Improving Post-Click User Engagement on Native Ads via Survival Analysis

Nicola Barbieri, Fabrizio Silvestri, Mounia Lalmas
Yahoo Labs London, UK
{barbieri,silvestr,mounia}@yahoo-inc.com

ABSTRACT

In this paper we focus on estimating the post-click engagement on native ads by predicting the dwell time on the corresponding ad landing pages. To infer relationships between feature of the ads and dwell time we resort to the application of survival analysis techniques, which allow us to estimate the distribution of the length of time that the user will spend on the ad. This information is then integrated into the ad ranking function with the goal of promoting the rank of ads that are likely to be clicked and consumed by users (dwell time greater than a given threshold). The online evaluation over live traffic shows that considering post-click engagement has a consistent positive effect on both CTR, decreases the number of bounces and increases the average dwell time, hence leading to a better user post-click experience.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Survival Analysis; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Design, Performance, Experimentation

Keywords

ad quality, dwell time, post-click experience, mobile advertising, survival analysis framework

1. INTRODUCTION

Many of today's free web services, such as social networks, search engines, and online news portals, are based on presenting users with advertisements (ads for short). These web services decide which ads to show to users according to the following protocol. First, a matching system retrieves all ads that are deemed "close enough" to the content of the considered context. Examples of context include a "user query" in sponsored search, a "user profile" in the case of personalized display advertising, and/or other generic features (e.g., geolocalization, device). Then eligible ads are ordered according to a function that combines the bid, which is the amount of money the advertiser is willing to pay for its ad to be shown, and the quality of the match. We denote the later as $(context, ad)$. Finally, ads with the highest scores are selected, sometimes on the condition that the score is higher than a specified minimum threshold.

This protocol departs from the classical auction-based matching, as top advertisement positions are not automatically allocated to advertisers who are willing to pay the highest price. This allows web services to maintain the quality of the ads they serve to users, including being able to control the experience of users when they click on an ad.

Both concepts are reflected by the name of one component of this scoring function, namely the *ad quality* score. The first component of such score is *relevance*, which focuses on estimating how relevant the ad is to the given context. The goal is to promote an alignment between the intent (or current interests, navigational context) of the user and the content of the ad, or product being advertised. The second component is the *click-through rate* (CTR), often used as proxy for relevance, although the two concepts are different.

While assessing the relevance of pairs $(context, ad)$ requires non-trivial editorial effort, the relationships between pairs $(context, ad)$ and the likelihood of being clicked by users can be *estimated* and *updated* by leveraging the huge information recorded in interaction logs and machine learning tools. Also, the ad quality score can take into account the quality of the *landing page* by promoting ads whose landing pages provide relevant, useful, original content *and* an engaging experience.

The final ranking function for a pair $(context, ad)$ is obtained by combining the ad quality score and the *bid*. This combination usually is formalized as:

$$score(context, ad) = bid \times \text{ad-quality}(context, ad) \quad (1)$$

CTR has usually the largest impact on the quality score and in this formula it counter-balances the effect of the bid. This allows an advertiser to win a higher position at a lower price, if the quality of its ad is higher than competitors.

In this paper, we report on results of an effort to improve the quality of native ads that are shown on a user mobile device. Native advertising refers to showing advertisements that are designed to be cohesive in both format and style to the actual content and context of the application. Our aim is to measure and predict user engagement on native ad

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883092 .

landing pages and integrate this information within the ad quality score.

Engagement can be measured in several ways. One can explicitly request users to provide feedback on ads. This way engagement can be measured as the number of positive votes over the total feedback received. This approach has the drawback of requiring active collaboration from the user base. We opt for a different, but equally effective, solution by analyzing what happens *after* the user clicks on an ad. There are two scenarios. The first one is when user immediately leaves the ad landing page and comes back to the service platform, i.e., the user *bounces*. The second one is when the user stays longer on the ad landing page and eventually enough to convert (e.g., purchasing an item, registering to a mailing list, or simply spending time on the site building an affinity with the brand). Simply put, we measure the level of user engagement with ads as the length of time user spend on the corresponding landing page, which is known as *dwell time*. A high dwell time suggests that the user found the content engaging enough to stay and "consume" it.

To compute the estimated dwell times on ad landing pages we leverage the enormous amount of data available in the interaction logs of mobile users on native ads. Our aim is to derive a way of computing $\Pr_\varphi = \Pr(DT > \varphi)$, which is the probability of dwell time of a given ad to be larger than a threshold φ. The computed probability is then used to weight the computation of the function used to rank the ads, $score(context, ad)$ from Equation 1. This means that ad-quality$(context, ad)$ is not defined simply on $\Pr(Click)$, which is the probability that the ad will receive a click, but in terms of a function

$$\text{ad-quality}(context, ad) = f(\Pr(Click), \Pr(DT > \varphi)) \quad (2)$$

that combines the probability that the ad is going to be clicked **and** the user is going to stay for more than the threshold φ. It is worth to point out that native ads are fundamentally different from search ads as in the former case there is no user query that can be used to identify the user need and match it to the ad. Therefore estimating for how long a user will stay on the ad landing page can only be done by means of landing page features themselves.

To infer relationships between the ad landing page and the corresponding dwell time, we resort to a survival analysis approach. Its main advantage over traditional regression and classification-based techniques relies on the fact that it provides simple tools for predicting the distribution over time-to-event rather than single point estimates. This makes survival techniques suitable for modeling our definition of post-click ad quality.

In this paper, we develop a tool for predicting the distribution over dwell time given a set of features extracted from the ad landing page. We build upon the research work done in [21], which was based on simple regression approaches, and propose a novel survival analysis based model to overcome some of the limitations of the previous approach.

The prediction model used is a *Random Survival Forest*, an ensemble of decision trees for *lifetime* data. We present and discuss both off-line (accuracy in predicting dwell time) and on-line (impact on advertising benchmarks) evaluation of our tool, applied to native ads served on mobile news-reading apps operated by a large Internet company.

The new model allows the setting of different thresholds for different classes of users/ads. This is important because not all the ads are comparable in terms of content nor all the users take the same amount of time to understand if an ad is interesting or not. With the previously proposed model this would require different models and therefore additional complexity at prediction time. We also show with our experiments that a simple regression on dwell time values perform poorly when compared to the survival analysis based solution. Furthermore, we include a new set of features that, as the feature importance analysis shows, are highly correlated with the target distribution to predict.

