_{k=1}^{l}\left[d_k\text{-}f(net_k)\right]^2 \tag{3-11}$$

ネットワーク入力誤差は各層の重みW_{jk}、V_{ij}の関数であるため、重みを調整することで誤差Eを変更することができる。重みを調整する原則は誤差を減少させることであるので、重みと誤差の勾配を比例させるべきである、すなわち

$$W_{jk}=\text{-}\eta\frac{\partial_E}{\partial W_{jk}} \qquad j=0,1,2,\cdots,\ m;\quad k=1,2,\cdots,\ l \tag{3-12}$$

BPNNの特徴は、関数自体とその導関数が連続しているため、処理が非常に便利で

あることであり、データの訓練速度は速く、分類の計算量は特徴数だけに関連しており、結果は解釈しやすく、連続性と分類性の変数に適している。

BPNN 法の手順を以下に示す。

①データ入力

② 訓練データと予測データの設定

③ 訓練サンプルデータの正規化処理

④ BP ニューラルネットワークの構築

⑤ ネットワークパラメータ配置（訓練回数、学習速度、訓練目標最小誤差など）

⑥ BP ニューラルネットワーク訓練

⑦ 試験サンプルの正規化処理

⑧ BP ニューラルネットワークの予測

⑨ 予測結果の逆正規化と誤差計算

⑩ 検証セットの真実値と予測値誤差の比較

6　アダブーストについて

アダブースト(AdaBoost)は、YoavFreund and Robert Schapireによって最初に提案された反復アルゴリズムである（Yoav and Robert、1996)。そのアルゴリズムのコアは、同じ訓練集に対して異なる分類器(弱分類器)を訓練し、これらの弱分類器を集めて、より強い最終分類器を構成することである。訓練集のサンプルは、

$$T=\{(x_1,\ y_1),(x_2,\ y_2),\ \ldots,(x_m,\ y_m)\},\quad w_{1i}=\frac{1}{m};\quad i=1,2,\cdots,\ m \qquad (3\text{-}13)$$

訓練セットのk番目の弱学習器での出力重みは、

$$D(k)=(w_{k1},\ w_{k2},\cdots,\ w_{km}) \qquad (3\text{-}14)$$

顧客流出と非流出は二元分類問題であり、出力が{-1, 1}であれば、訓練セット上のk番目の弱分類器G_k（x）の重み付け誤差率は、

$$e_k=P(G_k(x_i)\neq y_i)=\sum_{i=1}^{m}w_{ki}\mathrm{I}(G_k(x_i)\neq y_i) \qquad (3\text{-}15)$$

k番目の弱分類器G_k（x）の重み係数は、

$$\alpha_k=\frac{1}{2}log\frac{1\text{-}e_k}{e_k} \qquad (3\text{-}16)$$

k番目の弱分類器のサンプル集合重み係数が$D(k)=(w_{k1}, w_{k2}, ...\ w_{km})$であると仮定すると、対応するk＋1番目の弱分類器のサンプル集合重み係数は、

$$w_{k+1,\ i} = \frac{w_{ki}}{Z_k} exp\left(-a_k y_i G_k\left(x_i\right)\right) \tag{3-17}$$

AdaBoost分類は重み付け採決法を採用し、最終的な強分類器は、

$$f\left(x\right) = sign\left(\sum_{k=1}^{k} a_k G_k\left(x\right)\right) \tag{3-18}$$

AdaBoost の手順を以下に示す。

① サンプル重みの初期化、

② 弱分類器を用いて、学習誤差率（重み付けサンプル誤差率）を通じて、弱分類器の重み付けを得る、

③ 前の弱分類器の重みによるサンプル重みの更新

④ 上記の3ステップを繰り返し、すべての分類器が予測を完了するまで、いくつかの弱分類器は最後に簡単な加重によって強い分類器を得た。

AdaBoostは、高い検出速度を持ち、かつフィッティングされにくい現象を実現し、応用するための高度な組合せアルゴリズムである。反復のたびに、弱分類器を訓練すると、そのサンプルセットの各サンプルが対応し、各サンプルは多くの特徴を持つため、膨大な特徴の中から訓練して最適な弱分類器を得る計算量は大きい。

小括

本研究では、顧客セグメンテーションと流出予測の2つの重要な段階がある。どのようなセグメンテーションの手法とどのような流出予測の手法を採用するかは確定しなければならない重要な課題である。第Ⅰ章第2節では、いくつかの顧客セグメンテーションの手法のメリットとデメリットをそれぞれ説明し、第Ⅱ章第2節では、特徴変数選択の手法、データバランスの手法、および予測アルゴリズムとそのメリットとデメリットをそれぞれ説明した。本研究の目的とデータ特徴に基づいて、セグメント化の方法をk-meansとし、変数特徴選択の方法をランダムフォレストとして決定している。予測する手法をそれぞれロジスティック回帰、サポートベクターマシン、誤差逆伝播法によるニューラルネットワーク、アダブーストとする。また、流出予測アルゴリズムの選択については、多くの文献（Coussement et al.、2017, García et al.、2017, De Caigny et al.、2018）が同じデータセットを用いて複数のアルゴリズムの予測効果を比較することにより、最適アルゴリズムを決定し、通常、あるアルゴリズムを予測研究の基準アルゴリズムとしている。ロジスティック回帰は代表的且伝統的な機械学習アルゴリズムであるため、本研究では、基準アルゴリズムとしてロジスティック回帰を採用する。

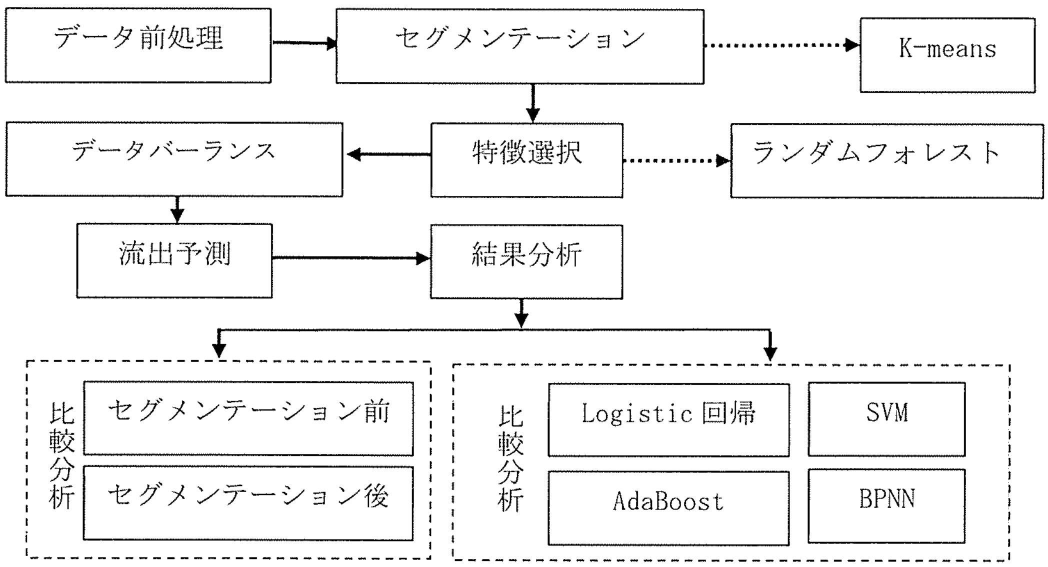
Ⅳ　顧客流出予測の実証研究

はじめに

本章では、B2Cイーコマース顧客の流出予測について実証研究を行い、顧客流出予測モデルを構築する。本研究の顧客とは、中国最大のB2Cイーコマース企業（天猫商城）の消費者をさす。消費者は「天猫商城」と契約する必要がない。つまり、企業と消費者の間に契約関係はない。これらの消費者の商品購入情報データは「TIANCHI天池」ビッグデータ科学研究プラットフォームが発表し、「TIANCHI天池」は学術界に向けてマスデータを開放し、中国ひいては世界数億人のユーザーにサービスを提供している。[天猫商城]の商品には、衣料品、バッグ、靴、日用雑貨、家電製品、健康用品などの各種消費製品が含まれている。アリグループが2022年に発表した財政年度報告によると、天猫商城の「登録ユーザー」数は9億3,900万人に達した。B2Cの顧客は、「天猫商城」から商品を購入するか、いつでも関係を終了することができる。そのため、顧客流出の定義や顧客流出の有無の予測が困難になる。

本章では、まず元のデータとデータの前処理を紹介し、顧客流出を定義している。本章は5節から構成され、第1節は原始データとデータの前処理、データ標準化、特徴変数の選択、第2節は顧客セグメンテーションとクラスタリング、第3節は予測変数の選択、第4節は評価指標、第5節は本研究の結果と分析である。 顧客流出予測の実証研究のフローを図Ⅳ-1に示す。

図Ⅳ-1 顧客流出予測の実証研究のフロー



出所：筆者作成。

1　データについて

1-1　原始データとデータの前処理

本研究で使われている原始データのテーブルの一覧を表Ⅳ-1 のとおり商品の種類、購入商品、購入時間などという 5 つ種類の消費者データが含まれている。

元のデータセットは Alibaba Cloud「TIANCHI 天池」プラットフォーム（2014）が公開した科学研究とビッグデータコンテスト用のデータセットである。このデータセットには、2017 年 11 月 25 日から 2017 年 12 月 3 日までの間にショッピング活動を持つ 987,994 ランダムユーザーのすべての行動データが含まれている。データセットには、User ID（ユーザーID）、itemid（商品 ID）、categoryid（商品種類）、behavior（消費者行動）および timestamps（時刻印）の 5 つの指標があり、このうち消費者行動は PV(項目の詳細ページのページ表示、アイテムクリックに相当する：Page view of an item's detail page、equivalent to an item click)、Buy(商品の購入：Purchase an item)、Cart(ショッピングカートに追加する：Add an item to shopping cart)、Fav(好きな商品を：Favor an item)の 4 種類がある。消費者行動のテーブルを表Ⅳ-2 に示す。

表Ⅳ-1 原始データのテーブル

テーブル名	説明
userid	ユーザーID
itemid	商品 ID
categoryid	商品種類
behavior	消費者行動
timestamps	時刻印

出所：筆者作成。

表Ⅳ-2 消費者行動のテーブル

テーブル名	消費者行動
pv	クリック数
buy	購買数
cart	カート数
fav	フェイヴァリット数

出所：筆者作成。

ユーザーIDには、消費者がタオバオ会員にログインする時使われている個人番号を入力することによりネットで店舗に来店した消費者数の集計と各消費者に属する消費者行動が集計される。商品 ID は消費者がネットで店舗に来店する際に、消費者行動の発生に伴い記録される商品番号を示す。商品の種類は、消費者行動が発生する時、記録される商品が属する商品分野を示す。behavior（消費者行動）には、表Ⅳ-2 のとおり消費者行動（pv、buy、cart、fav）が記録されている。pv というのは、消費者がネットで店舗に来店する際に、ある商品のページを閲覧すること、あるいはクリック数である。buy は、消費者が商品を購入することである。cart には、消費者が商品の購入を予定して、お金がまだ支払われていない状態である。fav とは、消費者がある商品のページを閲覧する際に、好きな商品のページをマークすること、あるいは商品をお気に入りに追加することである。時刻印とは、消費者行動が発生する際に作られた直接読めない時刻印である。12 日間のユーザーデータを整理し、消費者の原始データを主に扱っている、全てのデータをデータベース管理システム MySQL によるデータベースに格納し、統計解析ソフトウェア SPSS によりデータの抽出・可視化・分析を行なっている。

元のデータを前処理し、消費者の購買行動が発生した時間を 2 段階に分け、購買期間の前 6 日を観察期間、購買期間の後 6 日を検証期間とした。観察期間内に 1 回以上購入し、検証期間内に再び 1 回以上購入した顧客を非流出顧客と定義し、0 で表す。観察期間内に 1 回以上購入し、検証期間内に 0 回購入した顧客を流出顧客と定義し、1 で示す。まずユーザーID に基づいて顧客をグループ化し、各顧客の観察期間と検証期間での購入回数を計算し、フィルタリング条件に合致する顧客を保持し、最終的に 8,156 人の顧客の 95,388 件のデータを保持し、そのうち 7,576 人の流出顧客が 92.8%を占めた。580 人の非流出顧客が 7.2%を占めた。顧客のデータにバランシングが発生したため、後続のプロセスでデータのバランシング処理が必要となる。

1-2　データの標準化

複数のデータを比較する場合、平均値や標準偏差が大きく異なると比較することは難しくなり、また測定の単位が異なる場合も同様の問題が生じる。このような場合、データに標準化又は基準化と呼ばれる処理を行い、統一した基準で比較される。

観測値の標準化は、各観測値 X_{ij} に対して、平均値 $\bar{X}_j$ を差し引き、標準偏差 S_j で割ることにより行われる。標準化指標変量、平均値と標準偏差の計算式を以下に示す。

$$\overline{X}_{ij}=\frac{X_{ij}-\bar{X}_j}{S_j}\left(i=1,2,\ldots,n;\quad j=1,2,\ldots m\right) \tag{4-1}$$

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\left(i = 1,2,\ldots,n;\quad j = 1,2,\ldots m\right) \tag{4-2}$$

$$S_j = \frac{1}{n\text{-}1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2\left(i = 1,2,\ldots,n;\quad j = 1,2,\ldots m\right) \tag{4-3}$$

$\bar{X}_{ij}$：標準化指標変量

X_{ij}：n 個の対象についての、m 種の対象観測値を表わす組のデータ

$\bar{X}_i$：平均値；S_i：標準偏差

顧客流出予測では、データの事前処理が重要なステップである。まず、ソフトウェア「Navicat15」を使って、元のデータの各データに対するタイムスタンプを、時間フォーマットに適合する「年月日」、「時間分秒」に変換する。顧客ショッピング取引データには行動意図が隠されているため、行動は問題の発生において重要な役割を果たす可能性があるが、従来の顧客行動分析では、行動意図は潜在的な要因として弱められ、無視されている（Cao、2010, Stolfo et al.、2006)。そこで、顧客の異なる時間帯におけるショッピング行動の意図を深く研究するために、本研究ではショッピング行動の発生時間をさらに区分した。われわれの定義は、00:00-06:00は未明(Daybreak)、06:00-12:00は午前(AM)、12:00-18:00は午後(PM)、18:00-00:00は夜(Night)をあらわし、各顧客のこの4つの時間帯のショッピング行動を統計分析する。

データセットの行動データには、PV 数、Buy 数、Cart 数、Fav 数が含まれる。行動データをさらにセグメンテーションし、最終的に整理されたデータ型には 17 種類の変数が含まれ、具体的には商品種類(items of categories)、未明(Daybreak)PV、未明(Daybreak)Buy、未明(Daybreak)Cart、未明(Daybreak)Fav である。午前(AM)PV、午前(AM)Buy、午前(AM)Cart、午前(AM)Fav;午後(PM)PV、午後(PM)Buy、午後(PM)Cart、午後(PM)Fav;夜（Night）PV、夜（Night）Buy、夜（Night）Cart、夜（Night）Fav である。

2 顧客セグメンテーションと顧客クラスタ

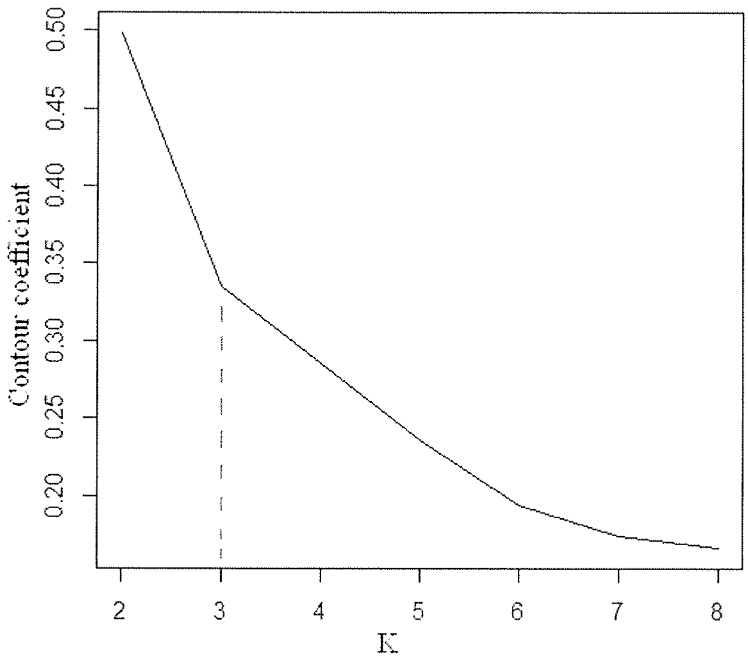
2-1 k-means と顧客セグメンテーション

顧客セグメンテーションに関する文献では、電気通信業界（Coussement et al.、2017, Alboukaey et al.、2020, Zhou et al.、2020)、金融業界（Nie et al.、2011, Abbasimehr and Bahrini、2021)、小売業界（Li et al.、2021, Christy et al.、2021)、B2Bイーコマース業界（Jahromi et al.、2014, Wu et al.、2020, Wu et al.、2021）の顧客セグメンテーション方法は、大規模なデータを処理し、計算が簡単で高い演算

