Time-Aware Subscription Prediction Model for User Acquisition in Digital News Media

Heidar Davoudi *
davoudi@cse.yorku.ca

Morteza Zihayat *
zihayatm@cse.yorku.ca

Aijun An *
aan@cse.yorku.ca

Abstract

User acquisition is one of the most challenging problems for online news providers. In fact, due to availability of different news media, users have a lot of choices in selecting the news source. To date, most of digital news portals have tried to approach the solution indirectly by targeting the user satisfaction through the recommendation systems. In contrast, we address the problem directly by identifying valuable visitors who are likely potential subscribers in the future. First, we suggest that the decision for subscription is not a sudden, instantaneous action, but is the informed decision based on positive experience with digital medium. As such, we propose effective engagement measures and show that they are effective in building the predictive model for subscription. We design a model that not only predicts the potential subscribers but also answers queries about the subscription occurrence time. The proposed model can be used to predict the subscription time and recommend accurately the "potential users" to the current marketing campaign. We evaluate the proposed model using a real dataset from The Globe and Mail which is a major newspaper in Canada. The experimental results show that the proposed model outperforms the traditional state-of-the-art approaches significantly.

1 Introduction

Digital media and online news providers are facing the user acquisition challenge as the pressing issue more than before. In fact, from a business point of view, successful user acquisition can be directly translated to huge profits and values. However, whilst around 45% of people pay for a printed newspaper at least once a week, it has been much harder to persuade readers to pay for the online news subscription [10].

News recommender systems are widely exploited to improve the user experience, and consequently user acquisition indirectly. However, such systems mainly focus on recommending items that coincide with user's interests (to maximize the user's satisfaction) and do not identify potential subscribers and predict the subscription time. Identifying potential subscribers and predicting their subscription time are of paramount importance for news websites since it allows them to launch a targeted marketing campaign in advance. To the best of our knowledge, this problem has not been explored directly in the digital news media domain from data mining/machine learning perspectives, but rather considered in marketing studies which need a lot of human efforts.

The problem of identifying potential subscribers for news media from the data mining/machine learning point of view is facing several challenges. First, a decision for subscription is under influence of many factors such as demographical, social, or cultural circumstances. For example, one might decide to subscribe as she/he was referred by her/his friend (e.g., word of mouth), or based on her/his good experience. Finding an appropriate set of predictors for identifying and recommending such users (i.e., potential subscribers) is a challenging problem. Second, domain knowledge is extremely limited for *"the decision to subscription"* process (i.e., the knowledge acquisition bottleneck). In other words, domain experts do not have a clear idea on who subscribes and why/when a subscription occurs. Third, subscription should be considered in combination with the time dimension. In fact, the predictive model should identify the potential subscribers in a right time (i.e., neither soon nor late) since targeting a user who is either not ready to subscribe yet or no longer interested in subscription (while was previously interested) by any marketing campaign results in no subscription.

In this paper, we propose an end-to-end solution to address the aforementioned challenges in the problem of identifying potential users prone to subscription in news portals. First, we argue that the subscription act is not an instantaneously sudden decision, but rather an informed decision based on previous positive experiences. Accordingly, we propose a set of engagement measures as subscription predictors. The engagement measures are quantified in fully data-driven fashion, so we do not rely on the domain expert knowledge for their calculation. Then, we propose a *Time-aware Subscription Prediction (TASP)* model that combines the time dimension with the suggested predictors. The proposed model not only identifies and recommends the users who are very likely to become subscribers but also is able to predict their subscription time. In the TASP model,

*Department of Electrical Engineering and Computer Science, York University, Canada.

we treat subscription time as a dependent random variable and utilize generalized linear model to combine all engagement measures (i.e., independent random variables). Then, we cast the problem into an optimization problem aiming to maximize the likelihood of the proposed model. The learning algorithm is designed and parameters of model are learned respectively. Our main contributions are as follows:

- We define the problem of *time-aware subscription prediction for user acquisition* in news portals and design an end-to-end data-driven solution based on the data which are usually available in news portals.
- We propose effective user engagement measures as the main component of the subscription prediction model and show that they have a good predictive power to model subscription occurrence/time.
- We argue that time is an important factor in user subscription prediction and develop a probabilistic model to recommend the trustworthy potential subscribers. The proposed model predicts the potential users prone to subscription before a given time. Moreover, it can predict when the subscription occurs.
- The conducted experiments on a real dataset show the effectiveness of the proposed framework and the developed model in solving the problem of time-aware subscription prediction for user acquisition.

The rest of paper is organized as follows. Section 2 discusses the proposed framework for user acquisition. In particular, we present our Time-aware Subscription Prediction Model (TASP) in section 2.3. We outline the empirical evaluation in section 3 and discuss related work in section 4. Section 5 concludes the paper and present the future work.

