

Web リスティング広告における基本広告データを用いた コンバージョン分析と予測

石川 善一郎[†] 奥 牧人[‡] 河野 崇^{†, ‡}

[†] 東京大学大学院工学系研究科 〒153-8505 東京都目黒区駒場 4-6-1

[‡] 東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

E-mail: {znsn0419, oku, kohno}@sat.t.u-tokyo.ac.jp

あらまし Web リスティング広告を用いた個人の広告主による広告出稿が広く行われているが、コンバージョン (CV) 予測に有用と考えられる検索ユーザーの個人 ID などの詳細なデータの獲得は非常に困難である。本研究では個人の広告主が得られる限られたデータ群と CV の関係を分析するとともに、それらを用いて CV の予測を行った。その結果、基本的なデータ群の中から CV 予測に大きく関与する指標の組み合わせが得られた。今後、本研究で得られた指標とテキストレベルの特徴を上手く組み合わせることが出来れば、さらに良い CV 予測が可能となると期待される。

キーワード Web リスティング広告, コンバージョン, Yahoo!プロモーション広告

1. 序論

1.1 研究背景

Web リスティング広告とは、ユーザーによる検索と連動して表示される広告である。各検索キーワードに対して広告主が入札を行い、オークションで掲載順位が決まる。広告がユーザーによってクリックされた際に広告主が課金されるクリック課金型が一般的である。広告主は、メーカーや広告代理店などの法人に限らず、個人による広告出稿も幅広く行われており、その広告費は市場で年間2000億円を超えている[1]。これまでこのWebリスティング広告の費用対効果・収益効果を向上させるための研究は膨大かつ詳細なデータの利用が可能なYahoo!やGoogle, サイバーエージェントといった大手IT系会社によるものが多かった[1]-[4]。

一方、個人の広告主が広告運用において得られるデータ群は限られており、例えば、検索ユーザーの個人ID などの詳細なデータを得るのは非常に困難である。このため個人の広告主が限られた情報のみから広告目的達成(CV:コンバージョン)を正確に予測するための手法はほとんど確立されていない。

そこで本研究では、筆者のうち一名が個人として Yahoo!プロモーション広告において広告運用する中で得られたデータのうちある特定の商品に関するデータを用い、CV(この場合は商品購入)にまで実際に繋がったまたは繋がらなかった検索キーワードに割り当てられた定量的な広告データにどのような分布傾向があるかをまず調査し、続いて機械学習を用いた分析で CV の有無に対する予測を行った。本研究は個人の広告主視点から CV 予測に重要な指標を明らかにし広告の収益増へと繋げるための知見を得ることを目的とする。

1.2 Web リスティング広告の概要

Web 上での広告出稿は様々な形式がある。近年では Twitter, Facebook などの SNS を媒体とした広告戦略も多数存在するが、本稿では Web 検索エンジンを媒体としたリスティング広告について説明する。

Webリスティング広告には、広告主、パブリッシャー (広告媒体)、ユーザーと呼ばれる三種類の主要なプレイヤーが存在する[2]。特にWebリスティング広告の場合、広告主は広告を出稿し、ユーザーを遷移URLによりリンク先ページへ誘導して製品購入につなげることを目指す。パブリッシャ

一は検索サイトのオーナーであり、検索結果の画面に広告を掲載して収益を得ることを目的とする。ユーザーは検索サイトで検索を行い、そこに掲載された広告を閲覧する、広告主の潜在的な顧客である。広告を閲覧したユーザーは広告のクリックやそれとは別ルートを通して広告主のサイトを訪れることがある。

ユーザーがオンライン広告に触れて広告主の顧客となるにはインプレッション、クリック、コンバージョンの三段階を経る[2]。特に Web リスティング広告の場合、インプレッションとは、ユーザーがある目的の商品に対するキーワードで検索を行い Web ページに掲載された広告を閲覧することを表す。クリックは、ユーザーが表示された広告文をクリックすることを指す。これによりユーザーは広告主があらかじめ指定したウェブページに遷移する。コンバージョンは、広告主が目標とする特定のユーザー行動である。例えばショッピングサイトと提携する個人の広告主の場合、ユーザーが提携サイトで商品を購入することがコンバージョンの例として挙げられる。本稿ではこの意味でコンバージョンを定義とする。ユーザーから見たインプレッション、クリック、コンバージョンの流れを図 1 に示す。なお、クリックの過程が 2 回登場するが本稿では全て最初の広告文をクリックしたときのクリック数を考える。

