International Journal of Applied and Advanced Multidisciplinary Research (IJAAMR)
Vol. 1, No. 3, 2023: 275-286

Targeted Display Advertising Using Machine Learning

Chaitali Khachane
AISSMS Institute of Information Technology
Corresponding Author: Chaitali Khachane chaitalikhachane13@gmail.com

A R T I C L E I N F O

Keywords: Problem formulation, Multi-stage Architecture, Large-scale Machine learning system, Experimental Validation.

Received : 25, September
Revised : 20, October
Accepted: 5, November

A B S T R A C T

This paper delves into the intricate challenges of problem formulation and data representation in the context of a large-scale machine learning system for targeted display advertising. Unlike traditional models, this system is not just conceptual but has been operational for years across thousands of advertising campaigns. Since obtaining ideal training data is cost-prohibitive, the data is sourced from related domains and tasks and then adapted for the target task. The paper outlines the architecture of this multi-stage transfer learning system, emphasizing the problem formulation aspects. Extensive experiments demonstrate the value of each transfer stage. Real-world results with diverse advertising clients from various industries showcase the system's performance. The paper concludes with valuable insights gained from over half a decade of work on this complex, widely deployed machine learning system.

INTRODUCTION

The advertising industry, a significant contributor to the U.S. GDP at approximately 2%, places great emphasis on precise ad targeting. Online display advertising, a subfield within this industry, presents both opportunities and complexities. It is promising due to the vast data available for ad targeting, yet challenging as it involves a convoluted ecosystem with multiple stakeholders. This paper primarily addresses the intricate realm of customer prospecting in online display advertising, targeting consumers who haven't interacted with a brand but are potential customers.

The rise of real-time bidding exchanges (RTBs) has revolutionized display advertising, offering efficient methods for advertisers to reach specific consumers with real-time auctions. Each ad view, referred to as an "impression," is auctioned off during webpage rendering. Advertisers receive bid requests containing user data, supplementing it with additional consumer and website information. With billions of daily auctions, advertisers require large-scale, high-speed systems for real-time decision-making.

This complexity naturally aligns with the integration of machine learning into ad optimization. It leverages massive consumer behavior data, brand-related actions, and real-time ad delivery. The paper explores the workings of a deployed machine learning system used by M6D for finding prospective customers and running targeted display ad campaigns.

The paper's key contribution to machine learning is its practical application, revealing how data characteristics and limitations translate into a complex problem formulation. It highlights that, for pragmatic reasons, the system must draw data from various sampling distributions to create the machine learning solution.

The core challenge is to identify prospective customers for diverse ad targeting campaigns automatically. Obtaining sufficient training data is prohibitively expensive and time-consuming, given the high dimensionality of the problem and low purchase probabilities. To address this, the system employs a two-level modeling approach. The first level utilizes abundant but biased data sources to handle sparsity and high dimensionality, while the second level combines and refines the outputs from the first level using data from the target distribution.

This paper aims to shed light on the design and operational choices of a massive-scale, real-world learning system, which is often overlooked in the machine learning literature. It emphasizes the importance of addressing data availability constraints, including working with non-ideal data distributions and rare outcomes. The system incorporates transfer learning and stacked ensemble classification techniques. Overall, the paper advocates for viewing most machine learning applications as instances of transfer learning, emphasizing the practicality of these techniques in real-world applications.

LITERATURE REVIEW

1. Background on M6D Display Advertising and Related Work

M6D, a significant player in the online display targeting industry, predominantly focuses on prospecting for over 100 brands, delivering millions of ad impressions daily. The system relies on cookies to maintain unique user identifiers, allowing the association of various events with the same consumer. They collaborate with data partners to track partial browsing histories and install tracking pixels on brand websites to record visits, purchases, and other meaningful interactions. This comprehensive data enables meaningful campaign evaluation, emphasizing post-view conversions as the primary metric for success.

M6D primarily delivers ad impressions through ad exchanges, evaluating the prospectiveness of consumers and submitting bids accordingly. Bid prices are determined by a separate machine learning process.

While M6D's system is not the only one in the advertising ecosystem, there is common ground in the challenges it faces, such as rare event rates, high-dimensional feature vectors, and the "cold start" problem of having no campaign data before a new campaign begins.

To address the rare event/high dimensionality problem, various solutions have been proposed. Agarwal et al. used hierarchical relationships for probability estimates. Chen et al. incorporated Laplacian smoothing into Poisson regression, while Pandey et al. and Dalessandro et al. augmented rare outcomes with correlated outcomes having higher occurrence rates. Transfer learning, specifically the use of alternative outcomes in classification models, has been explored.