2 Time-Aware User Acquisition in News Portals

Figure 1 shows an overview of the proposed framework for user acquisition in news portals. The framework consists of three main components: (1) *Data preparation*: most of news portals (e.g., The Globe and Mail[1]) use a data collection platform (e.g., Omniture by Adobe[2]) to capture the interactions with users. However, the captured data need to be preprocessed and aggregated before applying any learning algorithm (see §2.1). (2)

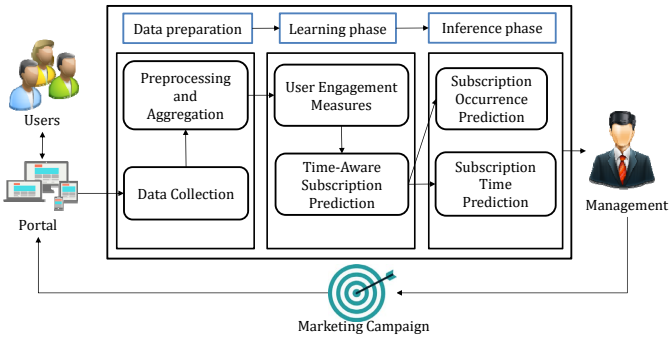


Figure 1: The proposed user acquisition framework.

Learning phase: given the preprocessed data, this component first finds a set of engagement measures (see §2.2) and then uses them to design the Time-aware Subscription Prediction (TASP) model (see §2.3). (3) *Inference phase:* as we learn the parameters of the proposed model, the interference models answer two type of questions: (i) time-aware subscription occurrence prediction: (i.e., what is the probability that a user becomes a subscriber by the given time t since the first visit?) (ii) subscription time prediction (i.e., when will a user become a subscriber since the first visit?). The inference outcomes can be utilized by the marketing campaign to boost user acquisition.

2.1 Data Preparation

In this section we present two main phases to prepare the data for user acquisition analysis: (1) Data collection, and (2) Preprocessing and aggregation.

2.1.1 Data collection: Every time a user reads an article, watches a video or generally takes an action in a news portal, the interaction is being tracked on the portal and is recorded as a *hit*. In data collection frameworks (e.g., Omniture), a hit simply shows a record in the data warehouse which contains rich information about the visitor and her/his actions. Typically, a hit contains information like date, time, user id (for a subscribed user), user environment variables (e.g., browser type, IP address), visited article, special events of interest like subscription, sign in, etc. Although the clickstream data are composed of billions of hits that tell what visitors have done in their visits, they contain a lot of noisy information needed to be cleaned properly.

2.1.2 Preprocessing and Aggregation: The data captured in any data collection platform contains a lot of low-level interactions (e.g., hits) mixed with a lot of noises. For example, spending a lot of time in a session does not necessarily mean that a user spends more time

[1] www.theglobeandmail.com
[2] https://my.omniture.com

on reading the clicked articles as the user might use multi-tab browser and is engaged in other activities. As such, we need to deal with both aggregation, and data cleansing as a part of data preparation.

Given that data are organized as hits, we roll-up the data from page view hits to visits and then to visitors. We refer to a *visit* as a set of page views in one "session" (a session is terminated if the data collection server does not hear from the same user for 30 minutes). We use cookie and the device's IP information which is anonymized and encoded in the data warehouse to detect the unique *visitors*.

Data collection platforms record a timestamp for each hit, so the difference between two consecutive page click timestamps can be utilized to calculate the time the user spent on an article. As usual in web analytics, the last article in a visit is ignored since we cannot estimate the time the user spent on it.

We filter out the unnecessary attributes which are not needed in calculation of user engagement defined in the next section (§2.2). We perform data cleaning by removing the outlier visitors whose engagement measures deviate more than 3 times the standard deviation from the mean of respective engagement measures in the data [5]. This helps us simply remove unreasonable values for the measures. Finally, all the engagement measures are normalized based on the z-score method.

2.2 User Engagement Measures As we suggest that user engagement have a close relationship with user acquisition, one important task in the proposed framework is to measure the user engagement. To understand the rationale behind the relationship, consider the scenario that we want to predict users prone to subscription based on the historical data stored as a clickstream collection. A reasonable assumption is that the user's decision on subscription is based on a long-term and short-term positive experiences rather than a sudden instantaneously thought. This is exactly related to the area of "user engagement" modeling. In fact, a well-known definition of engagement is based on "positive aspects" of user experience while interacting with an online application [7]. The positive aspects of experience are different among domains and applications and very hard to measure (e.g., visiting Twitter more frequently by a user in comparison to Facebook does not show essentially she/he has a better experience with Twitter due to differences in engagement patterns of these two social media). Moreover, other engagement measurement approaches such as self-reporting methods [8] (i.e., using questionnaires, surveys or interviews) and physiological methods [3] (i.e., utilizing observational methods such as facial expression or speech analysis) are based on a small number of users while assuming to be the representative of the whole population.