効率を持つK-meansを採用しており、他の業界（Runge et al.、2014, Li et al.、2021）でも広く使用されている。K-meansは教師なしクラスタリングの古典的アルゴリズムとして全自動分類に類似しており、クラスタ内が類似するほどクラスタリング効果が良く、簡単で実現しやすく、良好なクラスタリング効果があるため広く用いられている。そのため、本研究ではK-meansを用いて顧客セグメンテーションを行っている。

データの前処理が完了した後、顧客のセグメンテーションを流出予測の第一ステップとし、予測研究の第一段階において任務は、顧客群のショッピング行動に基づいて、顧客群を様々なタイプに区分し、どの顧客が価値のあるコア顧客であり、どれが忠実な顧客であり、どれが流出しやすい顧客であるかを明確にすることである。K-meansクラスタリングの変数は前述の17個の変数であり、与えられたタグのないサンプルデータセットに対してクラスタリングクラスタ数Kを事前に決定し、クラスタ内のサンプルをできるだけ密に分布させ、クラスタ間の距離をできるだけ大きくする。クラスタ数Kを2から8まで列挙することにより、輪郭係数とクラスタ数Kの関係を得た、図Ⅳ-2 において、横軸はKを表し、縦軸は輪郭係数を表し、k＝3の場合、明らかに曲線に変曲点が現れる。したがって、クラスタリング数は3、つまり顧客が3つに分類されていることを確認する。各クラスタリングクライアント群の流出と非流出のサンプル個数を観察し、各顧客のクラスタリング結果を表Ⅳ-3 に示す。

図Ⅳ-2　輪郭係数とｋの関係



出所：筆者作成。

表Ⅳ-3　K-means クラスタリング結果

顧客タイプ	非流出と流出	実際の顧客タイプ数	クラスタ顧客数
クラスタⅠ	0	484	4935
	1	4451	
クラスタⅡ	0	83	2697
	1	2614	
クラスタⅢ	0	13	524
	1	511	

(0: 非流出顧客;　1:流出顧客)

出所：筆者作成。

2-2　顧客クラスタと流出顧客の確定

顧客クラスタリング分析の目的は予測モデルの効果を検証するために、各クラスタの顧客流出数と顧客非流出数を統計している。

表 IV-3 において、[0]は非流出顧客を示す、すなわち、観察期間内に 1 回以上購入し、検証期間内に再び 1 回以上購入した顧客を非流出顧客とする。[1]流出顧客を示す、すなわち、観察期間内に 1 回以上購入し、検証期間内に 0 回購入した顧客を流出顧客とする。表Ⅳ-3 に k-means アルゴリズムを用いた顧客クラスタリングの結果を示す。

表Ⅳ-3 から見ると、クラスタ I の顧客は 4,935 人であり、8,156 人の顧客数の約 60.5%を占めている。そのうち非流出の顧客は 484 人、流出の顧客は 4,415 人である。クラスタ II の顧客は 2,697 人であり、8,156 人の顧客数の約 33%を占めいる。そのうち非流出顧客は 83 人、流出顧客は 2,614 人である。クラスタ III の顧客は 524 人であり、8,156 人の顧客数の約 6.4%を占めている。そのうち非流出顧客は 13 人、流出顧客は 511 人である。クラスタリング後の顧客タイプを観察すると、クラスタⅠ顧客の流出率は 90.2%、クラスタⅡ顧客の流出率は 96.9%、クラスタⅢ顧客の流出率は 97.5%である。これで、各クラスタ顧客の流出率からみるとクラスタⅠ顧客の流出率が一番低くて、注目しなければならない重要な顧客層と考える。

3　ランダムフォレストと予測特徴変数の確定

3-1　誤判別率 OOB の計算

顧客データセットには多くの顧客行動変数が含まれており、すべての変数が流出予測性能に寄与するわけではなく、データセットの冗長性と無関係の変数がモデルの予測性能を妨げる可能性がある(Verbeke et al.、2012)。流出予測に用いられる変数としては、データセットから直接的に導出される消費者購入情報や購入時間の情報などすべての変数（以下、一般変数と呼ばれる）と、データを集計する人にとって予測意味を持つ消費者買い物行動の情報である変数（以下、特徴変数と呼ばれる）の 2 種類がある。従来のイーコマース顧客のセグメンテーションと予測研究においては、[一般変数]を用いたものが多かったが、分類するが細かくなればなるほど、より代表的な[特徴変数]が必要となる。これらの分析結果を特徴変数として用いるためには多くの難点がある。第 1 に、計算量が膨大なものとなる。第 2 に、これらの特徴変数は B2C イーコマース顧客データのために用いられるものであり、他のタイプのイーコマース顧客データと金融業などのデータには適用できない。

そこで、次に特徴変数選択を行う。ランダムフォレストは有効な特徴変数選択アルゴリズムであり、分類精度が高く、ノイズや異常値に対して優れたロバスト性を有し、かつより強い汎化能力を有している（Breiman、2001)。そのため、ランダムフォレストは商業管理、経済金融、生物情報などの分野で広く応用されている。本研究の一般変数は 17 個あり、情報の煩雑化を招きやすいため、ランダムフォレストを用いて特徴変数を選択した。特徴変数の選択過程における重要な問題は、特徴変数の個数(M)をどのように選択するかであるため、誤判別率[20](Out-Of-Bag error rate：OOB error rate)を用いて特徴変数の個数を決定する（Breiman、1996)。ランダムフォレストで各ツリーを構築する際、訓練セットに対して異なる bootstrap sample を用い、サンプリングの特徴変数から OOB 誤判別率の計算結果を表Ⅳ-4 に示す。一般変数の数は 3 個、4 個、…、10 個、11 個で、計算された誤判別率はそれぞれ 0.080、0.081、…、0.104、0.104 である。

[20] 誤判別率（Out-Of-Bag 率 、OOB 率）とは各決定木の構築において、すべての訓練サンプルを用いるのではなくランダムに選ばれた N′ 個のサンプルを用いている。そして、決定木の構築に用いなかった残りの N-N′ 個のデータを Out-Of-Bag(OOB) サンプルとよび、これを用いることで Random Forests の性能を評価することができる。訓練サンプル中のあるサンプル s に着目する。T 本の決定木の中でサンプル s を学習に用いていない決定木をすべて抽出し、Random Forests のサブセットを作成する。この Random Forests のサブセットに対してサンプル s をテストサンプルとして識別を行い、正しく識別できるかを評価する。これをすべての訓練サンプルについて行い、誤答率の平均をとる。これを Out-Of-Bag 誤り率とよび、構築された Random Forests の識別能力を表す。『飯山将晃、(2015) 使える！統計検定・機械学習-IV　-Random Forests を用いたパターン」認識システム/制御/情報、Vol.59 No.2 、p.71-76。』

表Ⅳ-4 誤判別率(OOB error rate)

一般変数の数	3	4	5	6	7	8	9	10	11
OOB error rate	0.080	0.081	0.083	0.089	0.097	0.098	0.099	0.104	0.104

出所：筆者作成。

3-2　変数重要度の算出と特徴変数の確定

ランダムフォレストの具体的な実装は、ランダムフォレスト関数を用いて行われる。パラメータ ntree は森の中の木の数を表し、木の数はデフォルトで 500 で、mtry は木ごとに使用される変数の数を表し、n 個の変数の中からランダムに mtry 個の変数を抽出して木を構築するたびに、木のランダム性を増加させることができ、さらに良い変数データにフィッティングすることができ、データの最適な分類を実現する。

ランダムフォレストを構築するもうひとつの重要な問題は、最適な特徴個数（m）をどのように選択するかであり、この問題を解決する方法は主に OOB error（誤判別率）の計算を完了しなければならない。ランダムフォレストを使用して木を構築する場合、われわれは訓練データセットに対して異なる bootstrap sample を使用し、ランダムフォレストサンプリングの特徴に基づいて、データの訓練に OOB 推定の計算を完了させることができ、以上の計算によって、OOB error を得ることができる。

ランダムに選択される各一般変数の数を変えると、得られる誤判別率の差は極めて小さく、一般変数の数 M の影響数は大きくない。選択する一般変数数が 4 のとき、誤判別率は比較的小さいので、パラメータ mtry が 4 のランダムの木を確立し、表Ⅳ-5 に示すように変数重要度を出力する。

ランダムフォレストにおける平均減少 Gini 指標（ジニ係数）[21]は一般変数の重要度を判別することができ、平均減少 Gini 指標の数値を計算することにより各一般変数の重要度を判断し、平均減少 Gini 数値が大きいほど一般変数の重要度が強いことを示す（Goldstein et al.、2011)。表Ⅳ-5 に示すように一般変数重要度を出力する。

表Ⅳ-5 から見ると、[一般変数]は 17 個あり、この 17 個の変数の中で、[平均減少 Gini]の計算結果の大きさに基づいて、順序よく並べる。明らかに、「夜 Buy」、「午後

[21] ランダムフォレストは、決定木を大量に作成して、最後に決定木で多数決を取り、最終的な結果を予測する。決定木は不純度という指標を使い、特徴量の領域を分割する。不純度はジニ係数、エントロピーの 2 つの指標があるが、ここでは「ジニ係数」を使う。ジニ係数は分割領域に正解ラベルが混在した状態だと 0.5 に近く、正解ラベルが分割された状態だと 0 に近くなる。ジニ係数については、『Breiman, L.、(1996) Bagging predictors, Mmachine learning、24:123-140.』を参照。

Buy」、「夜 PV」、「午後 PV」、[午前 PV]の 5 個の変数の[平均減少 Gini]は前に並んで、それぞれ「135.2431 」、「120.5565」、「114.7848」、「110.5329」と[102.2129]である。これら 5 つの変数の値は 100 より大きい。一方、 [午前 PV]という変数の[平均減少精度]は-6.71931 であることを考慮したため、[平均減少 Gini]と[平均減少精度]の 2 つの要素を総合的に考慮して、われわれは流出予測の特徴変数として「夜 Buy」、「午後 Buy」、「夜 PV」、「午後 PV」の 4 つ変数を選択している。

以上の一般変数重要度の計算により、顧客流出予測に用いられる特徴変数が決定された。

表IV-5 ランダムフォレスト変数の重要度

順番	一般変数	平均減少 Gini (MDG)	平均減少精度（MDA）
1	夜 Buy	135.2431	23.645784
2	午後 Buy	120.5565	26.333866
3	夜 PV	114.7848	-2.905496
4	午後 PV	110.5329	-5.03892
5	午前 PV	102.2129	-6.711931
6	午前 Buy	87.24348	21.823613
7	商品種類	68.46213	16.58795
8	未明 PV	47.48254	-2.350555
9	夜 Cart	41.98615	-3.251536
10	未明 Buy	37.84351	12.998789
11	午後 Cart	31.62181	-2.414922
12	午前 Cart	25.06281	-1.886356
13	夜 Fav	15.16813	-4.411997
14	午後 Fav	13.76148	-0.308831
15	午前 Fav	11.68928	-2.350264
16	未明 Cart	10.74105	0.3483816
17	未明 Fav	6.106746	1.8525728

出所：筆者作成。

3-3　データの不均衡処理

非流出顧客と流出顧客の数(580非流出、7,576流出)にバランスがとれていないという問題があるため、本段階では不均衡データを処理する。不均衡データを処理する最も一般的な技術の1つは、オーバーサンプリングやアンダーサンプリングなどのサンプリングである（Drummond and Holte、2003)。本研究では、オーバーサンプリング法を用い、オーバーサンプリングの割合は1:1であり、データセットをSMOTE(Chawla et al.、2002）によりバランスしている。バランスしたデータセットを表Ⅳ-6 に示す。

表Ⅳ-6　バランスしたデータセット

データ名	流出	非流出	正クラス：負クラス
Sample dataset	7576	580	1:13
SMOT balance	3788	3788	1:1
Cluster Ⅰ	4451	484	1:9.2
SMOT balance	2225	2225	1:1
Cluster Ⅱ	2614	83	1:31.5
SMOT balance	1307	1307	1:1
Cluster Ⅲ	511	13	1:39.3
SMOT balance	256	256	1:1

出所：筆者作成。

4　流出予測モデルの評価指標

4-1　混同行列について

予測モデルの性能評価方法は、主に混同行列を生成し、その上でモデル予測後の精度（Accuracy)、再現率(Recall)、適合率（ Precision）の 3 つの指標をそれぞれ計算し、その後、これら 3 つの評価指標を用いて予測モデルの効果と性能を評価し、ROC曲線を描いた後、曲線の下の面積[Area Under the receiver operating Curve：AUC]

を用いてモデルの総合評価を行う(Provost、1999, Fan and Ke、2010)。モデル評価の混同行列を表IV-7 に示す。

流出予測のモデル性能を評価するにあたって、正しく予測した場合の数と間違って予測した場合の数を定量化する必要がある。この定量化には、混同行列が用いられている。陽性・陰性のラベルが既知のデータを機械学習のモデルに予測させ、陽性を正しく陽性と予測した場合の数(TP)、陽性を間違って陰性と予測した場合の数(FP)、陰性を正しく陰性と予測した場合の数(TN)、陰性を間違って陽性と予測した場合の数(FN)をまとめると、次のような2×2の混同行列にまとめることができる。

表IV-7　予測モデル評価の混同行列

<table>
<tr><td colspan="2" rowspan="2">混同行列</td><td colspan="2">予測結果</td></tr>
<tr><td>陽性</td><td>陰性</td></tr>
<tr><td rowspan="2">事実
(ラベル)</td><td>陽性</td><td>TP; True Positive
真陽性 (True Positive)に予測した</td><td>FN; False Negative
偽陰性(false negative)に予測した</td></tr>
<tr><td>陰性</td><td>FP; False Positive
偽陽性(False Positive)に予測した</td><td>TN; True Negative
真陰性(True Negative)に予測した</td></tr>
</table>

出所：筆者作成。

4-2　精度･再現率･適合率と ROC 曲線

(1)精度(Accuracy)

精度は、すべての予測が正確な正･負のサンプルの数とサンプルの合計数の割合である。主にモデル全体予測の正確な状況を評価する。精度の計算式を式4-4に示す。

$$\text{Accuracy}=\frac{TP+TN}{TP+FN+FP+TN} \tag{4-4}$$

(2) 再現率(Recall)

再現率は、正サンプルを正しく予測する数がすべての実際の正サンプル数に占める割合であり、主にモデルのカバレッジを説明する。再現率の計算式を式4-5に示す。

$$\text{Recall}=\frac{TP}{TP+FN} \tag{4-5}$$

(3)適合率 (Precision)

適合率は、「すべての予測が正のサンプルである」数に占める「正のサンプルを正しく予測する」数の割合である。主に「予測プラスサンプル」の正確さを説明する。適合率の計算式を式 4-6 に示す。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{4-6}$$

以上の 3 つの評価指標において、TP、TN、FP、FN の意味は、

TP と TN は、それぞれどの予測結果が正しい (真陽性 (TP) と真陰性 (TN)) かを表す。FP と FN は、それぞれどの予測結果が誤り (偽陽性 (FP) と偽陰性 (FN) かを表す。

(4) ROC 曲線と AUC

ROC 曲線 (Receiver Operating Characteristic Curve : 受信者動作特性曲線) は、流出予測モデルの精度の評価に用いられ、どの範囲でカットオフポイント (cut-off point) を取るかを示すものである。カットオフポイントをどこに取るかで、予測モデルの予測能力を視覚的に示すことが可能となる。ROC は予測モデルの感受性と特異性を評価する総合的な指標である。

基本手法はモデルの敏感度と特異性の相互関係を構図法で明らかにすることである。連続変数を複数の異なる臨界値に設定することで、一連の敏感性と特異性を計算し、敏感度 (sensitivity) を縦軸、特異度 (specificity) を横軸として ROC 曲線を描画する。曲線下の面積が大きいほど、予測精度が高くなる。ROC 曲線では、座標図の左上に最も近い点は 敏感度と特異度のいずれも高い臨界値である。予測モデルの優劣を判定する場合は、この曲線がより左上方に位置するほど 優れていると判断する。ROC 曲線下の面積を AUC として計算することができ、AUC が 1 に近づくほどモデルの性能が良いとされる。