The main contributions of this work can be summarized as follows:

- We discuss the application of survival analysis techniques to predict the distribution of dwell time on native advertisements. As we shall describe further in the paper, survival analysis based models are a natural fit for dwell time prediction problems [12].
- We improve the feature set that we proposed in a previous work [21] by including important features that help improving the quality and generalizability of the prediction model.
- We test and validate our solutions on two separate benchmarks: an offline one measuring the quality of our produced models using AUC, and an online one using A/B testing to show the increase of average dwell time (as well as of average CTR) of a system using our post-click ad quality score in production.

2. RELATED WORK

Online advertising has been extensively studied in the context of display advertising [2, 30] and sponsored search [4, 31, 32]. Studies have mostly focused on predicting how an ad will perform according to various effectiveness measures, mostly click-through rate (CTR) which is the number of times the ad was clicked out of the number of times it has been shown (number of ad impressions). The higher the CTR the better the ad is considered to perform; it attracts the users to click on it. Furthermore, only recently the focus has been shifted towards mobile advertising [21] and originally studies were conducted mostly considering desktop users.

To optimize for CTR, most efforts have been around improving the matching between web queries (in sponsored search) or web pages (in display advertising) and the ad textual content (creative and/or bid phrases and title) [5, 4, 7, 22, 25]. The context of the ad landing pages have been used to enhance matching algorithms [5, 4, 8, 16]. We also exploit landing page features to predict the quality of ads, but with respect to the post-click experience and focusing on its quality.

These works have mostly focused on the short-term revenue, i.e., optimizing for as many clicks as possible, with little or no regards at all for long-term effects. This comes from using as the main success criteria relevance metrics such as CTR. However, this approach does not account for the quality of the advertising experience either on the creative or its landing page. Our focus is with the latter, the quality of the landing page and its effect on the post-click experience.

Recent work has studied the long-term effect of ad quality on the revenue of a big search engine company [14]. Inspired by the work of [20], a novel methodology was proposed to predict long-term effects on various metrics (e.g., revenue,

CTR) using short-term variables. The results showed that optimizing for immediate revenue is an obvious choice but one that is detrimental in the long run, and in addition, optimizing for ad quality has a positive effect on long-term revenue. Similar findings in the context of native advertising were reported in [21], who showed that users with positive post-click experience were more likely to click on an ad in the near future.

A good proxy of the post-click experience is the time a user spends on the ad site before returning back to the publisher site: "the longer the time, the more likely the experience was positive". The two most common measures used to quantify time spent on a site are *dwell time* [35] and *bounce rate* [31]. These measures have been used as proxies of post-click experience in online advertising and organic search, e.g., to improve the performance of ranking algorithms [18], as well as in recommender systems, e.g., to estimate the relevance of an item to a user [34]. Both dwell time and bounce rate have been validated as good proxies of post-click experience in the context of native advertising, both for desktop and mobile [21].

Users spend an increasing amount of their time online through their mobile devices, presenting unique opportunities for advertisers interested in promoting their products beyond the desktop. Previous studies have investigated the degree in which mobile advertising is accepted by users [6] and how users perceive display advertisements on mobile [11]. Efforts have been put in building models to predict when to show an ad [27, 28]. Our work has a different focus, that of how users experience landing pages in the mobile context, measured using dwell time as our post-click ad quality score.

This is particularly important in the context of native advertising. As highlighted in [13, 29], the dominance of the feed-based structure on mobile makes pop-ups and banners impractical, whereas native ads provide an optimal format as they are seamlessly incorporated to the main feed, thereby promoting relatively similar experience across all ads. This is not the same for the ad landing pages, as advertisers have total freedom in how they design them.

In the context of mobile advertising, whether the landing page of an ad is mobile-optimized,or not, was shown to affect post-click experience [23]. Our research adds to this body of work by analysing other features of landing pages for mobile advertising, and how these help predicting user post-click experience. We exploit the set of landing page features experimented with in [21] (e.g. whether the landing page contain images or click-to-call actions) and extend them with two new sets, one related to the readability of the text of the landing page, and the other concerned with the structure and layout of the landing page.

Finally, this work considerably expands our previous effort in measuring the post-click quality of an ad [21]. The novelty lies primary in the choice of the machine learning framework used to predict the quality of the post-click experience.

3. POST-CLICK AD QUALITY

In the context of online advertising (and sponsored search in particular), click through rate (CTR) has traditionally been considered the most important benchmark to assess the performances of ad campaigns, or the quality of $(ad, context)$ matches. CTR is a good indicator for the volume of traffic, but a poor one for its quality. Indeed, CTR can be heavily influenced by accidental taps which in a mobile scenario may include up to 40% of the overall number of clicks [1].

The quality of ads becomes especially important in the context of native advertising where ads are embedded within the editorial content of the publisher; ads should be friendlier and deliver an engaging experience to users, the latter comparable to the experience offered by the publisher displaying the ads. While maximizing CTR and boosting the volume of traffic is a good investment in the short term, exposing users to low quality ads is likely to make the platform lose audience in the long term. High CTR is not a reflection of an engaging post-click experience, and the focus should lie on driving engagement in addition to volume.

As a measure of engagement for ads we consider dwell time, defined as the actual length of time that a user spends on the ad landing page before leaving it. Compared to CTR, dwell time is a more reliable indicator for the quality of the ad and it can be easily used to discriminate good ads from clickbaits. This choice is supported by recent studies [3, 21, 34] both in the advertising and non-advertising contexts.

Several choices are equally suitable to define an ad quality score that takes into account dwell time. Since our goal is to mainly filter out ads with poor level of engagement, we formulate the ad quality score for each pair ad-context (a, c) as the joint probability that the match will lead to a click and the probability that the dwell time is higher than a given threshold $\varphi > 0$:

$$\text{ad-quality}(a, c) = \Pr(click, DT > \varphi | a, c, \Theta)^1 \quad (3)$$

This is further decomposed, by applying the chain rule, as:

$$\Pr(click | a, c, \Theta_{context}) \cdot \Pr(DT > \varphi | a, click, \Theta_{dwell}) \quad (4)$$

where $\Theta_{context}$ is the set of parameters of the click model,[2] and Θ_{dwell} refers to the parameters of the post-click engagement model. This formulation allows the two models to be learnt independently, and depends on a single parameter: the threshold used to separate good from low quality ads.