Alternatively, as we aim to have a fully data-driven framework, we propose the following simple but effective web analytics measures, inspired by [7], to quantify the user engagement and show that they have predictive power for subscription prediction in digital news media domain.

Total Number of Page Wall: In news portals which provide subscribed services, there is a restriction on the number of articles that a non-subscriber can read in a period of time. For example, in The Globe and Mail this period is one month. That is, as a visitor tries to read more articles, she/he is directed to a page asking for subscription (or login). This page is referred as a page wall. In our proposed approach, this interaction is used as an indicator of a user's interest in subscription. We calculate the total number of page walls each user hits in all of her/his visits.

Average Number of Page Walls per Visit: This measure is calculated by normalizing the total number of page walls by the number of visits.

Total Article Read: This measure is simply defined as the number of articles read by the user. There is difference between page visit and this measure. While in page visit we consider all of the pages (e.g., navigational or search pages), in this measure we only count article pages since they may better show the interest of users in contents and could be more close to the real user engagement, considering situations where, e.g., we count the number of page visits when a user visits a lot of navigational pages while looking for a single article.

Average Number of Articles per Visit: This measure is the number of articles read by the user normalized by the number of visits.

Average Spent Time per Article: The time a user spent on each article is calculated based on the method described in §2.1.2. The average time spent per article is calculated by dividing the total time that the user spent on articles by the number articles she/he visited. This measure roughly shows how much a user is interested in articles.

Average Spent Time per Visit: This measure is defined as the time that the user spent on visits divided by the number of visits. Each visit time is calculated based on the sum of time that the user spent on all articles during the respective visit.

Total Spent Time: The total spent time is measured as the sum of time that a visitor spent on each article during all her/his visits.

Although these measures are the indirect proxy of real engagement our experimental results show their effectiveness for user subscription prediction.

2.3 Time-aware Subscription Prediction Model (TASP): Given the set of engagement measures, in this section, we first outline the problem statement; then, in subsequent sections we describe our proposed Time-aware Subscription Prediction (TASP) model in details. We utilize the generalized linear model as the building block of the model. By assuming an underlying distribution for subscription time (i.e., Weibull), we cast the problem into the maximum likelihood optimization. Finally, we derive the solution to learn the parameters of the model.

2.3.1 Problem Statement: Given the processed data for all the users, we refer to the time period of this data set as "exploration period". We first remove the users who subscribed before the exploration period. The remaining users either subscribed during the exploration period (i.e., *subscribers*) or never subscribed either before or during the period (i.e., *non-subscribers*). Note that we do not consider the users who subscribed before the exploration period since we do not have their information before their subscription and our targeted problem is to build a model to predict how and when the unsubscribed users turn to subscribed ones.

DEFINITION 2.1. *(Subscription Occurrence Time): The subscription time $\widetilde{t_i}$ is defined as the time that passed since the first visit of user i until her/his subscription. Thus, given the absolute subscription time t'_i and the first visit time t_{f_i} for user i, $\widetilde{t_i}$ is computed as follows:*

$$\widetilde{t_i} = t'_i - t_{f_i} \tag{2.1}$$

The absolute subscription time refers to the timestamp that is recorded for each subscription. In our analysis, all timestamps are in day scale.

For non-subscribers, we define the possible subscription period as follows:

DEFINITION 2.2. *(Possible Subscription Period): We define $(\bar{t_i}, \infty)$ as the possible subscription period for user i, where $\bar{t_i}$ is defined as:*

$$\bar{t_i} = t_{li} - t_{f_i} \tag{2.2}$$

where t_{li} is the last visit time in the exploration period for a non-subscriber. Alternatively, $\bar{t_i}$ might be considered as the time that subscription might occur afterward since the first visit for the user i. Please note that if the subscription occurs we know the exact time of subscription ($\widetilde{t_i}$), whereas in the case that the subscription does not occur, all we know is that the subscription time exceeds $\bar{t_i}$.

The training set for the subscription time prediction problem is defined as follows:

$$L = \{(X_i, t_i, I_i) | i = 1, 2, \ldots, n\} \tag{2.3}$$

where $X_i = [x_{ji}]_{m\times 1}$ is the engagement measure vector for the user i (x_{ji} is the j'th engagement measure calculated for the user i, see §2.2). We calculate the user engagement measures for subscribers based on the visits before the subscription time and for non-subscribers based on the first visit till the last visit in the exploration period. For simplicity, the vector of X_i is append by 1 to address the bias in the linear system. I_i is defined as the indicator function which specifies whether user i subscribed during the exploration period or not:

$$I_i = \begin{cases} 1 & \text{if user i is a subscriber} \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

and t_i is defined as $\widetilde{t_i}$ for subscribed users (i.e., $I_i = 1$) and $\bar{t_i}$ for non-subscribed user (i.e., $I_i = 0$). We refer to this arrangement in §2.3.4 as we want to formulate the optimization problem.

