Factoring Past Exposure in Display Advertising Targeting

Neha Gupta[‡], Abhimanyu Das[†], Sandeep Pandey[†], Vijay K. Narayanan[†]

‡ University of Maryland, College Park, MD 20742, USA
† Yahoo! Labs, Santa Clara, CA 95054, USA
neha@cs.umd.edu | {abhidas | spandey | vnarayan}@yahoo-inc.com

ABSTRACT

Online advertising is increasingly becoming more performance oriented, where the decision to show an advertisement to a user is made based on the user's propensity to respond to the ad in a positive manner (e.g., purchasing a product, subscribing to an email list, etc). The user response depends on how well the ad campaign matches the user's interest, as well as the amount of the user's past exposure to the campaign – a factor shown to be impactful in controlled experimental studies. Past exposure builds brand-awareness and familiarity with the user, which in turn leads to a higher propensity of the user to buy/convert on the ad impression. In this paper we propose a model of the user response to an ad campaign as a function of both interest match and past exposure, where the interest match is estimated using historical search/browse activities of the user.

The goal of this paper is two-fold. First, we demonstrate the role played by the user interest and the past exposure in modeling user response by jointly estimating the parameters of these factors. We test this response model over hundreds of real ad campaigns. Second, we use the findings from this joint model to identify more relevant target users for ad campaigns. In particular, we show that on real advertising data this joint model identifies better target users compared to conventional targeting models.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms

Algorithms, Experimentation

Keywords

Advertising, Targeting, User Modeling, Latent Factors

1. INTRODUCTION

Online advertising is growing at a rapid pace with more and more advertisers using the internet to reach out to their potential customers. This is not surprising given that users are spending an unprecedented amount of time performing online activities ranging from sending emails, searching and browsing, to dating and shopping. The advertisers (or intermediaries such as ad networks, online exchanges) use these past online activities of the users to identify those who are most likely to respond positively to their advertising campaigns. This is called behavioral targeting [15, 12, 7, 19], whereby models are built based on past user behavior to target likely potential customers. Behavioral targeting benefits both the parties involved: it allows advertisers to spend their money effectively, while it helps users by providing them more relevant ads.

Effective targeting requires deeper understanding of the factors influencing a user response to an ad. Broadly speaking, for a given advertising campaign this involves: (a) understanding user interests and identifying whether she is likely to be interested in the products/services offered by the advertiser (called "interest match"), and (b) making the user "aware" of the advertiser's product by repetitively exposing her to the ad so as to elicit a positive response (called "past exposure"). Previous work has shown how each of these factors can influence the likelihood of a user to pursue the ad. While the effect of the "interest match" factor is fairly natural and intuitive (e.g., the more relevant the ad campaign is to the user, the more likely she is to pursue it), the effect of past exposure is more complex. Controlled experiments have been done by economists and social scientists to identify the effect of past exposure [16, 13, 9]. While the results from these studies differ in some aspects, largely it has been shown that with increased past exposure the user is more likely to pursue the ad. Intuitively, "past exposure" to an ad might help in several ways such as increased brand-awareness and familiarity with the advertised product, or even an increased probability of the user to notice the ad. At the same time, some studies have also shown an "ad fatigue" effect where users might tire of an ad if it is displayed too often [1].

Our goal in this paper is to improve user targeting by accounting for the hidden interplay of the interest match and past exposure factors in a real world advertising environment. This is challenging for several reasons. First, when we receive a positive response from a user for an ad campaign, it is unclear how much of this is due to repeated past exposure of the ad, as opposed to the user interests matching with the campaign objectives. Most previous work has been limited to investigating these factors individually in a controlled experimental setting. Second, these past exposure and interest match factors themselves are complex and difficult to characterize precisely. For example, as discussed in [17], inferring user interests from their profiles requires dealing with millions of features and is challenging technically as well as computationally. Third, the contribution of these factors to a positive user response might be different for different advertising campaigns. For exam-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'12, August 12–16, 2012, Beijing, China.

Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$15.00.

ple, past exposure might play a positive role for certain campaigns, be unrelated to user response for other campaigns, and might even lead to ad fatigue for a third set of campaigns. Lastly, in our study we define a positive user response in terms of "conversions", which represent desired user actions as specified by the advertiser in the form of purchases and product information request. While conversions are advertiser-defined and thus are clearly more tangible indicators of user interest compare to clicks, time spent or other implicit feedback, conversions are much rarer as well. Thus, we have to take necessary precautions while performing inference or learning from such sparse data.

We deal with the above mentioned issues by using mathematical models that jointly account for the two factors involved: past exposure and interest match. Our models use three sources of information for a given user and advertising campaign: (a) the number of previous ads that the user has seen from the campaign in consideration, (b) a rich user profile in terms of her past online activities that include, but are not limited to, search queries issued, pages browsed and the associated timestamps and (c) users who have responded well to the campaign in the past (called positive users). The number of previously viewed ads is used to model the past exposure factor, while the user profile is leveraged for gauging interest match with the advertiser. By jointly accounting and optimizing for past exposure and interest match with respect to the given positive and negative users, we are able to separate their influence [16, 13, 9]. Intuitively speaking, given a set of users with similar interest match but with different past exposure, the role of past exposure can be estimated and vice versa. However, doing so makes the model for characterizing interest match to depend on the given model for past exposure and vice versa, thus requiring a joint optimization.

Our work can have significant implications for display advertising and user targeting. We show how the two factors, past exposure and interest match, can be modeled jointly to perform more accurate prediction of potential converters. Current state of the art in targeting is largely focused on matching interested users with advertisers, but does not give deserved attention to past exposure. Through our experiments on hundreds of real advertising campaigns, we show how past exposure, along with interest match, can substantially improve the targeting performance. Also, understanding the effect of past exposure allows advertisers to allocate and spend their advertising budget more judiciously. For example, if their campaigns require large amount of past exposure, then it makes sense for them to find interested users and pursue them for a reasonable period before expecting a positive response. On the other hand, advertisers who only need small past exposure should shuffle through users quickly and avoid spending much resources on any particular user (since if the user is likely to convert, she would convert promptly for such advertisers). On our dataset we find, through our modeling approach, that different campaigns have sufficiently different needs in terms of past exposure and interest match.

Contributions. We make the following contributions in this paper:

- We study the effect of past exposure and interest match in terms of eliciting positive user response in a real world advertising environment. To the best of our knowledge, this is the first study that investigates the hidden interplay of the two factors in a non-controlled setting.
- We propose mathematical models that jointly account for both the factors. We give several variants to capture different kinds of advertising campaigns and user conversion models.
- Using several sources of data (involving user profiles and ad campaigns), we validate our models and surface the role played by past exposure for different campaigns.
- We use our models to show how past exposure, along with interest match, can help in significantly improving the targeting systems. From empirical analysis on 200 advertising campaigns, we show substantial improvement over the conventional targeting methods.