Liu et al. introduced transfer learning in the context of online display advertising with a multi-task learning approach, where data from multiple tasks are pooled, and parameters are estimated across a joint feature space. However, cross-campaign transfer is not applied by M6D to avoid using one brand's data to optimize a competing brand's campaign, which is undesirable.

The transfer learning approach presented in this paper extends beyond the standard campaign and utilizes source domains not typically considered. This paper is the first to describe such an application of transfer learning in advertising, particularly one that conducts transfer learning across numerous source tasks at scale. Additionally, it's the first to detail a functional display advertising system that combines multiple models via (stacked) ensemble learning.

2. Transfer Learning for Display Advertising

The paper's central focus is on transfer learning across different tasks, which necessitates precise definitions to discuss the concept thoroughly. Transfer learning involves learning from a task that differs from the target task in terms of sampling distribution, features, label, or functional dependence between features and the label, and then applying this knowledge to enhance learning in the target task.

A task consists of a domain and a mapping, where the domain includes an example space, a sampling distribution on that space, and a featurization for the examples. Importantly, users may be sampled and featurized differently from the target distribution to augment training data.

A target task is the ultimate goal, with its own domain and mapping. Transfer learning aims to improve the learning of the target task by leveraging knowledge from one or more source tasks. Each source task has its domain and mapping, distinct from the target task. For the M6D system, the target task is to identify internet users likely to make their first purchase shortly after seeing an advertisement. The target sampling distribution, featurization, and outcome are precisely defined.

Drawing data from the target task is expensive and impractical due to the need to purchase random impressions, the large feature space, the scarcity of positive examples, and the inefficiency of random ad targeting. Advertisers require campaigns to meet their goals rapidly, and thus, the M6D system addresses this by using existing data collected over time, involving different sampling distributions and actions related to the target outcome. Transfer learning is essential for leveraging this alternative data effectively.

2.1 Possible Mappings/Labels for Targeted Advertising

To increase the number of positive examples for estimation and make transfer learning more effective, various liberal definitions of labels (Y) can be considered. The primary target label, "purchase after being exposed to an ad," is a rare event that requires costly impressions. Alternative labels (YS) can include:

1. Clicking on an ad (still requires showing impressions).
2. Any purchase, not necessarily the first time, after exposure to an ad.
3. Any purchase, with or without exposure to an ad.
4. Any other brand action, with or without exposure to an ad.

The number of positively labeled internet users is larger for the alternative actions, with option 4 being a superset of 3, and 3 being a superset of 2. For effective knowledge transfer, the estimated function $fS(\cdot)$ should be closely related to the function of interest, $fT(\cdot)$. Consequently, the outcomes YS and YT should be strongly related. In essence, this implies that the fundamental behavioral drivers for YT should also reasonably influence YS.

2.2 Domains and Features of a Users's Online Activity

As defined earlier, a domain (D) comprises three key components: the example space (E), the sampling distribution (P(E)), and the featurization (X(E)). The example space generally represents internet users or online consumers, but these users are sampled in various ways, resulting in substantial heterogeneity across different source and target tasks.

The sampling events during which M6D interacts with users include:

1. General internet activity: Users visiting sites/URLs with which M6D has data partnerships.
2. Bid requests from exchanges/bidding systems.

3. Showing ad impressions, whether targeted or untargeted.
4. Clicking on ads.
5. Making purchases at a campaign's brand's site.
6. Engaging in other online brand-related actions that can be tracked, like visiting the brand's homepage or store locator page.

The main distinctions between populations collected through these sampling events lie in the differences in their sampling distributions (P(E)). In this paper, the source domain for stage-1 experiments is based on the union of all these events, although in practice, M6D builds separate source-domain models for different events.

Furthermore, sampling events can be used to label examples, and this can lead to the creation of modeling datasets by sampling one population and assigning labels from a different event. For example, users who were shown an ad might represent the population, while those who subsequently purchase from the brand's website are the positively labeled consumers.