Let T be a non-negative continuous random variable representing the waiting time for subscription occurrence since the first visit. We assume $f_T(t)$ be the probability density function and $F_T(t) = P(T < t)$ (p.d.f.) be the cumulative distribution function (c.d.f.) of subscription occurrence by time t.

Now we define the problem of user subscription time prediction as follows. Given training data L (Eq. 2.3), we want to estimate the cumulative distribution function $F(t) = P(T < t)$ for any subscription time t.

2.3.2 Generalized Linear Model: In order to make a connection between subscription time (i.e., variable of interest) and engagement factors, we first develop a generalized linear model. The generalized linear model bridges the gap between the probability distribution of subscription time and engagement factors calculated for each user and parameterize our model from observed data. Once the connection (i.e., a model) is established, we can predict the subscription time from the engagement behaviors.

Given vector X_i as the engagement measure vector (i.e., exploratory variables), subscription time observation for the user i is modeled as follows:

$$T_i = B^\intercal X_i + \epsilon \tag{2.5}$$

where ϵ is a stochastic residual coming from exponential family. The main idea is to model the expectation of subscription time as a function (i.e., link function) of

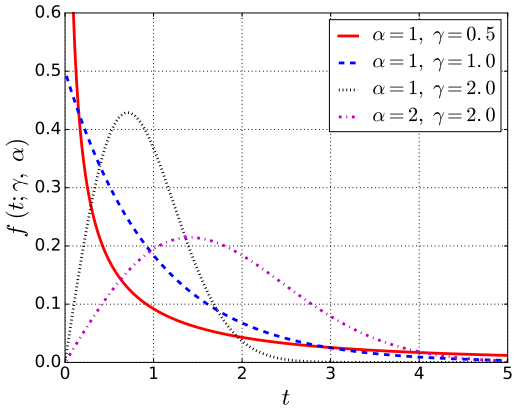


Figure 2: Weibull Distribution.

linear combination of engagement measures. So,

$$E[T_i] = g^{-1}(B^\intercal X_i) \tag{2.6}$$

To ensure the strict positivity of $E[T_i]$, we assume g is the exponential function:

$$E[T_i] \propto \exp\left(-B^\intercal X_i\right) \tag{2.7}$$

This assumption also helps us simplify the objective function introduced in §2.3.4. Please note that if we choose Gaussian or Bernoulli distribution, the model will be reduced to linear regression or logistic regression respectively.

2.3.3 Underlaying Distribution for Subscription Time: As our goal is to model relationship between user engagement and subscription time, we need to find the proper distribution for predicting the subscription time. The Weibull distribution has the flexibility to model right-skewed, left-skewed or even symmetric distributed data. Thus, we chose to use it in our model. It has been used in different domains to model the waiting time of an event [6]. The Weibull probability distribution for subscription time is as follows:

$$f_{T_i}(t;\gamma,\alpha) = \frac{\gamma}{\alpha_i}\left(\frac{t}{\alpha_i}\right)^{\gamma-1}\exp\left\{-\left(\frac{t}{\alpha_i}\right)^{\gamma}\right\} \tag{2.8}$$

where α_i and γ are scale and shape parameters respectively. The shape parameter γ can be learned to model the waiting time where the rate of event (i.e., hazard function) decreases ($\gamma < 1$), increases ($\gamma > 1$), or is constant ($\gamma = 1$) with time. Increasing the value of scale parameter (α_i) while holding shape parameter (γ) constant has the effect of stretching out the probability density function. Figure 2 shows the Weibull distribution for different parameters. The expectation of Weibull distribution is expressed as:

$$E[T_i] = \alpha_i\Gamma(1+\frac{1}{\gamma}) \tag{2.9}$$

where γ is the Gamma function. Given (Eq. 2.7), we can assume that:

$$\alpha_i = \exp\left(-B^\intercal X_i\right) \tag{2.10}$$

The cumulative distribution function is written as follows:

$$F_{T_i}(t) = P(T_i \le t) = 1 - \exp\left\{-\left(\frac{t}{\alpha_i}\right)^{\gamma}\right\} \tag{2.11}$$

Note that the distributions in (Eq. 2.8) have the same shape parameter γ, but different expectation values via parameter α_i. In fact, the basic assumption is that each value of a random variable T_i is drawn from a distribution indicated in (Eq. 2.8) where the expectation of distribution depends on the data point in (Eq. 2.9 and 2.10).