Let $\mathcal{D}$ denote a log of events in the form $e = \langle a, dt \rangle$, where each event e corresponds to a click on the landing page of the native ad a and dt measures the actual length of time that the user spent on the landing page before closing it, that is the dwell time associated with the click. Moreover, each ad a can be associated with a feature vector $\vec{x}(a)$, computed by analyzing the corresponding landing page.

Our goal is to analyze historical data recording the interactions of users with ads and develop predictive tools that can be used to predict the strength of such interactions on new ads, in the form:

$$\Pr(DT > \varphi | \vec{x}(a), \Theta_{dwell}) \quad (5)$$

If the threshold φ is known a-priori, this learning problem can be approached by using standard binary classifiers: positive examples are observations having $dt > \varphi$, while negative examples are the remaining ones. However, by performing this initial discretization step we might lose information. All examples are treated equally without considering the actual gap (positive or negative) between their corresponding dwell time and the threshold. Moreover, a new model must be used each time we need to consider a different threshold.

[1] Or equivalently $\Pr(click, BR < \varphi | a, c, \Theta)$, where BR represents the bounce rate.

[2] In the remainder of the paper we consider the click model as the one used in production for Gemini Native Ads.

Approaching this problem from a survival analysis perspective allows us more flexibility. Indeed, the model is not bound to a fixed threshold value, which can in turn be specified at serving time. For instance, the survival probabilities at different timestamps can be computed by using the same model and then used as features for other prediction tasks (rather than using a single point estimate).

In the context of web search system, Dupret and Lalmas [12] defined a framework that uses survival analysis to study how user interaction within a session with a search system affects future engagement with the system itself. They were able to confirm known behaviors and found unexpected ones. One of the strength of the research work was that it enabled the explanation of factors affecting engagement without the need for building different model for the various factors under study.

4. DWELL TIME PREDICTION

We summarize the main concepts in survival analysis and then describe a state-of-the art, based on an ensemble tree method, for predicting the distribution of the dwell time.

4.1 Survival Analysis in a Nutshell

Survival analysis[3] is a branch of statistics that deals with time-to-event data (also known as *lifetime data*), i.e. data in which the outcome variable is the time it takes for an event to occur. In our setting, the event of interest corresponds to the return of the user to the main page (e.g. search engine result page, stream of news), after he or she visited the landing page associated with an ad. Hence, the survival time corresponds to the length of time that the user spent on the landing page of the ad, which is the dwell time.

Let T be a non negative continuous random variable representing the survival times of individuals in some population (dwell time observed for each event in the click-log), and let $f(t)$ be its probability density function. The cumulative distribution function (CDF), defined as $F(t) = \Pr(T \leq t) = \int_0^t f(t)\ dt$, represents the probability that the event has occurred before time t. The probability that the event of interest has not happened by time t is given by the complement of CDF

$$S(t) = \Pr(T > t) = 1 - F(t) \tag{6}$$

and is usually referred to as *survival function* in the survival analysis literature. An alternative characterization of the distribution of T can be provided by introducing the *hazard function*, which gives the instantaneous rate of occurrence of the event at time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{7}$$

That is, the rate of occurrence of the event at time t is expressed as the ratio between the density of events at t and the probability of surviving t without experiencing the event. The survival and hazard functions provide alternative but equivalent characterizations of the distribution of T. In fact, $S(t)$ and $f(t)$ can be derived in terms of $h(t)$ by exploiting the following relationships:

$$h(t) = -\frac{d}{dt}\log\left(S(t)\right) \tag{8}$$

$$S(t) = e^{-H(t)} \tag{9}$$

$$f(t) = h(t)e^{-H(t)} \tag{10}$$

where by definition

$$H(t) = \int_0^t h(u)du \tag{11}$$

is the *cumulative hazard* at time t. Depending on the considered case, it is easier to model (or to make assumptions on) the hazard function rather than the survival, or vice-versa. The above relationships allow us to focus on modeling one component and deduce the other automatically.

Different types of survival models can be obtained by making different assumptions on the form of the hazard or survival function. For instance we can use *Weibull* distribution to model the belief that the risk is a monotone function (either increases/decreases or it remains constant over time). These approaches are known as *parametric models* and they provide smooth (and more robust to noisy data) estimates for $S(t)$ and $h(t)$ by performing the maximum likelihood estimate over the considered lifetime data. Conversely, *non-parametric models*, such as the *Kaplan-Meier* [17] and *Nelson-Aalen* [26] estimators, do not make any assumption on the distribution of survival times and they are generally easier to estimate.

One of the most interesting setting in survival analysis is to assess the effect of particular circumstances or characteristics on increasing, or decreasing, the probability of survival. Several models have been proposed to predict the survival of a new observation given its features (e.g. proportional hazard models [19]). In the next section, we describe Survival Random Forest, a non-parametric state-of-the art method that is able to exploit non-linear interactions between features of each observation and the corresponding survival time.

4.2 Survival random forest

Given an ad, characterized by a set of features, what is the probability that a user will spend a given amount of time on its corresponding landing page? What is the expected value of such dwell time and the overall shape of its distribution? What are the features that generally lead to high dwell time? To address such questions we need a predictive model that is able to infer the relationships between a given set of features and a response variable that represents the time needed to observe an event (e.g. the user leaving the ad landing page). We use Survival Random Forest for this purpose.

Survival Random Forest [15] is an ensemble tree method for the analysis of right-censored[4] lifetime data. It provides an ensamble non-parametric estimate for the cumulative hazard (Equation 11), which can in turn be used to estimate the empirical survival function according to Equation 9.

The procedure for building a survival random forest is summarized in Algorithm 1. Each of the M survival trees is built independently. Survival trees are binary trees grown by recursively splitting the tree nodes. Each tree starts at the root node, containing a *bootstrap* of samples from the

[3]The interested reader can refer to [24] for extensive discussions on this topic.