The target featurization (XT(E)) includes a consumer's browsing history and other user information. In any domain or event sample, a user is characterized by a set of features {x1i, x2i, ..., xKi}, which capture various aspects of the event, the user, and the user's browsing history. Features can include binary indicators of visiting specific URLs or real-numbered values reflecting browsing frequency and recency. The system anonymizes URL data by hashing it to maintain user privacy. Appendix B provides specific definitions of the target and source tasks used in the experiments in section 4. Figure 1 illustrates the relationships between user events, the target task, and the two-stage transfer learning tasks

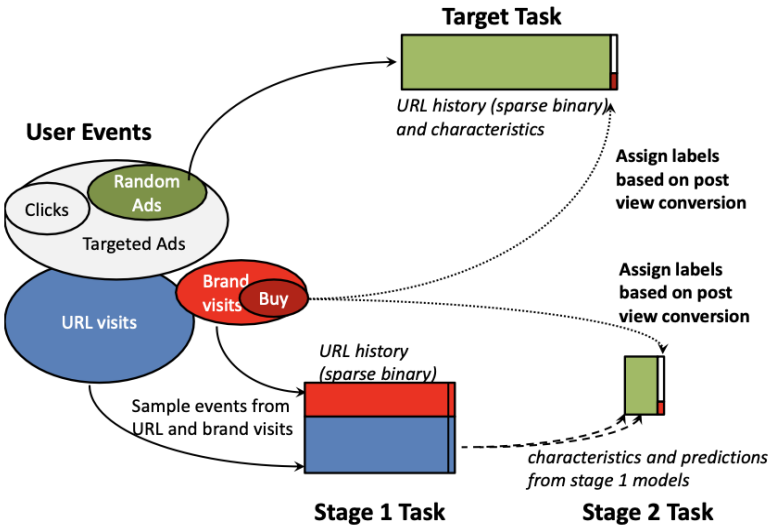

Fig. 1 Conceptual overview of the different events and tasks, their feature sets and how they are sampled from browser events. The colors identify different event types; the italic text describes the featurization for the different tasks; dotted arrows and bold text specifies the process of assigning labels; regular arrows indicate the event sampling; the dashed arrows show the stage 1 to stage 2 transition where the predictions of models from stage one form the new featurization for stage 2.

2.3 Two-Stage Transfer Learning

To achieve the ultimate goal of predicting which users are most likely to purchase a product after being exposed to an ad, the system employs a two-stage transfer learning approach. Instead of selecting a single source learning task, the system leverages multiple source learning tasks, each with its own domain and mapping. The first stage aims to significantly reduce the target feature set (XT) so that in the second step, learning can occur based on the target sampling distribution (PT).

In the first stage, multiple parallel source learning tasks are considered, and each task estimates a function (fs(X)) to approximate the label (YS). In the second stage, the system learns how to transfer the set of predictions from the first stage by weighting individual inputs using a learned linear classifier. The distinctions between source and target tasks are rooted in different events, leading to varying sampling distributions and labels, as illustrated in Figure 1.
An interesting aspect of the system is that the "correct" target learning task, which is whether a consumer purchases after an ad impression, is not always used in the production system for certain campaigns. Budget constraints or issues with tracking pixels on the brand's website may make it unrealistic to serve enough impressions to observe sufficient conversions. In such cases, the system uses the next best outcome, often a visit to the brand's website following an ad impression, as the target learning task. In practice, using a site visit as the training outcome can outperform using a purchase as the training outcome when predicting purchases. Therefore, the paper combines purchases and site visits as the target label, and the focus in this paper lies primarily on sampling distributions (P(E)) and how site visits/purchases are used as labels.

METHODOLOGY

In our study, we have extensively addressed the intricate challenges in targeted display advertising through a carefully defined problem formulation. The fundamental obstacle we tackled is the cost and scarcity of training data from the target sampling distribution. To mitigate this, we introduced a two-stage transfer learning approach that harnesses models trained on surrogate domains and learning tasks and subsequently transfers this knowledge to the target task. Our empirical findings have underscored the remarkable value of different transfer stages in enhancing system performance. From these findings, several critical insights have emerged for the broader machine learning community. These include the significance of deliberate data definition, the ability of transfer learning to combat cold-start problems, the importance of pragmatic constraints and data cost in decision-making, the efficacy of progressive dimensionality reduction, and the prevalence of transfer learning in diverse real-world applications. Overall, our study underscores the transformative potential of explicit transfer learning considerations in solving complex real-world challenges and guiding the development of automated systems.