2.3.4 Optimization Problem: Assuming that observations (i.e., data points) are statistically independent and drawn from the distribution (Eq. 2.8), the log-likelihood of the model is formulated as follows:

$$\log\ell = \sum_{i=1}^{n}\{I_i \log f_{T_i}(t_i;\gamma,\alpha_i) + (1-I_i)\log P(T_i > t_i)\} \tag{2.12}$$

where t_i is the subscription occurrence time ($\widetilde{t_i}$) for the subscriber i (i.e., $I_i = 1$) and the start of possible subscription period ($\bar{t_i}$) for the non-subscriber i (i.e., $I_i = 0$). The basic idea is that subscribers contribute to the log-likelihood by the probability density function f_{T_i} while non-subscribers contribute to log-likelihood by the probability $P(T_i > t_i)$. If we plug in the probability density function in (Eq. 2.8) and the cumulative distribution function in (Eq. 2.11) into the log-likelihood function (Eq. 2.12), we can simplify the log-likelihood of model in vector format as follows:

$$\begin{aligned}\log\ell =& I^\intercal(\log(\gamma)\mathbb{1} + (\gamma-1)\log(T_s)) + \\ & \gamma\ I^\intercal\mathbf{X}B - \mathbb{1}^\intercal\exp\{\gamma\ (\log(T_s) + \mathbf{X}B)\}\end{aligned} \tag{2.13}$$

where $I = [I_i]_{n\times 1}$ is the indicator vector whose components are defined in (Eq. 2.4), $\mathbb{1} = [1]_{n\times 1}$ is the identity vector (i.e., all components are 1), $T_s = [t_i]_{n\times 1}$ is the vector of subscription time defined in (Eq. 2.3), $\mathbf{X} = [X_i]_{n\times m}$ is the matrix of engagement measures, where each row is X_i defined in (Eq. 2.3), and γ (scaler) and $B = [\beta_i]_{m\times 1}$ (vector) are parameters.

Algorithm 1 TASP Learning Algorithm

1: Initialize $B_1^{(0)}, B_2^{(0)}, \dots B_m^{(0)} randomly$
2: Initialize $\gamma^{(0)} = 1$
3: $t := 0$
4: **while** *not converge* **and** t $<$ *max iterations* **do**
5: $\quad B^{(t+1)} := B^{(t)} + \eta \nabla_B [\log \ell(B^{(t)}, \gamma^{(t)})]$
6: $\quad \gamma^{(t+1)} := \gamma^{(t)} + \eta \nabla_\gamma [\log \ell(B^{(t+1)}, \gamma^{(t)})]$
7: $\quad t := t + 1$
8: **end while**
9: **return** B, γ

2.3.5 Learning Algorithm: We use the gradient ascending method to maximize the log-likelihood and learn the parameters of the proposed model. First, we derive the gradient of log-likelihood of model (Eq. 2.13) with respect to γ and B. The gradient of model with respect to B is specified as follows:

(2.14)
$$\nabla_B [\log \ell(B, \gamma)] = \gamma \mathbf{X}^\intercal I - \gamma \mathbf{X}^\intercal \exp\{\gamma \ (\log(T_s) + \mathbf{X}B)\}$$

and gradient of log-likelihood with respect to the γ is derived as follows:

$$\nabla_\gamma [\log \ell(B, \gamma)] = I^\intercal \{(1/\gamma)\mathbb{1} + \log T_s + \mathbf{X}B\} -$$
(2.15)
$$(\log(T_s) + \mathbf{X}B)^\intercal \exp\{\gamma(\log(T_s) + \mathbf{X}B)\}$$

The overall procedure is outlined in Algorithm 1. We use the coordinate ascending method [14] to learn the B and γ iteratively. In step 5, we update the parameter B based on the gradient derived in (Eq. 2.14), then in step 6, keeping B fixed, the parameter γ is updated according to the gradient in (Eq. 2.15).

2.4 Inference models: After the parameters of the model (i.e., γ and B) are learned, inference with the model is straightforward. Particularly, we are interested in answering two types of questions: (1) what is the probability that a user be subscriber by the given time t since the first visit? (time-aware subscriber prediction) (2) when will a user be a subscriber since the first visit? (subscription time prediction).

2.4.1 Time-aware Subscription Occurrence Prediction: To find the users who will be subscriber by time t since the first visit, we need to estimate the $P(T \leq t)$. Given the user $\widehat{u}$ has a engagement vector $X_{\widehat{u}}$, we calculate the scale parameter $\alpha_{\widehat{u}}$ using (Eq. 2.10):

(2.16)
$$\alpha_{\widehat{u}} = \exp\left(-B^\intercal X_{\widehat{u}}\right)$$

The desired probability is calculated as follows:

(2.17)
$$F_T(t) = P(T \leq t) = 1 - \exp\left\{-\left(\frac{t}{\alpha_{\widehat{u}}}\right)^\gamma\right\}$$

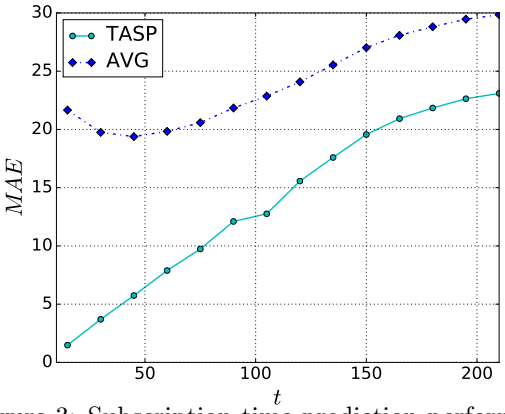


Figure 3: Subscription time prediction performance.

