3C1-4

行動経済学的な知見を用いた消費者の情報探索行動の予測とレコメンデーション法の開発

Prediction of consumer's information search behavior from a behavioral–economics point of view and development of a recommendation method

高畑圭佑 *1 Keisuke Takahata　星野崇宏 *2 Takahiro Hoshino　柳博俊 *2 Hirotoshi Yanagi　渋谷友磯子 *3 Yukiko Shibuya

*1慶應義塾大学大学院経済学研究科 Graduate School of Economics, Keio University

*2慶應義塾大学経済学部 Faculty of Economics, Keio University

*3株式会社プロトコーポレーション PROTO CORPORATION

This paper reports a result of the analysis of goo-net data, which is one of the largest used car sales sites in Japan. In addition to aggregated variables which represent consumer behavior, we introduce HHI, Herfindahl-Hirschman Index, into the model to consider qualitative differences of consumer behavior. We show that there are significant differenes in the distribution of some variables between consumers who do CV and those who do not. Also we show that AUC of some learnig models which consider HHI is higher than that of models which do not consider. We suppose those results may be useful both to understand consumer behavior and to enhance performance of e-commerce websites.

1. はじめに

本稿では，web サイト上における利用客のページ閲覧行動から CV(コンバージョン: 会員登録や商品購入など，EC サイトにおける成果目標) 達成に関わる要因を抽出して成果向上を目指すマーケティング研究の一つのケースとして，大手中古車販売サイトであるグーネットを対象とした分析結果を報告する．今回は CV を「web サイトからの中古車見積もり発注」と定義した上で，利用客のサイト訪問回数・滞在時間・商品ページ閲覧回数などのサイト内における行動データや，後述する HHI といった指標を変数として，CV する利用客とそうでない客を識別することを試みる．また，これらの変数の分布の違いを分析することを通して，積極的にマーケティングを行うべき利用客層を特定するための示唆を得ることを目標とする．

2. データおよび変数について

2.1 使用するデータ

CV する利用客を識別するための説明変数として，Google Analytics より取得したサイト内行動に関する以下のデータを用いる：訪問回数，滞在時間，直帰回数，ページ閲覧回数．

また，データ全体 (2016 年 11 月から 2017 年 1 月までにサイトを訪問した利用者の中である条件を満たす群) に対して CV した利用客に関するデータの件数が小さいため，推定に困難が伴う場合がある．よって，「ページ閲覧数が平均以上である利用客」は CV する可能性のある集団であると断定し，これを前述の変数を用いて予測することも試みる．なお，これにより定義される集団の中で実際に CV した利用客は，CV した利用客の全体のうちのおよそ半数に当たる．

2.2 HHI について

各ページの閲覧数や滞在時間といった量的変数のみでは，利用客の行動の質的な多様性を表現するのには不十分である．そこで経済学において特定産業のシェアや消費者の購買商品のシェアなどを一次元の指標として用いられる HHI(ハーフィンダール・ハーシュマン・インデックス) を閲覧した車の価格帯および車種に関して計算し，これも説明変数として加える [新美・星野 15，新美・星野 17]．本研究では閲覧したページのうち何割が特定の価格帯・車種に集中しているかという観点から利用客の閲覧パターンを分類する指標としてこれを用いる．

連絡先: 星野崇宏，email: hoshino@econ.keio.ac.jp

3. 分析結果

3.1 変数の分布

図 1 から図 6 は，CV した利用客とそうでない利用客に関して，前述の説明変数の分布の違いを表したものである．まずサイト滞在時間に関しては，CV しなかった利用客の分布は最頻値が 0 付近となる指数型の分布になるのに対して，CV した利用客の分布は最頻値が右にシフトした裾の重い分布となった．つまり CV した利用客の方が平均的に長くサイトに滞在していることを意味するが，これは直観に沿うものであろう．

一方で直帰回数の分布に関しては，CV の有無による差は生まれなかった．ページ閲覧数およびサイト訪問回数については共に，CV した利用者の分布は CV しなかった利用者の分布の裾を重くしたものとなったが，分布の概形には大きな違いは生まれなかった．

次に車種と価格帯に関する HHI の分布を見てみると，両者に共通する傾向が見てとれた．CV しなかった利用客は HHI が 5000 となる領域か，もしくはそれより小さい領域に集中するのに対して，CV した利用客は比較的 HHI が大きい領域に集中していた．すなわち，CV した利用客は閲覧する対象が車種・価格帯ともにある程度限定されているのに対し，逆に CV しなかった利用客 (購入以外の目的で閲覧している場合も含まれる) は幅広い対象を閲覧している傾向があると言える．よって HHI は，潜在的に CV する可能性のある利用客を特定し，集中的に広告介入を行う際の一つの判断材料とすることが可能であると考えられる．