敏感度は、陽性のデータを正しく陽性と予測した割合である。敏感度の計算には、陰性データの予測結果が含まれていない。そのため、すべてのデータを陽性と判定すれば (FN = 0)、敏感度を 100%にすることができる。敏感度の計算式を式 4-7 に示す。

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{4-7}$$

特異度は、陰性のデータを正しく陰性と予測した割合である。特異度の計算には、陽性データの予測結果が含まれていない。そのため、全てのデータを陰性と判定すれば (FP=0)、特異度を 100% にすることができる。特異度の計算式を式 4-8 に示す。

$$\text{TNR}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{4-8}$$

本研究の結果と分析において、以上の各指標を用いて予測モデルの性能を評価し、比較分析を行っいる。

5 結果と分析

データの不均衡処理が完了した後、処理後のデータをそれぞれLR、SVM、BPNNとAdaBoostに入力して予測し、収束するまで複数回を反復した。予測実験は10折交差検証法を用いて、データを10部に作成し、訓練セットとして9部、テストセットとして1部、実験10回で得られた予測実験結果の平均値をLR、SVM、BPNNとAdaBoostの最終評価の結果とした。LR、SVM、BPNNとAdaBoostモデルを試験セットデータに適用して、混同行列を導出した。

本研究の目的の1つは、セグメンテーション前とセグメンテーション後の予測効果を比較し、セグメンテーションに基づく流出予測方法の有効性を検証することである。具体的な結果は、(1) 顧客セグメンテーション前の予測結果、(2) 顧客セグメンテーション後の予測結果。

5-1 顧客セグメンテーションと流出予測実験の結果

流出予測モデルの評価指標における計算式に基づいて、式4-4で精度を計算し、式4-5で再現率を計算し、式4-6で適合率を計算し、精度、再現率、適合率、AUC値の計算結果を混同行列に入れ、ROC曲線を描画する。表Ⅳ-8～表Ⅳ-11は、セグメンテーション前にLR、SVM、BPNN、AdaBoostの予測アルゴリズムで得られた混同行列を示す。表Ⅳ-12～表Ⅳ-15は、セグメンテーション後にLR、SVM、BPNN、AdaBoostのの予測アルゴリズムで得られた混同行列を示す。

表Ⅳ-8 セグメンテーション前のロジスティック回帰（LR）混同行列

	Predicted		Accuracy	Recall	Precision	AUC
	Predicted Positive (0)	Predicted Negative (1)				
Actual Positive	753	1	0.9065	0.9984	0.8149	0.938
Actual Negative	140	618				

出所：筆者作成。

表IV-9 セグメンテーション前のサポートベクターマシン(SVM)混同行列

<table>
<tr><td rowspan="2"></td><td colspan="2">Predicted</td><td rowspan="2">Accuracy</td><td rowspan="2">Recall</td><td rowspan="2">Precision</td><td rowspan="2">AUC</td></tr>
<tr><td>Predicted Positive (0)</td><td>Predicted Negative (1)</td></tr>
<tr><td>Actual Positive</td><td>753</td><td>1</td><td rowspan="2">0.9081</td><td rowspan="2">0.9990</td><td rowspan="2">0.8175</td><td rowspan="2">0.942</td></tr>
<tr><td>Actual Negative</td><td>138</td><td>620</td></tr>
</table>

出所：筆者作成。

表IV-10 セグメンテーション前の BP ニューラルネットワーク(BPNN)混同行列

<table>
<tr><td rowspan="2"></td><td colspan="2">Predicted</td><td rowspan="2">Accuracy</td><td rowspan="2">Recall</td><td rowspan="2">Precision</td><td rowspan="2">AUC</td></tr>
<tr><td>Predicted Positive (0)</td><td>Predicted Negative (1)</td></tr>
<tr><td>Actual Positive</td><td>749</td><td>5</td><td rowspan="2">0.9082</td><td rowspan="2">0.9937</td><td rowspan="2">0.9229</td><td rowspan="2">0.951</td></tr>
<tr><td>Actual Negative</td><td>134</td><td>624</td></tr>
</table>

出所：筆者作成。

表IV-11 セグメンテーション前の AdaBoost 混同行列

<table>
<tr><td rowspan="2"></td><td colspan="2">Predicted</td><td rowspan="2">Accuracy</td><td rowspan="2">Recall</td><td rowspan="2">Precision</td><td rowspan="2">AUC</td></tr>
<tr><td>Predicted Positive (0)</td><td>Predicted Negative (1)</td></tr>
<tr><td>Actual Positive</td><td>718</td><td>36</td><td rowspan="2">0.9446</td><td rowspan="2">0.9524</td><td rowspan="2">0.9372</td><td rowspan="2">0.994</td></tr>
<tr><td>Actual Negative</td><td>48</td><td>710</td></tr>
</table>

出所：筆者作成。

表IV-12 セグメンテーション後の LR 混同行列

		Predicted Positive (0)	Predicted Negative (1)	Accuracy	Recall	Precision	AUC
Cluster I	Actual Positive	10	2	0.9050	0.8496	0.9176	0.963
	Actual Negative	4	47				
Cluster II	Actual Positive	107	0	0.9098	1	0.8728	0.955
	Actual Negative	33	228				
Cluster III	Actual Positive	635	0	0.9050	1	0.7697	0.904
	Actual Negative	102	343				
Average				0.9066	0.9498	0.8533	0.940

出所：筆者作成。

表IV-13 セグメンテーション後の SVM 混同行列

		Predicted Positive (0)	Predicted Negative (1)	Accuracy	Recall	Precision	AUC
Cluster I	Actual Positive	11	1	0.9256	0.9184	0.9298	0.981
	Actual Negative	4	48				
Cluster II	Actual Positive	107	0	0.9158	0.9982	0.8820	0.965
	Actual Negative	31	231				
Cluster III	Actual Positive	635	0	0.9053	0.9998	0.7706	0.932
	Actual Negative	102	343				
Average				0.9156	0.9721	0.861	0.959

出所：筆者作成。

表IV-14 セグメンテーション後の BPNN 混同行列

		Predicted		Accuracy	Recall	Precision	AUC
		Predicted Positive (0)	Predicted Negative (1)				
Cluster I	Actual Positive	11	2	0.9287	0.8764	0.9404	0.940
	Actual Negative	3	48				
Cluster II	Actual Positive	105	2	0.9164	0.9766	0.8916	0.963
	Actual Negative	28	233				
Cluster III	Actual Positive	633	2	0.9052	0.9973	0.9940	0.913
	Actual Negative	101	344				
Average				0.9167	0.9501	0.942	0.938

出所：筆者作成。

表IV-15 セグメンテーション後の AdaBoost 混同行列

		Predicted		Accuracy	Recall	Precision	AUC
		Predicted Positive (0)	Predicted Negative (1)				
Cluster I	Actual Positive	11	1	0.9604	0.9145	0.9719	1.00
	Actual Negative	1	50				
Cluster II	Actual Positive	100	7	0.9663	0.9346	0.9792	0.999
	Actual Negative	5	256				
Cluster III	Actual Positive	601	34	0.9399	0.9457	0.9303	0.990
	Actual Negative	31	414				
Average				0.9555	0.9316	0.9604	0.996

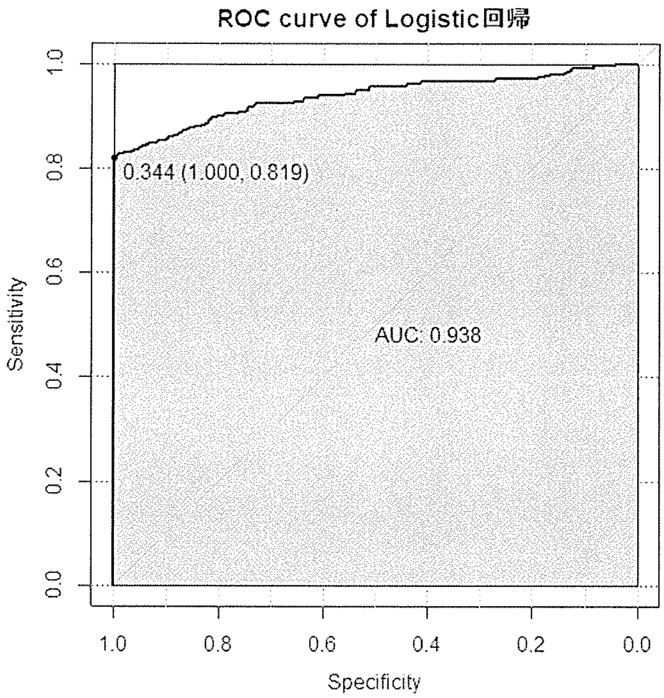
出所：筆者作成。

LRとAdaBoostを例に、表IV-8と表IV-11から見ると、セグメンテーション前のLRの精度、再現率、適合率は0.9065、0.9984、0.8149であったが、AdaBoostの精度、再現率、適合率は0.9446、0.9524、0.9372であり、結果からみると、AdaBoostアルゴリズムの精度値と適合率値はLRアルゴリズムの評価指標値より高い、また、精度値は0.0381より高い、適合率値は0.1123より高い。

表IV-12と表IV-15から見ると、セグメンテーション後のLR精度、再現率、適合率は、「ClusterI」は0.9050、0.8496、0.9176であり、「ClusterII」は0.9098、1、0.8728であり、「ClusterIII」は0.9050、1、0.7697である。AdaBoostの精度、適合率、再現率は、「ClusterI」は0.9604、0.9145、0.9719であり、「ClusterII」は0.9663、0.9346、0.9792であり、「ClusterIII」は0.9399、0.9457、0.9303である。また、AdaBoostアルゴリズムの精度値と適合率値はLRアルゴリズムの評価指標値よりも高い。セグメンテーション前の結果とセグメンテーション後の結果を比較すると、セグメンテーション後の各指標値が上昇していることがわかる。

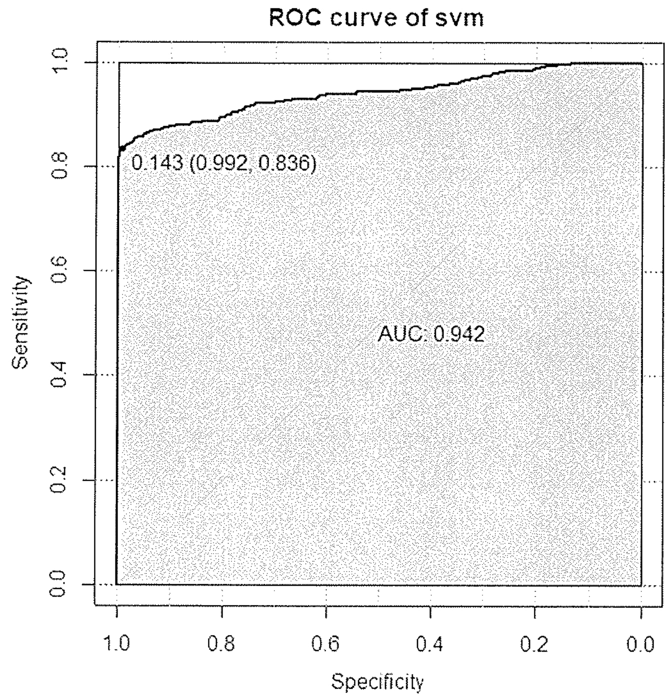
一方、セグメンテーション前とセグメンテーション後のROC曲線とAUC値は、明らかに異なっている。図IV-3～図IV-6は、セグメンテーション前にLR、SVM、BPNN、AdaBoostで得られた各アルゴリズムのROC曲線及び曲線下の面積AUC値を示している。図IV-7～図IV-10 は、セグメンテーション後にLR、SVM、BPNN、AdaBoostで得られた各アルゴリズムのROC曲線及び曲線下の面積AUC値を示す。

図IV-3 セグメンテーション前の LR の ROC 曲線



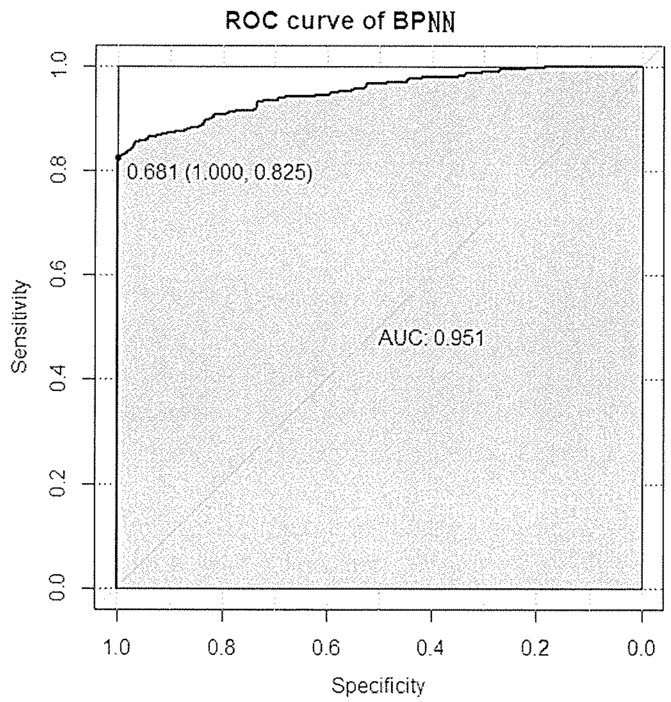
出所：筆者作成。

図IV-4 セグメンテーション前の SVM の ROC 曲線



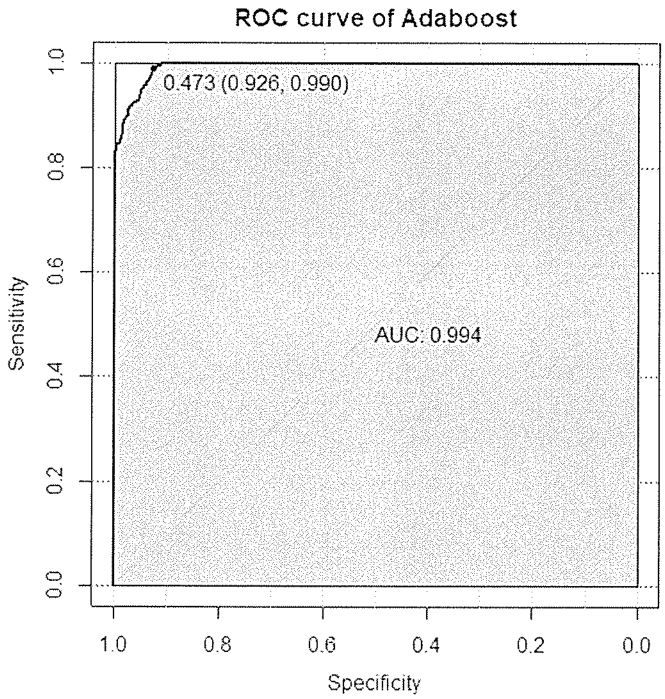
出所：筆者作成。

図IV-5 セグメンテーション前の BPNN の ROC 曲線



出所：筆者作成。

図Ⅳ-6 セグメンテーション前の AdaBoost の ROC 曲線



出所：筆者作成。

新規イーコマース顧客のデータ（未知データ）処理と流出予測を行う場合、予測モデルの汎化性能は非常に重要である。予測モデルの汎化性能とは、訓練学習時に予測モデルで与えられた訓練データだけでなく、未知の新たなデータの両方をうまく予測できる性能である。未知のデータに対する予測性能は予測モデルの優劣の評価によく使われている。AUC値も同じで予測モデルの汎化性能を評価する時に用いられる重要な指標である。

予測モデルの優劣をROC曲線で判別する場合は、図の左上の臨界値と曲線下の面積AUC値を基準としてモデルの汎化性能を判別する。

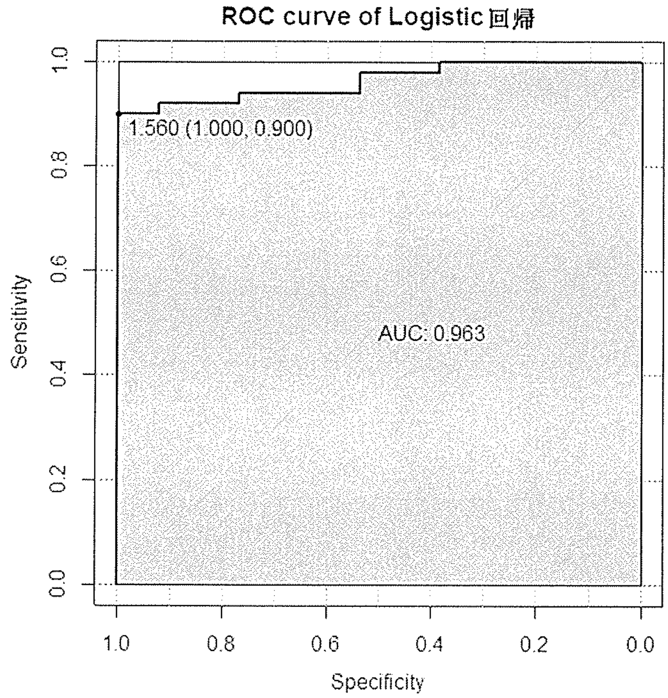
図Ⅳ-3から見ると、セグメンテーション前のLRの臨界値は0.344（1.000, 0.819）であり、AUC値は0.938である。