We consider $F_T(t) = P(T \leq t) \geq 0.5$ as the subscription occurrence.

2.4.2 Subscription Time Prediction: For the subscription time prediction, as the final distribution can be skewed, we propose to use the median as prediction time. This measure is less susceptible to outliers and extreme values and empirically performs better in our experiments. Given user $\widehat{u}$ with $X_{\widehat{u}}$ as the engagement vector, the subscription time t for the user $\widehat{u}$ is calculated as follows:

(2.18)
$$t_{\widehat{u}} = \alpha_{\widehat{u}} \log(2)^{\frac{1}{\gamma}}$$

where $\alpha_{\widehat{u}}$ is estimated using (Eq. 2.16).

3 Empirical Evaluation

In this section, we evaluate our proposed Time-aware Subscription Prediction (TASP) model and compare it with the state-of-the-art techniques as the baselines. We compare our model with Logistic Regression (LR), Random Forest (RF), Decision Tree (J48) and Naive Bayes (NB). We use the Mean Absolute Error (MAE) and F_1-Measure as performance measures for "subscription time" and "subscription occurrence prediction" accordingly. All the experiments in this section are based on the 10-fold cross validation. All the time values in the experiments are in day scale. We use The Globe and Mail dataset in our experiments. We set the learning rate η and maximum number of iterations (i.e., *max iterations*) in Algorithm 1 to 0.01 and 1000 respectively.

3.1 Dataset The Globe and Mail is the major news paper in Canada. In this news portal, interactions with users are captured using the Omniture data collection platform. The original data repository contains about 2 billions of hits (see §2.1.2) ranging from article reading behavior to video watching. We use the data from

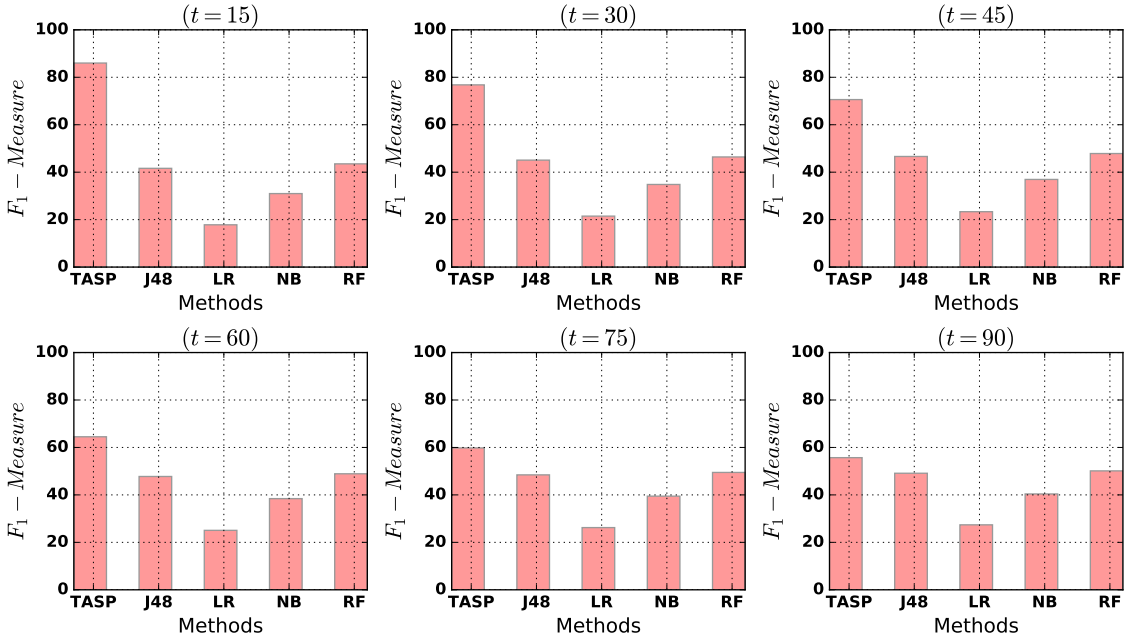


Figure 4: Subscription occurrence prediction performance (all time values are in days).

2014-01 to 2014-08 as the exploration period in our experiments. Since the data contains a lot of irrelevant information, the original data is processed to keep only the necessary information needed to calculate the proposed engagement measures. Then, we aggregate and preprocess the data based on the proposed steps described in (§2.1.2). The dataset used in the experiments contains 17,009 subscribers and 71,639 non-subscribers. Note again that the subscribers are the ones who subscribed during the exploration period.