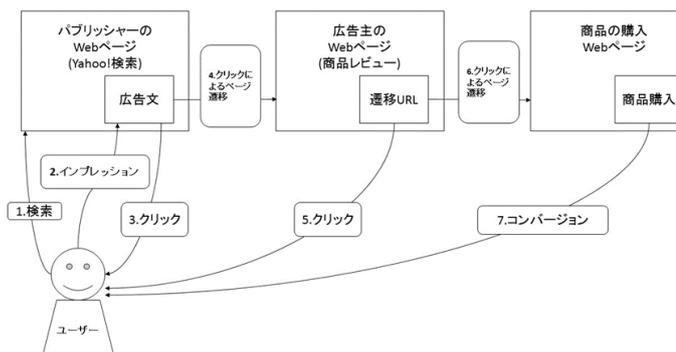


図 1. Web リスティング広告の模式図

(1) ユーザーはパブリッシャーの Web ページで検索し(2) ページ上に掲載された広告を閲覧する。(3) 興味を持ったユーザーが広告をクリックすると、(4) 広告主の商品レビューなどが記載された Web ページに遷移する。(5) ユーザーが広告主のページ上でさらに興味を持てば、(6) 商品購入を行える Web ページへと遷移し、(7) 実際に商品購入を行うとコンバージョ

ンとなる。

1.3 関連研究

まず、田頭らや McMahan らが広告のクリック率(CTR)について同時期に研究されており[4][5]、続いて田頭らはコンバージョン率(CVR)をデータの素性をロジスティック回帰を用いることで評価することで予測しており、また Web リスティング広告についての詳しい説明を行っている[2]。これを元に、宮西らはユーザーの行動履歴と広告属性を NMF(Nonnegative Matrix Factorization)を用いることでレコメンド手法に関する提案を行っている[1]。また原らも広告主のサイトでのユーザーのアクセスログを用いて、コンバージョンに近いユーザーを自動的に抽出する手法について提案している[4]。Web リスティング広告と類似したディスプレイ広告における CTR や CVR の予測については Chapelle らや山本らが行っている[7][9]。有坂は広告運用における数理モデルを提案している[6]。

また、本研究では扱っていないが Web 広告において重要なテキストレベルの解析についても藤田らによって行われており[10]、本研究との関連を強めるアプローチを測ればより効率的な広告運用に関する知見が得られると考えられる。

2. 使用データと解析手法

2.1 使用データ

本節では本研究において使用したデータの種類とその説明およびコンバージョンの有無ごとの見解を述べる。

個人での Web 検索エンジン上での広告出稿は 2012 年までは Google AdWords を利用した Google の検索エンジン上のものが主流であったが、以降仕様の変更に伴い審査基準が厳しくなり主に法人からの広告出稿のみとなった。これに伴い 2017 年現在、個人の広告出稿は Yahoo! プロモーション広告を利用した Yahoo! の検索エンジン上に限定されている。

本稿では、ある 1 つの商品に関する広告グループの 2016 年 8 月 1 日から 10 月 31 日までのデータ(n=253)を扱う。なお、インプレッションが 0 のものは除いている。データの中に含まれる情報は Yahoo! プロモーション広告で広告管理している全てのユーザーが獲得可能であり、運用管理の上で指標となりうる以下の基本広告データを状況によって組み合わせることで解析を行った。なお、これらはその広告グループの複数

登録された検索キーワードごとに設定されている。解析に用いた商品および期間は、季節性や局所的なトレンドを含んでおらず検索ボリュームがバースト的に増減していないようなものを選んだ。

基本広告データ

- a)インプレッション数
- b)クリック数
- c)クリック率
- d)入札価格
- e)検索1ページ目に必要な入札価格
- f)検索1ページ目上部に必要な入札価格
- g)平均CPC(Cost per Click)
- h)品質インデックス
- i)平均掲載順位
- j)合計コスト
- k)CVの有無(有: CV=1 , 無: CV=0)