[4]*Right censoring* is a form of missing data. It happens when it is only known that the survival time for one individual is above a certain value, e.g. a subject does not experience the event before the study ends.

original data (line 3). Each node is split by using a survival criterion (line 13) on a randomly chosen set of features (line 12) and such split produces two children (line 14). Nodes are recursively processed until the termination criterion is met (line 8). Next, we analyze each phase in detail.

Bootstrapping. Each bootstrap is built on about 63% of the available data, leaving the remaining approximately 37% for validation. Validation samples are referred to as *Out Of Bag* (OOB) samples. The size of each bootstrap is n, which corresponds to the size of the original data, and samples are drawn *with replacement*. This means that observations from the original data set may occur multiple times (or, equivalently, have a weight > 1).

Termination criterion and lifetime estimation. The recursive splitting procedure stops when no new child can be formed from the current node, because the number of *unique* events is less than a specified threshold (3 is the default value). The node becomes a terminal node and should provide an estimator of the cumulative hazard (line 9), which is used in the prediction phase. Let $\mathcal{D}(h) = \{(\cdot, t_1(h), \delta_1(h)), \cdots, (\cdot, t_{n(h)}(h), \delta_{n(h)}(h))\}$ be the set of samples in the terminal node (h) (omitting the feature values for notational convenience), ordered by increasing survival time. In this case, $t_i(h)$ denotes the survival time for the i-th sample, $\delta_i(h)$ is a censoring binary indicator (1 if the sample is right-censored, 0 otherwise), and $n(h)$ is the number of samples in the node h. The cumulative hazard estimate for the node h ($\hat{H}_h$) is computed as the Nelson-Aalen estimator [26]:

$$\hat{H}_h(t) = \sum_{\substack{t_i \in \mathcal{D}(h) \\ t_i \leq t}} \frac{d_i(h)}{n_i(h)} \tag{12}$$

where $d_i(h)$ is the number of samples in $\mathcal{D}(h)$ with survival time exactly equal to t_i, $n_i(h)$ is the number of samples in $\mathcal{D}(h)$ with survival time greater than t_{i-1} (this includes observation right-censored at t_i).

Node Splitting. The tree growth is regulated by a greedy splitting procedure. The data in each non-terminal node must be split into two populations, such that the difference between their survival distributions is maximal. Let $L_{f,c}(\mathcal{D})$ be a function that measures the survival difference between the populations obtained by splitting data $\mathcal{D}$ on feature f and value c.

At each node h, the procedure randomly picks a set of features $\mathcal{F}(h)$, (where $|\mathcal{F}(h)| = F$) as candidate for splitting. For each candidate feature f, let $\mathcal{V}_f(h)$ be the set of possible values on f of samples in $\mathcal{D}(h)$. The best split $(f*, c*)$ is found by applying a *locally optimal* feature/split decision, by analyzing all selected features $f \in \mathcal{F}(h)$ and split values c and picking the one that maximizes:

$$(f^*, c^*) = \arg \max_{\substack{f \in \mathcal{F}(h) \\ c \in \mathcal{V}_f(h)}} L_{f,c}(\mathcal{D}(h)) \tag{13}$$

As measure of goodness of a split, random survival forest adopts the log-rank splitting rule:

$$L_{f,c}(\mathcal{D}(h)) = \frac{\sum_{t_i} \left(d_i^L(h) - n_i^L(h) \frac{d_i(h)}{n_i(h)} \right)}{\sqrt{\sum_{t_i} \frac{n_i^L(h)}{n_i(h)} \left(1 - \frac{n_i^L(h)}{n_i(h)}\right) \left(\frac{n_i(h) - d_i(h)}{n_i(h) - 1}\right) d_i(h)}} \tag{14}$$

where $d_i^L(h)$ and $n_i^L(h)$ are computed on the left-child data obtained by splitting $\mathcal{D}(h)$ on feature f and value c, and the summatories go over distinct event-times in $\mathcal{D}(h)$. Once the split (f^*, c^*) has been found, the observations in the current node are split into two populations: the left child contains observations such that $x_{f*} \leq c^*$, and the right child contains the remaining ones.

Algorithm 1 Building a Random Survival Forest

Require: M (number of trees in the forest),
F (number of features to consider at each split),
lifetime data $\mathcal{D} = \{(\vec{x}_1, t_1, \delta_1) \cdots (\vec{x}_n, t_n, \delta_n)\}$

1: $forest = Forest()$
2: **for all** trees $m = \{1, \cdots, M\}$ **do**
3: $\mathcal{D}_m \leftarrow bootstrap(\mathcal{D})$;
4: $root_m = SurvivalNode(\mathcal{D}_m)$
5: $queue = \{root_m\}$
6: **while** $\neg queue.isEmpty$ **do**
7: $curr\_node = queue.pop()$
8: **if** $checkTerminal(curr\_node)$ **then**
9: $curr\_node.buildLifetimeEstimator()$
10: **else**
11: $curr\_data = curr\_node.getData()$
12: $f\_curr = selectFeatures(F)$
13: $\langle f\_id, splitValue \rangle = getBestSplit(curr\_data, f\_curr)$
14: $\langle D_l, D_r \rangle = split(curr\_data, f\_id, splitValue)$
15: $curr\_node.setSplit(f\_id, splitValue)$
16: $left = Node(D_l)$
17: $curr\_node.setLeftChild(left)$
18: $right = Node(D_r)$
19: $curr\_node.setRightChild(right)$
20: $queue.enqueue(left)$, $queue.enqueue(right)$
21: **end if**
22: **end while**
23: $forest.addTree(root_m)$
24: **end for**

Ensamble prediction. Given a sample with features $\vec{x}$, we drop it down each tree in the forest. On each tree the sample eventually reaches a terminal node that provides a cumulative hazard estimator according to Equation 12. Let $\hat{H}_m(t|\vec{x})$ be the cumulative hazard estimate at time t for the sample with features $\vec{x}$ in the m-th tree. The ensemble cumulative hazard estimator of a forest $\mathcal{M}$ is computed as

$$\hat{H}_{\mathcal{M}}(t|\vec{x}) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \hat{H}_m(t|\vec{x}) \tag{15}$$

Given an ad with features $\vec{x}(a)$, the survival random forest $\mathcal{M}$ predicts the likelihood that its dwell time will be greater than φ as:

$$\hat{S}(\varphi|\vec{x}(a), \mathcal{M}) = \exp\{-\hat{H}_{\mathcal{M}}(\varphi|\vec{x}(a))\} \tag{16}$$

The above equation can hence be used to compute the post-click ad quality score, which was defined in Equation 5.