図Ⅳ-4から見ると、セグメンテーション前のSVMの臨界値は0.143（0.992, 0.836）であり、AUC値は0.942である。

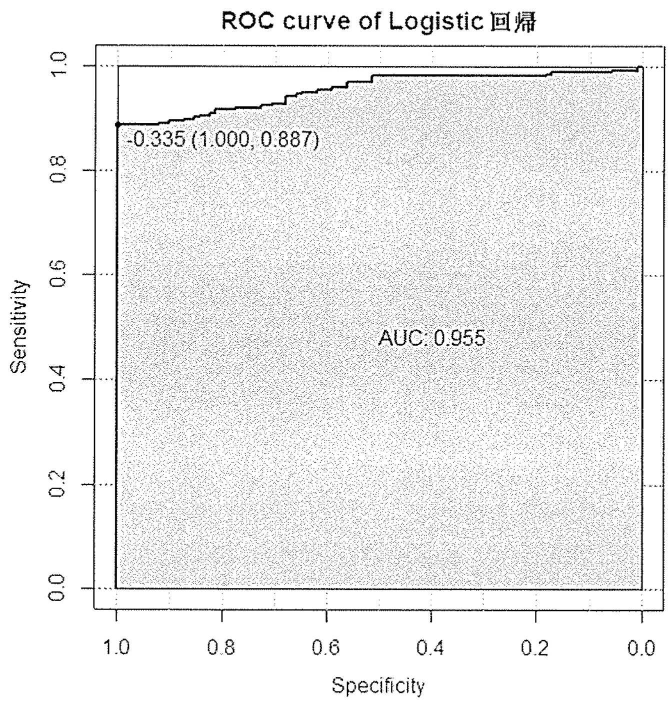
図Ⅳ-5から見ると、セグメンテーション前のBPNNの臨界値は0.681（1.000, 0.825）であり、AUC値は0.951である。

図Ⅳ-6から見ると、セグメンテーション前のAdaBoostの臨界値は0.473（0.926, 0.990）であり、AUC値は0.994である。

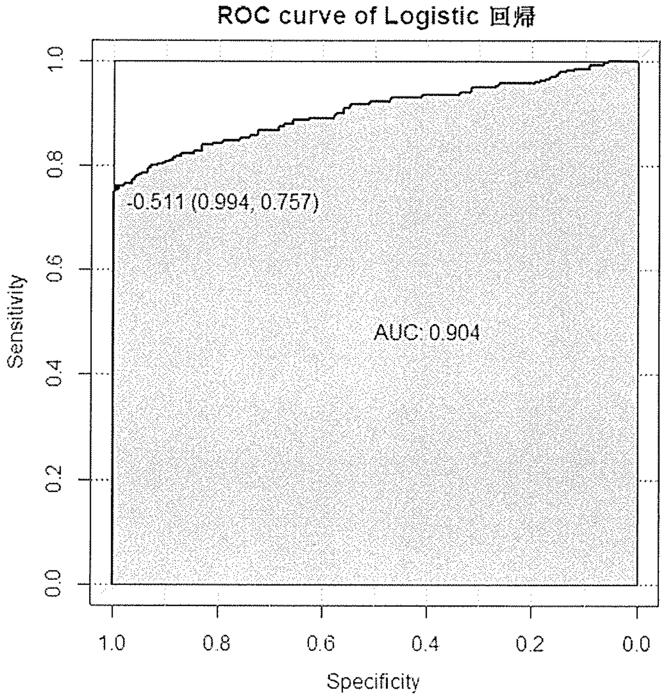
図IV-7 セグメンテーション後の LR の ROC 曲線



(a) Cluster I



(b) Cluster II



(c) ClusterⅢ

出所：筆者作成。

図Ⅳ-7 のセグメンテーション後のLRの臨界値とAUC値：

(a) ClusterⅠの臨界値は1.560 (1.000, 0.900) であり、AUC値は0.963である。

(b) ClusterⅡの臨界値は-0.335 (1.000, 0.887) であり、AUC値は0.955である。

(c) ClusterⅢの臨界値は-0.511 (0.994, 0.757) であり、AUC値は0.904である。

図Ⅳ-8のセグメンテーション後のSVMの臨界値とAUC値：

(a) ClusterⅠの臨界値は0.270 (1.000, 0.980) であり、AUC値は0.992である。

(b) ClusterⅡの臨界値は0.161 (1.000, 0.913) であり、AUC値は0.966である。

(c) ClusterⅢの臨界値は0.071 (0.983, 0.786) であり、AUC値は0.927である。

図Ⅳ-9 のセグメンテーション後のBPNNの臨界値とAUC値：

(a) ClusterⅠの臨界値は0.917 (0.846, 0.980) であり、AUC値は0.940である。

(b) ClusterⅡの臨界値は0.259 (0.961, 0.928) であり、AUC値は0.963である。

(c) ClusterⅢの臨界値は0.352 (0.976, 0.788) であり、AUC値は0.913である。

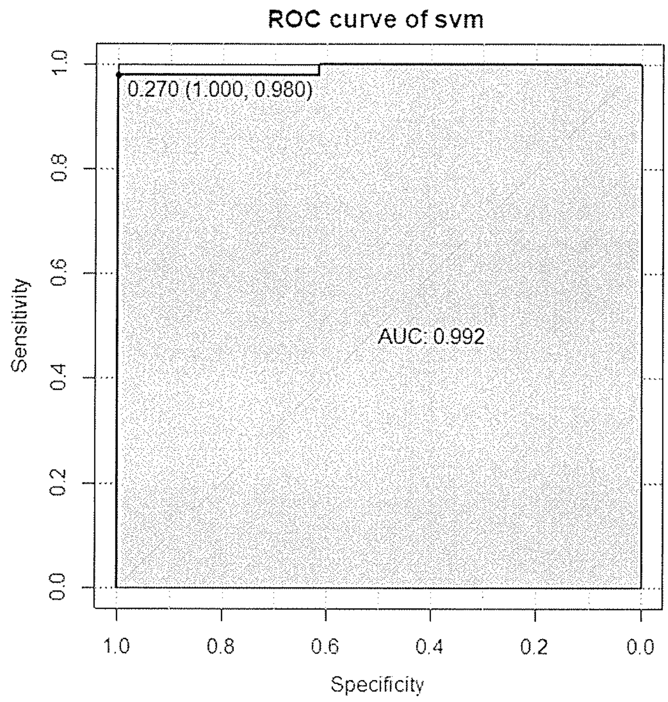
図Ⅳ-10 のセグメンテーション後のAdaBoostの臨界値とAUC値：

(a) ClusterⅠの臨界値は0.513 (1.000, 1.000) であり、AUC値は1.000である。

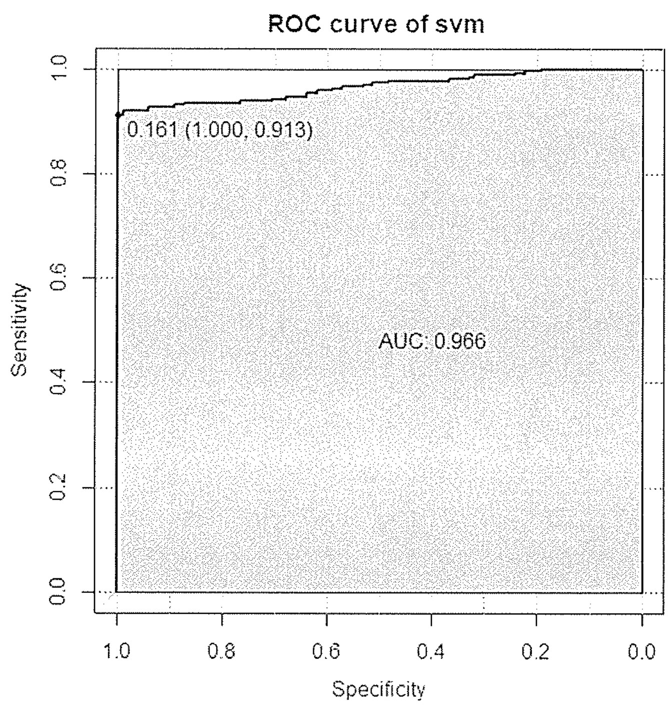
(b) ClusterⅡの臨界値は0.518 (0.981, 0.977) であり、AUC値は0.999である。