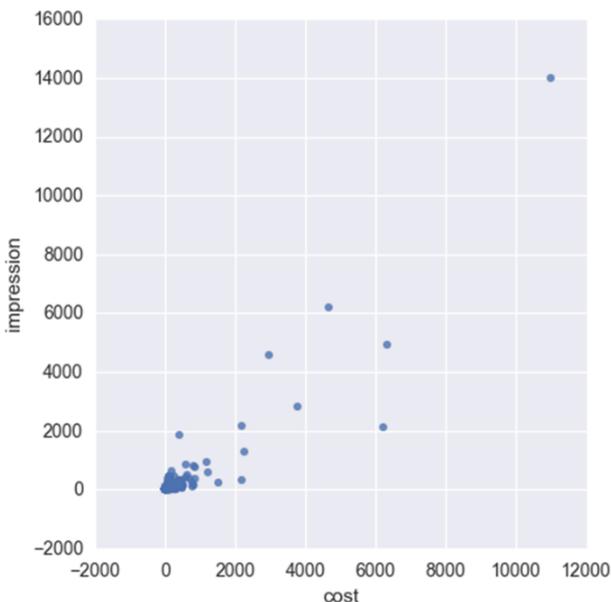


図 2 合計コストとインプレッション数の相関関係 (相関係数 : 0.91782)

2.2 解析手法

I. 決定木

II. ランダムフォレスト

III. ロジスティック回帰

IV. K 近傍

本研究では上記 4 種類の教師ありの機械学習の手法を用いた。決定木は Yes または No の二分岐の繰り返しによって表現されるため人間が最も解釈しやすい手法であり、ランダムフォレストはその決定木の手法を応用したものである。ロジスティック回帰は計算が簡単で大規模データにも適用しやすく宮西ら[1]が CV 予測モデルのなかで用いている。K 近傍も計算が簡単な手法だが、データ全体ではなく一部のデータだけを使うため外れ値の影響を受けにくいという特徴がある。

CV 達成の識別を複数の手法で行う理由としては、偏った手法を扱うことでデータの恣意性を避けることや、解析結果の要因の追求に難易度の差があるためである。例えば、決定木であれば木構造を見ることで結果の要因は可視化できるが、その応用のランダムフォレストになると結果の精度は上がりやすいものの結果の要因を追求するのは難しくなる。

3.教師あり学習による CV ラベル予測

前章で述べたデータと解析手法を用いて個人広告主が広告運用をする上で役に立つ知見を得ることを目指し解析を行った。以下、検索キーワードにおける「商品名 ロコミ」などの「ロコミ」の部分を検索フレーズと称する。

3.1 概要

個人広告主が実際にYahoo!プロモーション広告上で得られる基本広告データは前章の(a)~(k)であるが、これらは各検索キーワードに割り当てられている。「商品名 ロコミ」のような消費者が頻繁に検索するキーワードは多くのデータが大きな値を示すがときにはCVまで繋がらない、つまり(k)が0となることもある。一方、その逆で(a)のインプレッションが小さい値にも関わらずCVにまで繋がる、つまりCV=1となるケースも見られる。さらに、ある一定期間で必ずCVまで繋がる検索フレーズというのは存在しないことから、本稿ではその検索キーワードごとに割り当てられた(a)~(j)のデータのなかで、CV=1とラベルが当てはまるために必要なデータの組み合わせはどれが最適かを、教師あり学習の各手法(決定木、

ランダムフォレスト, ロジスティック回帰, K近傍)をそれぞれ用いることで解析した。

CV予測の性能を評価するため, 表1の混同行列をもとにした, 機械学習でよく用いられる4つのスコアを用いた。以下その式と意味を示す。

表 1 CVの有無のカウントに関する混同行列

		予測モデル	
		CV = 1	CV = 0
実際のデータ	CV = 1	TP	FN
	CV = 0	FP	TN

TP: True Positive, FP: False Positive,

FN: False Negative, TN: True Negative

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_{score} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

適合率(Precision)と再現率(Recall)はトレード・オフの関係にあるため, 一概にいずれかの値が1に近いからといって, 組み合わせた(a)-(j)のデータセットが良かったものとは限らない。そこで(a)-(j)のうち少なくとも1つを含む組み合わせの合計1023通りを全て調べ, F値(F_{score})が高い値を示すケースを探し, 以下にまとめた。F値は適合率と再現率が共に良い場合のみ高い値をとる指標である。