5. EVALUATION

We analyze the performances of our approach in two ways:

- *Predictive accuracy*: Through an off-line evaluation, we assess the accuracy of the survival random forest model on the task of predicting the dwell time of ads.
- *Online performances:* Through an on-line evaluation, we focus on standard benchmarks for evaluating the success of online ad campaigns by analyzing the difference between a bucket and a control segment.

5.1 Off-line evaluation

The predictive accuracy of our implementation of survival random forest is assessed with an off-line test. We

Table 1: Features considered for the off-line evaluation task. This set extends features discussed in previous work [21].

Domain	Name	Description
Document Object	NumberOfExternalLinks	No. of links pointing to external domains
	NumberOfInternalLinks	No. of links pointing to the same domain as the landing page
	NumberOfLinks	Sum of the previous two features
	ExternalInternalLinksRatio	Ratio of External vs. Internal links
	ExternalTotalLinksRatio	Percentage of external links
	InternalTotalLinksRatio	Percentage of internal link
	TextSizeExternalLinksRatio	Text per external links ratio
	TextSizeInternalLinksRatio	Text per internal links ratio
	TextSizeLinksRatio	Text per total number of links ratio
	MainTextSizeExternalLinksRatio	Main Text (without boilerplate text) per External links ratio
	MainTextSizeInternalLinksRatio	Main Text (without boilerplate text) per Internal links ratio
Readability	MainTotalTextSizeRatio	Main Text (without boilerplate text) per total links ratio.
	TotalTextSize	Text size
	MainTextSize	Main Text size
	FlashKincaidTitleReadability	Readability of the title
	FlashKincaidAbstractReadability	Readability of the abstract
	tokenCount	Number of tokens
	summarizabilityScore	Summarizability of the text
	FlashKincaidMainTextReadability	Readability of the main text.
Mobile Optimized	clickToCall	Is there a click to call button?
	iPhoneButton	Is there an iPhone button?
	viewPort	Is viewport available?
	windowSize	Size of the window
Media	imageHeight	Height of the rendered landing page
	imageWidth	Widht of the rendered landing page
	numberImages	Number of images contained in the landing page
	media	Is there a media (e.g., video) on the landing page?
Input	numberClickable	Number of clickable objects int he landing page.
	numberDropdown	Number of dropdown elments.
	numCheckbox	Number of checkbox
	numInputString	Number of Input Strings
	numRadio	Number of radio buttons
Semantic Similarity Landing-Page & Creative	nouns	Detected nouns in the landing page
	numConceptAnnotation	Number of concepts detected in the landing page
	similarityNoun	Jaccard between the set of nouns in the title/abstract and the main text
	similarityWikiIds	Jaccard between the set of wiki entities in the title/abstract and the main text
History	impressions	Number of impressions
	clicks	Number of clicks
	ctr	Observed CTR
	clicks_dwell	Number of clicks corresponding to dwell times greater than 10 seconds
	histDwellTime	Average dwell time from historical data
	histBounceRate	Bounce rate from historical data

base such evaluation on a dataset with $46,914$ observations $(ad, dwellTime)$, which refer to $2,438$ ads provided by over 850 advertisers. We perform a 80/20 training/test split. For each ad we extracted 42 features, which are listed in Table 1. Many of these features were experimented in previous work [21]. We added to them two sets, namely "document object" and "readibility". Each feature can be associated with a feature domain, defined as follows:

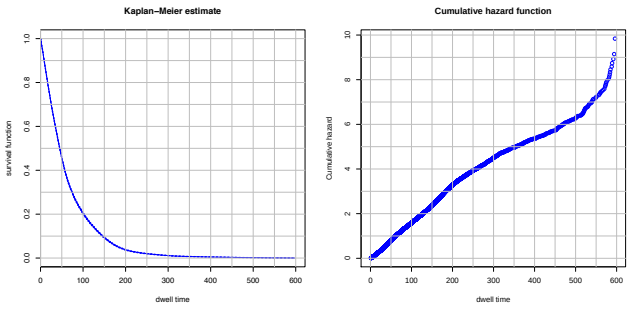
- *Document Object*: is a set of features representing the landing page content of "informative" text versus other information (e.g., hyperlinks to pages that are hosted on domains different from the landing page one).
- *Readability*: is a set of features representing the readability of the various ad components: title, description, and landing page content.
- *Mobile Optimized*: represents features that can identify if a landing page is mobile-optimized or not (in a previous project we have built a classifier using these features to detect if a landing page is mobile-optimized or not).
- *Media*: represents information about the multimedia content of the landing page.
- *Input*: is a set of features indicating what and how many input elements exist in the landing page.
- *Semantic Similarity Landing-Page & Creative*: are features evaluating the similarity between the ad creative and its landing page.
- *History*: represents how the advertisement has performed in the past in terms of engagement metrics.

The dwell time observed on these observations is summarized by the survival function in Figure 1, where the dwell time is measured in seconds. Roughly 80% of the observations have dwell time less than 100 seconds; the median is 45 seconds, while the average is approximately 65 seconds.

We measure the prediction effectiveness as in a classical binary classification task. We pick 6 thresholds on the dwell time, ranging from the minimum value of 10 to 100 seconds,[5] and for each threshold we compute the survival of observations in the test set. Effectiveness is measured as the area under the *tpr-fpr* curve (AUC). We compare survival random forest with linear models (logistic and cox regression [9]) and two random forest methods. Cox regression is a survival analysis model that assumes that the effect of features on the hazard rate is multiplicative:

$$h(t|\vec{x}) = h_0(t)\exp\{\vec{x}'\vec{\beta}\},$$

[5]The value used in production at the evaluation period was 39 seconds.