1.1 Related Work

Behavioral targeting systems for display advertising are rapidly increasing on the Internet. Utilizing historical online user information to target users with display ads has been shown to be highly effective, resulting in an increase in the performance of ad-campaigns [19]. Several optimization techniques have been suggested [15, 12, 7] which use different user features such as demographics, past views, past searches, pseudo social-networks, etc. to differentiate between converters and non-converters.

Different statistical models can be applied to capture the user behavior effectively. Regression models have been introduced in [3, 4, 10] in this domain. Bayesian factor-based models [2, 6] have also been proposed in literature. In [2] the authors used *ad factors* whose parameters are derived from different types of Markov models to generalize user behavior patterns for ad-campaigns. The goal of this study was on providing additional insight about the users to the advertisers. In [6] the authors use latent factor models similar to LDA for targeting [5], but do not consider the effect of past user exposure. Also, there is related work on user browsing models (UBM) which, for organic and sponsored search, separates the click-through rate into an "examination" probability and perceived relevance probability. However, these models apply for search results and do not consider any effect due to repeated exposure [18].

The effect of user exposure has been studied in [9, 13, 16] in a controlled setting, where it has been shown that multiple exposure to an advertisement can have varying effects on the purchasing decision of the users. In [16] the authors study the effect of temporal spacing between ads and show that at a purchase occasion, the probability of a product purchase increases if its past ads are spread apart rather than bunched together. In [13] the authors show that the purchase probability of a user varies as a function of banner advertising exposure. They also show that the number of websites and number of pages on which a user is exposed to advertisements can have an effect on the purchase probabilities.

2. TARGETING WITH PAST EXPOSURE AND INTEREST MATCH

Before we explain our modeling approach, we next describe the problem setup in detail.

2.1 Problem Formulation

The focus of our study is *performance based* display advertising campaigns, which are set up with conversions goals, where conversions are advertiser specified actions (e.g., email subscription, form fill, product purchase) that show positive user intent. For each campaign, the goal is to identify and target users who are most likely to convert, also known as *converters*. In the following section, we formulate this task mathematically.

We use the following sources of information for a given campaign:

- Past ad impressions x_i: the number of previous ads that user i, denoted by u_i, has seen from the campaign.

- User profile vector $\mathbf{f}_i$: representing the past online activities from user u_i. These activities include page visits and search queries, which are analyzed using established text processing techniques to construct a feature vector. More details are given in Section 4.1.

- Seed users $\mathcal{S}$: users who have been targeted for the campaign in the past along with their response values. In other words, for these seed users, we know whether they converted on the ad campaign or not.

We denote the user response using binary variable y_i. A y_i value of 1 represents a converter, and 0 represents a non-converter.

Let us denote the probability of user u_i converting by the time she has been served x_i ad impressions by $\mathrm{P}[y_i = 1|\mathbf{f}_i, x_i]$. Then the goal of ad targeting is to identify users with the largest values of $\mathrm{P}[y_i = 1|\mathbf{f}_i, x_i]$. There are several technical questions involved in doing so. For instance, it is worth investigating the gain achieved by accounting for the number of adviews x_i versus ignoring it (see Section 2.2). How does the user profile vector $\mathbf{f}_i$ interact with x_i? As described in Section 2.4, a naive solution is to treat x_i as another feature in the profile vector, but this leads to over-simplification and does not model the effect of x_i correctly. Note that, in our setting, the dimension of $\mathbf{f}_i$ is extremely large since it represents sparse user activities. Furthermore, the number of positive examples (i.e., y_i =1) in the given seed set $\mathcal{S}$ is typically small since conversions are rare events, making the problem harder.

Next we describe our proposed factor model for user targeting that accounts for both past exposure and interest match.

2.2 Proposed Factor Model (EFM)

We believe that the user conversion for a campaign comes from a two-stage process. In the first stage, the user is repeatedly exposed to the ad to get her attention. According to past work, this may lead to increased user *awareness* about the product/brand, making her more familiar with the ad offering and increasing her trust in the advertisement. Of course, an excess of repeated exposure to the ad can negatively affect the user sentiments about the advertiser. Once the user has become aware, denoted by the *awareness* random variable ϕ_i (defined precisely in Section 2.2.1), she moves to the second stage.

In the second stage, the user evaluates the offering and decides whether to pursue the ad or not. This decision is based on several other variables such as whether the user is interested in the offered product, whether the timing and price is appropriate,etc. We call this the *intrinsic conversion* probability and denote it by $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$. Thus, our model factorizes the probability of user conversion into the product of two probabilities, the awareness probability ($\mathrm{P}[\phi_i = 1|x_i]$) and intrinsic conversion probability ($\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$). In other words,

$$\begin{aligned}\mathrm{P}[y_i = 1|\mathbf{f}_i, x_i] &= \mathrm{P}[\phi_i = 1|x_i] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1] + \\ &\quad \mathrm{P}[\phi_i = 0|x_i] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 0] \\ &= \mathrm{P}[\phi_i = 1|x_i] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]\end{aligned}$$

where the last equation arises since the user cannot convert unless she is aware of the ad.

Note that our model decouples the effect of previous ad impressions x_i and profile vector $\mathbf{f}_i$, whereby x_i influences the awareness probability while the profile vector is used for modeling the intrinsic conversion probability. While we can envision the awareness probability to be a function of the profile vector as well (i.e., different users may build awareness in different manners) and the intrinsic probability to depend on the number of impressions, this could make the two stages fairly similar and mathematically unidentifiable. Hence, for the ease of model interpretation we keep the two stages decoupled as described above.

We refer to this model as the *Exposure-based Factor Model* for User Targeting (EFM). Next we describe the two factors in more detail.

2.2.1 Modeling Awareness Probability

For a given advertising campaign and a user, the awareness probability captures the likelihood with which the user is aware of the campaign and considers it for evaluation (i.e., moves to the second stage). We represent this as a function of the number of previous ad impressions x_i of the campaign that the user has been exposed to.

Below we describe two possible distributions that can be used to model the awareness probability.

1. *Geometric Process*: Let us assume that each time the user is shown the ad, she notices the ad with probability α. If each of these Bernoulli trials are independent, then the probability of the user becoming aware by noticing the ad on the i^{th} impression is a geometric distribution. In other words, $\mathrm{P}[\phi_i = 1|x_i] = \sum_{j=1}^{x_i}(1-\alpha)^{j-1}\alpha = 1 - (1-\alpha)^{x_i}$. We denote this EFM model that uses this Geometric process for modeling awareness probability as EFMG.