次項からの結果は訓練データとバリデーションデータによりF値が高くなるようなチューニングを施し, そのときの基本広告データの組み合わせを元にしてテストを同チューニング下で行った。

3.2.2 データセット A

データセットAでは, 8月を訓練データ (n=129), 9月をバリ

デーションデータ (n=67), 10月をテストデータ (n=57) とし, その結果を図3に示す。

結果, バリデーションはF値が約70~86%の高いスコアを出力できた。使用した基本広告データの種類や組み合わせについて, 特に80%を超える結果を示したランダムフォレストとK近傍に共通するのは (b) クリック数, (e) 1ページ目に必要な入札価格, (i) 平均掲載順位, (j) 合計コストの頻度が多いことである。

しかし, テスト結果において著しく各スコアが下がってしまった。低下の原因としては, 今回の8~10月の使用データは季節性によるトレンドやバーストな増減が含まれていない定常性のあるデータを用いていることが前提にあったが実際はそうではなかったことが考えられる。



図 3 データセット A 結果

3.2.3 データセット B

データセット B ではデータの定常性を図るためにデータ総数 n=253 を一度シャッフルし, CV=1 および 0 の割合が均一なデータにリサンプリングを行った。具体的には訓練データを n=151 (n_{CV=1} = 24, n_{CV=0} = 127), バリデーションデータとテストデータを n=51 (n_{CV=1} = 8, n_{CV=0} = 43) と設定した。

結果の定常性は図れたものの, 高スコアを示した決定木と

ランダムフォレストにおいて使用した基本広告データの重みに偏りが見られた。

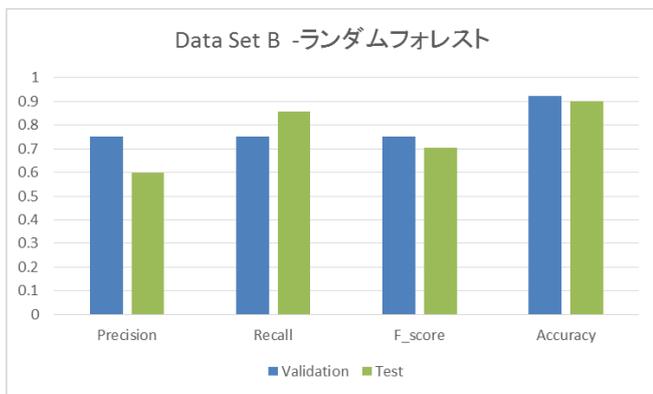
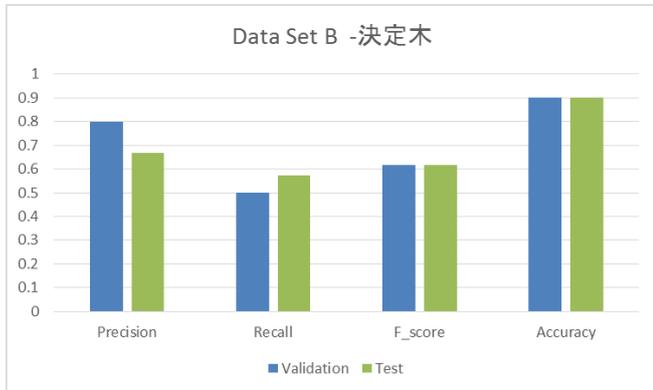


図 4 データセット B 結果

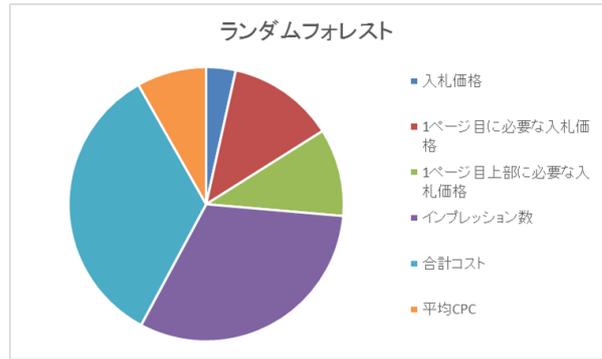
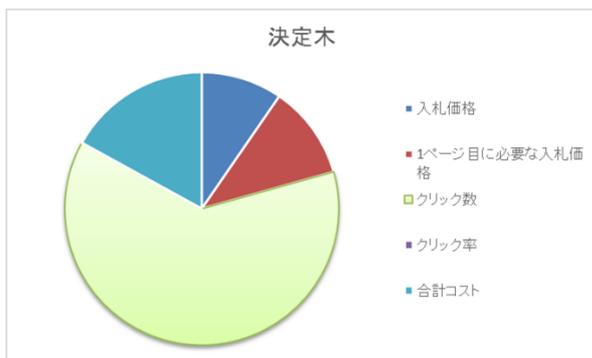