3.2 Subscription Time Prediction: Figure 3 shows the results of subscription time prediction for the proposed model (TASP) and Average Time (AVG) as the baseline. Each point in the figure shows the MAE between the predicted subscription time and actual subscription time for users who subscribe before time t (all time values are in days). For the AVG model we calculate the average subscription time of visitors who subscribed before time t in the training set. Then, the MAE is calculated based on the difference between the actual subscription time of users in the test set and the respective average time value. As observed, MAE for the proposed method is much less than the AVG method for different t. In particular, for small values of t, the proposed model performs better than bigger time values, which means that the proposed method works better in short-time subscription time prediction than it does in longer term prediction although it performs better than the AVG method in both short term and long term.

3.3 Subscription Occurrence Prediction: Figure 4 shows the performance of TASP compared to the other baselines for different values of t. Each figure shows the performance of the different models in predicting the subscription occurrence before time t. Note that the proposed model (TASP) considers the time in the training stage and answers the queries about subscription with respect to the time (i.e., probability of subscription before given time t). Figure 4 shows that the proposed model outperforms the baselines for different t values. Moreover, it can be seen that the TASP model performs better in short-time subscription prediction. Among the baselines, tree-based models (i.e., J48 and Random Forest) perform the best and the Logistic Regression has the worst performance.

3.4 Imbalanced Sensitivity Analysis: In this section, we study the performance sensitivity of the proposed model under different portions of non-subscribers to subscribers as the training data. As such, we vary the portion of non-subscribers to subscribers by down sampling the non-subscribers. Figure 5 shows the performance of the proposed model (TASP) as well as the baselines for different portions of non-subscribers to subscribers (n_{ns}/n_s) where n_{ns} and n_s are the number of non-subscribers and subscribers respectively in the training set. The performance of the proposed model in

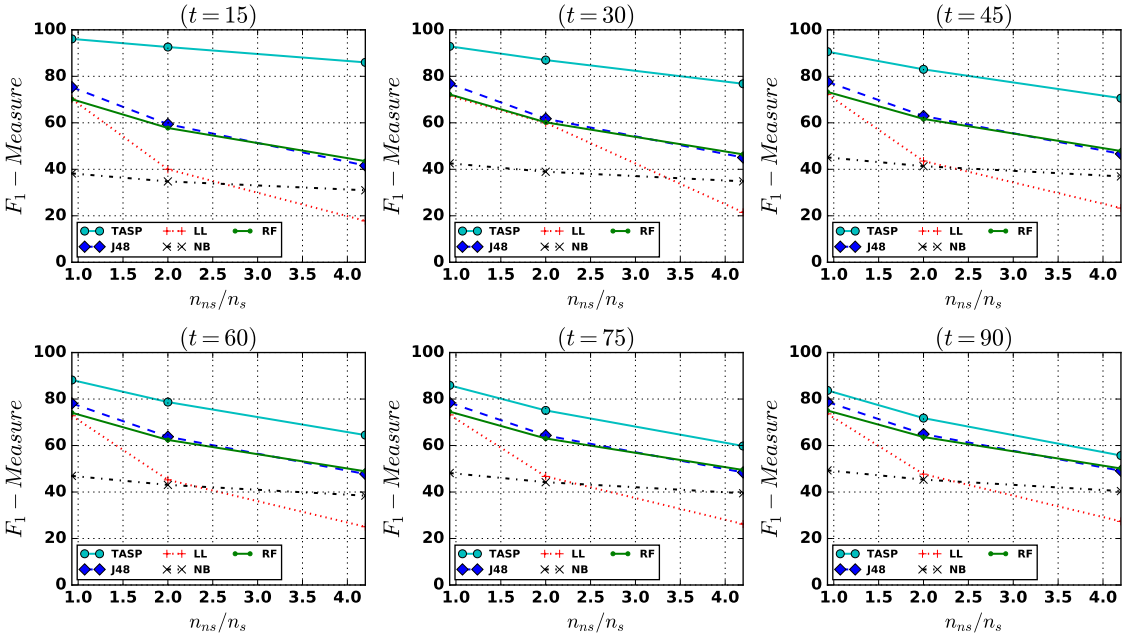


Figure 5: The subscription occurrence prediction performance sensitivity with respect to the number of non-subscribers to subscribers (n_{ns}/n_s).

predicting the subscriber is better when the dataset is balanced and consistently better than the baselines for different portions. As our model performance is better in the case that the dataset is balanced, we are aiming to embed a mechanism in our model to deal with imbalanced data as the future work. Figure 6 shows the performance sensitivity of the proposed model in predicting the subscription time. As it can be seen the MAE has a small sensitivity to portion of non-subscribers to subscribers in the training set.