2. *Poisson Process*: Here we model the number of times a user notices the ad as a Poisson process, with an average rate of $\lambda = \alpha \cdot x_i$ (where α is the probability of noticing an ad impression). Under the Poisson process, the probability of k events is $\frac{(\alpha \cdot x_i)^k e^{-\alpha \cdot x_i}}{k!}$. Hence, the probability of the user noticing the ad at least once and being aware of it is $\mathrm{P}[\phi_i = 1|x_i] = 1 - e^{-\alpha \cdot x_i}$.

   We denote this model by EFMP.

Both the above models are intuitive in nature and capture the fact that with increased exposure, the awareness probability increases (as found in previous studies). In other words, the more ads the user is exposed to, the higher is her awareness probability. While this is largely true, it is possible that with excess exposure the user gets annoyed and does not consider the ad for evaluation (i.e., the second stage). More specifically, the awareness may decrease due to negative sentiments.

We account for such negative sentiments by proposing a *Multinomial* distribution for awareness probability where $\mathrm{P}[\phi_i = 1|x_i] = \{\pi_1, \pi_2, \ldots\}$. Here π_x denotes the awareness probability after x number of previous ad impressions and is learned from the data during joint optimization (parameter estimation is described in Section 3). Hence, if there is excess exposure which leads to negative sentiments, π_x would start decreasing beyond a certain number of ad impressions (x).

2.2.2 Modeling Intrinsic Conversion Probability

As describe earlier, the intrinsic conversion probability for user u_i represents the likelihood with which the user converts on the ad when she evaluates it. The probability depends on many user attributes such as demographics, her interests (short-term and long-term), online shopping tendency. We model this using a logistic function of the features from the user profile vector:

$$\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1] = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}}$$

In other words, $logit(\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]) = \mathbf{w}^T\mathbf{f}_i$ where $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$ denotes the unknown weight vector. The weight vector is campaign specific since it models the targeting constraints of a campaign and needs to be learned separately for each campaign.

Note that the choice of logistic regression for modeling intrinsic conversion probability is not central to our factor model and can be easily substituted by other approaches. Methods for identifying user interests is a research topic by itself and has been focus of many previous work on advertising and other applications such as news browsing, personalized search, etc (see related work). Many models such as Logistic regression, SVM, Naive-Bayes, Nearest-neighbor have been studied in this context and can potentially be used for our work.

2.3 Incorporating Delay Factor in the Model

In our factor model we described how the user converts on a campaign in a two step process — in the first step the user becomes aware of the ad and the advertiser, while in the second step she evaluates the ad and decides whether to convert on it or not. Another dimension that is worth investigating is the time delay between the first and the second step. For example, say the user is exposed to the ads from a cruise line and after a few exposures she becomes aware of it. Since she is not planning any vacation for the next few months, she decides to not react or evaluate the ad at the time. But when she finally gets to the vacation period, the awareness might have faded away. We make this more formal next.

Suppose the user has been targeted with a display advertisement for the campaign x_i times so far. Let the timestamp for the x-th adview for this user be $t_i(x)$, and suppose we know that the user has converted by time θ_i.[1] Let us denote the probability of this conversion event by $\mathrm{P}[y_i = 1|\mathbf{f}_i, x_i, \theta_i]$. Our generative model for the conversion is as follows: first, the user becomes aware at the time of the x-th adview (where $1 \leq x \leq x_i$). After that she waits for a time interval t (where $0 \leq t \leq \theta_i$), before deciding whether to convert or not convert. In other words, the probability of conversion is the product of the following three terms - the probability of the user becoming aware at the x-th adview (denoted by $\mathrm{P}[\phi_{i,x} = 1]$),[2] the probability of the user incurring a delay of time t in making her decision (denoted by $\mathrm{P}[\Delta_i = t]$), and the intrinsic conversion probability of the user given that she has noticed the ad and is ready to make a decision (denoted by $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$). Thus, we have

$$\begin{aligned} \mathrm{P}[y_i = 1|\mathbf{f}_i, x_i, \theta_i] &= \sum_{x=1}^{x_i} \sum_{t=0}^{\theta_i - t_i(x)} (\mathrm{P}[\phi_{i,x} = 1] \\ &\quad \cdot \mathrm{P}[\Delta_i = t] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]) \end{aligned}$$

Assuming that the impressions are shown to users uniformly over time, we can approximate the time interval $\theta_i - t_i(x)$ in terms of the difference in the ad impression counts, that is, $\theta_i - t_i(x)$ is proportional to $x_i - x$. Hence,

$$\begin{aligned} \mathrm{P}[y_i = 1|\mathbf{f}_i, x_i] &= \sum_{x=1}^{x_i} \sum_{t=0}^{x_i - x} (\mathrm{P}[\phi_{i,x} = 1] \\ &\quad \cdot \mathrm{P}[\Delta_i = t] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]) \\ &= \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1] \cdot \\ &\quad \sum_{x=1}^{x_i} \sum_{t=0}^{x_i - x} (\mathrm{P}[\phi_{i,x} = 1] \cdot \mathrm{P}[\Delta_i = t]) \end{aligned}$$

If it is the case that the user, after becoming aware, immediately decides whether to convert (without waiting for future ad impressions), i.e., $\mathrm{P}[\Delta_i = 0] = 1$, then $\mathrm{P}[y_i = 1|\mathbf{f}_i, x_i]$ simplifies to $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1] \cdot \mathrm{P}[\phi_i = 1|x]$. Thus, we obtain the model described in Section 2.2 as a special case of this delay model.

More generally, if the probability $\mathrm{P}[\Delta_i = t]$ decays obeys an exponentially decaying probability distribution, i.e., $\mathrm{P}[\Delta_i = t] = \gamma e^{-\gamma t}$, where γ is the exponential parameter, then we get:

$$\begin{aligned} \mathrm{P}[y_i = 1|\mathbf{f}_i, x_i] &= \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1] \cdot \\ &\quad \sum_{x=1}^{x_i} \sum_{t=0}^{x_i - x} (\mathrm{P}[\phi_{i,x} = 1] \cdot \gamma e^{-\gamma t}) \end{aligned}$$

We denote this model by EFMD.

2.4 Alternative Approaches for User Targeting

We start by describing the conventional targeting strategy which does not take exposure into account [3, 4]. In other words, this approach only accounts for the intrinsic conversion probability, $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$. The probability is typically modeled using a SVM or Logistic function of the user features including geographic, demographic, behavioral and/or social attributes (see Section 4.1 for more details about the type of features used). We refer to this baseline model by LR (for logistic regression). Note that this is a special case of our factor model from Section 2.2 where the bernoulli probability α is set to 1.