図 5 データセット B : 重み値

3.2.4 データセット C

データセット B を NMF(Nonnegative Matrix Factorization)を用いて 6 次元まで圧縮し同様の計算を行った。表 2 で強調した特徴量は、高スコアを示す中で使用頻度及び重み値が高かった複合した基本広告データである。

表 2 データセット C : 複合した基本広告データ

特徴量名	複合した基本広告データ
feature1	品質インデックス
feature2	入札価格
feature3	平均CPC
feature4	1ページ目に必要な入札価格, 1ページ目上部に必要な入札価格
feature5	インプレッション数, クリック数, 合計コスト
feature6	平均掲載順位

4.考察

本研究では、Yahoo!プロモーション広告への個人の広告出稿において得られる基本的な広告データを中心に扱い、広告主が効果的な運用を行うための定量的な指標を明確にしておく試みを行った。3章における実験で明らかになったことは、重要な基本広告データは、(a)インプレッション数、(b)クリック数、(e)1ページ目に必要な入札価格、(f)1ページ目上部に必要な入札価格、(h)品質インデックス、(j)合計コストの6種類ということである。しかし、3.3節で既に述べたように(a), (b), (j)については直接CVの達成へと繋がることは考えにくい。そこで、3.4節で行ったNMFを再度この6種類の基本広告データのみに対して行いより重要な特徴量を feature7, feature8の2つまで削減した。

その結果、feature7が品質インデックス、feature8が1ページ目に必要な入札価格と1ページ目上部に必要な入札価格を複合したものとなり(a), (b), (j)は重要な特徴量として算出

されなかった。この新たな2つの特徴量を各軸に設定しCVの有無ごとに色分けして散布図としてプロットしたものが図7である。すると、CV=1を表す赤点の割合がCV=0を表す青点に対して多い領域が2箇所見られた。

集合(A)は品質インデックスが大きい値を示すときに1ページ目に掲載するための入札価格が下がりCV達成が多く発生していることを、集合(B)は(A)に比べて品質インデックスが低いために1ページ目に掲載するための入札価格が高騰しているなかでCV達成が多く発生していることを示している。実際のWebリスティング広告では運用効率向上のため集合(A)のような状態を目標にするべきであるが、品質インデックスの値は広告主が任意で調整できる特徴量ではなくクリック率の変化によって変動する値だとされている[12]。そのクリック率をよくするためにはユーザーの注意を引きやすくするために作成した広告文が1ページ目及びその上部に掲載されることは必須であり、さらに競合している他社の広告文と差をつけるため、テキストレベルといった基本広告データ以外の要因の改善も必要とされる。

掲載順位が上位でありかつ品質インデックスの値が高い状態は、それに伴うCPCは下がっていくため広告運用に好循環が生まれる。そこに今回結果としてCV達成が成果として現れているものが集合(A)である。ここに分布する検索キーワードの特徴は、商品名の打ち間違いが多く見られた。一方、集合(B)は商品名の正式名称に多く見られたがインプレッション数も他の検索キーワードに比べ多いことからユーザーによる検索頻度の高いものとなっていると考えられる。このような場合、広告主同士が競合する可能性が高く、1ページ目およびその上部に掲載するための価格も高騰しがちだが、CV達成のためには積極的に入札することが有用であると集合(B)から窺える。

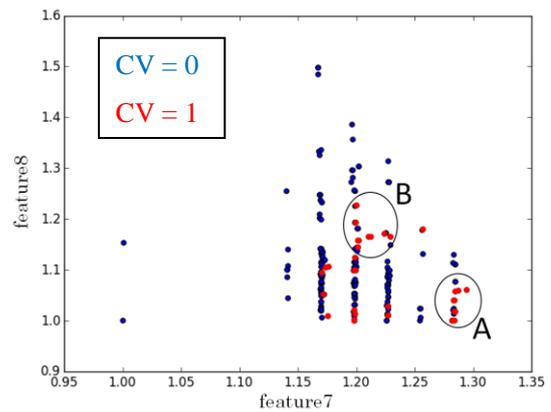


図6 6種類の基本広告データをNMFで二次元に変換した場合の散布図
: 集合(A)と集合(B)