以上の変数は分布としては類似しているものもあるが，相関をみると必ずしも高くはないので，これらを次節の予測に用いることにする．

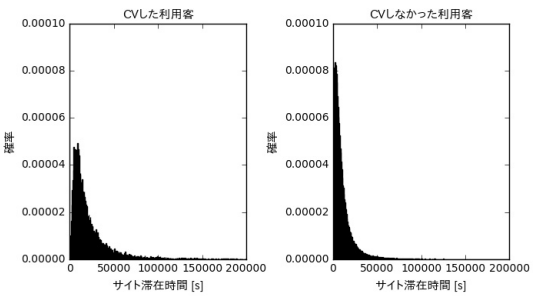

図 1: 滞在時間の分布の違い

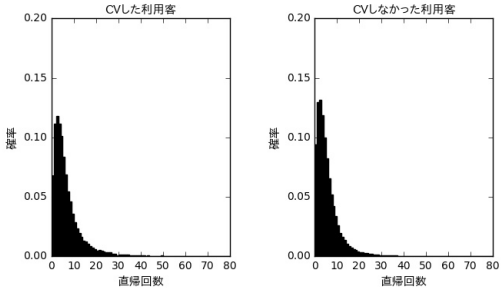

図 2: 直帰回数の分布の違い

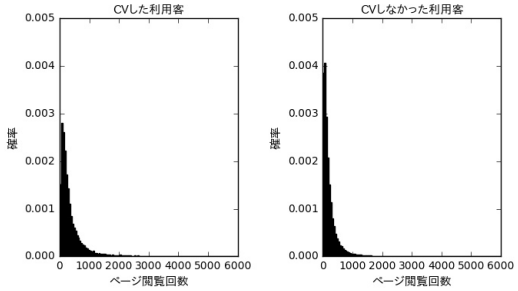

図 3: ページ閲覧回数の分布の違い

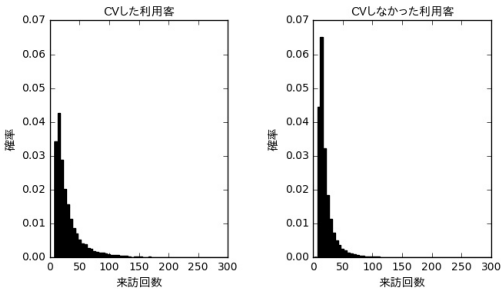

図 4: 来訪回数の分布の違い

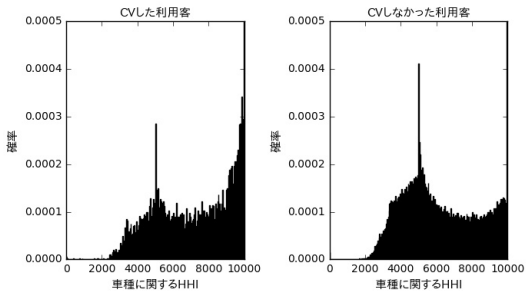

図 5: 車種に関する HHI の分布の違い

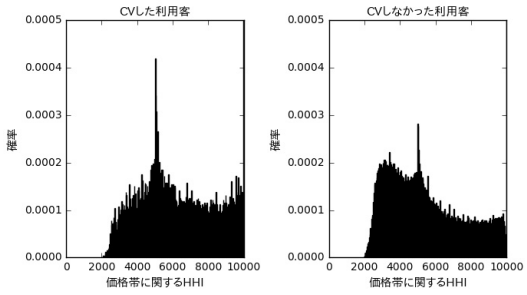

図 6: 価格帯に関する HHI の分布の違い

3.2 CV する利用客の予測

いくつかの機械学習手法を用いて CV する利用客を識別した結果を以下に示す．今回は識別手法としてロジスティック回帰 (LR)，ランダムフォレスト (RF) および XGBoost を用いた．説明変数は前述のサイト訪問回数，滞在時間，直帰回数，ページ閲覧回数，車種に関する HHI および価格帯に関する HHI を用いた．

まず HHI 以外の 4 つの変数で学習を行うと，結果は表 1 のようになった (なお以下で示す AUC の結果はすべて $k = 5$ で交差検証をしたときの平均値である)．AUC の意味で相対的に最も精度が高かったのは XGboost であったが，前述のとおりデータ数に対して CV したデータが少ないため，必ずしも学習に成功したとは言えない．

HHI を含めた 6 つの説明変数を用いて学習を行った結果も同様に表 1 に示してあるが，これを見るとすべての手法において HHI を導入した場合のほうが AUC が高くなっていることが分かる．つまり，利用客の閲覧パターンの多様性を考慮することにより，学習の精度が高まるということが分かった．

表 1: CV した利用者の推定結果

手法	AUC (HHI なし)	AUC (HHI あり)
LR	0.727	0.728
RF	0.624	0.671
XGBoost	0.745	0.779

3.3 潜在的に CV する可能性のある利用客の予測

表 2「ページ閲覧回数が平均以上である利用客」を潜在的に CV する可能性のある客と定義し，それを学習させた場合の結果である (この分析ではページ閲覧回数は説明変数から削除している)．AUC の値から，実際に CV した利用客そのものを識別するよりも精度が高くなっていることが分かる．またすべての学習手法において，HHI を説明変数に加えた場合の方が AUC が高くなっている．よってこの結果からも HHI の有効性が確認された．ただし，今回の定義では実際に CV した利用客のうちの約 48 %しかカバーできていないため，今後はさらに定義を精緻化する必要がある．

表 2: 潜在的に CV する可能性のある利用者の推定結果

手法	AUC (HHI なし)	AUC (HHI あり)
LR	0.945	0.950
RF	0.894	0.932
XGBoost	0.948	0.953

4. 今後の課題

今回はデータの大まかな予測を行うにあたり，閲覧に関する代表的な指標のみを用いたが，今後は閲覧パターンの時系列的な変化を考えるなど，より詳細な分析を行おうと考えている．また既存の機械学習手法を適用するに留まらず，行動経済学的な知見に基づいた利用客の行動メカニズムを導入し，より精度の高いモデルを構築することを試みる．

参考文献

[新美・星野 15] 新美潤一郎，星野崇宏: ユーザ別アクセス・パターン情報の多様性を用いた顧客行動の予測とモデリング, 応用統計学, Vol. 44, No. 3 (2015).

[新美・星野 17] 新美潤一郎，星野崇宏: 顧客行動の多様性変数を利用した購買行動の予測–Deep Learning を応用した実店舗・Web・モバイルの多面的な分析– , 人工知能学会論文誌, Vol. 32, No. 2 (2017).