We propose another baseline method which takes exposure into account but in a rather simplistic manner. In particular, we can think of the past impression count x_i as another feature in the profile vector $\mathbf{f}_i$. Hence, instead of doing logistic regression over just the profile features as in LR, now we regress over both x_i and $\mathbf{f}_i$. We denote this by LR+ and study its performance in the experiment section. As we will see in the results section, while leveraging past exposure benefits LR+ over LR, it is substantially outperformed by our factor models.

3. PARAMETER ESTIMATION FOR FACTOR MODEL

We now present optimization methods for estimating the model parameters (α and $\mathbf{w}$) for our factor model. Note that the parameters α and $\mathbf{w}$ are specific to each campaign and are estimated separately.

We provide two different optimization approaches for this problem: an Alternate Maximization method and an Expectation Maximization method.

3.1 Alternate Maximization

For the ease of explanation, we consider the zero-delay factor model from Section 2.2. Also, let us assume $\mathrm{P}[\phi_i = 1|x]$ to be a

[1] Note that the precise time of a user conversion is difficult to determine in practice since the conversion, unlike click, does not happen right after the ad impression; the user may evaluate the ad, like it and then buy the product few days later.

[2] Note that awareness probability $\mathrm{P}[\phi_i = 1|x] = \sum_{k=1}^{x} \mathrm{P}[\phi_{i,k} = 1]$. For the Geometric model described in Section 2.2.1, $\mathrm{P}[\phi_{i,x} = 1] = (1-\alpha)^{x-1}\alpha$.

Geometric process with parameter α and $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1] = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{f}_i}}$.

Given a set of users $U = \{u_i, \ldots, u_n\}$ along with corresponding labels $Y = \{y_i, \ldots, y_n\}$, feature vectors $F = \{\mathbf{f}_i, \ldots, \mathbf{f}_n\}$ and past ad impression counts $X = \{x_i, \ldots, x_n\}$, our goal is to find the unknown logistic parameters $\mathbf{w}$ and the Geometric Parameter α. Using $\Theta = \{\mathbf{w}, \alpha\}$ to represent the model parameters, $D = \{Y, F, X\}$ to represent the observed data, and using the expression for $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$ from Section 2.2, we can express the log-likelihood function for the data as:

$$\begin{aligned}
L[\Theta; D] &= \sum_{i=1}^{n}(y_i \log(\mathrm{P}[\phi_i = 1|x_i] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]) + \\
&\quad (1 - y_i)\log(1 - \mathrm{P}[\phi_i = 1|x_i] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]) \\
&= \sum_{i=1}^{n}(y_i \log((1 - (1 - \alpha)^{x_i}) \cdot \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}}) + \\
&\quad (1 - y_i)\log(1 - (1 - (1 - \alpha)^{x_i}) \cdot \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}}))
\end{aligned}$$

To solve this maximum-likelihood optimization problem, we use an alternating maximization algorithm consisting of a sequence of iterations, each comprising of two steps: in the first step (α-step), we optimize for α while keeping $\mathbf{w}$ constant, and in the second step ($\mathbf{w}$-step), we optimize for $\mathbf{w}$ while keeping α constant.

- α-step: In this step, the value of $\mathbf{w}$ (and hence $\frac{1}{1+e^{-\mathbf{w}^T\mathbf{f}_i}}$) is kept constant. We denote these constant terms $\frac{1}{1+e^{-\mathbf{w}^T\mathbf{f}_i}}$ by c_i. The optimization step is therefore:

$$\begin{aligned}
\underset{\alpha}{\operatorname{argmax}}\, g(\alpha) &= \sum_{i=1}^{n}(y_i \log((1 - (1 - \alpha)^{x_i}) \cdot c_i) + \\
&\quad (1 - y_i)\log(1 - (1 - (1 - \alpha)^{x_i}) \cdot c_i)).
\end{aligned}$$

  While the above equation does not lead to a closed form solution, maximizing $g(\alpha)$ can be efficiently solved using grid search since it contains only one unknown $\alpha \in [0, 1]$.

- $\mathbf{w}$-step: In this step, we use the value of α obtained from the previous step. Hence, the $1 - (1 - \alpha)^{x_i}$ terms, which we denote by a_i, are kept constant. The optimization problem is then:

$$\begin{aligned}
\underset{\mathbf{w}}{\operatorname{argmax}}\, f(\mathbf{w}) &= \sum_{i=1}^{n}(y_i \log(a_i \cdot \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}}) + \\
&\quad (1 - y_i)\log(1 - (a_i \cdot \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}}))).
\end{aligned}$$

While the above function $f(\mathbf{w})$ is not concave, we can approximate it by another function that is concave:

$$\begin{aligned}
f'(\mathbf{w}) &= \sum_{i=1}^{n}(y_i \log(a_i \cdot \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}}) + \\
&\quad (1 - y_i)\log(1 - \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{f}_i}})^{a_i}).
\end{aligned}$$

Essentially, we are approximating $(1 - \mathrm{P}[\phi_i = 1|x_i] \cdot \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1])$ by $(1 - \mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1])^{\mathrm{P}[\phi_i = 1|x_i]}$ which is known to be a good approximation if $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$ is small. Since this approximation is performed only in the case of those users for which $y_i = 0$, we expect $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]$ to indeed be small for such users.

We can now solve the optimization problem $\operatorname{argmax}_{\mathbf{w}} f'(\mathbf{w})$ efficiently, since $f'(\mathbf{w})$ is a concave function of $\mathbf{w}$. Specifically, we performed L-BFGS gradient descent using the MALLET package [14].

LEMMA 3.1. *$f'(\mathbf{w})$ is a concave function of* $\mathbf{w}$.

Proof. Let $l_i(w) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{f}_i}}$. Hence

$$\begin{aligned}
f'(\mathbf{w}) &= \sum_{i=1}^{n}(y_i \log(a_i \cdot l_i(w)) + (1 - y_i)\log(1 - l_i(w))^{a_i}) \\
&= \sum_{i=1}^{n}(y_i \log(a_i \cdot l_i(w)) + (a_i - a_i y_i)\log(1 - l_i(w))) \\
&= \sum_{i=1}^{n}(y_i(1 - a_i)\log(l_i(w))) + \\
&\quad \sum_{i=1}^{n}(a_i(y_i \log l_i + (1 - y_i)\log(1 - l_i(w))))
\end{aligned}$$

The second term is the Logistic Regression log-likelihood function, which is concave. For the first term, it suffices to show that $-\log l_i(w)$ is convex.

Now $\frac{\partial(-\log l_i(w))}{\partial w_p} = -x_p(1 - l_i(w))$. Therefore $\frac{\partial^2(-\log l_i(w))}{\partial w_p^2} = x_p^2 l_i(w)(1 - l_i(w))$, and $\frac{\partial^2(-\log l_i(w))}{\partial w_p w_q} = x_p x_q l_i(w)(1 - l_i(w))$.