この集合(A)と(B)を目標状態とし、それらを実際の広告運用で達成するための手法を図8と9にそれぞれ模式的に示した。どちらにおいても重要な事は広告出稿時に設定する入札価格を高め設定することである。これにより、まず集合(A)においては、掲載順位が上がる。すると、クリック率が高まるため品質インデックスも向上する。これにより、入札価格に対して実際の課金額が下がるので、より低い入札価格で1ページ目およびその上部への掲載が可能になる。さらに、広告文を改善しユーザーの注意を惹きやすく検索フレーズとともに関連性が高くなるようにすることで、クリック率と品質インデックスを向上させることが出来る。1ページ目及びその上部に必要な入札価格が下がれば、最初に高めに設定した入札価格に対しての割合が大きくなるためより掲載順位が上位に来る可能性が上がりCV=1へと繋がると考えられる。次に、集合(B)に対応する検索フレーズはユーザーによる検索頻度が高いため得られるインプレッション数は非常に大きい、対するクリック数の割合が少ないためにクリック率の向上は広告文の改善だけでは難しい。よって集合(A)と同様な循環になることは困難であると考えられるが、CVが発生しやすいキーワード群であるため初期段階からより積極的な入札価格の設定を行うことで掲載順位を上上げる必要があると考えられる。

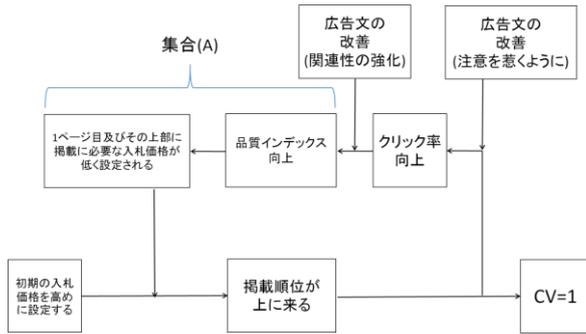


図 7 集合(A)を達成するための模式的な手法



図 8 集合(B)を達成するための模式的な手法

また、図 7 の別の注目すべき点として図 10 で示す集合(C)と(D)があげられる。集合(C)は低い品質インデックスに対して 1 ページ目及びその上部に必要な入札価格が低い箇所であり、その原因を各点ごとに追求することは困難であった。しかし、ここで CV=1 を示す赤点の検索フレーズは「ロコミ」という広告運用で特に効果的とされているものに偏っていた。集合(D)は 1 ページ目及びその上部に必要な入札価格が非常に高いが CV が発生していない箇所である。ここで見られる検索フレーズは広告運用で効果的とされている「効果」、「公式サイト」、「amazon」などに偏りを見せたが、これにはある一定以上のインプレッション数が見られることから設定した入札価格が低すぎたことが原因ではなく、この検索フレーズそのものがユーザーの嗜好性に沿っていなかったことが考えられる。

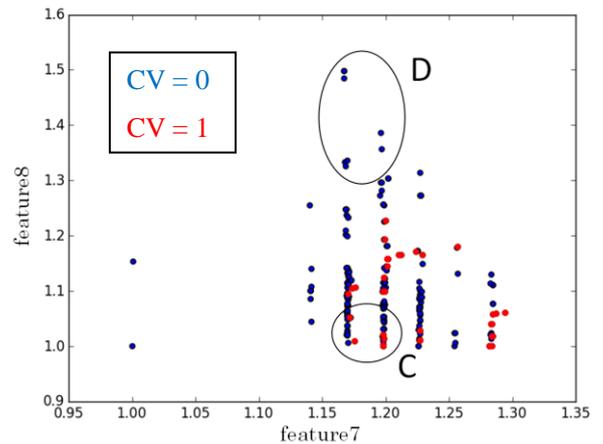


図 9 集合(C)と集合(D)