(a) (b)

Figure 1: (a) Survival and (b) cumulative hazard function over the considered dataset.

where $h_0(t)$ is the baseline hazard and $\vec{\beta}$ specifies the regression coefficients.

The first random forest method is a suite of random forests for classification, where a different forest is built for each threshold. The second is a random forest model in a regression setting, which aims at minimizing the error between the actual dwell time and the predicted one. We set the number of trees at 100 and the number of features to be selected at each split at $\sqrt{\#features} \approx 6$.

Performances are reported in Figure 2. Survival analysis models achieve the same accuracy as their counterparts (suite of logistic regression models versus Cox Regression and suite of Random Forests versus Survival random forest) and non-linear models outperforms linear ones. Also, survival random forest outperforms consistently the random forest model trained in a regression setting. This shows that the survival model has the flexibility of a regression model (not requiring a predetermined threshold to identify high dwell time ads) with a high prediction accuracy on all the threshold values.

The importance of each variable in predicting the distribution over dwell time can be assessed by using the variable Breiman-Cutler permutation technique [33]. As shown in Figure 3, historical features (such as the current estimate of dwell time for the considered ad) are the most important, followed up by document object and readability features. It is important to remark again that document object and readability features are those introduced in this research work and their importance confirms our initial hypothesis that dwell time is indeed influenced by how much "actual" content is present within a landing page, and what is the quality of this content. We return to this when we discuss future work.

Finally, we study the prediction accuracy when varying the number of trees in the forest. This test was run on a 2.7 Ghz Intel Core i7 with $4Gb$ of Ram and the learning procedure uses 6 threads. In Table 2 we report how the OOB error and learning time vary with the number of trees. The OOB error is an internal estimate of the generalization error of a random forest; on this dataset it converges to a steady state after 100 trees have been built. The learning time increases linearly. The best fit between the number of trees and the learning time is given by the function $Time = \#Trees \cdot 0.22 - 1.65$ with an adjusted R^2 coefficient of 0.96.

5.2 On-line evaluation

To measure the online performances of our proposed approach, we run an evaluation through A/B testing on mobile web traffic of Yahoo. This traffic is randomly split into two buckets: *Adquality*, which ranks ads according to Equation 4, and *Control* on which the ranking function does not account for post-click engagement. Our analysis focuses on a time window of eight weeks. The goal of this evaluation is to assess the impact of our proposed approach on standard benchmarks for assessing the performances of online ad campaigns, namely CTR and the impact on the degree of engagement of users. The engagement in these experiments is assessed by analyzing the dwell time and bounce rate values as reported by the A/B testing system. In this context, we define bounce rate as the percentage of users that left the ad landing page within 5 seconds (we use the same criteria and motivations as in [21]).

In Figure 4 we report a comparison between the average benchmark values recorded, whereas in Table 3 we show the correlation between those values. As expected, Adquality shows a consistent uplift in dwell time over the control bucket. Overall the dwell time on ads served by integrating the post-click component is 13.3% higher than the dwell time recorded on the control bucket. The increase of dwell time has a positive effect on both CTR (average uplift of 6.8%) and bounce rate (average decrease of 10.3%). We also see that an increase of dwell time is correlated with increase of CTR, and the reverse is observed with bounce rate. This again suggests that serving ads of higher quality, in terms of the post-click experience, has a positive effect on users; more ads are being clicked.

Table 2: Building time and OOB error when varying the number of trees.

$\#Trees$	OOB	$Time(min)$
10	0.4439	2.14
50	0.4180	9.7
100	0.4094	20
150	0.4093	28
200	0.4094	47

Table 3: Correlation between % uplift/losses on benchmarks.

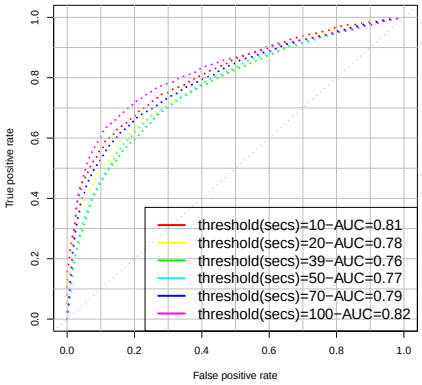
	$\uparrow DT$	$\downarrow BR$
$\uparrow CTR$	0.854	0.986
	$\uparrow DT$	0.808

6. CONCLUSIONS AND FUTURE WORK

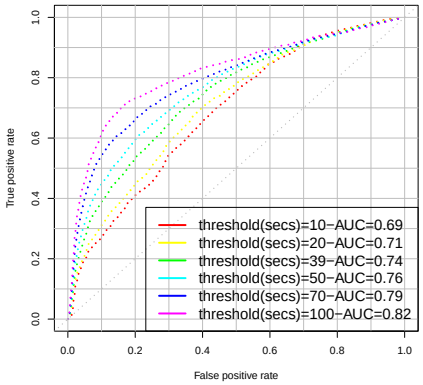
As users are spending increasingly more time on mobile, providing them with the best possible experience is important for both the owner of the platform and advertisers. In this paper, we describe an approach to estimate the post-click engagement on mobile native ads, where the latter is measured as the dwell time. To infer relationships between ads and dwell time, we resort to the application of survival random forest, an ensemble of decision trees for lifetime data.

Survival random forest based model not only slightly outperforms all the other competing model (including a suite of classification random forest) but, more importantly, it allows to compute the survival at different thresholds. Other methods require, in fact, setting the dwell time threshold before the model is actually built.