Thus, the Hessian $\bigtriangledown(-\log l_i(w)) = l_i(w)(1 - l_i(w))(\mathbf{x}\mathbf{x}^T)$ where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$. This is clearly a positive-semidefinite matrix, and hence $-\log l_i(w)$ is convex.

Hence, $f'(\mathbf{w})$ is concave. ■

3.2 Expectation-Maximization

We can also perform model estimation using the Expectation Maximization framework. We use the same definitions for U, Y, F, Θ and X as before. In addition, for each user u_i we have an associated hidden (or latent) variable z_i that takes a value of 1 if the user is aware of the ad (with probability $\mathrm{P}[\phi_i = 1|x_i]$), and 0 otherwise. We define $\mathbf{Z} = \{z_1, z_2, \ldots, z_n\}$. For ease of notation, we denote $\mathrm{P}[y_i = 1|\mathbf{f}_i, \phi_i = 1]) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{f}_i}}$ by p_i, and $\mathrm{P}[\phi_i = 1|x_i]$ by q_i.

Given the hidden variables $Z = \{z_i\}$ and data $D = \{Y, F, X\}$, we first compute the log-likelihood of the model $L[\Theta; D, Z]$. For each user, clearly $\mathrm{P}[y_i = 1|z_i = 1] = q_i p_i$, $\mathrm{P}[y_i = 0|z_i = 1] = q_i(1 - p_i)$, $\mathrm{P}[y_i = 1|z_i = 0] = 0$ and $\mathrm{P}[y_i = 0|z_i = 0] = 1 - q_i$. Thus,

$$\begin{aligned}
L[\Theta; D, Z] &= \sum_{i=1}^{n} y_i \log(z_i q_i p_i) + \\
&\quad (1 - y_i)\log(z_i q_i(1 - p_i) + (1 - z_i)(1 - q_i)) \\
&= \sum_{i=1}^{n} y_i \log(G_i) + (1 - y_i)\log(H_i),
\end{aligned}$$

where we use $G_i = z_i q_i p_i$ and $H_i = z_i q_i(1 - p_i) + (1 - z_i)(1 - q_i)$.

Then, in the E-step we consider the expected value of the likelihood w.r.t. the conditional distribution of Z given the data and model parameters, i.e., $E_{Z/\Theta,D}[L[\Theta; D, Z]]$. Thus,

$$\begin{aligned}
&E_{Z/\Theta,D}[L[\Theta; D, Z]] = \textstyle\sum_{\mathbf{Z}} \mathrm{P}[Z|\Theta, D] \cdot L[\Theta; D, Z] \\
&= \textstyle\sum_{\mathbf{Z}} \mathrm{P}[Z|\Theta, D] \cdot \sum_{i=1}^{n} (y_i \log(G_i) + (1 - y_i) \log(H_i)) \\
&= \textstyle\sum_{i:y_i=1} \sum_{z_i} \mathrm{P}[z_i|\Theta, y_i = 1] \log(G_i) \\
&\quad + \textstyle\sum_{i:y_i=0} \sum_{z_i} \mathrm{P}[z_i|\Theta, y_i = 0] \log(H_i) \\
&= \textstyle\sum_{i:y_i=1} \log(q_i p_i) + \sum_{i:y_i=0} \mathrm{P}[z_i = 1|\Theta, y_i = 0] \log(q_i(1 - p_i)) \\
&\quad + \mathrm{P}[z_i = 0|\Theta, y_i = 0] \log(1 - q_i)
\end{aligned}$$

where, the last equality follows since $\mathrm{P}[z_i = 1|\Theta, y_i = 1] = 1$ and $\mathrm{P}[z_i = 0|\Theta, y_i = 1] = 0$. Also, we have

$$\mathrm{P}[z_i = 1|\Theta, y_i = 0] = \tfrac{q_i \cdot (1-p_i)}{q_i \cdot (1-p_i) + (1-q_i) \cdot 1} = \tfrac{q_i \cdot (1-p_i)}{1 - q_i \cdot p_i}.$$

Thus, we can further simplify:

$$\begin{aligned}
E_{Z/\Theta,D}[L[\Theta; D, Z]] &= \sum_{i:y_i=1} \log(q_i p_i) \\
&\quad + \sum_{i:y_i=0} \frac{q_i \cdot (1 - p_i)}{1 - q_i \cdot p_i} \log(q_i(1 - p_i)) \\
&\quad + \sum_{i:y_i=0} (1 - \frac{q_i \cdot (1 - p_i)}{1 - q_i \cdot p_i}) \log(1 - q_i) \\
&= \sum_{i:y_i=1} \log(q_i p_i) + \sum_{i:y_i=0} \log(1 - q_i) \\
&\quad + \sum_{i:y_i=0} \frac{q_i \cdot (1 - p_i)}{1 - q_i \cdot p_i} \cdot \log(q_i \frac{1 - p_i}{1 - q_i})
\end{aligned}$$

Thus we obtain the **E-step** expression:

$$\begin{aligned}
E_{Z/\Theta,D}[L[\Theta; D, Z]] &= \sum_{i:y_i=1} \log(q_i p_i) + \sum_{i:y_i=0} \log(1 - q_i) \\
&\quad + \sum_{i:y_i=0} \frac{q_i \cdot (1 - p_i)}{1 - q_i \cdot p_i} \cdot \log(q_i \frac{1 - p_i}{1 - q_i})
\end{aligned}$$

In the M-step, we aim to estimate the model parameters α and $\mathbf{w}$ that maximizes the above expression. Notice that p_i is a function of $\mathbf{w}$ and q_i is a function of α. Hence, in the M-step we can again use an alternating maximization algorithm consisting of a sequence of iterations, each comprising two steps: in the first step (α-step), we optimize for α while keeping $\mathbf{w}$ constant, and in the second step ($\mathbf{w}$-step), we optimize for $\mathbf{w}$, keeping α constant.

- α-step: In this step, the value of $\mathbf{w}$ (and hence p_i) is kept constant. Then the above expression can be efficiently maximized using grid search over the unknown parameter $\alpha \in [0, 1]$.

- $\mathbf{w}$-step: In this step, we use the value of α obtained from the previous step. Hence, the $q_i = 1 - (1 - \alpha)^{x_i}$ terms are kept constant. The optimization problem is then

$$\begin{aligned}
\underset{\mathbf{w}}{\operatorname{argmax}} \quad &\{ \sum_{i:y_i=1} \log(p_i) \\
&+ \sum_{i:y_i=0} \frac{q_i \cdot (1 - p_i)}{1 - q_i \cdot p_i} \cdot \log(q_i \frac{1 - p_i}{1 - q_i}) \}
\end{aligned}$$

While the above function $f(\mathbf{w})$ is not concave, we use a gradient descent approach to reach a local maxima and obtain a reasonable estimate of the $\mathbf{w}$ parameter, by running the gradient descent solver using MALLET with multiple starting points [14].