4 Related Work

User acquisition traditionally is studied under area of Customer Relation Management (CRM) [12] where the main goal is to understand the customer behaviors and maximize the customer value to the organization in the long term. However, to date, most of efforts have focused on *user attraction*, *retention* and *churn management* rather than user acquisition.

Ng et al. [11] performed one of the first attempts on using data mining techniques for user retention in imaginative telecommunication domain (the real domain was unanimous due to privacy issue). They identified the objective indicators and used a decision tree induction method for the prediction purpose. In [1], authors proposed a rule-based evolutionary algorithm and applied it to predict churns in a telecommunication domain. They argued that interpretability was important in this problem, and their suggested rule-based method could address the issue by uncovering interpretable churn patterns.

In another work, Mozer et al [9] considered the problem of churn prediction for a major carrier company. They utilized features (overall 134 variables) such as call details records, billing information, application for service to predict the users churns. Three classes of predictive models: (i.e., decision tree, logit regression, and non linear neural network) were exploited and compared for the user churn prediction.

Kim et al. [4] conducted a research to measure the attractiveness (click values) of individual words for users. Assuming some words significantly induce more clicks than others, they proposed a generative model which jointly modeled headlines, contents of news articles as well as the clickstream data. The model was an extension to Latent Dirichlet Allocation (LDA) [2] whereas topic-specific click values of each word and clicked words were modeled using beta and binomial distributions respectively.

Customer Life Time Value (LTV) analysis is another related area to user acquisition. Customer LTV is usually defined as the total net income that a company expects from its customers. For example, Rosset et al. [13] calculated the current customer LTV based on three factors: customer value over time, length of service, and discounting factor. However, they estimated

the effects of retention campaigns on Lifetime Value and did not investigate how a visitor (e.g., non-subscriber) becomes a customer (subscriber).

In this paper, we consider the problem of user acquisition in the digital news portal domain. To best of our knowledge this is the first work that considers this problem in the digital news media domain and provides end-to-end solution for it. In particular, our proposed predictive model considers subscription time as a main component in both learning and inference process which has not been tackled before.

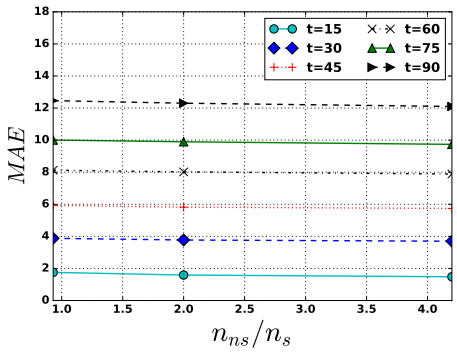


Figure 6: The subscription time prediction performance sensitivity with respect to the number of non-subscribers to subscribers (n_{ns}/n_s).

5 Conclusion and Future Work

User acquisition for digital news portals are one of the most pressing issue as the users are exposed to many available news sources. In this paper, we addressed the problem by predicting users who are prone to subscription in a given period of time. One important challenge is to define the measures that have enough power to predict the subscription (since the subscription is a complex decision depending on many factors). We simply showed that engagement measures had the good capability in predicting the subscription. The intuition is that the engagement as a positive experience has a direct impact on subscription. We proposed a time-aware prediction model that not only could predict the subscription in a given period of time, but also the subscription time. The empirical study on a real dataset showed that the proposed model performed well compared to the baseline models. In the future, we plan to improve and embed a mechanism in the model to deal with imbalanced data (for the situation that the number of subscribers to non-subscribers are very low). We will also investigate the capability of the proposed model in other domains.

Acknowledgements

We would like to thank The Globe and Mail for providing the dataset used in this research.

References

[1] W.-H. Au, K. C. Chan, and X. Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE transactions on evolutionary computation*, 7(6):532–545, 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661, 2008.

[4] J. H. Kim, A. Mantrach, A. Jaimes, and A. Oh. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2016.

[5] H.-P. Kriegel, P. Kröger, and A. Zimek. Outlier detection techniques. In *Tutorial at the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[6] C.-D. Lai, D. Murthy, and M. Xie. Weibull distributions and their applications. In *Springer Handbook of Engineering Statistics*, pages 63–78. Springer, 2006.

[7] M. Lalmas, H. O'Brien, and E. Yom-Tov. Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4):1–132, 2014.

[8] I. Lopatovska and I. Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*, 47(4):575–592, 2011.

[9] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, 11(3):690–696, 2000.

[10] N. Newman, D. A. Levy, and R. K. Nielsen. Reuters institute digital news report 2016. *Available at SSRN 2619576*, 2016.

[11] K. Ng and H. Liu. Customer retention via data mining. *Artificial Intelligence Review*, 14(6):569–590, 2000.

[12] E. W. Ngai, L. Xiu, and D. C. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[13] S. Rosset, E. Neumann, U. Eick, and N. Vatnik. Customer lifetime value models for decision support. *Data mining and knowledge discovery*, 7(3):321–339, 2003.

[14] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.