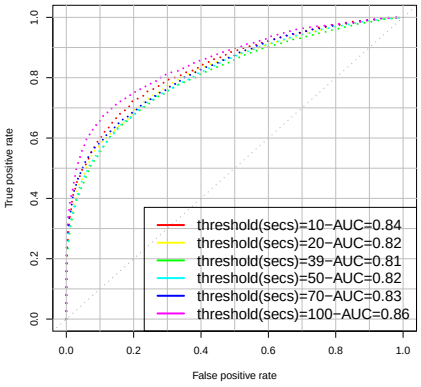
Our approach was deployed in a large-scale online setting, with the goal of promoting ads that are likely to be clicked



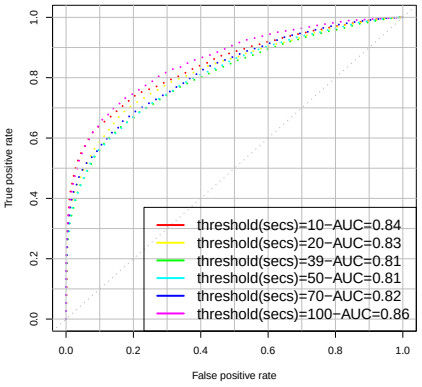
(a) Logistic Regression



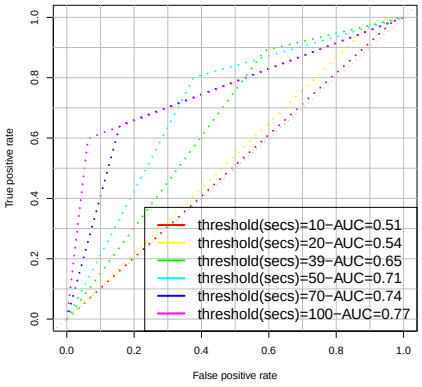
(b) Cox Regression



(c) Survival Random Forest



(d) Suite of Random Forests: Classification



(e) Random Forest: Regression

Figure 2: Prediction accuracy of models based on random forest.

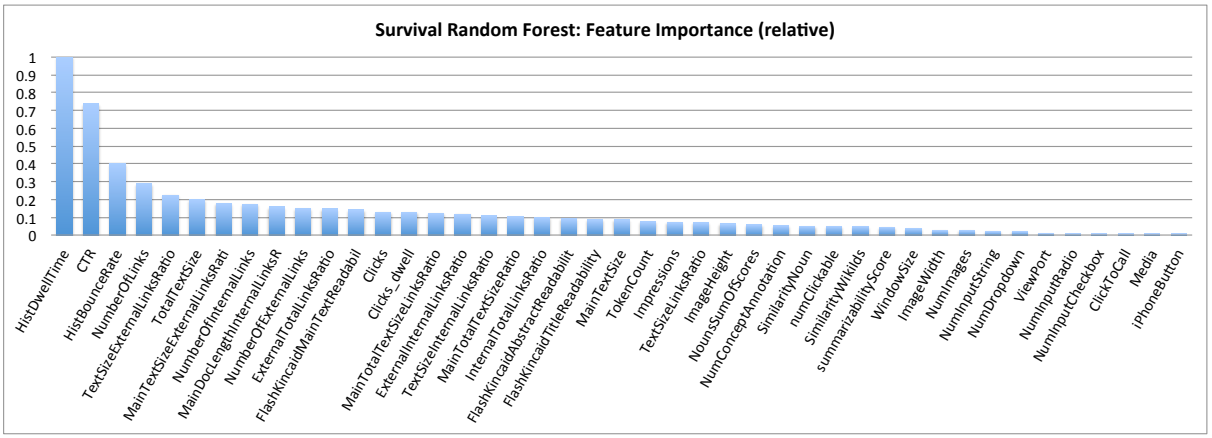


Figure 3: Variable importance on survival random forest.

and offer high level of engagement with users. The online evaluation over live traffic shows that considering post-click engagement has a consistent positive effect on CTR. In addition, it decreases the number of bounces and it increases the average dwell time, hence leading to a better post-click experience.

This work can be extended in several directions. The first one is to consider more features of the landing pages and study their correlation with dwell time. Features extracted from the domain object and features related to the readability of the content are among the most important predictors for the distribution of dwell time. We also expect that visual features, such as the quality of the images on the landing pages, have a consistent impact on the degree of user engagement on ads. Works on computational aesthetics,

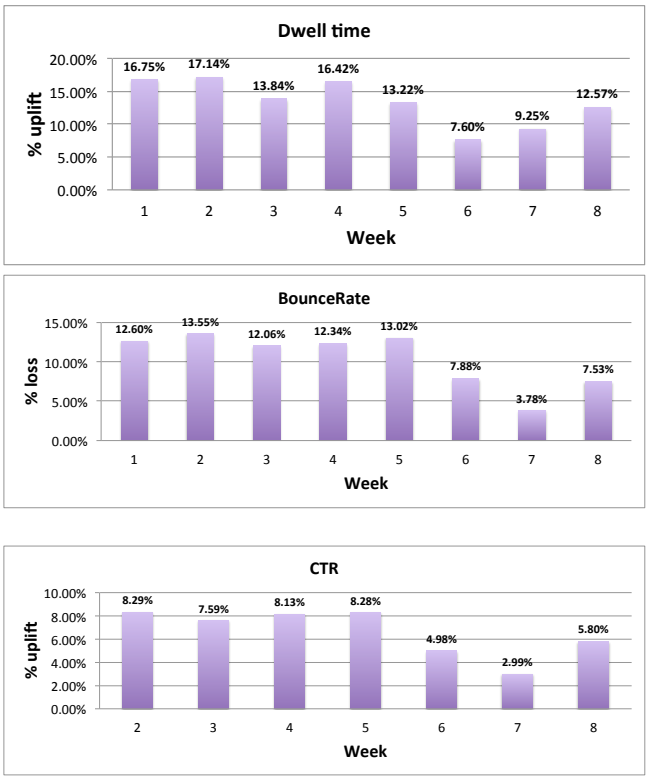


Figure 4: Online performances of AdQuality vs Control bucket.

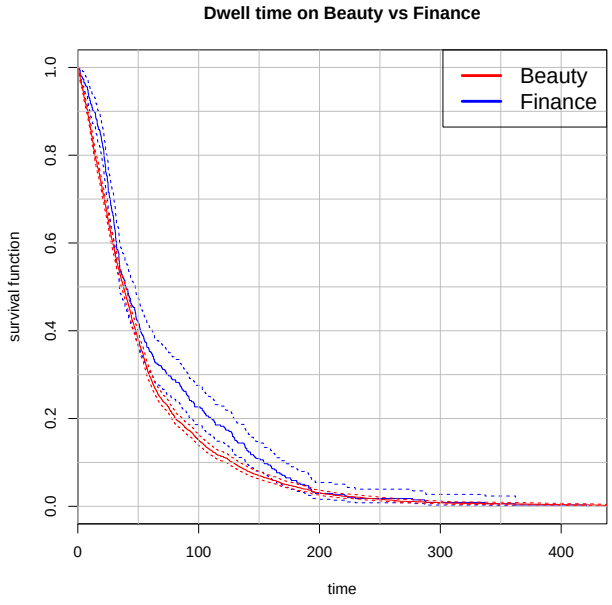


Figure 5: Survival functions associated with two separate ad categories: beauty and finance.

for instance [10], may bring additional perspectives in what makes a positive versus negative post-click experience.

