Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

On Machine Learning towards Predictive Sales Pipeline Analytics*

Junchi Yan[123], Chao Zhang[2], Hongyuan Zha[14], Min Gong[2], Changhua Sun[2], Jin Huang[2], Stephen Chu[2], Xiaokang Yang[3]

{yanjunchi,xkyang}@sjtu.edu.cn, zha@cc.gatech.edu, {bjzchao,gminsh,schangh,huangjsh,schu}@cn.ibm.com
[1]Software Engineering Institute, East China Normal University, Shanghai, 200062, China
[2]IBM Research – China, Shanghai, 201203, China
[3]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[4]College of Computing, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA

Abstract

Sales pipeline win-propensity prediction is fundamental to effective sales management. In contrast to using subjective human rating, we propose a modern machine learning paradigm to estimate the win-propensity of sales leads over time. A profile-specific two-dimensional Hawkes processes model is developed to capture the influence from seller's activities on their leads to the win outcome, coupled with lead's personalized profiles. It is motivated by two observations: i) sellers tend to frequently focus their selling activities and efforts on a few leads during a relatively short time. This is evidenced and reflected by their concentrated interactions with the pipeline, including login, browsing and updating the sales leads which are logged by the system; ii) the pending opportunity is prone to reach its win outcome shortly after such temporally concentrated interactions. Our model is deployed and in continual use to a large, global, B2B multinational technology enterprize (Fortune 500) with a case study. Due to the generality and flexibility of the model, it also enjoys the potential applicability to other real-world problems.

Introduction

Business-to-business (B2B) selling has evolved considerably over the last five decades from the in-person pitches depicted in the television series, to email and user profile-based deals, to customer relationship management (CRM) systems (Linoff and Berry 2011), and to the emerging trend of automatic sales analytics that allows the optimization of sales processes (Kawas et al. 2013).

Therefore, companies are adopting more systematic and digitalized sales management systems to support the sales process. The common pipeline operation model (Kawas et al. 2013) can be described as follows: As new sales leads are identified, the seller enters these leads into the sales opportunity pipeline management system. These leads are further evaluated and some are qualified into opportunities. A sales opportunity consists of a set of one or more products or services that the salesperson is attempting to convert into an actual client purchase. All open opportunities are tracked, ideally culminating in a "won" deal that generates revenue.

By collecting the up-to-date information about the pipeline, analytics approaches can be used to streamline the sales pipeline management. From the management perspective, the resource owner can reallocate their resources based on the pipeline quality assessment in comparison with their sales target or quota, which in turn, can also be dynamically adjusted based on the updated assessment result. From the individual salesperson perspective, assessment can further provide actionable advise to field sellers. By predictively scoring the quality of each lead at hand, it allows field sellers to better prioritize their personal resources and actions, in face of a relatively large number of ongoing leads within a tight period. These two situations are especially pronounced for companies having large and global client-facing sales teams dealing with increasingly complex portfolios of ever-changing products and services.

The fundamental building block to pipeline quality assessment is the lead-level win-propensity scorer. In fact, machine learning currently has not been widely applied to the B2B sales pipeline environment, or little technical work have been released from the business side. In practice, many internal pipeline systems, including the company that will be studied in our case study, typically ask the field seller to enter his subjective rating towards each of the leads that he owns. Then these fine-grained evaluations would be aggregated by accompanying with other factors to facilitate the decision making at different management levels.

However, such a subjective approach would unavoidably introduce noise. From our observation to the referred company, on one hand, many sellers intentionally manipulate the ratings in two ways: i) some leads are underrated by the seller in order to avoid the attention and competition from other sellers who may also have the channel to touch the clients behind the leads; ii) in contrast, some leads are overrated because the sellers suffer pressure from their leaders, who set different subtle performance metrics in a process oriented management fashion, not only for the final won revenue. Another drawback is different sellers may have biased personal expectations to similar leads. This fact is also common for human rating in many information retrieval applications, and is typically solved by asking pairwise comparison instead of entering a global score. However, such interfaces

*The work is partially supported by NSF IIS-1116886, NIH R01 GM108341, NSFC 61129001 and 61025005/F010403.