4. EXPERIMENTS

In this section we describe the experimental setup and the data used to test our user targeting models proposed in Section 2.1.

4.1 Data Description

Our goal in these experiments is to build a conversion model for online display advertisement campaigns. Essentially, this amounts to training a discriminative classifier for identifying potential converters versus non-converters, using a training dataset comprising of previous converters and non-converters (similar to [3, 4]). We collect 4 weeks of advertisement data for a set of 200 campaigns from a historical time period. The campaigns were chosen in a non-uniform manner to ensure adequate representation of campaigns across the full range of conversion volumes, where the conversion volume of a campaign denotes the number of conversions obtained by the campaign over a fixed time period (more details below).[3] The data consists of the online activity of users (such as page views, search queries) over a period of 4 weeks, along with information about whether a user converted on a campaign or not, and how many times (within the 4 week period) that the user was exposed to the ads for that campaign.

The activities of the users are a sequence of events collected from server logs. The user's feature vector comprises both raw and categorized event counts for that user for the following types of events:

- Pages visited: Features include a unique identifier of the page and the category of the page derived using an existing hierarchical page categorizer into a category taxonomy.

- Search queries: Searches issued, clicks on search links, clicks on search advertising links. Features include the category of the query, the click information and the unigrams/bigrams in the query.

For each campaign, we therefore have a dataset of users (u_i) along with their user feature vectors ($\mathbf{f}_i$) , the number of times the user was targeted with the campaign ad (x_i), and whether the user converted or not (y_i).

Since the total dimension of the above feature space can be very high (of the order of millions), we use a Mutual-Information filter to select the top $30,000$ user features in each campaign for the subsequent modeling experiments. We further down-sample the total number of users (i.e., training and test instances) used by our models to around $100,000$ users per campaign.

4.2 Evaluation Methodology

For a thorough empirical analysis of our models, we run our experiments over the entire set of $\sim$ 200 display advertising campaigns described in subsection 4.1, using an efficient Map-Reduce implementation [8]. We bin each campaign into one of 6 buckets (based on the conversion volume in the campaign). Instead of showing performance over individual campaigns, to avoid cluttering we present the average performance of models for campaigns in each bucket. The buckets correspond to an increasing log-scale in the number of conversions, and span a few orders of magnitude in conversion volume. Figure 1 plots the histogram of the number of campaigns in each bucket.

[3] We note here that users that opt out of targeting are not profiled and are therefore not included in the experiments or in the actual campaigns.

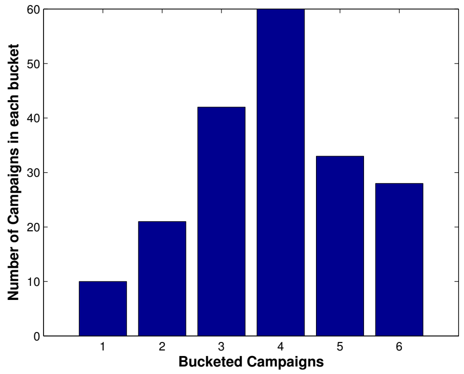

Figure 1: Number of campaigns in each conversion-volume bucket (the buckets correspond to an increasing log-scale in terms of the number of conversions).

In all our experiments, 66% of the data from each campaign is used for training the models, and the remaining 34% is used for testing, with 2-fold cross-validation. We train and evaluate the performance of the models independently for each campaign, and also estimate the average performance across all the 200 campaigns.

We use the area under the Receiver Operating Characteristic (ROC) curve to evaluate the ranked list of users produced by the different targeting models. The Area Under Curve (AUC) gives the probability that the targeting model assigns a higher score to a random positive example than a random negative example (i.e., probability of concordance) [11]. So, a purely random selection method will have an area under the curve of exactly 0.5. An algorithm that achieves AUC of 0.6 can distinguish a positive user from a negative user with 60% probability, and is thus 20% better than a random method.

An alternative metric could be to measure precision/recall at a certain rank in the list. Note that different campaigns may have different requirements in terms of precision and recall. For example, a small campaign whose reach is limited would prefer higher recall, while a large campaign that reaches out to many users might prefer higher precision. Consequently, selecting a rank at which to evaluate precision such that it would be suitable for all campaigns, is not possible. Hence, we use the AUC as a performance metric since it summarizes the prediction performance over all ranks in a single number.

4.3 Results

In this section, we show and discuss the AUC performance of the various models described in Section 2.1. This includes the baseline approaches such as logistic regression without and with the number of past impressions as a feature (denoted by LR and LR+ respectively). Our model variants include the geometric factor model (EFMG), the poisson factor model (EFMP) and the delayed factor model (EFMD) with alternate maximization being used for model inferencing. Lastly, EFMG_{EM} is the geometric model learned using the EM based inference method. We do not report the multinomial model from Section 2.1 since it did not perform well (primarily due to too many parameters being estimated). Table 1 provides a high-level overview of the average AUC performance of each of these methods over the full set of 200 campaigns. In the remainder of this section, we discuss the results of each of these methods separately, and also analyze their performance and model parameters for each stratified bucket of campaigns.

LR	LR+	EFMG	EFMP	EFMD	EFMG_{EM}
0.67	0.687	0.716	0.699	0.717	0.675

Table 1: Average AUC for the various models over the full set of 200 campaigns.

4.3.1 *Using Past Exposure as a feature*

We first obtain a baseline by using the Logistic Regression (LR) model (specified in Section 2.4), that uses $L1$-regularized logistic regression over the user feature vectors to model conversions for each campaign (ignoring information about past exposure of the user to the ad campaign in consideration). Figure 2 plots the AUC performance of this model for each stratified bucket of campaigns. The average AUC for LR over all the 200 campaigns is 0.67, and as seen from the figure, the AUC increases as the number of conversions available for training increases.

Next, we compare the performance of the LR+ model that additionally uses the number of past impressions (x_i) as another feature in the user feature vector $\mathbf{f}_i$, which is then modeled using logistic regression. As we see from Figure 1, using this simple method alone increases the AUC by 2.5% to 0.687. This demonstrates that past user exposure clearly has some discriminative signal for predicting user conversions, and suggests scope for exploiting this signal better by using a richer model.