(c) ClusterⅢの臨界値は0.454 (0.908, 0.998) であり、AUC値は0.990である。

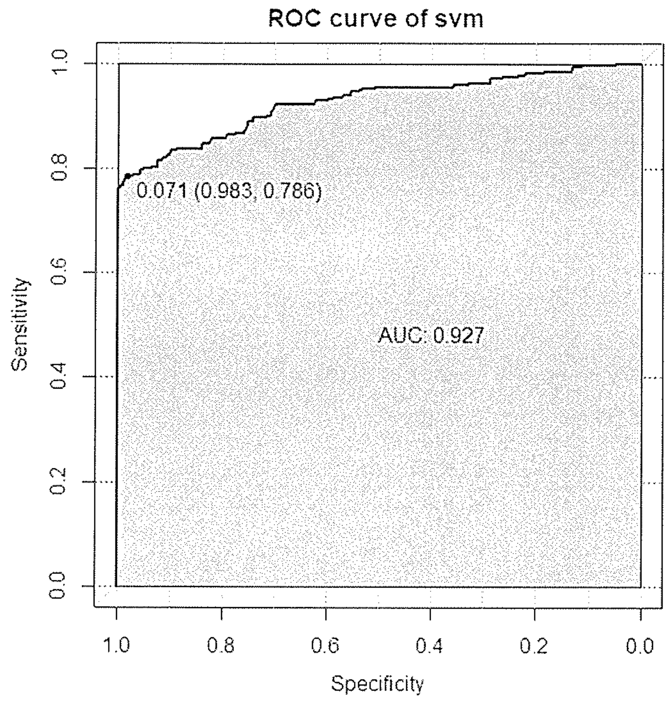
図Ⅳ-8 セグメンテーション後の SVM の ROC 曲線



(a) Cluster Ⅰ



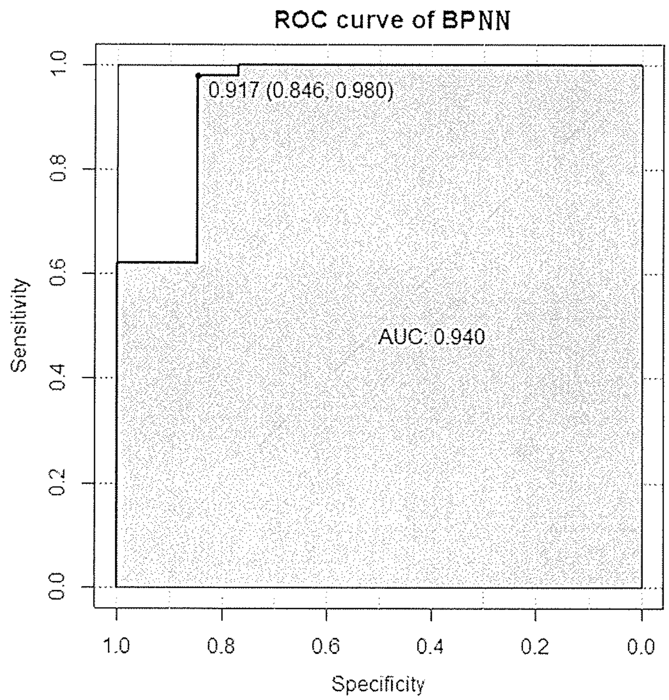
(b) Cluster Ⅱ



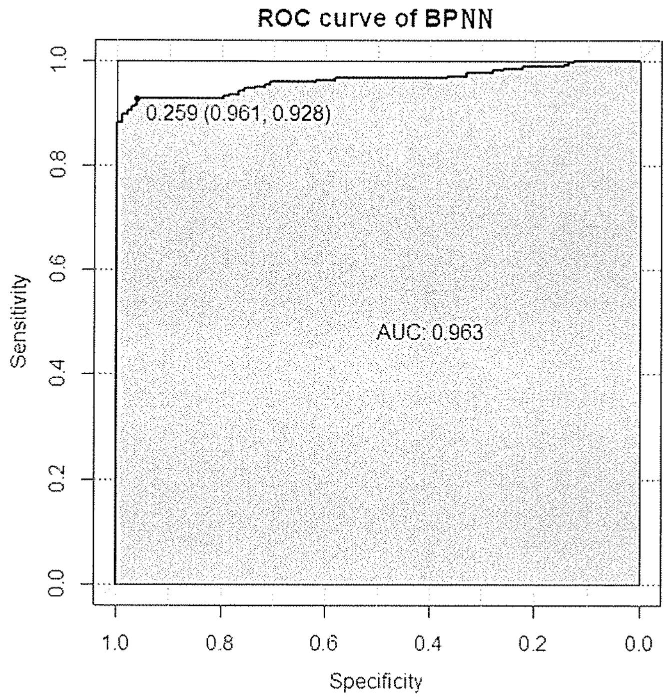
(c) ClusterⅢ

出所：筆者作成。

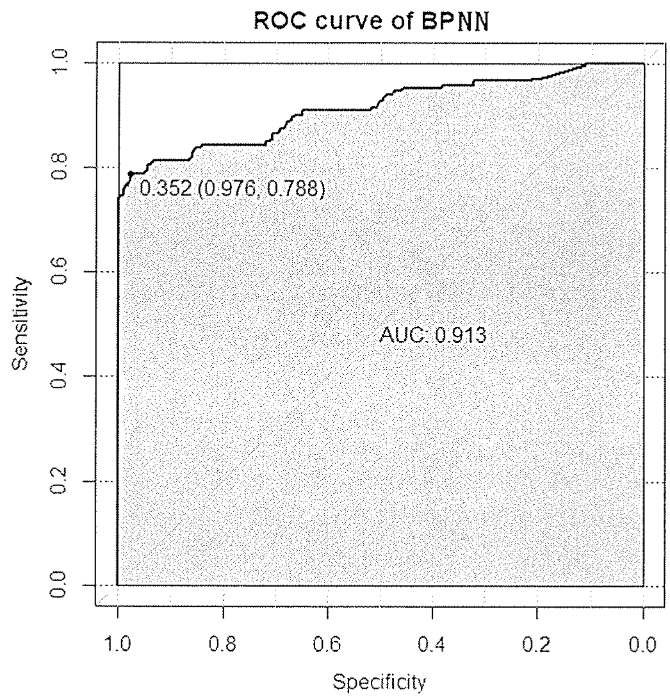
図Ⅳ-9 セグメンテーション後の BPNN の ROC 曲線



(a) ClusterⅠ



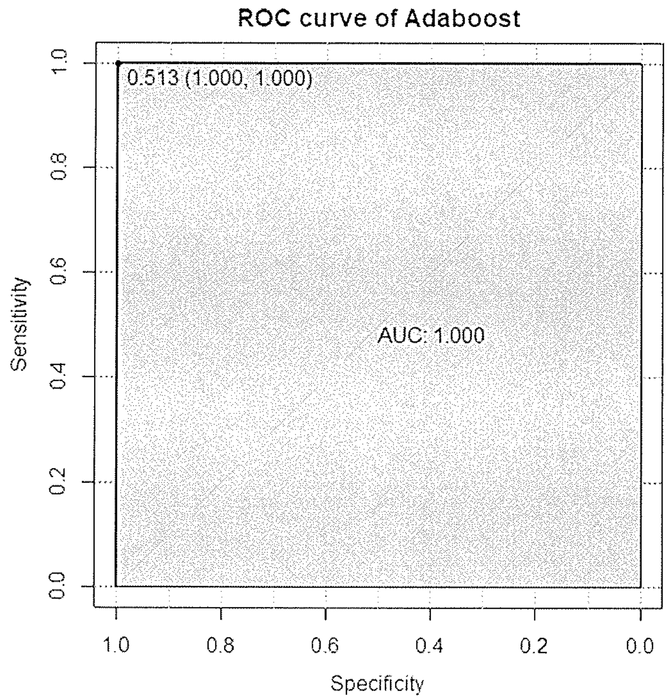
(b) Cluster Ⅱ



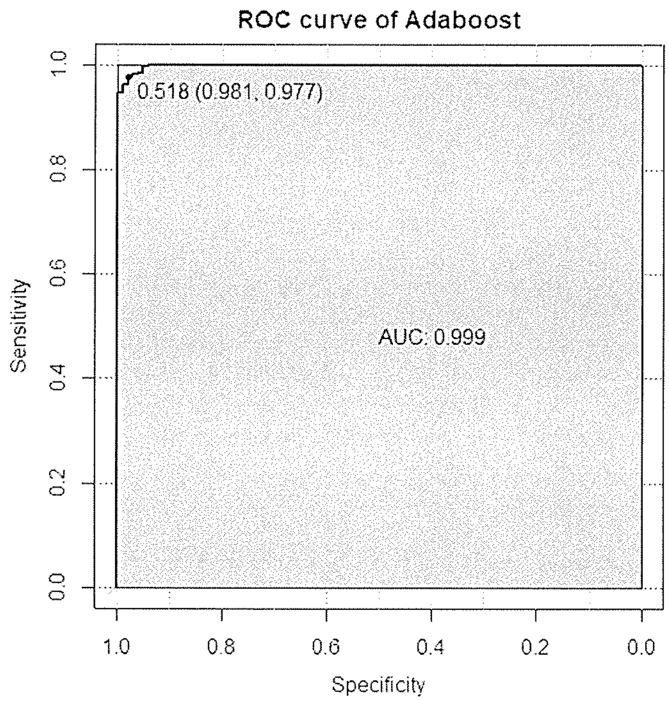
(c) Cluster Ⅲ

出所：筆者作成。

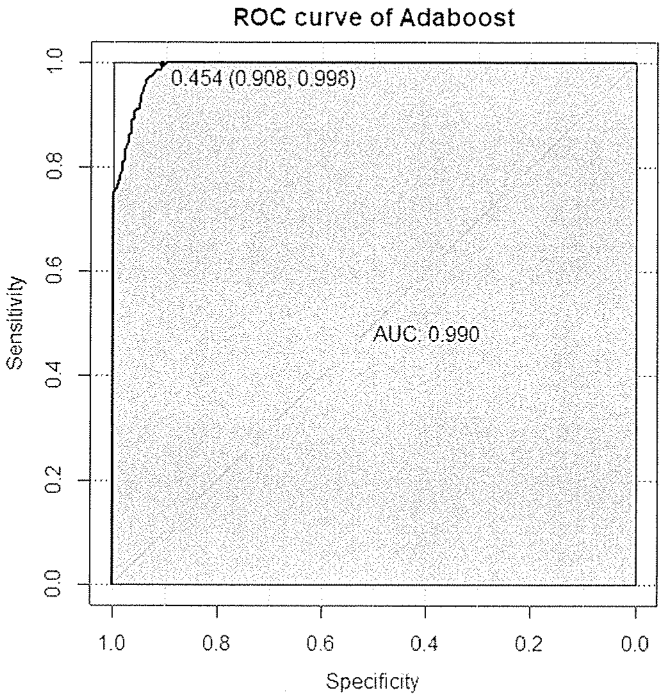
図IV-10 セグメンテーション後の AdaBoost の ROC 曲線



(a) Cluster I



(b) Cluster II



(c) ClusterⅢ

出所：筆者作成。

セグメンテーション前のROC曲線とセグメンテーション後のROC曲線を比較すると、セグメンテーション後の図の左上の臨界値がセグメンテーション前の臨界値よりも高い。また、AUC値も細分化前のAUC値よりも高い。LRのAUC値を基準として、SVM、BPNN、AdaBoostの3種類のAUC値を観察し、LRのAUC値（ClusterⅠ: 0.963、Cluster II : 0.955、Cluster III : 0.904）と比較すると、SVM、BPNN、AdaBoostのAUC値の上昇幅は、

SVMのAUC値の上昇幅：

(a) ClusterⅠの上昇幅は3.01%である。

(b) ClusterⅡの上昇幅は1.15%である。

(c) ClusterⅢの上昇幅は2.54%である。

BPNNのAUC値の上昇幅：

(a) ClusterⅠの上昇幅は-2.38%である。

(b) ClusterⅡの上昇幅は0.83%である。

(c) ClusterⅢの上昇幅は0.99%である。

AdaBoostのAUC値の上昇幅：

(a) ClusterⅠの上昇幅は3.84%である。

(b) ClusterⅡの上昇幅は4.60%である。

(c) ClusterⅢの上昇幅は10.5%である。

上記4つのアルゴリズムのセグメンテーション前とセグメンテーション後の評価指標を比較するために、表 IV-16 と表 IV-17 の結果を示す。表IV-16 は、セグメンテーション前の 4 種類の予測アルゴリズムの精度、再現率、適合率及び「ROC 曲線下の面積 AUC」の比較である。表IV-17 は、セグメンテーション後の 4 種類の予測アルゴリズムの精度、再現率、適合率及び「ROC 曲線下の面積 AUC」の比較である。

表IV-16 セグメンテーション前の評価指標の結果の比較

評価指標 / アルゴリズム	Accuracy (精度)	Recall (再現率)	Precision (適合率)	AUC (曲線下の面積)
LR	0.9065	0.9984	0.8149	0.938
SVM	0.9081	0.9990	0.8175	0.942
BPNN	0.9082	0.9937	0.9229	0.951
AdaBoost	0.9446	0.9524	0.9372	0.994

出所：筆者作成。

顧客流出予測モデルの 3 つの評価指標(精度、再現率、適合率)は、通常、精度（正解率：Accuracy）値が高いほど、モデルの予測効果が良いと考えられている。(Hamel、2009, Majnik and Bosnic、2013, Norton and Uryasev、2019)。精度と再生率という 2 つの指標は相互間に制約されているため、再現率（Recall）値が低いほど、予測モデルが「流出顧客」を識別するのがより正確であると考え、適合率（Precision）値が高いほど、予測モデルの「正確な予測」の能力が高いと言われている。表IV-16 と表IV-17 により、AdaBoost の総合性能が高くて、3 つの指標値はいずれも LR、SVM、BPNN の 3 つの指標値より優れている。

セグメンテーションした結果を比較すると、3 つのクラスタの精度、再現率、適合率の結果に不一致があり、主に「ClusterI」、「ClusterII」、「ClusterIII」の不均衡データがあるため、これらの不均衡データは流出予測モデルの性能を低下させた。

表Ⅳ-17 セグメンテーション後の評価指標の結果の比較

評価指標 / アルゴリズム		Accuracy (精度)	Recall (再現率)	Precision (適合率)	AUC (曲線下の面積)
ClusterⅠ	LR	0.9050	0.8496	0.9176	0.963
	SVM	0.9256	0.9184	0.9298	0.992
	BPNN	0.9287	0.8764	0.9404	0.940
	AdaBoost	0.9604	0.9145	0.9719	1.000
	Average	**0.929**	**0.889**	**0.939**	**0.973**
ClusterⅡ	LR	0.9098	1	0.8728	0.955
	SVM	0.9158	0.9982	0.8820	0.966
	BPNN	0.9164	0.9766	0.8916	0.963
	AdaBoost	0.9663	0.9346	0.9792	0.999
	Average	**0.927**	**0.977**	**0.906**	**0.970**
ClusterⅢ	LR	0.9050	1	0.7697	0.904
	SVM	0.9053	0.9998	0.7706	0.927
	BPNN	0.9052	0.9973	0.9940	0.913
	AdaBoost	0.9399	0.9457	0.9303	0.990
	Average	**0.913**	**0.985**	**0.866**	**0.933**

出所：筆者作成。

表Ⅳ-16より、セグメンテーション前の精度指標では、LRの精度とAUC値は低い。AdaBoostの精度とAUC値が最も高く、0.9446と0.993であり、LR、SVM、BPより明らかに高い。表Ⅳ-14のセグメンテーション後の精度とAUC値を見ると、3種類の顧客（ClusterⅠ、ClusterⅡ、ClusterⅢ）のAdaBoostの精度とAUC値もLR、SVM、BPより明らかに高い。また、セグメンテーション前とセグメンテーション後を比べると、表Ⅳ-17から見ると、セグメンテーション後の各指標が大きく向上することがわかる。

5-2　顧客セグメンテーションの分析

顧客セグメンテーションは、企業がコア顧客に対して製品マーケティング案を向上させ、ロイヤル顧客とのコミュニケーション戦略を再作成し、製品を顧客の好みと一致させることができ、製品の再計画と納品の加速に役立つ（Brito et al.、2015)。本研究では、k-meansを使って、整理した原始データを基にセグメンテーションし、顧客を3つのクラスタに分けている。表Ⅳ-3から見ると、「ClusterI」は、顧客4,935人で、非流出顧客484人、流出顧客4,451人である。「ClusterI」の顧客流出率は約90.2%に達するが、非流出率は約9.8%である。「ClusterⅡ」は顧客2,697人で、非流出顧客83人、流出顧客2,614人である。「ClusterⅡ」の顧客流出率は約96.9%に達するが、非流出率は約3.1%である。「ClusterⅢ」は顧客524人で、非流出顧客13人、流出顧客511人である。「ClusterⅢ」の顧客流出率は約97.5%に達するが、非流出率は約2.5%である。「ClusterⅡ」と「ClusterⅢ」と比較すると、「ClusterI」の非流出率は一番高い。これから見ると、「ClusterI」の顧客は企業のロイヤル顧客であると考えられ、企業はマーケティング戦略の策定とCRMの面で重視しなければならない。また、k-meansを採用して顧客のセグメンテーションに用いられる有効性も示していて、B2Cイーコマース企業のデータ分析と顧客流出モデリングに価値があると考える。

5-3　予測モデル性能

本研究はLR 、SVM、BPNNとAdaBoostモデルの比較実験を行った。Confusion matrixに基づいて、データセットの3つのカテゴリのパフォーマンスを評価するために、各カテゴリのAccuracy、Recall、Precision値を計算した。表Ⅳ-17には、顧客セグメンテーション後のLR、SVM、BPNNとAdaBoostモデル予測の実験結果を示している。そして、表Ⅳ-17から見ると、AdaBoostは3つの顧客クラスタに対するAdaBoostの予測精度AccuracyがLR、SVMとBPNNの予測精度と比べて高くて、AdaBoostの予測効果は良好である。しかし、Accuracyだけでモデルの性能と予測の効果を確認することには誤導が生じやすい場合がある（Sturm and Bob、2013)。そのため、モデルの性能と予測の効果を評価する際に、Accuracyだけでなく、RecallやPrecisionも確認する必要がある。これで、予測モデルの性能はAccuracy、RecallとPrecisionという3つの評価指標で総

合的に決定する。本研究の実験結果(Xiahou and Harada、2022.03, Xiahou and Harada、2022.06) により, 顧客セグメンテーション後、各アルゴリズムの予測の指標の平均値は以下のように示す。

Accuracy：AdaBoostは0.9555、LRは0.9066、SVMは0.9156、BPNNは0.9167。

Recall：AdaBoostは0.9316、LRは0.9498、SVMは0.9721、BPNNは0.9501。

Precision：AdaBoostは0.9604、LRは0.8533、SVMは0.861、BPNNは0.942。

述のように, AdaBoostのAccuracyとPrecisionは一番高い、Recall は一番低い。従って、AdaBoostはLR、SVM、BPNNよりも優れていると考えられる。

一方、予測性能の優劣を判別する重要的な根拠として、受信者動作特性曲線(Receiver Operating Characteristic Curve：ROC)とそのROC曲線下の面積(Area Under the ROC Curve：AUC)を利用してモデルの汎化能力を評価し、ROCは任意の閾値が学習器の汎化性能に及ぼす影響を容易に検出できる(Ma and Huang、2005, Song and Ma、2010)。図Ⅳ-10はセグメンテーション後のAdaBoostアルゴリズムについての顧客流出予測のROC曲線である。図Ⅳ-10から見ると、ClusterI、ClusterII、ClusterIIIの閾値が0.513、0.518、0.454のとき、敏感度(sensitivity)は1.0、0.977、0.998、特異度(specificity) は1.0、0.981 、0.908、AUCは1.0、0.999、0.99である。同じように、図Ⅳ-7～図Ⅳ-9はLR、SVM、BPの各アルゴリズムの性能指標を示す。これらの実験データはAdaBoostが良好的な汎化能力を有し、予測性能が良好であることを実証している。また、AdaBoostとLR、SVM、BP比較を通して、B2Cイーコマースにおける顧客流出予測でAdaboostを推奨して、良い効果が期待される。

小括

本章では、「TIANCHI天池」データプラットフォームの非契約型取引の顧客のショッピングデータを用いての顧客について流出予測モデルの作成を行った。まず、顧客のショッピングの時間帯について4つの時間帯（午前、午後、夜、未明）に分ける。そして、データ前処理及びデータの標準化を行い、消費者行動データ（商品の種類、商品の購入、ショッピングカートの追加、好きな商品）を17個の一般変数に整理した。データ前処理と標準化が完了した後、顧客セグメンテーションが行われ、k-meansアルゴリズムを用いて「ClusterI」、「ClusterII」、「ClusterIII」の3種類の顧客クラスタに分けて、流出顧客の数が確認された。これは本研究の重要なステップであり、「セグメンテーション・ファースト」と言われる。

次に、混同行列の構築とROC曲線の作成を通じて、LR、SVM、BPNN、AdaBoostの4つの予測アルゴリズムについて精度、再現率、適合率、ROC曲線下の面積(AUC)の4つの指標を算出した。SVM、BPNN、AdaBoostを基に予測モデルの汎化性能を考察するために、LRアルゴリズムを基準として、SVM、BPNN、AdaBoostのAUC値とLRのAUC値を比較

し、SVM、BPNN、AdaBoostのAUC値の上昇幅を計算した。また、セグメンテーション前とセグメンテーション後の評価指標の結果を詳細に比較し、評価結果は構築された「セグメンテーション・ファースト」と言われる予測モデルが有効であることを示した。最後に、構築されたイーコマース顧客流出の4つのモデルについて結果分析と比較を行って、AdaBoostを基に構築した顧客流出予測モデルの性能が最も高くて、B2Cイーコマースに対する流出予測に効果があると考える。

V　考察

本研究では、機械学習を用いてアリババグループの「TIANCHI天池」ビッグデータプラットフォームが発表した公開データを利用して、B2C非契約型のイーコマースの顧客に対して流出予測の研究を行った。データ前処理段階では、データ変数の整理と統計を行い、顧客セグメンテーションの変数として新しい行動変数を採用した。そして、ランダムフォレストアルゴリズムを使って、一般変数をスクリーニングし、変数の重要度に基づいて流出予測のための特徴変数を決定した。これで確定した特徴変数をAdaBoostなど4つの予測アルゴリズムに代入し、流出予測の結果を算出した。以下、流出予測の有用性、顧客セグメンテーションと手法および予測モデルの3つの面から考察する。

1　顧客流出予測の有用性について

顧客流出予測は顧客流出管理における重要な構成部分として、顧客関係管理(CRM)に対して重要な役割を果たしている。ビッグデータや機械学習技術の発展に伴い、企業の CRM の内容も進んでいる。CRM の概念から、CRM は情報技術の援用、すなわちデータを用いて顧客を識別し、顧客に対して顧客ごとに合ったダイレクト・メッセージを配信するなどで満足させ、顧客維持につなげていくものとされている（藪野祥太、2020)。従来の CRM のセグメンテーション法は統計学的な方法を用いて顧客セグメンテーションを行うものであり、この方法で使用される顧客行動変数は変数の属性と数量の影響を受け、これらの変数は顧客消費行動の中の有用な情報を無視し、それによって顧客セグメンテーションの結果に影響を与える。例えば、Rachid ら(2018) と Gattermann -Itschert ら(2021)は RFM 法を細分化して使用しており、これらの RFM モデルの変数は R、F、M の 3 つだけであり、この変数分割の方法は横断的な時間性（直近の購入日）だけを考慮しており、この時間変数（直近の購入日）は消費者の買い物行動の縦断的な時間性、各顧客についての午前の購買行動や午後の購買行動、夜の購買行動などの時間性を考慮していない。本研究における予測変数は縦断的な時間変数を含み、消費行動の習慣に合致し、流出予測の結果は消費行動を正確に反映することができる。これは流出予測の有用性のひとつと考える。 K-means は代表的な教師なしアルゴリズムとして、顧客セグメンテーションの流れの中で特徴変数の影響を受けることが少ない、あるいは、k-means は変数の種類が沢山ある場合で相関する複数の変数を簡潔に表現して顧客セグメンテーションを行うことができる。多変数は、流出に影響を与える有効な情報または原因である可能性があり、顧客セグメンテーションの効果を高めることができる。