RESEARCH RESULT

In the subsequent sections, we present the results obtained from the different stages of our transfer learning system. These experiments aim to address the questions posed earlier and assess the impact of training on various source tasks in stage 1, as well as the combination and weighting of models in stage 2. For our evaluation, we employ tasks characterized by the appropriate sampling distribution PT(ET), representing the target task, which consists of random and untargeted users who can be exposed to an ad and have not previously engaged in any brand actions. These tasks utilize the same featurization as the training data. It's important to note that our stage 1 and stage 2 models, in sequence, provide a mapping for the complete feature set of the target featurization XT, which includes browsing history (Xbinary) and user characteristics (Xinfo). Furthermore, positive instances in these tasks are users who perform a brand action within seven days of encountering the ad.

The Benefits of Stage-1 Transfer

This section explores the results of our transfer learning system's different stages and aims to answer the questions posed earlier. The experiments focus on using a convenient sampling distribution (PS(E)) and labeling scheme to maximize positive examples, even if they don't perfectly reflect the actual target task, often yielding better results than consistently using the target distribution (PT(E)). From a transfer learning perspective, we demonstrate that the estimation of function fS(·) often serves as a better predictor of YT (target label) than the estimation of fT(·).

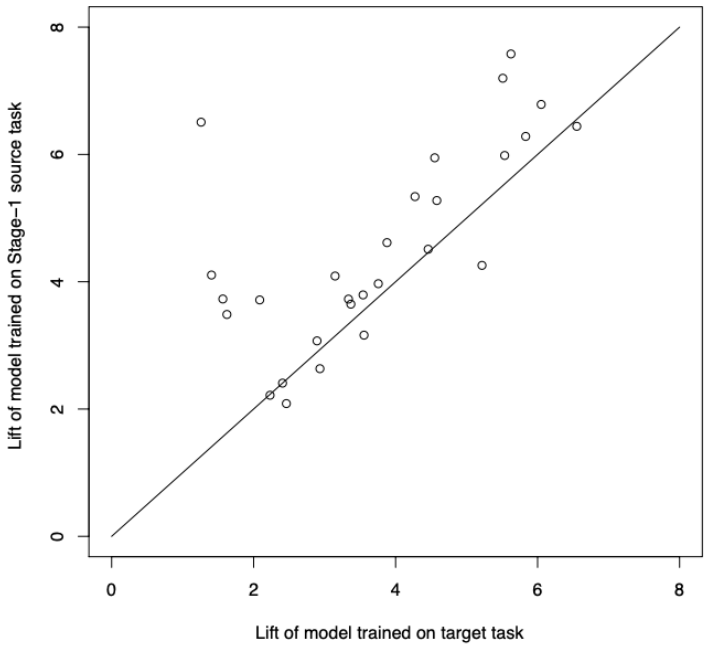

Fig. 2 Comparison of model performance between learning on the stage-1 source task and on learning on the target task (default learning parameters) when evaluated on the target task. Every point is a campaign. Points above the identity line indicate that the models trained on the source perform better on the target task than the models trained on the target task.

To empirically confirm the significant differences between source and target tasks, we conducted tests comparing the sampling distributions PT(E) and PS(E). A classifier was built using binary URL indicators as features to distinguish users sampled from these distributions, demonstrating measurable differences between the two. The out-of-sample AUC achieved by this model further supports the disparities between the populations.

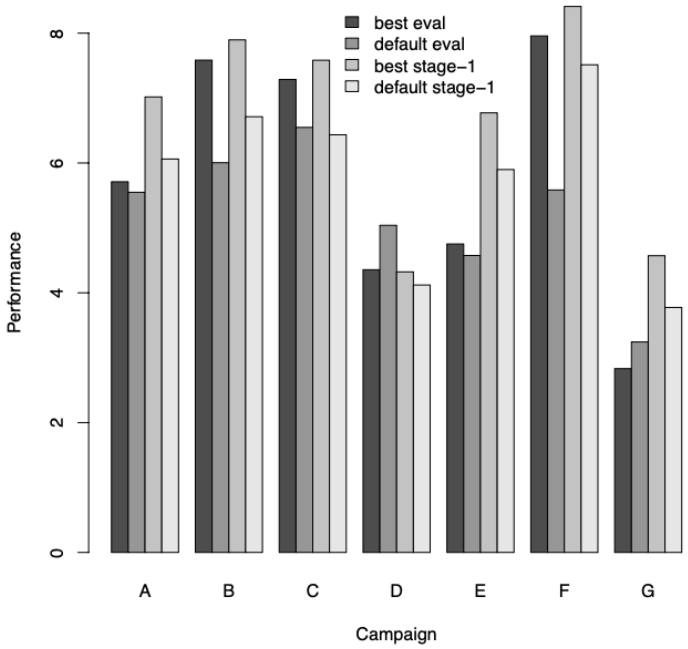

Fig. 3 Comparison of the performance on the target task of stage-1 and target-task (eval) models for a seven campaigns (A-G) after tuning the learning parameters. The "best" model is the one that performed best in cross-validation on the training set.