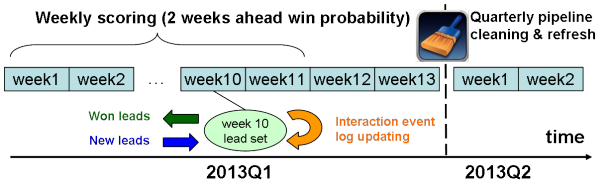

Figure 1: Business scenario illustration for the proposed scoring model. Win propensity for the next two weeks is issued on a weekly basis within one quarter, until week 11. Then new weekly scoring procedure starts as new quarter begins, by a pipeline cleaning step.

are not available to current pipeline systems.

From the domain application perspective, there is an extensive literature (Linoff and Berry 2011) in the field of marketing science in which various selling strategies are characterized and optimized, but the focus is on the business-to-consumer (B2C) domain rather than business-to-business (B2B). In fact, quantitative sales analytics in the B2B selling has recently been an emerging and active topic in both industry and research community (Lawrence et al. 2010; Varshney and Singh 2013). However, the above analysis is mostly performed in a *retrospective* data mining and knowledge discovery fashion. Specifically, (Lawrence et al. 2010) describes two practical solutions deployed in IBM tailored for identifying whitespace clients, i.e. OnTARGET and Market Alignment Program, by analyzing the data from external market. (Kawas et al. 2013) addresses the problem of efficient sales resource optimization in two steps. The first step involves using training samples of historically won sales opportunities, to estimate the sales response function between the salesforce's full-time equivalent (FTE) effort and the expected revenue or profit (Varshney and Singh 2013); The second step involves finding the optimal salesforce allocation subject to business constraints. To our surprise, few work has been done or released for involving predictive modeling in sales pipeline analytics, especially estimating the lead-wise win propensity.

This paper attempts to score the lead-level win-propensity for a given forward time window, by using the static profiles and dynamic clues from the sales pipeline. To this end, we propose a *profile-specific two-dimensional Hawkes processes model* tailored to the problem of estimating the lead-level win-propensity within a forward time window. Our model is able to incorporate the static profile features such as lead revenue size, product offering, client industry etc., as well as to capture the the dynamic influence from seller to lead, which is observed from their interactions activities including browsing, updating the client-visiting log. The model is implemented and deployed to a real sales pipeline business environment in a multinational Fortune 500 technology company across different products lines, and generated direct revenue impact which is estimated up to $43.2 million via internal evaluation in year 2013. The research team received the Research Accomplishment Award in year 2013 due to the recognition from the business side.

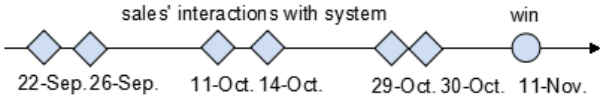

Figure 2: A seller-lead interaction event sequence ending by a win event. The interaction exhibits temporal clustering patterns and is won after a short period.

Learning the Proposed Model

We specify our problem and practice under the referred company, whereby the goal is to weekly issue and update the win likelihood of each lead, within the time window of the next two weeks since current week. In particular, the life-cycle of a sales lead is regarded as confined in a business cycle, which is one quarter in this paper. During one quarter (13 weeks), the lead is monitored and scored by the model until week 11. Those "non-won" leads in the end of current quarter would be treated in two ways: some of them are identified as garbage leads that would be removed in the beginning of next quarter; the rest would be refreshed as new leads together with those newly created ones in the next quarter.

Motivation and Problem Formulation

For this weekly updated scoring model, the input features on one hand consist of the static profile information such as deal size, geography, sector, product and other attributes as exemplified in Table 1. On the other hand, there is an additional dynamic clue which is in the form of a seller-lead interaction event sequence associated with each lead within a censored time window. This time window is usually set to the end of a recent previous quarter when building a training dataset, and up to now when performing model scoring on testing data. We would elaborate in more details for how to build the training and testing datasets in the rest of the paper.