一方、本研究は顧客流出率を予測する流れの中に縦断的な時間変数とその他の変数

（クリック、追加購入、コレクションなど）を採用して、顧客セグメンテーションを行って、各顧客クラスタの流出の割合を計算した。顧客そして、流出顧客のデータに基づいて、流出に対する変数の影響重みを計算し、顧客が流出した具体的な原因を明らかにする。これにより、企業が顧客関係管理レベルを高めると同時に、顧客流出管理をより細かく行うことができ、顧客関係管理効率を高め、管理財務コストを減らすと考えられる。また、マーケティングの面でも積極的な影響を与えることも期待されている。

2　変数の選択と顧客セグメンテーションについて

本研究の第Ⅳ章の実証研究では、B2Cイーコマースの顧客行動データの縦断的な時間性と多様な行動情報(クリック数、購買回数、カートに入れ数、お気に入り数)の特徴を重要な変数として、消費者行動を4つの時間帯(未明、午前、午後、夜)に基づいて細分化することが本研究の新しい見解のひとつである。

顧客セグメンテーションモデルでは、時間情報を変数、すなわち時間変数として用いる伝統的なRFMモデルが広く用いられている。顧客セグメンテーションと顧客流出予測に及ぼす時間変数の影響を調べるために、Changら(2004)は時間変数を拡張し、RFMに基づいてLRFMモデルを提案し、Lを時間周期(number of time periods(such as days))として定義した。その研究結果は時間周期(L)が顧客流出予測の重要な変数であり、顧客ロイヤルティの評価に利用できることを提示した。Rachidら(2018)はオンライン小売業者のデータセット(電子、ファッション、家電、児童用品)を用いてLRFMモデルを研究し、時間周期を流出予測の重要な変数とした結果、顧客の異なる時間周期(日数)でのショッピング行動に大きな違いがあることを示した。Wuら(2016)の文献では、イーコマースの顧客のショッピング時間帯を昼、夜、深夜の3つの時間帯(「時間」に細分化されていない)に分けており、その結果、異なる時間帯の顧客流出率が著しく異なることが明らかになった。Alboukaeyら(2020)は、時間変数が毎日の動態的な行動(毎月の動態的な行動ではない)として定義され、毎日の動態的な行動変数および他の変数によって多変数の時間の系列を構成するRF-Dailyと呼ばれるモデルを提案した。その結果、モデルの流出予測効果が良好であり、毎日のモデルが月モデルより明らかに優れていることを明らかにした。Chenら(2015)はRFMモデルを拡張し、時間変数Lが取引間隔時間(time between transactions)であるLRFMPモデルを提案し、その結果、取引間隔時間が顧客流出に大きな影響を及ぼすことを示した。実証実験では、時間(期間)変数の観点から、時間に関する変数を「時間」と定義される。予測データセットでは商品種類変数に加えて、データ変数は時間変数(未明、午前、午後、夜)と4つの動態的な変数(クリック数、購買回数、カートに入れ数、お気に入り数)と

組合せで、RF-PACVモデル（夏侯賢城・原田良雄、2022）[22]と呼ばれている。そして、われわれはこのモデルについてセグメンテーションを行って、3つの顧客クラスタを分けられている。第Ⅳ章の予測実験結果により、夜変数と午後変数は流出予測に大きな影響を及ぼし、表Ⅳ-5から「夜Buy」と「午後Buy」の変数重要性が上位2位にランクされていることが分かった。われわれの研究結果は、ショッピング時間変数が流出予測の重要な変数であることを示すことができた。これは上記文献（Chang et al.、2004, Chen et al.、2015, Rachid et al.、2018, Alboukaey et al.、2020）の結果とほぼ一致している。

実際に見ると、マーケティングマネージャは、時間を細分化した時間変数がマーケティングマネージャに操作可能な情報を提供することができるため、顧客がどの時間帯に買い物をするかということに特に興味を持っている。マーケティングマネージャは、これらの操作可能な時間情報に基づいて、これらの時間帯にリアルタイムで顧客にカスタマイズされた商品プッシュ活動を提供し、顧客を説得して維持することができる。例えば、企業が「ClusterI」が重点的に注目するコア顧客であることを確定した後、企業は適切なアルゴリズムを採用して「ClusterI」顧客に対して流出予測を行い、予測評価指標に基づいて「ClusterI」顧客の流出傾向を判断し、その後、企業は適切なマーケティング戦略を策定することができる、企業がこのような顧客をタイムリーに判断できない場合、有効な顧客保持活動を展開できないため、「ClusterI」の一部の顧客が流出する可能性がある。

本研究の方法は顧客の保持という要因を十分に考慮することによって、実用性を体現しようとする。これも本研究の重要な価値のひとつである。例えば、中国では年に1度11月11日にピークを迎える「天猫ダブルイレブン」ショッピングフェスティバルは、グローバルサプライチェーンを活用して、新しいブランドや新商品に対する消費者の需要の高まりに応える。「アクティブなコア顧客」を的確に判断できれば、企業に大きなビジネス利益をもたらすことができる。

3 予測モデルについて

イーコマースの拡大とIT技術の発展に伴い、B2C非契約型のイーコマース顧客の流出に対して、予測モデルの構築については様々な研究方法がある。

---

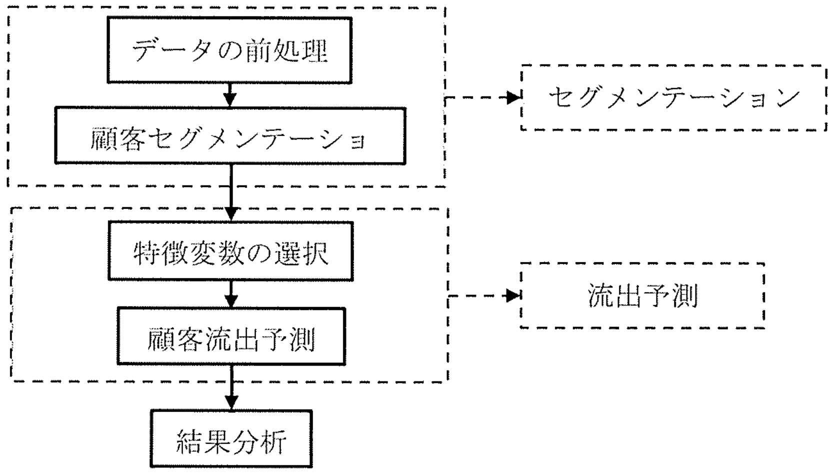
[22] RFM 法に基づいて、直近購買日（Recency : R)、累積購買回数（Frequency : F)、クリック数（Print of View : P)、カートに入れ数（Add to Cart : A)、商品種類(Category : C)、お気に入り数（Favourite : V）という 6 つの変数に増加した。そして、Calinski-Harabasz(CH)基準を導入する上に、k-means を使って効果を検討し、新しい消費者セグメンテーションの RF-PACV モデルを提案した。この研究成果は、2022 年度人工知能学会全国大会(第 36 回 京都 2022 年 6 月 14 日)で発表した。『夏侯賢城、原田良雄（2022）K-means を用いてネットビジネスユーザデータの顧客セグメンテーションの研究、1G1-GS-10-04。』

従来のイーコマース顧客流出予測モデルの構築方法は、消費行動変数(Chen et al.、2012)を直接的に機械学習のアルゴリズムに代入して、予測モデルを構築する。あるいは、伝統的な統計学の方法に基づいて顧客セグメンテーションを行って、各クラスタのデータを機械学習のアルゴリズムに代入して顧客流出率を予測する。Chenら（2012）はRFMセグメンテーションモデルを改善し、消費金額、購入数量、毎月の購入数量を行動変数とし、顧客の人口属性変数を静態的な変数とし、顧客流出を予測した。しかし、モデル構築の視点から、予測用のアルゴリズムの選択についてまだ様々な議論がある。例えば、Nieら（2011）はLRアルゴリズムが決定木より優れていると考え、Mozerら（2000）はAdaBoostが他のアルゴリズムより最も良いと考えられる。Masandら（1999）は線形回帰、k近傍法、決定木とBPNNを比較し、BPNNの予測効果が最も良いと考えている。

イーコマースのデータ量が膨大であり、消費者行動データの多様性（多変数、縦断的な時間性など）があるため、予測性能が高い予測モデルが必要であると考える。また、予測モデルの予測性能は通常、特定のサンプルデータに基づいて、同じ予測モデルは異なる研究において異なる精度と効果を持つと考えられる。しかし、従来のモデル構築方式を採用すれば、予測アルゴリズムの性能には限界がなる。従って、B2Cイーコマースにおいて、顧客流出の予測モデルは顧客消費行動の実際の状況を全面的に反映すべきであり、消費行動の多変数の状況を十分に反映すれば、モデルの予測精度と効果を高めると考えられる。

第Ⅳ章の実証研究では、顧客流出予測モデルを「データ前処理」、「顧客セグメンテーション」、「特徴変数の選択」、「顧客流出予測」、「結果分析」という5つの部分に分けて構築している。モデルの枠組みは図Ⅴ-1のように示す。

図Ⅴ-1「セグメンテーション・ファースト」モデルの枠組み



出所：筆者作成。

その流れにおいて、まず、k-meansを使って顧客セグメンテーションを行う。そして、一般変数をランダムフォレストに代入して一般変数の重要度について特徴変数をスクリーニングする。最後に、特徴変数を4つの機械学習アルゴリズムLR、SVM、BPNN、AdaBoostに代入し、セグメンテーション前後の予測性能の比較と各アルゴリズムの予測性能の比較を行った。また、従来の研究によると、通常LRを基準のアルゴリズムとして、予測性能を比較する。例えば、Gordiniら（2017）はLRを基準にSVM、動径基底関数（RBF）[23]、BPNNなど4つのアルゴリズムを用いて流出予測を行っている。これによって、LRもわれわれの研究の基準のアルゴリズムとして、他のアルゴリズムと比較する。また、アルゴリズムの予測性能を評価する際には、通常いくつかの性能指標を用いて評価する。例えば、Limaら（2011）はAccuracy、Sensitivity（敏感度）、Specificity（特異度）、AUCを評価基準として用いる。Coussementら（2017）は、AUCとリフト(Lift)[24]指標を用いてLRモデルの効果を評価する。Vafeiadisら（2015）はAccuracy、Recall、Precision、F-measure(F-尺度)指標[25]を用いてSVM、LR、決定木、ナイーブベイズ、BPNN及びSVMの統合アルゴリズム、BPNNの統合アルゴリズム、決定木の統合アルゴリズムのモデル効果を比較した。Holtropら（2016）はGini係数とLiftを評価指標として用いる。Jahromiら（2014）はLiftとAUC指標を用いてLR、決定木、AdaBoostの予測効果を評価した。Deら（2011）はAccuracy、AUC、Liftの3つの指標を用いてモデル効果を評価した。夏ら（2008）はAccuracy、Precision、Recall、Liftの4つの指標を用いてSVM、BPNN、Logit回帰、ナイーブベイズの分類効果を評価した。われわれの研究ではAccuracy、Recall、Precisionの3つの指標を用いており、これは他の文献の評価指標とほぼ同じである。

結果から見ると、AdaBoostのセグメンテーション後の予測の確率はセグメンテーション前の確率より高いだけではなく、セグメンテーション後の予測の確率は他のアルゴリズムよりも高い。その結果、AdaBoostを用いて構築している顧客流出の予測モデルが優れていると考える。つまり、B2Cイーコマースにおいて、AdaBoostは適用すると考えられる（Xiahou and Harada、2022）[26]。また、本研究で構築している「セグメ

[23] 動径関数は「基底のように」使われることがしばしばあり、そのような文脈では動径基底関数（Radial Basis Function：RBF）と呼ばれることがある。「基底のように」とは、(1) 同じ種類の動径関数を集めてきて、(2) それらの線形和を取ることでいろいろな関数を近似するというイメージである。

[24] リフト値とは、ある特定の商品（X）を購入した消費者が、別の商品（Y）を一緒に購入する確率はどの程度か相関性を示す指標であり、商品購入の組み合わせ分析などを行うバスケット分析などで用いられる。本研究では、消費者がある特定の商品を購入することは指定されていないため、このリフトは評価指標として採用されていない。https://retailguide.tokubai.co.jp/store/12438/を参照。

[25] F-measure(F-尺度)とは、適合率(precision)と再現率(recall)という、正確性の総合的な評価の際に利用される尺度のこと。適合率、再現率は互いにトレードオフの関係なので、F-尺度を高くするように出来れば、バランス良く両方の値が高いというのを測る感じだと思われる。SVM、BPNN、決定木などのアルゴリズムを統合する場合、F 指標が採用される。フリー百科事典『ウィキペディア（Wikipedia)』（2022/08/25 06:24 UTC 版)を参照。

[26] この研究成果は、「American Journal of Industrial and Business Management」で発表した。『Xiahou and Harada、(2022) Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-Commerce Industry、12, p.277-293.』

ンテーション・ファースト」予測モデルは、二つの機械学習アルゴリズムを利用して、顧客流出について流出予測を行った。その結果、各アルゴリズムのセグメンテーション後の予測性能はセグメンテーション前より高い。これで、本研究で構築している「セグメンテーション・ファースト」モデル（Xiahou and Harada、2022）[27]は有効であると考えられる。

[27] この研究成果は、「Journal of Theoretical and Applied Electronic Commerce Research」で発表した。『Xiahou and Harada、(2022) B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM、Vol.17 (2)、p./458 -475.』

結論

顧客流出予測はイーコマースにおいて極めて重要である。市場競争力を維持するためにB2C企業は顧客関係管理の中で機械学習技術を十分に利用して流出する可能性のある顧客を予測し、予測結果に基づいて新しいマーケティング戦略と顧客保留措置を提示し、効率的で正確な流出予測モデルを構築することはすでにイーコマース企業の重要な課題となっている。

本研究では、B2Cイーコマース企業の顧客行動データを用いて、LR、SVM、BPNNとAdaBoostモデルの予測能力をテストした。このモデルの予測性能を評価するために、まずk-meansアルゴリズムを用いてクラスタリングを細分化し、3種類の顧客を分類し、これら3種類の顧客を予測し、Accuracy、Recall、Precision、AUCをそれぞれ計算した。われわれの研究方法は3つの目的がある。

第1の目的は、顧客のショッピング行動の縦断時間性と変数の多次元性に基づいて、新しい顧客セグメンテーションモデル「RF-PACV」モデルを提案することである。従来の研究は伝統的なRFMモデルをよく利用して、直近購買日変数を使って顧客流出予測を行っている。本研究は、各顧客の一日中の時間帯（午前、午後、夜、未明）に基づいて、顧客行動を直近購買日（Recency,R），累積購買回数（Frequency,F），クリック数(Print of View, P)，カートに入れ数(Add to Cart, A)，商品種類(Category,C)，お気に入り数（Favourite,V）という6つの特徴変数に縦断的に分けて、顧客セグメンテーションを行う。実証研究の結果から見ると、顧客が3つのクラスタに分けられて、「RF-PACV」モデルが有効であると考えられる。

第2の目的は、顧客流出予測モデルを構築する時、二つの機械学習のアルゴリズムを採用するが、ひとつの機械学習アルゴリズムで顧客セグメンテーションを先にして、その結果をもう一つの機械学習アルゴリズに代入して顧客流出予測を行うこと。つまり、「セグメンテーション・ファースト」モデルである。これで、セグメンテーション前とセグメンテーション後のモデル予測結果の比較により、顧客セグメンテーションの効果を測定する。予測結果によると、k-meansクラスタリングを用いて顧客のセグメンテーション後の各予測指標が明らかに向上し、セグメンテーションが必要であることを証明した。これで、モデルの構築から見ると「セグメンテーション・ファースト」モデルが有効であると考えられる。

第3の目的は、「伝統的な統計分析のLRに基づく予測」、「統計的学習理論のSVMとBPNNに基づく予測」、「アンサンブル分類器のAdaBoostに基づく予測」により、三つの手法の効果を比較することである。その結果、「アンサンブル分類器のAdaBoostに基づく予測」の精度が他の3つのモデル予測の精度よりも高いことを証明し、B2Cイーコマースにおいて、AdaBoostが適当であると考えられる。

これらの研究結果はB2Cイーコマース企業の顧客関係管理に非常に積極的な意義がある。

1　本研究の学術的な貢献と現実の意義

本研究はB2Cイーコマースの顧客流出予測モデルの開発と企業の顧客関係管理に啓示がある。例えば、イーコマース企業が顧客関係管理を展開する際に、顧客をどのように細分化し、予測アルゴリズムをどのように選択するかは、本研究の成果を参考にすることができる。