In our analysis, we define the source population (ES) as all active internet users within our system, with the sampling distribution (PS(ES)) representing a composite of various sampling events. The source label (YS) indicates whether a user has visited the marketer's website in the past. These models are compared against models trained directly on the target task, where the target population (ET) comprises users who could potentially win ad auctions, and the target label (YT) represents a brand action following an ad.

The results indicate that the models trained on the stage-1 source task consistently outperform those trained on the target task, with a notable advantage in learning from the extensive, high-dimensional URL featurization. In cases where we conducted an extensive parameter search for target training, models trained on the source task still proved to be more effective. This counterintuitive result suggests that, in scenarios with scarce positive examples and different training distributions, the bias introduced by the source task can be outweighed by the increased positive-class signal it provides. These findings highlight the practicality of using biased initial sampling schemes in real-world applications, where positive-class data are limited or expensive.

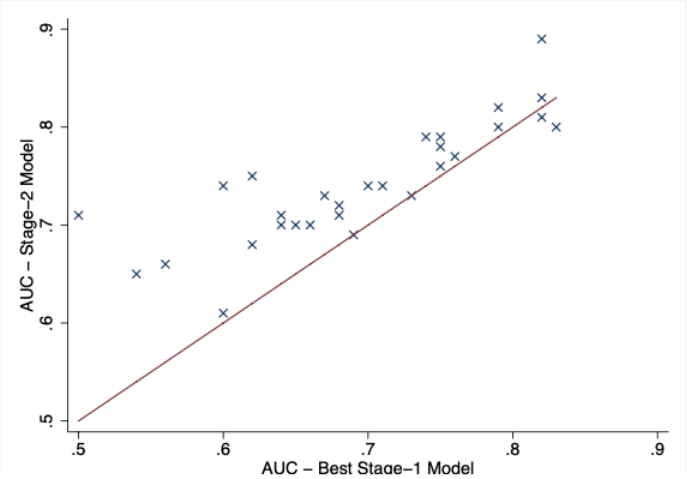

Fig. 4 Performance of the stage-2 stacked ensemble trained on the target task data compared to the best stage-1 model for each campaign. Performance is reported as the areas under the ROC curve (AUC) for each campaign and model. The stage-2 models consistently outperform even the best stage-1 models. (Here, to be conservative, the "best" stage-1 models are chosen based on their performance on the test data.)

Stage-2 Ensemble Model

In this section, the performance of the second stage (stage-2) in our transfer learning process is evaluated by comparing it to the constituent stage-1 models. The primary aim is to assess whether the adjustment to the target task, achieved through the stage-2 ensemble, offers improvements over solely using one of the source models without any target task adjustment.

For these experiments, we collected 30 days of randomly targeted users from PT(ET) as the basis for the target distribution. The data sets had varying numbers of positive examples ranging from 50 to 10,000, along with a large number of negative examples. The stage-2 featurization involved approximately 50 features, including stage-1 model scores specific to the campaign and user, along with various user characteristic features (Xinfo) such as browser type, cookie age, and geo-location information.

The stage-2 model is a logistic regression classifier trained using elastic net regularization, combining L1 and L2 regularization. The experimental results are presented across 29 different campaigns, representing recurring advertising tasks. The performance comparison is based on the area under the ROC curve (AUC) of the stage-2 model against the AUC of the best-performing stage-1 model. All performance evaluations were conducted on an out-of-time hold-out set, ensuring a proper assessment of both stages.

The results demonstrate the significant improvements achieved by combining source models and integrating information about the target task in the stage-2 ensemble. The median and average AUC improvements across different campaigns were 0.0375 and 0.0411, respectively. Notably, the enhancement is even more pronounced when the best stage-1 model exhibits relatively poor performance. Cases where the best stage-1 model falls in the lower 50% of campaigns showed median and average improvements of 0.056 and 0.061, respectively. Any potential "negative transfer" is effectively managed by the learning procedure, where poorly performing stage-1 models receive low or negative weights in the ensemble.