There are several ways to transform the referred business problem into a machine learning problem. One straightforward way is using supervised binary classification. Given one quarter historical weekly snapshot data, one can define the labeled training dataset by lead profile features $f_1, f_2, \ldots, f_{11}$ including its past interaction by sellers, where the subscript denotes week number, and the corresponding "win" or "no-win" outcome within the corresponding two weeks ahead: $o_{34}, o_{45}, \ldots, o_{1213}$. Then Logistic classification model or other models can be applied.

This approach suffers several limitations: i) it truncates the observation window to an ad-hoc period which induces the label; ii) the binary classifier is not a dynamic model, and unable to capture the dynamics of the lead life-cycle flexibly. To improve this baseline approach, one way is to use a censored classification model e.g. (Shivaswamy, Chu, and Jansche 2007), or survival analysis model like Cox model (Cox and Oakes 1984) under the point process framework.

In this paper, we are motivated by the specific observation that a more indicative pattern comes from the interactions between sales and pipeline, where the interactions refer to different activities logged by the pipeline system when the seller visits the pipeline web portal, i.e. which lead he

Table 1: Exemplary features and data types of sales leads.

Profile	type	remark or examples
geography	categorical	Greater China, Southeast Asia
deal size	categorical	expected deal size in USD
sector	categorical	general business, industry clients
industry	categorical	health-care, energy and utility
product	categorical	Sub-brands of the main brand

is browsing/updating. More concretely, we find sellers usually focus on one or few certain leads thus (s)he may actively interact with them frequently within a short time period. Furthermore, as shown in Fig.2, such temporal clustering activities would also trigger "win" shortly. Thus it is appealing to suppose the interactions are prone to occur repeatedly shortly after a recent interaction event, so for the "win" event. Based on the above observations, it is desirable to capture the dynamic pattern of sales leads over time, preferably by a parsimonious parametric model to make the modeling interpretable and efficient. Note the conventional Cox model or its time-varying variants does not incorporate such recurrence pattern in the model, thus lacks of the flexibility in coping with such interaction event sequences.

Seller-pipeline Interaction Modeling

In the following, we will show how to model the win outcome's dependency on the interaction sequences using a two-dimensional point process model, see (Daley and Vere-Jones 1988) and the references therein. Specifically, we adopt the Hawkes process model (Hawkes 1971) to capture the temporal clustering dynamics. We will start with a brief description of one-dimensional Hawkes processes, and then extend it to the multi-dimensional case. In particular, our main work lies in proposing a profile-specific two-dimensional Hawkes processes model and a tailored alternating optimization algorithm to learn the model parameters. For mathematical tractability, the exponential kernel is used to model our seller-lead interaction modeling problem.

For the general Hawkes processes model, in its basic form as a one-dimensional point process, its conditional intensity can be expressed as (Hawkes 1971): $\lambda = \mu + a\sum_{i:t_i<t} g(t-t_i)$, where μ is the base intensity and t_i the time of events in the process before time t. $g(t)$ is the kernel to mimic the influence from the previous events. Given an event sequence $\{t_i\}_{i=1}^n$ observed in [0,T], its log-likelihood estimator is

$$\mathcal{L} = \log \frac{\prod_{i=1}^n \lambda(t_i)}{\exp\left(-\int_0^T \lambda(s)ds\right)} = \sum_{i=1}^n \log\lambda(t_i) - \int_0^T \lambda(t)dt$$

Extending the above equation to the U-dimension case, a multi-dimensional Hawkes process is defined by a U-dimensional point process, and its conditional intensity for the u-th dimension is (Zhou, Zha, and Song 2013b)

$$\lambda_u(t) = \mu_u + \sum_{i:t_i<t} a_{uu_i} g_{uu_i}(t-t_i)$$

where λ consists of a base intensity term μ_u and an accumulative exciting term $\sum_{i:t_i<t} a_{uu_i} g_{uu_i}(t-t_i)$. It can be interpreted as the instant probabilities of point occurrence, depending on the previous events across different dimensions.