まず、B2Cイーコマース業界に関する顧客のセグメンテーションと流出予測モデルに関する研究の発展に積極的な貢献をした。本研究では、現在のオンラインショッピングサイトの各種情報の中から4つの消費者のショッピング行動に直接影響する変数をデータ変数として選び、新しい顧客流出予測モデルを開発した。その際、モデルの変数は消費者のショッピング行動と具体的なショッピング時間とした。B2Cイーコマース環境の下でどのように変数を選択するか、どのように顧客のセグメンテーションを行うか、どのように変数の数を減して次元を下げるか、新しい見解を提示した。ショッピングサイトが大量に登場する状況で、すべての変数が顧客の流出に同じ影響を及ぼすわけではない。

先行研究（Chang et al.、2004, Chen et al.、2015, Wu and Meng、2016, Rachid et al.、2018, Alboukaey et al.、2020）を分析する中で、ショッピング時間(期間)が重要な予測変数であり、従来のR変数およびF変数が重要な予測要因ではないことを示すことができた。他社で最も消費が可能な顧客にとって、夜と午後期間の[購買]と「クリック」は、顧客の流出に最も関連する変数である。従来の顧客流出に関する多くの研究（Buckinx and Van、2005, Migueis and Van、2012, Miguéis et al.、2013）は、モデルの予測性能に集中していた。B2Cイーコマースにおいて顧客流出の原因についての理解は限られているが、これによる時間変数を含む顧客行動変数(P、A、C、V)に対する影響を研究することは、伝統的なRFMモデルの普及に対する根拠を提供することができる。

また、本研究では「セグメンテーション・ファースト」という予測モデルの構築によって、B2Cイーコマースにおいて顧客流出に対する新しい予測方法を提案した。これも本研究のもうひとつの貢献である。実践の観点から見ると、本研究の結果は重要な意義を持っている。企業の実際の顧客流出予測モデルを開発することは企業の顧客関係管理に明らかなメリットがあり、企業は消費特徴の重要性に基づいて顧客流出の原因を洞察することができる。本研究はB2Cイーコマース管理者が本研究で提案した新しい予測モデル(伝統的な流出予測モデルではなく)の結果を利用し、企業のマーケ

ティング戦略と顧客を維持する商品プッシュ活動を最適化することを推奨する。従来の流失予測モデルは「顧客最大化を保持する」というビジネス目標に完全に合致することはできない。すなわち、企業は「商品推薦システム」を通じて顧客の流出を最小限に抑えることが難しい。また、多くのモデルは全体的な結果に注目している。つまり、顧客の流出状況を重視している。「実際の状況」における「顧客を保留すること」という要因をよく無視している。しかし、顧客は自分の好みや意思決定を持っているため、企業にとって、いくつかの問題が生じる可能性がある。企業が顧客の好みや意思決定に適切に対応なければ、消費欲がある顧客が競合他社に移ることになるかもしれない。

従来のモデルでは、「顧客流出数」と「顧客を保留すること」の関連性が無視されることが多く、「消費欲がある顧客」の一部を「保留活動」から除外している。例えば、顧客は夜に買い物をよくしているが、「商品プッシュ」は夜に実施されていない。この問題は本研究が提案している顧客セグメンテーション方法と予測アルゴリズム、つまり、顧客タイプを分類する方法や予測アルゴリズムの改善によって解決することができる。Gattermann -Itschertら(2021)は、現場実験の構築を通じて顧客流出数と顧客保留活動の関連性を研究し、流出予測に基づく顧客保留活動の有効性を実践証拠で証明した。本研究で提案したモデルを採用すると、企業はターゲット顧客を確認することを容易にするだけではなく、顧客流出を減らすこととマーケティングコストを下げることもできる。結果から見ると、顧客セグメンテーシと顧客流出予測の中に二つの異なる機械学習のアルゴリズムを採用して顧客流出予測を行う方法は、B2Cイーコマース企業が顧客流出に対応するための実行可能な方案であることを明らかにした。

一方、本研究によると、「セグメンテーション・ファースト」モデルはSVMaucモデル(Gordini and Veglio、2017)、セグメンテーションに基づくモデリング方法(Caigny et al.、2021)、マルチスライス技術(Gattermann-Itschert et al.、2021)などの他の方法より優れていることを証明した。これによって、われわれは企業が「セグメンテーション・ファースト」モデルを使用することを推奨する。既存のマーケティングに関する研究によると、顧客グループの中には異なる買い物行動を示すものがよくあるため、k-meansを採用して顧客セグメンテーションを行うことは様々な状況の中でも価値があるツールと考えられる。例えば、Sood and Kumar (2017) によれば、時間に基づく顧客セグメンテーション手法と企業収益性(営業額など)に基づく顧客セグメンテーション手法によって、顧客流出率が大きく異なっている。われわれの研究では、予測の第1ステップとして「セグメンテーション・ファースト」を提唱している。これは、企業の顧客の維持に実際の指導的意義を持っている。本研究では、「セグメンテーション優先」という手法を予測の第1歩として提唱し、これは企業の「顧

客を保留すること」に対して現実的な意義を持っていると考える。

本研究で提案された予測方法は信頼性と操作性が高い方法として、顧客の流出の可能性を早めに判断することができる。結果によると、管理者は「セグメンテーション・ファースト」モデルを通じて「ターゲット顧客」を容易に探索し、企業の「顧客保持戦略」と「商品プッシュ活動」を最適化できることを明らかにした。また、B2Cイーコマース会社にとって商品推薦システムにおいて機械学習ツール（技術）を充分に活用ことは難しくないため、管理者は機械学習ツール（技術）の利用を通じて、企業の管理効率を高めて、経営コストを下げることができる。

モデルの演算効率と企業技術チームの能力から考えて、われわれの方法は従来の文献（Duan et al.、2019；Dwivedi et al.、2021）の見解と一致している。会社の技術チームと管理者は、「予測技術」と「機械学習技術運用の管理者」の能力と知識レベルを同時に考慮する必要がある。要するに、企業の視点から考えると、データ処理と顧客のセグメンテーション、「モデルの最適化」と「モデルの性能評価」にかかわらず、人間と技術の協力が必要である。

2　今後の課題

本研究にはまだいくつかの限界がある。データ分析の面から見ると、本研究は8,156人のデータを含むデータセットを利用したが、987,994人の原始データから見ると、データの分析量は小さい部分を占めているにすぎない。つまり、ユーザーデータのサイズと顧客流出の予測範囲をさらに拡大する必要がある。

また、顧客セグメンテーションの結果はモデルの予測性能に対して大きな影響を及ぼすため、適切な機械学習アルゴリズムを採用して顧客セグメンテーションを行うことが重要である。本研究はk-meansに基づいて顧客流出を予測したが、k-meansの適用性について議論していない。よって、2つ以上の機械学習アルゴリズムを利用してセグメンテーションの効果を比較することにより、モデルに最適な機械学習アルゴリズムを検証し、予測モデルの性能をより高めることができるかもしれない。

今後の課題として，ユーザーデータのセグメンテーションの効果をもっと上げるような方法を考案するために、ユーザーデータのサイズはもっと拡張する必要がある。また，モデルの構築の面から見ると、予測に対して最適な機械学習アルゴリズムを選んで検証効果を上げることも重要である。様々なデータセットやアルゴリズムに対して，それらの可能性について解析を行って行きたい。

参考文献

Bi, Q.Q. (2019) Cultivating loyal customers through online customer communities: A psychological contract perspective, Journal of Business Research, 103: 34-44.

Maria, O., Bravo, C., Verbeke,W., et al. (2017) Social network analytics for churn prediction in telco: Model building, evaluation and network architecture, Expert Systems with Applications,85: 204-220.

Roberts, J.H. (2000) Developing new rules for new markets, Journal of the Academy of Marketing Science, 8:31-44.

Reichheld, F.F., Sasser, E.W. (1990) Zero defeofions: Quoliiy comes to services, Harvard Business Review, 68(5):2-8.

Jones, T. and Sasser, W. (1998) Why satisfied customers defect, IEEE Engineering Management Review, 26(3):16-26.

Nie,G.L., Rowe,W., Zhang,L.L., Tian,Y.J.,Shi, Y.,(2011) Credit card churn forecasting by logistic regression and decision tree,Expert Systems with Applications, 38(12) : 15273-15285.

Gordini, N. and Veglio, V. (2017) Customers churn prediction and marketing retention strategies: An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry, Industrial Marketing Management, 62:100-107.

中村博·熊倉広志·生田目崇(2011)「顧客セグメンテーションのための分析手法とその効果— ID-POS データをもとにした事例 —」『商学研究所報』42(5):1-28.

佐藤由将·大竹恒平·生田目崇(2017) 「ゴルフのEC サイトにおけるリピート顧客の特徴の分析」『情報処理学会第79回全国大会』5K-08.

Kotler, P. and K. Keller (2008) Marketing Management: International Edition, Prentice Hall.

夏侯賢城· 原田良雄 (2022)「K-meansを用いてネットビジネスユーザデータの顧客セグメンテーションの研究」『2022年度人工知能学会全国大会(第36回) 論文集』1G1-GS-10-04。

Bradley, P.S. and U.M. Fayyad (1998) Refining Initial Points for K-means Clustering, Proceedings of the 15th International Conference on Machine Learning, 91-99.

田中豊·垂水共之(1995)『Windows版統計解析ハンドブック多変量解析』『共立出版』。

Wendell, R. Smith (1956) Product Differentiation and Market Segmentation as Alternative Marketing Strategies, Journal of Marketing, 21(1):3-8.

Lazer, W. (1963) Life style concept and marketing, toward scientific marketing, in stephen greyser(Ed.),Toward Scientific Marketing,Chicago:American Marketing Assn.,130.

Wells,W.and Tigert,D.(1971)Activities,Interests And Opinions, Journal of Advertising Research, 11(4):27-35.

Mitchell, A. (1983) The nine American lifestyles: who we are and where we' re going. New York: Warner.

劉英姿·呉昊 (2006) 「顧客細分方法研究综述」『管理工程学報』20(1):53-57。

劉昱涛 (2013) 『市場營銷実務』『電子工業出版社』2:105-106。

向小玲 (2015)「基于人口統計因素的住宅市場細分研究—以德陽市為例」『商務營銷』1:61-63。

Hughes, Arthur M.(1994)Strategic Database Marketing,Chicago:Probus publishing.

Marcus,Claudio(1998)A practical yet meaningful approach to Customer Segmentation, Journal of Consumer Marketing, 15(5):494-504.

Cheng,C.H. and Chen,Y.S. (2009) Classifying the segmentation of customer value via RFM model and RS theory, Expert Systems with Applications, 36: 4176-4184.

Hsu, F.M., Lu, L.P., Lin,C.M. (2012) Segmentating customers by transaction data with concept hierarchy, Expert Systems with Applications, 39:6221-6228.

曽小青·徐秦·張丹·林大瀚 (2013)「基于消費数据挖掘的多指標客户細分新方法」『計算机応用研究』30:2944-2947。

蔡玖林·張磊·張秋三 (2015)「一种基于数据挖掘的零售業客户細分方法研究」『重庆工商大学学報(自然科学版)』32: 43-48。

楊倩倩·生佳根·趙海田 (2015)「K-means 聚类算法在民航客户細分中的応用」『電子設計工程』12:25-27。

盧海明·劉向東(2016)「電力客户細分及增值服務系統研究与応用」『华北電力技術』10:8-13。

趙铭·李雪·李秀婷 (2013)「基于聚类分析的商業銀行基金客户的分类研究」『管理评论』 7:38-44。

徐翔斌·王佳强·涂欢 (2012)「基于改进RFM模型的電子商务客户細分」『計算机応用』32(5): 1439-1442。

Brito, P.Q., Carlos, S., Sérgio, A., Mote, A., Byvoet, M.(2015) Customer segmentation in a large database of an online customized fashion business, Robotics & Computer Integrated Manufacturing,36:93-100.

Verhoef,Peter C.and Donkers,Bas (2001) Predicting customer potential value an application in the insurance industry, Decision Support Systems, 32:189-199.

Hwang, J. and Su,H.(2004) An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry, Expert Systems with Applications,26(2):181-188.

Kim,S.Y. and Jung, T.S.(2006) Customer segmentation and strategy development based on customer lifetime value: A case study, Expert System with Applications, 31:101-107.

綦欣德 (2013)「客户細分方法新視角」『商業時代』26:31-33。

趙萌·齐佳音(2014)「基于購買行為RFM及评论行為RFMP模型的客户终身価値研究」『統計与信息論坛』9:91-98。

劉寅·关志新·王缠(2015)「基于"数量·質量·效益"的金融産品客户細分-一个改进的RFM模型」『海南金融』12:21-24。

李静·朱金福 (2015)「公務航空客户価値评价与客户細分方法」『中国民航飞行学院学報』27: 28-36。

藪野祥太(2020)「データ視点からのCRM(顧客関係管理)の再考」『経営研究』71(3):87-107。

蘇朝晖 (2010)『客户関係管理:客户関係的建立与維护(第2版)』『清华大学出版社』。

Ramendra,T. and Letty,W.(2016) Customer portfolio management (CPM) for improved customer relationship management (CRM): Are your customers platinum, gold, silver,or bronze, Journal of Business Research,69(10).

Baydar,C.M. (2002) One-to-One Modeling and Simulation:A New Approach in Customer Relationship Management for Grocery Retail,Proceedings of SPIE-The International Society for Optical Engineering.

伍京华 (2017)『客户関係管理』『人民郵電出版社』。

Yuen, F. T. (2014) Exploring customer relationship management using simulation modelling in the retail sector, PhD Thesis, University of Warwick.

沢登秀明 (2000)「図解でわかるeCRMマーケティング : 見込み客を顧客に変えるeマーケティング手法!」『日本能率協会マネジメントセンター』5。

徐兰娇 (2011)「网络環境下的客户関係管理探析」『市場營銷』5:31-32。

Verma, D. and Verma, D.S. (2013) Managing Customer Relationships through Mobile, CRM In Organized retail outlets, International journal of engineering trends & technology, 4(5).

Soltani , Z. and Navimipour,N.J. (2016) Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research, Computers in Human Behavior, 61:667-688.

中田祐誠·村本直樹·山本岳洋·藤田澄男·大島裕明(2021) 「ウェブ検索ログからのカメラのオンライン購買行動予測」『人工知能学会論文誌 WI2-C』36(1):1-10.

安田大誠·吉野孝·松山浩士(2019)「POS データを用いた需要予測手法の事前自動判別

に関する基礎検討」『2019年度情報処理学会関西支部支部大会』B03.
新美潤一郎(2017)「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2017年度博士学位論文』。
北澤晃樹·生田目崇·大竹恒平(2020)「ゴルフポータルサイトにおけるユーザ行動データを用いた新規顧客の定着要因の特定」『情報処理学会第82回全国大会』2ZG-02。
佐藤千尋·高久雅生(2020)「Webページ閲覧履歴を用いた情報収集行動の振り返り支援」『情報知識学会誌』30(2):220-229.
中里見祐介·生田目崇·大竹恒平(2020)「ロイヤルティプログラムが与える消費者行動への影響の分析」『情報処理学会第82回全国大会』2ZG-04。
川名純平·諏訪竜也·関庸一(2018)「食品スーパーの多様性と顧客購買行動との関係分析」『日本オペレーションズ・リサーチ学会和文論文誌』61:23-44。
中村綾乃·酒井航太·吉野孝·松山浩士·貴志祥江·大西剛(2020)「ID-POS データを用いた客動線分析方法の検討」『2020年度情報処理学会関西支部支部大会』D-06。
松嵜祐樹·三川健太·後藤正幸(2020)「マルコフ潜在クラスモデルに基づくECサイトにおける施策実施効果分析に関する一考察」『情報処理学会論文誌』58(12): 2034-2045。
李銀星·照井伸彦(2018)「大規模集計POSデータの高次元スパースモデリング」『統計数理』66(2):235-247。
Amin,A., Anwar,S., Adnan,A., et al.(2017) Customer churn prediction in the telecommunication sector using a rough set approach,Neurocomputing,237: 242-254.
Jahromi,A.T., Stakhovych,S., Ewing,M.(2014) Managing B2B customer churn, retention and profitability,Industrial Marketing Management,43(7):1258-1268.
Migueis, V.L., Van den Poel, D., Camanho, A.S., Cunha, J.F.(2012) Modeling partial customer churn: On the value of first productcategory purchase sequences, Expert Systems with Applications, 39(12):11250-11256.
Huang,Y.and Kechadi,T.(2013) An effective hybrid learning system for telecommunication churn prediction, Expert Systems with Applications, 40(14):5635-5647.
Coussement,K., Lessmann,S., Verstraeten,G.(2017)A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry,Decision Support Systems, 95:27-36.
張玮·楊善林·劉婷婷(2014)「基于CART 和自适应Boosting算法的移動通信企業客户流失预测模型」『中国管理科学』22(10):90-96。
Larivière,B. and Van Den Poel D.(2004) Investigating the role of product features

in preventing customer churn, by using survival analysis and choice modeling: The case of financial services, Expert Systems with Applications, 27(2):277-285.