検索キーワードごとに割り当てられた10種類の基本広告データは同一キーワードであっても時期によって値が異なり CVも発生するときとそうでないときが見られる。これは広告掲載された商品に季節性の要素が含まれていたり、各種メディアや著名人による外部からの影響を強く受けたりと原因は様々であり、その特定を検索キーワードごとに行うのは難しい。また一時的なインプレッション数の増加によりクリック数が増えることだけがCVに繋がるとは限らないということも広告運用を行う中で分かっている。したがって、実際に CVにつながる定量的な要素は基本広告データ10種類の中でどの組み合わせに注目すべきかという観点において本研究で得られた知見は有用だと考えられる。

5.まとめと今後の課題

本研究では、Yahoo!プロモーション広告への個人の広告出稿において得られる10種類の基本広告データのなかで広告主に定量的な指標を明確に与えるデータについての議論を行った。結果、品質インデックスと1ページ目及びその上部に必要な入札価格が特に広告主が注目すべき特徴量だということが明らかになり、またその2つの散布図からCV=1となっている集合には検索キーワードの特徴に偏りがあった。さらに、実際の広告運用の観点から効果的にCVを達成するための具体的な手法をこれらの結果に基づき示した。

今後の課題として、まずデータ数や商品数を増やすことがあげられる。本稿ではある 1 つ商品に関する3ヶ月分のデー

タを解析したが、他の期間のデータや商品ごとにラベルを定義し他の商品も合わせればよりはっきりした結果が得られるかもしれない。また、本研究では4種類の機械学習の識別法を用いてCV達成の有無の二値分類を行ったが、検索キーワードごとのCVの達成数をカウントしたデータを追加できればCVR(Conversion Rate)の分析および予測も行える。

次に、検索キーワードや広告説明文の CV に対する影響の解析が挙げられる。これらに焦点を当てた研究はあまり前例がない。実際に広告運用を行った経験上、これはかなり複雑であると考えられる、例えば検索キーワードに含まれる1つの商品名をひらがなからカタカナに変更したり打ち間違った表記にするだけで途端に CV の有無が大きく変化するケースも頻繁に見られた。しかし、本稿で扱った基本広告データと検索キーワードなどのテキストレベルの特徴をうまく組み合わせることができれば、より良い CV 予測が可能となりより個人広告主が効率的で明確な運用を行えると考えられる。

参考文献

- [1] 宮西一徳, 高野雅典, 吉田岳彦. 大規模リワード広告システムにおける行動履歴と広告属性を利用したコンバージョン予測モデルの構築. 人工知能学会全国大会論文集, 2015, 29: 1-4.
- [2] 田頭幸浩, 小野真吾, & 田島玲. (2014). オンライン広告における CVR 予測モデルの素性評価. 第6回データ工学と情報マネジメントに関するフォーラム, DEIM.
- [3] 原淳史, 高野雅典, 川端貴幸. インターネット広告におけるコンバージョンに近いユーザの抽出方法の検討. 人工知能学会全国大会論文集, 2015, 29: 1-4.
- [4] TAGAMI, Yukihiro, et al. CTR prediction for contextual advertising: Learning-to-rank approach. In: Proceedings of the Seventh International Workshop on Data Mining for Online Advertising. ACM, 2013. p. 4.
- [5] MCMAHAN, H. Brendan, et al. Ad click prediction: a view from the trenches. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013. pp. 1222-1230.
- [6] 有坂昌樹. インターネットにおける広告戦略モデル, 2007, 東京大学大学院情報理工学系研究科修士論文
- [7] CHAPELLE, Olivier, MANAVOGLU, Eren, ROSALES, Romer. Simple and scalable response prediction for display advertising. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 5. 4: 61.
- [8] 櫻井祐子, 岩崎敦, 横尾真. エージェント適切な掲載数を決定するキーワード広告オークションプロトコルの提案. コンピュータ ソフトウェア, 2008, 25. 4: 60-67.
- [9] 山元浩平, et al. 一般画像認識手法を応用したディスプレイ広告のユーザ属性推定とクリック率予測. 電子情報通信学会技術研究報告 2015, 114. 454: 179-184.
- [10] 藤田篤, 幾島 克洋, & 佐藤 理史. (2011). 検索連動型広告の自動生成と集客効果の測定—飲食店ドメインを例題に. 情報処理学会論文誌, 52(6), 2031-2044.
- [11] 電通. 2011年日本の広告費(2011).
- [12] Yahoo!プロモーション広告. 公式ラーニングポータル.