Now we show how to formulate the problem into a specific machine learning paradigm. Suppose we have m samples, i.e. m independent event sequences $\{c_1, ..., c_m\}$ from the multi-dimensional Hawkes process, where each sample, in the form of c_s=$\{(t_i^s,u_i^s)\}$(i=1,...,n_s), is an event sequence of length n_s, occurring during the observation time window [0, T_s]. Each pair corresponds to an event occurring by dimension u_i^s at time t_i^s. We use the following formula for the log-likelihood of general multi-dimensional Hawkes processes whose parameters can be estimated via maximum likelihood estimation (Rubin 1972; Ozaki 1979)

$$\mathcal{L} = \sum_{s=1}^m \left(\sum_{i=1}^{n_s} \log\lambda_{u_i^s}(t_i^s) - \sum_{u=1}^U \int_0^{T_s} \lambda_u(t)dt \right)$$

By specifying the multi-dimensional Hawkes model for the intensity function, we obtain the following objective function (Liniger 2009) where $G_{uu_j^s}(t) = \int_0^t g_{uu_j^s}(t)dt$.

$$\mathcal{L}(\boldsymbol{\mu}, \mathbf{a}) = \sum_{s=1}^m \left(\sum_{i=1}^{n_s} \log(\mu_{u_i^s}(t_i^s) + \sum_{t_j^s<t_i^s} a_{u_i^s u_j^s} g_{u_i^s u_j^s}(t_i^s - t_j^s)) \right.$$
$$\left. -T_s \sum_{u=1}^U \mu_u - \sum_{u=1}^U \sum_{j=1}^{n_s} a_{uu_j^s} G_{uu_j^s}(T_s - t_j^s) \right)$$

Here we collect the parameters into vector-matrix formats, $\boldsymbol{\mu} = (\mu_u)$ for base intensity, and $\mathbf{a} = (a_{uu'})$ for the mutually exciting coefficients for dimension u affected by u'.

Learning Profile-specific Hawkes Processes

In this paper, we have U=2 processes to model: $u = 1$: interaction sequences and $u = 2$: the outcome event. Under this context, the base term incorporates the inherent interaction intention of salespeople to the leads - promising leads typically receive more attention from sales, and also better chance to win even no sellers' interaction is observed. The exciting term is used to properly account for the contributions from much earlier interaction event which may trigger subsequent interaction events that eventually lead to win. Specifically, the exciting effects are modeled to be decaying over time as $g_{ij}(t-t_0) = w_{ij}e^{-w_{ij}(t-t_0)}$.

Furthermore, our problem at hand bears several more *specific characters* to explore: i) the mutual influence is only one-way from the interaction dimension to outcome dimension rather than two-way, thus $a_{12} = 0$ and $w_{12} = 0$; ii) the self exciting phenomenon only exits for the interaction events since the outcome event is one-off, thus $a_{22} = 0$ and $w_{22} = 0$; iii) for the given lead s, the base intensity is assumed to be associated with its intrinsic attributes $\mathbf{x}^s = [x_1^s, x_2^s, \ldots, x_K^s]^T$ including deal size, channel, age, sales stage and other related profiles, which can be encoded by a parameter vector $\theta_u = [\theta_{u0}, \theta_{u1}, \ldots, \theta_{uK}]^T$ where θ_{u0} is a constant term. Therefore, the training set of leads with different profiles shall be heterogenous regarding with the base intensity. We chose the widely used Logistic function by a scaled coefficient μ_u^0 i.e. $\mu_u^s = \frac{\mu_u^0}{1+\exp(-\theta_u^T \mathbf{x}^s)}$ for both the interaction process (u=1) and the final outcome process (u=2). For two dimensions, the parameters of base intensity can be different as modeled by $\{\mu_1^0,\theta_1\}$ and $\{\mu_2^0,\theta_2\}$ respectively. Note this parametrization only assumes different

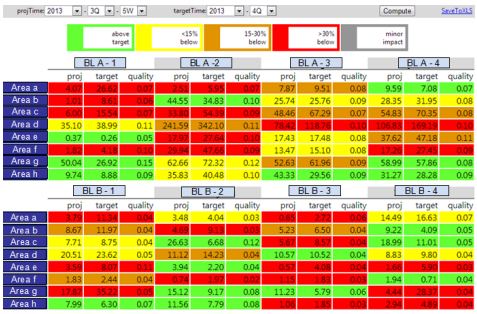

Figure 3: Pipeline quality and gap analysis web portal.

leads may have different base intensity, but it still assumes the base intensity is constant over time for a given lead.