Hadden, J., Tiwari, A., Roy, R, et al.(2007) Computer assisted customer churn management : State-of-the-art and future trends, Computers & Operations Research, 34(10) : 2902-2917.

Datta,P., Masand,B., Mani,D.R., et al. (2000) Automated cellular modeling and prediction on a large scale, Artificial Intelligence Review,14(6):485-502。

Lima,E., Mues,C., Baesens,B. (2011) Monitoring and backtesting churn models, Expert Systems with Applications, 38(1):975-982.

Maaten, L.v.d. and Hinton,G.,(2008) How to Use t-SNE Effectively,Journal of Machine Learning Research,9:2579-2605.

Renjith, S. (2015) An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn, International Journal of Engineering Trends and Technology, 27(3):152-157.

De Caigny, A., Coussement, K., De Bock, K. W. (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, European Journal of Operational Research, 269 (2):760-772.

Pmar,K. and Topcu,Y.I.(2011) Applying Bayesian Belief Network approach to customer chum analysis :A case study on the telecom industry of Turkey, Expert Systems with Applications,38(6):7151-7157.

Farquad, MAH., Ravi,V., Raju,S.B.(2014) Churn prediction using comprehensible support vector machine:An analytical CRM application,Applied Soft Computing,19:31-40.

Tian,L.,Qiu, H.,Zheng,L.(2007) Telecom chum prediction modeling and application based on neural network,Computer Applications,27(9):2294-2297.

Yu, R., An, X., Jin, B., et al.(2018) Particle classification optimization-based BP network for telecommunication customer churn prediction,Neural Computing and Applications,2:707-720.

Neslin,S.A., Gupta,S., Kamakura,W.,et al. (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models,Journal of Marketing Research, 43:204-211.

Zhang,Y.(2015) A Customer Churn Alarm Model based on the C5.0 Decision Tree-Taking the Postal Short Message as an Example, Statistics & Information Forum,30(1):89-94.

Abbasimehr, H., Setak, M., Tarokh, M.J. (2014) A comparative assessment of the performance of ensemble learning in customer churn prediction, The

International Arab Journal of Information Technology, 11(6):599-606.

Vafeiadis,T., Diamantaras,K.I., Sarigiannidis,G., et al. (2015) A comparison of machine learning techniques for customer churn prediction, Simulation Modelling Practice and Theory, 55:1-9.

Gordini,N. & Veglio,V.(2013) Using neural networks for customer churn prediction modeling: preliminary findings from the italian electricity industry, Proceedings delX° Convegno Annuale della Società Italiana Marketing: "Smart Life. Dall' Innovazione Tecnologica al Mercato", 3-4 Ottobre 2013 obre. Milano, Italy: Università degli Studi di Milano-Bicocca.,1-13.

Xie,Y.Y. and Li,X.(2008) Churn prediction with linear discriminant boosting algorithm, 2008 International Conference on Machine Learning and Cybernetics, Kunming, China,228-233.

Wu, X. and Meng, S.(2016) E-commerce Customer Churn Prediction based on Customer Segmentation and AdaBoost, In Proceedings of the International Conference on Service Systems and Service Management (ICSSSM), Kunming, China, 24-26 June 2016.

Ji, H., Ni, F., Liu, J. (2021) Prediction of telecom customer churn based on XGB-BFS feature selection algorithm, Computer technology and development, 31(5):21-25.

Ahmed, A.A.Q. and Maheswari, D.(2019) An enhanced ensemble classifier for telecom churn prediction using cost based uplift modeling, International Journal of Information Technology, 11:381-391.

Ying, W., Lin, N., Xie, Y., et al. (2010) Research on the LDA boosting in customer churn prediction, Journal of Applied Statistics & Management, 29(3):400-408.

Zhang,W.,Yang,S.,Liu,T.(2014)Customer churn prediction in mobile communication enterprises based on CART and Boosting algorithm, Chinese Journal of Management Science, 22(10):90-96.

張玮·杨善林·劉婷婷 (2014)「基于CART和自适应Boosting算法的移动通信企業客户流出预測模型」『中国管理科学』22(10):90-96。

Masand, B, Datta, P, Mani, D.R., et al. (1999) CHAMP:A prototype for automated cellular churn prediction, Data Mining and Knowledge Discovery, 3(2):219-225.

丁君美·劉贵全·李慧(2015)「改进随机森林算法在電信業客户流出预测中的応用」『模式識別与人工智能』28(11):1041-1049。

羅彬·邵培基·羅尽尧 (2011)「基于粗糙集理論一神经网络一蜂群算法集成的客户流出研究」『管理学報』8(2):265-272。

Larivi, B. and Van Den Poel, D. (2004) Investigating the role of product features in preventing customer churn by using survival analysis and choice modeling: The case of financial services, Expert Systems with Applications, 27(2): 277-285.

応維雲·覃正·趙宇·李兵·李秀(2007)「SVM方法及其在客户流失预测中的応用研究」『系统工程理論与実践』27(7):105-110。

Yu, X.B., Guo, S.S., Guo, J., et al.(2011)An extended support vector machine forecasting framework for customer churn in e-commerce, Expert Systems with Applications, 38(3):1425-1430.

Wu, J., Shi, L., Lin, W.P., Tsai, S.B., Xu, G.S. (2020) An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm, Mathematical Problems in Engineering: Theory, Methods and Applications, 2020(Pt.42)-8884227.1-8884227.7.

Wu, J., Shi, L., Yang, L.P., Niu, X.X., Li, Y.Y., Cui, X.D., Tsai, S.B., Zhang, Y.B.(2021) User Value Identification Based on Improved RFM Model and K-Means++ Algorithm for Complex Data Analysis, Wireless Communications and Mobile Computing, 9982484, 1-8.

Li, Y.; Chu, X.Q.; Tian, D.; Feng, J.Y.; Mu, W.S.(2021) Customer segmentation using K-means clustering and the adaptive, Applied Soft Computing, 113(B): 107924.

Christy, A.J., Umamakeswari, A., Priyatharsini, L., Neyaa, A.(2021) RFM ranking-An effective approach to customer segmentation, Journal of King Saud University -Computer and Information Sciences, 33(10):1251-1257.

Abbasimehr, H., Bahrini, A.(2022) An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation, Expert Systems with Applications, 192:116373.

Alboukaey, N., Joukhadar, A., Ghneim, N.(2020) Dynamic behavior based churn prediction in mobile telecom, Expert Systems with Applications, 162:113779.

Zhou, J., Zhai, L.L., Pantelous, A.A.(2020) Market Segmentation Using High-dimensional Sparse Consumers Data, Expert Systems with Applications, 145:113136.

Li, M.; Wang, Q.W.; Shen, Y.Z.; Zhu, T.Y. (2021) Customer relationship management analysis of outpatients in a Chinese infectious disease hospital using drug-proportion recency-frequency-monetary model, International Journal of Medical Informatics147: 104373.

元田浩·栗田多喜夫·樋口知之·松本裕治·村田昇 監訳, C.M.ビショップ 著(2007)「パターン認識と機械学習(原タイトル: Pattern recognition and machine learning)」

3:Springer。

Birodkar,V., Mobahi,H., Bengio.S.(2019) Semantic redundancies in image-classification datasets: The 10% you don’t need, ArXiv.

Onoda,T.,Sakai,M.,Yamada,S.(2011)Experimental Comparison of Clustering Results for k-means by using different seeding methods, The 25th Annual Conference of the Japanese Society for Artificial Intelligence.

金明哲(2009)「統計的テキスト解析 (15) –テキストの分類②–」『統計と情報の専門誌「エストレーラ」』5 (182) :44-49。

Lee,S., Lee,H., Abbeel,P. et al.(2006) Efficient L1 regularized logistic regression, Proc. of the 21st national conf. on Artificial intelligence, 1:401-408.

Minka,T. P. (2004) A comparison of numerical optimizers for logistic regression, Technical Report (Mathematics),1-18.

Vapnik, V.N. (2000) The Nature of Statistic Learning Theory (Second Edition), Springer, New York.

Scholkopf, B., Smolla, A. (2002) Learning with kernels-Support Vector Machines, Regularization, Optimization and Beyond, MIT press.

Yoav, F. and Robert,E. (1996) Schapire, Experiments with a New Boosting Algorithm, Machine Learning, Proceeding of the thirteenth international conference, 1-9.

García, D. L.,Nebot,À.& Vellido, A.(2017) Intelligent data analysis approaches to churn as a business problem: A survey. Knowledge and information systems, 51(3):719-774.

Alibaba Cloud Tianchi Data Sets.Available online: https://tianchi.aliyun.com/datase (accessed on 17 March 2021).

Cao, L. (2010) Behavior Informatics and Analytics: Let Behavior Talk, Proc. of 2008 IEEE Inter. Conf. on Data Mining Workshops,15-19.

Stolfo,S., Hershkop,S., Hu,C., et al. (2006) Behavior-based modeling and its application to Email analysis, ACM Transactions on Internet Technology, 6(2):187-221.

Li, M., Wang, Q.W., Shen, Y.Z., Zhu, T.Y.(2021) Customer relationship management analysis of outpatients in a Chinese infectious disease hospital using drug-proportion recency-frequency-monetary model,International Journal of Medical Informatics,147:104373.

Runge, J., Peng, G., Garcin, F., Faltings,B. (2014) Churn prediction for high-value players in casual social games,IEEE Conference on Computational Intelligence and Games(CIG),Dortmund, Germany. 1-8.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B. (2012) New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. Eur. J. Oper. Res. 2012, 218, 211-229.

Breiman, L. (2001) Random forests, Mmachine learning, 45:5-32.

Breiman, L. (1996) Bagging predictors, Mmachine learning, 24:123-140.

Goldstein, B. A., Polley, E. C., Briggs, F. B. S. (2011) Random Forests for Genetic Association Studies, Stat. Appl. Genet. Mol. , 10:32.

Drummond, C., Holte, R. C. (2003) C4. 5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling, In Proceedings of Workshop on Learning from Imbalanced Datasets II, ICML, Washington, DC, USA, 21 August.

Chawla, N. V. ; Bowyer, K. W. ; Hall, L. Q. (2002) Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res., 16: 321-357.

Provost, F. (1999) Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In Proceedings of the International Conference on knowledge Discovery and Data Mining (KDD), San Diego, CA, USA, 15-18. August 1999, 43-48.

Fan, X., Ke, T. (2010) Enhanced maximum AUC linear classifier. In Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Yantai, China, 10-12 August 2010, 1540-1544.

Majnik, M., Bosnic, Z. (2013) ROC analysis of classifiers in machine learning: A survey, Intelligent Data Analysis, 17(3): 531-558.

Hamel, L. (2009) Model assessment with ROC curves, Encyclopedia of Data Warehousing and Mining, Second Edition, Pennsylvania, USA: IGI Global, 1316-1323.

Norton, M., Uryasev, S. (2019) Maximization of AUC and buffered AUC in binary classification, Mathematical Programming, 174(1-2): 575-612.

Sturm, A. and Bob, L. (2013) Classification accuracy is not enough, Journal of Intelligent Information Systems, 41: 371-406.

Xiahou, X. C and Harada, Y. O. (2020) B2C E-Commerce Customer Churn Prediction based on k-means and SVM, Journal of Theoretical and Applied Electronic Commerce Research, 17(2):458-475.

Xiahou, X. C and Harada, Y. O. (2020) Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-commerce Industry, American Journal of Industrial and Business Management, 12(3):277-293.

Ma, S. and Huang, J. (2005) Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data, Bioinformatics, 21 : 4356-4362.

Song, X. and Ma, S. (2010) Penalized Variable Selection with U-Estimates, Journal of Nonparametric Statistics, 22(4):499-515.

Chang, H. H. and Tsay, S. F. (2004) Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation, Journal of Information Management, 11:161-203.

Rachid, A.D., Abdellah, A., Belaid, B., Rachid, L.(2018) Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context, International Journal of Electrical and Computer Engineering, 8(4):2367-2383.

Chen, Z.Y. , Fan, Z.P., Sun, M.H(2012) A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data, European Journal of Operational Research, 223:461-472.

Mozer, M.C., Wolniewicz, R., Grimes, D.B., et al. (2000) Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry, IEEE Transactions on Neural Networks, 11(3):690-696.

Holtrop, N., Wieringa, J.E., Gijsenberg, M.J., et al. (2016) No future without the past? Predicting churn in the face of customer privacy, International Journal of Research in Marketing, 34(1):154-172.

De Bock, K.W. and Van Den Poel, D.(2011) An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction, Expert Systems with Applications, 38(10):12293-12301.

夏国恩、金炜東（2008）基于支持向量机的顧客流失预测模型，系统工程理论与实践，28(1): 71-77.

Chen, K., Hu, Y.H., Hsieh, Y.C.(2015) Predicting customer churn from valuable B2B customers in the logistics industry: A case study, Information Systems and e-Business Management, 13:475-494.

Caigny, D.A., Coussement, K., Verbeke, W., Idbenjra, K., Phan, M.(2021) Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach, Industrial marketing management, 99: 28-39.

Buckinx, W. and Van den Poel, D. (2005)Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, Eur. J. Oper. Res. 164:252-268.

Miguéis, V.L., Camanho, A., Cunha, J.F.(2013) Customer attrition in retailing: An application of Multivariate Adaptive Regression Splines, Expert Systems with Applications, 40:6225-6232.

Gattermann-Itschert, T. and Thonemann, U.W.(2021) How training on multiple time slices improves performance in churn prediction, European Journal of Operational Research, 295(2):664-674.

Sood, A. and Kumar, V. (2017) Analyzing client profitability across diffusion segments for a continuous innovation, Journal of Marketing Research, 54(6):932-951.

Duan, Y., Edwards, J.S., Dwivedi, Y.K. (2019) Artificial intelligence for decision making in the era of big data-Evolution, challenges and research agenda, International Journal of Information Management, 48:63-71.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwived, R., Edwards, J., Eirug, A., et al. (2021)Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy, International Journal of Information Management, 57: 101994.1-101994.47.

謝 辞

本論文内の研究を遂行するにあたり，指導教員である大阪産業大学大学院経営・流通学研究科の原田良雄教授より多大なるご指導・助言を頂きました。原田先生は私が博士前期課程修了後，博士後期課程進学を希望してご相談した際にも温かく迎え入れてくださりました。本研究を進めるにあたっても非常にきめ細かいご指導を頂き，内容を深めてことができました。この場をお借りして厚く感謝申し上げます。本研究は、筆者が博士後期課程において行った研究成果をまとめたものです。 本論文作成にあたり、原田先生には、本研究の構想からデータ分析、論文作成に至るまで、終始一貫しての暖かいご指導、ご鞭撻を賜り、心から厚く御礼を申し上げます。

また、研究成果の取りまとめにあたって多くのご教示を賜りました大阪産業大学大学院経営・流通学研究科の中村徹教授、藤岡芳郎教授に深甚の謝意を表します。本研究の遂行にあたって、ご助言を頂きました経営・流通学研究科の各教官の方々に御礼を申し上げます。

最後に、私の研究生活を様々な面で支えてくれた数多くの先輩、友人、知人に改めて感謝いたします。

「Journal of Theoretical and Applied Electronic Commerce Research」,「American Journal of Industrial and Business Management」, 「International Journal of Computer Trends and Technology」の匿名の査読者の先生方にも感謝を申し上げます。査読プロセスを経験することにより，研究に深みをあたえることができました。

私の研究生活を温かく見守り続けてくれた両親に深い感謝の気持ちを捧げます。

研究業績一覧

学術論文

[1] **Xiancheng Xiahou, Yoshio Harada,** “B2C E-Commerce Customer Churn Prediction based on k-means and SVM”, *Journal of Theoretical and Applied Electronic Commerce Research*, 2022, Vol.17 (2), pp.458-475. (SSCI Impact Factor : 5.318 ; Scopus CiteScore : 3.1)

[2] **Xiancheng Xiahou, Yoshio Harada**, “Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-commerce Industry”, *American Journal of Industrial and Business Management*, 2022, Vol.12 (3), pp. 277-293.

[3] **Xiancheng Xiahou, Yoshio Harada,** "K-Medoids Clustering Techniques in Predicting Customers Churn: A Case Study in the E-Commerce Industry", *International Journal of Computer Trends and Technology*, 2022, Vol.70 (2), pp.22-28.

学会口頭発表

夏侯賢城･原田良雄, 「K-meansを用いてネットビジネスユーザデータの顧客セグメンテーションの研究」, 2022年度人工知能学会全国大会(第36回, 京都, 2022年6月14日), 1G1-GS-10-04。