It's important to note that the variance in AUC across campaigns in both stages is due to the diverse nature of clients and brands involved. While some brands yield highly discriminative models, others, particularly mass-market brands, face more challenges in building discriminative models. Therefore, the absolute AUC values are less significant compared to the relative improvements demonstrated across methods. These results underscore the effectiveness of the stage-2 ensemble in the transfer learning process.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, this paper offers valuable insights and practical lessons derived from a real-world, large-scale machine learning system for targeted display advertising. The system addresses the challenges of limited data availability by employing a two-stage transfer learning approach, leveraging different source sampling distributions and training labels before transferring the knowledge to the target task. Explicit consideration of the nuances in defining events (E), sampling distributions (P(E)), and labels (Y) can significantly enhance machine learning outcomes. Employing data from distributions and labels that differ from the target task can lead to performance improvements, highlighting the need for results adjustment to the target distribution. Transfer learning serves as a practical solution to the "cold-start" problem, especially when insufficient training data is available for the target task. The flexibility to add new modeling methods easily and adapt to evolving requirements makes this multi-stage approach highly attractive for production settings.

Acknowledging the expense and difficulty of obtaining data from the ideal data-generating distribution, it is crucial to adapt by collecting more cost-effective data, even if it may not be optimal but still serves the intended purpose. In many cases, a larger quantity of data from suboptimal data distributions can outperform a smaller amount of data from the ideal distribution. Careful evaluation of the cost-benefit trade-off for acquiring data from various source tasks is essential. Practical constraints often make training on the ultimate target outcome, such as purchases, sub-optimal. Using alternative outcomes or proxies, like site visits, can lead to more effective models in such situations. Building an automated system that learns multiple models simultaneously, updates them continuously, and maintains scalability requires decisions that benefit the majority of models without significantly harming any of them.

The concept of progressive dimensionality reduction, creating lower-dimensional models in stages, is beneficial in various contexts, particularly when data is available at different resolutions. These lessons are intended to be a valuable resource for other practitioners in the field of machine learning, particularly when they aim to develop automated systems with minimal human intervention. The study highlights that learning from distributions and labels that don't precisely match the target task is a common practice in real applications, often leading to substantial improvements. This paper demonstrates the tangible benefits of transfer learning and emphasizes that

explicit consideration of transfer aspects can yield more improvement than relying solely on modeling intuition.

REFERENCES

Agarwal,D.,Agrawal,R.,Khanna,R.,Kota,N.:Estimatingratesofrareeventswithmultiplehierarchiesthroughscal- able log-linear models. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 213–222 (2010)

Attenberg, J., Ipeirotis, P., Provost, F.: Beat the machine: Challenging workers to find the unknown unknowns. In: Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence (2011)

Attenberg,J.,Provost,F.:Whylabelwhenyoucansearch?strategiesforapplyinghumanresourcestobuildclassifica- tion models under extreme class imbalance. In: KDD (2010)

Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Y. Lechevallier, G. Saporta (eds.) Pro- ceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), pp. 177–187. Springer, Paris, France (2010). URL http://leon.bottou.org/papers/bottou-2010

Breiman,L.:Stackedregressions.Machinelearning24(1),49–64(1996)

Chen, Y., Pavlov, D., Canny, J.: Large-scale behavioral targeting. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 209–218. ACM (2009)

Dalessandro,B.,Hook,R.,Perlich,C.,Provost,F.:Evaluatingandoptimizingonlineadvertising:Forgettheclick,but there are good proxies. NYU Working Paper CBA-12-02 (2012)

Evgeniou, T., Pontil, M.: Regularized multi–task learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 109–117. ACM (2004)

Fawcett,T.,Provost,F.:Adaptivefrauddetection.Dataminingandknowledgediscovery1(3),291–316(1997)

Stitelman, O., Dalessandro, B., Perlich, C., Provost, F.: Estimating the effect of online display advertising on browser conversion. Data Mining and Audience Intelligence for Advertising (ADKDD 2011) p. 8 (2011)

TheNewYorkTimes:Youronlineattention,boughtinaninstant(2012)

Weinberger,K.,Dasgupta,A.,Langford,J.,Smola,A.,Attenberg,J.:Featurehashingforlargescalemultitasklearning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1113–1120. ACM (2009)

Weiss,G.,Provost,F.:Learningwhentrainingdataarecostly:Theeffectofclassdistributionontreeinduction.J.Artif. Intell. Res. (JAIR) 19, 315–354 (2003)

Xue, Y., Liao, X., Carin, L., Krishnapuram, B.: Multi-task learning for classification with dirichlet process priors. The Journal of Machine Learning Research 8, 35–63 (2007)