4.3.2 *Factor Models for Past Exposure*

We now move on to experiments using our factor models for user exposure: EFMG (that uses a Geometric prior for $\text{P}[\phi_i = 1|x_i]$) and EFMP (that uses a Poisson prior for $\text{P}[\phi_i = 1|x_i]$). As seen in Table 1, EFMG yields an average AUC value of 0.716, and outperforms LR, LR+, and EFMP by around 7%, 4.2% and 2.5%, respectively. The fact that the EFMG factor model gives a substantial improvement over the LR+ and even the EFMP model, even though all of them use the past user campaign exposures, suggests that the Geometric model for modeling awareness might be a more accurate framework for accounting the effect of user exposure on conversion propensity.

While the EFMP model also improves over LR and LR+ by around 4.3% and 1.5% respectively, it does not perform as well as EFMG on average.

Figure 2 shows the average AUC performance for EFMG and EFMP for each campaign bucket. Interestingly, we see that the difference in performance among the models is more pronounced in campaigns with a smaller number of conversions (smaller bucket indices). In bucket 6 all the models perform nearly the same (except EFMP), while in bucket 1 EFMG has about a 20% higher AUC compared to the LR.

4.3.3 *Effect of Past Exposure on Different Campaigns*

Figure 3 shows the average value of the optimal Geometric parameter α (which represents the Bernoulli probability of noticing an ad) computed by the Alternate Maximization algorithm in the EFMG model, and the optimal Poisson parameter α in the EFMP model, in different buckets. Except for bucket 2 in the EFMP model, campaigns with a larger number of conversions (larger bucket indices) generally have higher α values. Recall from Section 2.2.1 that as α goes to 1, $\text{P}[\phi_i = 1|x_i]$ moves closer to 1 regardless of x_i. Hence, in such cases $\text{P}[y_i|\mathbf{f}_i, x_i]$ is relatively unaffected by the past impression count x_i, and all the user exposure models perform similarly in this situation (including the LR model that does not include user exposure).

We believe that the α parameter for each campaign is a measure

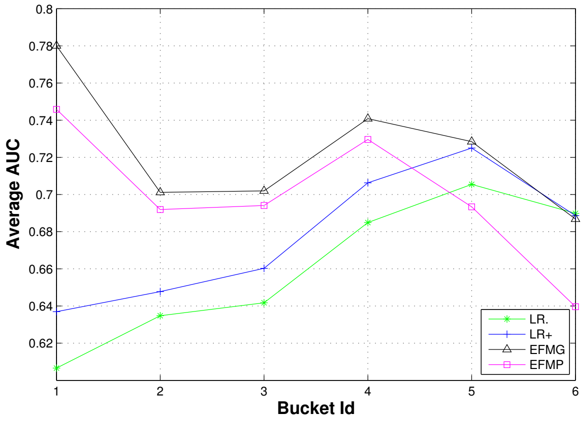

Figure 2: Average AUC of the campaigns in each conversion volume bucket for the LR, LR+, EFMG **and** EFMP **models.**

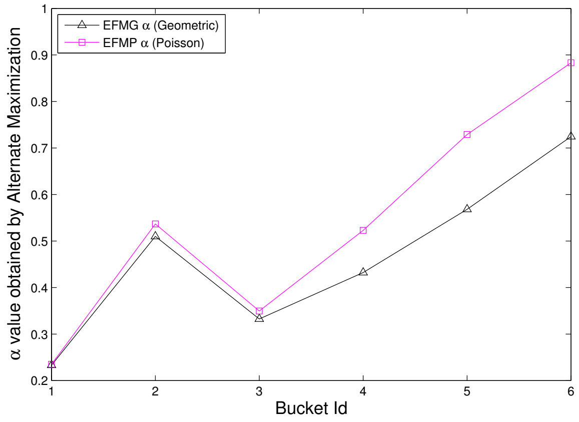

Figure 3: Average probability of noticing the ad (α parameter) for campaigns in each conversion volume bucket, for the EFMG **(Geometric Distribution) model and** EFMP **(Poisson) models.**

of the required past exposure and is unrelated to how well the campaign matches the user interest. The α parameter (noticeability) of the ad campaign depends on many hidden campaign characteristics such as brand awareness of the campaign, the product trustworthiness as well as other observable campaign characteristics such as the ad image attributes (e.g., size, color, text), the context where the ad is displayed, the location on the page where the ad is displayed (e.g. at the top of the page, above/below the fold). Modeling the ad noticeability as a function of these non-user attributes is a subject for future research.

4.3.4 Factor Model with Delay

We repeat the above experiment for the EFMD model, which incorporates an additional exponentially decaying delay factor with parameter γ in the conversion model. As described in Section 2.3, we build this delay model on top of the Geometric distribution with parameter α, and solve the resulting maximum likelihood estimation problem using alternate maximization. Table 1 shows that the performance of EFMD (AUC of 0.717) is almost identical to that of the EFM model (AUC of 0.716). Figure 4 also shows that the average AUC of the EFMD model and the EFM models are very similar over all campaign buckets. Figure 5 shows the delay parameter γ across all buckets, and this variation is non-monotonic. We believe that the delay period represents a phase when a user is considering whether to convert or not, and that our EFMD model without using any additional data source may not be rich enough to effectively model this consideration phase. Figure 5 shows the geometric parameter α (noticeability) for the EFMD model, and as with the case for EFMG and EFMP models, α increases with the number of conversions.

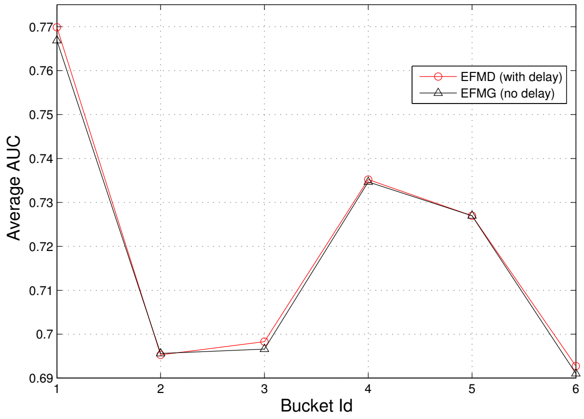

Figure 4: Average AUC of the campaigns in each conversion volume bucket for EFMG **and** EFMP **models, with and without the additional delay factor (drawn from an exponential distribution) between the user noticing an ad and deciding whether or not to convert on it.**