The second direction of future work focuses on determining the threshold that separates low-quality ads from high quality ones. Variations in dwell time might be explained by considering the category of the ad (e.g. finance, health, entertainment). To support this we have plotted in Figure 5 the dwell time distribution, as estimated using Kaplan-Meier estimator, of two different ad categories: beauty and finance. As it can be immediately observed the two dwell time cumulative functions are totally different. Users, in fact, tend to spend more time on finance ads rather than beauty ads. As the confidence interval curves also hint this difference is also significant. Of course, one could compute different models corresponding to multiple threshold but this solution would not be flexible enough (the need to store multiple models) nor elegant.

Ads with different content are likely to perform in a different way: some contents are inherently more engaging than others or require a different time to lead to a conversion. To take into account this phenomenon, the low-quality ads should be considered as the ones for which the mass of the predicted distribution of dwell time is shifted towards low values, if compared to the one of "similar" ads. Similarly, different categories of users may have different levels of engagement with advertising. These considerations suggest that it is important to consider several thresholds rather than a global one, where each threshold is personalized at the user/ad level.

Finally, from a business oriented perspective, as the quality of an ad might not be necessarily correlated with the bid, it is important to dynamically balance the quality of the served ads with other benchmarks indicators related to the maximization of revenue. The survival random forest framework presented in this paper can be easily adapted and integrated with other prediction modules to address all these scenarios.

7. ACKNOWLEDGMENTS

We thank Tamar Lavee and Neetai Eshel from Yahoo Tel Aviv for the deployment of the model for the online A/B testing evaluation and for the guidance on how to analyze the data logged by the production system.

References

[1] Measuring the Fat Fingers Problem. `http://www.emarketer.com/Article.aspx?R=1009470`, 2012. [Online; accessed 23-July-2015].

[2] J. Azimi, R. Zhang, Y. Zhou, V. Navalpakkam, J. Mao, and X. Fern. Visual appearance of display ads and its effect on click through rate. In *CIKM*, 2012.

[3] M. Barris. Dwell time on mobile native ads twice as long as on desktop. `http://www.mobilemarketer.com/cms/news/research/20522.html`, 2015. [Online; accessed 23-July-2015].

[4] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: Quantifying the impact of landing pages in web advertising. In *CIKM*, 2009.

[5] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. Context transfer in search advertising. In *SIGIR*, 2009.

[6] K. E. Boudreau. Mobile advertising and its acceptance by american consumers. Bachelor thesis, 2013.

[7] A. Z. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR*, 2007.

[8] Y. Choi, M. Fontoura, E. Gabrilovich, V. Josifovski, M. Mediano, and B. Pang. Using landing pages for sponsored search ad selection. In *WWW*, 2010.

[9] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, volume 3953 of *Lecture Notes in Computer Science*, pages 288–301. Springer Berlin Heidelberg, 2006.

[11] M. de Sa, V. Navalpakkam, and E. F. Churchill. Mobile advertising: evaluating the effects of animation, user and content relevance. In *CHI*, 2013.

[12] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, pages 173–182, 2013.

[13] T. Foran. Native advertising strategies for mobile devices. `http://www.forbes.com/sites/ciocentral/2013/03/14/native-advertising-strategies-for-mobile-devices/`, 2013.

[14] H. Hohnhold, D. O'Brien, and D. Tang. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1849–1858, 2015.

[15] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

[16] A. Kae, K. Kan, V. K. Narayanan, and D. Yankov. Categorization of display ads using image and landing page features. In *LDMTA*, 2011.

[17] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[18] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *WSDM*, 2014.

[19] D. G. Kleinbaum. Survival analysis, a self-learning text. *Biometrical Journal*, 40(1):107–108, 1998.

[20] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 786–794, 2012.

[21] M. Lalmas, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users. In *KDD'15 Industry Track*. ACM, 2015.

[22] J.-H. Lee, J. Ha, J.-Y. Jung, and S. Lee. Semantic contextual advertising based on the open directory project. *ACM TWEB*, 2013.

[23] H. Liu, W.-C. Kim, and D. Lee. Characterizing landing pages in sponsored search. In *LA-WEB*, 2012.

[24] R. G. Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[25] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy-channel approach to contextual advertising. In *ADKDD*, 2007.

[26] W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972. ISSN 00401706. URL `http://www.jstor.org/stable/1267144`.

[27] R. J. Oentaryo, E.-P. Lim, J.-W. Low, D. Lo, and M. Finegold. Predicting response in mobile advertising with hierarchical importance-aware factorization machine. In *WSDM*, 2014.

[28] A. Penev and R. K. Wong. Framework for timely and accurate ads on mobile devices. In *CIKM*, 2009.

[29] L. Ritzel, C. V. der Schaar, and S. Goodman. *Native Advertising Mobil*. GRIN Verlag GmbH, 2013.

[30] R. Rosales, H. Cheng, and E. Manavoglu. Post-click conversion modeling and analysis for non-guaranteed delivery display advertising. In *WSDM*, 2012.

[31] D. Sculley, R. G. Malkin, S. Basu, and R. J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *KDD*, 2009.

[32] E. Sodomka, S. Lahaie, and D. Hillard. A predictive model for advertiser value-per-click in sponsored search. In *WWW*, 2013.

[33] L. B. Statistics and L. Breiman. Random forests. In *Machine Learning*, 2001.

[34] X. Yi, L. Hong, E. Zhong, N. Liu, and S. Rajan. Beyond clicks: Dwell time for personalization. In *RecSys*, 2014.

[35] P. Yin, P. Luo, W.-C. Lee, and M. Wang. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *KDD*, 2013.