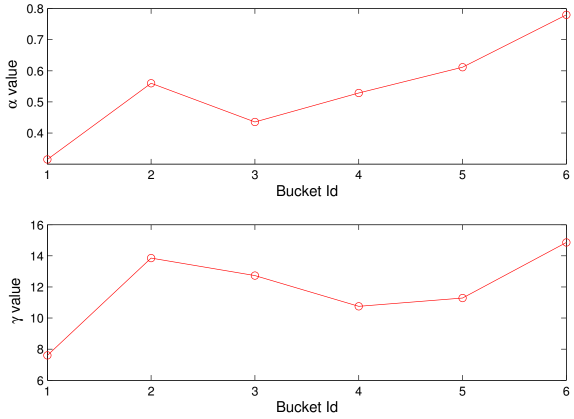

Figure 5: Top: Average probability of noticing the ad (α parameter) for campaigns in each conversion volume bucket, for the EFMD **delay model. Bottom: Average value of the parameter (γ) in the exponential distribution of the delay between the user noticing an ad and making a conversion decision.**

4.3.5 Effect of the Model Estimation Method

All the previous results for the factor models used Alternate Maximization (AM) for estimating the unknown model parameters. Next we explore whether the use of other optimization techniques such

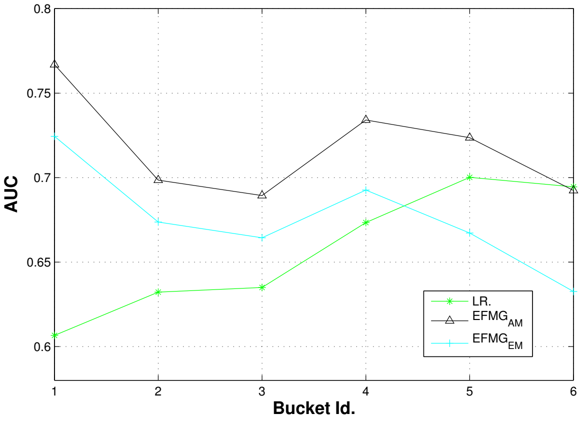

Figure 6: Average AUC of the campaigns in each conversion volume bucket for the EFMG models, with parameters estimated using Alternate Maximization ($\mathrm{EFMG}_{\mathrm{AM}}$) and Expectation Maximization ($\mathrm{EFMG}_{\mathrm{EM}}$), compared with the LR model

as Expectation Maximization (EM) affects the performance of the models. For this experiment we first learned the geometric factor model using the alternate maximization (denoted by EFMG) and then using the expectation maximization (denoted by $\mathrm{EFMG}_{\mathrm{EM}}$).

Figure 6 compares the AUC performance of $\mathrm{EFMG}_{\mathrm{EM}}$ and EFMG across all the buckets, on campaigns that converged successfully (a few models in $\mathrm{EFMG}_{\mathrm{EM}}$ did not successfully converge). We observe that EFMG significantly outperforms $\mathrm{EFMG}_{\mathrm{EM}}$ in all the buckets: the average AUC for $\mathrm{EFMG}_{\mathrm{EM}}$ is 0.675 compared to 0.716 for EFMG. Further, in our experiments with $\mathrm{EFMG}_{\mathrm{EM}}$, we observed that EM was much slower to converge than the AM approach and often encountered numerical stability issues with the M-step iterations. The reason for this poor performance with EM is likely due to the fact that we use a gradient-descent method for the M-step iterations even though the optimization objective is not convex.

5. DISCUSSION AND FUTURE WORK

In this paper we have presented an ad-exposure based factor model framework for capturing the joint effect of past ad exposure and user interest on conversion propensity for display ads. Our experimental results over a large set of 200 real-world online ad campaigns provide insights into the role that a user's past exposure to an ad campaign plays in predicting her response to the campaign. Our EFM models obtain more than a 4 to 7% gain in AUC over traditional targeting schemes that either do not consider past exposure at all, or use it naively as yet another user feature. This suggests that our models can indeed effectively factor out the separate components in terms of the interest-match component and the past-exposure component.

We can leverage this capability in two ways: first, this allows our models to target potential converters more accurately, as evident in the AUC performance. More interestingly, the α parameter that we obtain from the joint modeling of two factors in EFMG and EFMP, gives us a single metric for capturing the "effectiveness" of different ad campaigns. Advertisers can potentially use this information to determine aspects of their campaign strategy such as finding the right amount of past exposures needed by an ad to effectively target an interested user.

6. REFERENCES

[1] Zoë Abrams and Erik Vee. Personalized ad delivery when ads fatigue: an approximation algorithm. WINE'07, pages 535–540, Berlin, Heidelberg, 2007. Springer-Verlag.

[2] N. Archak, V. S Mirrokni, and S. Muthukrishnan. Mining advertiser-specific user behavior using adfactors. In *WWW*, pages 31–40, 2010.

[3] Abraham Bagherjeiran, Andrew O. Hatch, and Advait Ratnaparkhi. Ranking for the conversion funnel. In *SIGIR*, 2010.

[4] Abraham Bagherjeiran, Andrew O. Hatch, Advait Ratnaparkhi, and Rajesh Parekh. Large-scale customized models for advertisers. In *ICDMW*, 2010.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[6] Ye Chen, Michael Kapralov, Dmitry Pavlov, and John F. Canny. Factor modeling for advertisement targeting. In *NIPS*, 2009.

[7] Ye Chen, Dmitry Pavlov, and John F. Canny. Large-scale behavioral targeting. In *KDD*, 2009.

[8] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51:107–113, January 2008.

[9] Xiang Fang, Surendra Singh, and Rohini Alhuwalia. An examination of different explanations for the mere exposure effect. *Journal of consumer research*, 34(1):97–103, 2007.

[10] Thore Graepel, Joaquin Quinonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in bing search engine. In *ICML*, 2010.

[11] J. A. Hanley. *Receiver Operating Characteristic (ROC) Curves*. John Wiley & Sons, Ltd, 2005.

[12] Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Yun Jiang, and Zheng Chen. Exploitation and exploration in a performance based contextual advertising system. In *KDD*, 2010.

[13] Puneet Manchanda, Jean-Pierre Dube, Khim Yong Goh, and Pradeep K. Chintagunta. The effect of banner advertising on internet purchasing. *Journal of Marketing Research*, 43(1):98–108, 2006.

[14] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet, 2002.

[15] Foster Provost, Brian Dalessandro, Rod Hook, Xiaohan Zhang, and Alan Murray. Audience selection for on-line brand advertising: Privacy-friendly social network targeting. In *KDD*, 2009.

[16] Navdeep Sahni. Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. To appear.

[17] Xiaoxiao Shi, Kevin L. Chang, Vijay K. Narayanan, Vanja Josifovski, and Alex Smola. A compression framework for user profile generation. In *SIGIR Workshop on Feature Generation and Selection for Information Retrieval*, 2010.

[18] Ramakrishnan Srikant, Sugato Basu, Ni Wang, and Daryl Pregibon. User browsing models: Relevance versus examination. In *KDD*, 2010.

[19] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Jianchang Mao, and Rong Jin. How much can behavioral targeting help online advertising? In *WWW*, 2009.