The two above facts decouple the mutual influence to a reduced parameter space, while the third character addresses the heterogenous property of sales leads. Now we formulate the profile-specific decoupled two-dimensional exciting process as follows (by letting $h^s_{\theta_u} \triangleq h_{\theta_u}(\mathbf{x}^s) = \frac{1}{1+\exp(-\theta_u^T \mathbf{x}^s)}$)

$$\begin{aligned}
\mathcal{L} =& \mathcal{L}_1(\mu_1^0, \theta_1, a_{11}, w_{11}) + \mathcal{L}_2(\mu_2^0, \theta_2, a_{21}, w_{21}) \\
\mathcal{L}_1 =& \sum_{s=1}^{m} \left(\sum_{i=1}^{n_s-1} \log \left(\mu_1^0 h^s_{\theta_1} + \sum_{t_j^s < t_i^s} a_{11} g_{11}(t_i^s - t_j^s) \right) \right. \\
& \left. - T_s \mu_1^0 h^s_{\theta_1} - \sum_{j=1}^{n_s-1} a_{11} G_{11}(T_s - t_j^s) \right) \\
\mathcal{L}_2 =& \sum_{s=1}^{m} \left(\log \left(\mu_2^0 h^s_{\theta_2} + \sum_{t_j^s < t_{n_s}} a_{21} g_{21}(t_{n_s} - t_j^s) \right) \right. \\
& \left. - T_s \mu_2^0 h^s_{\theta_2} - a_{21} G_{21}(0) \right) \qquad (1)
\end{aligned}$$

Here the event associated with a lead consists of $n_s - 1$ interactions and the n_s-th event i.e. final outcome. Note that the above formulation decouples the first four terms from the remaining four terms regarding parameter $\mu_1^0, \theta_1, a_{11}, w_{11}$ (for self-exciting interaction sequence) from $\mu_2^0, \theta_2, a_{21}, w_{21}$ (for interaction's effect to outcome and its base intensity), thus we are seeking for estimating all the parameters rather than like the previous work (Zhou, Zha, and Song 2013a) that assumes the triggering kernel w is known and imposing additional regularization for the matrix of a_{ij} to be low rank and sparse. Below we show how to maximize the above $\mathcal{L}(\boldsymbol{\mu}^0, \boldsymbol{\theta}, \mathbf{a}, \mathbf{w})$ using an alternating optimization algorithm. Since the two terms $\mathcal{L}_1(\mu_1^0, \theta_1, a_{11}, w_{11})$ and $\mathcal{L}_2(\mu_2^0, \theta_2, a_{21}, w_{21})$ can be decoupled during optimization, in the following we give a strict derivation for the first term and a similar procedure is done to the second term.

Solving for a, w and $\boldsymbol{\mu}^0$ by fixing $\boldsymbol{\theta}$ $\mathcal{L}$ can be surrogated by its tight lower bound based on Jensen's inequality, which allows for the Majorize-Minimization (MM) algorithm (Hunter and Lange 2004) on the surrogate function:

$$\begin{aligned}
\mathcal{L}_1(\boldsymbol{\mu}^0, \mathbf{a}) \geq \sum_{s=1}^{m} \left(\sum_{i=1}^{n_s-1} (p_{ii}^s \log \frac{\mu_1^0 h^s_{\theta_1}}{p_{ii}^s} + \sum_{j=1}^{i-1} p_{ij}^s \log \frac{a_{11} g_{11}(t_i^s - t_j)}{p_{ij}^s}) \right. \\
\left. - \left(T_s \mu_1^0 h^s_{\theta_1} + \sum_{j=1}^{n^s-1} a_{11} G_{11}(T_s - t_j^s) \right) \right)
\end{aligned}$$

In the k+1-th iteration, we have $p_{ii}^{s\,(l+1)}$, $p_{ij}^{s\,(l+1)}$

$$p_{ii}^{s\,(l+1)} = \frac{\mu_1^{0(l)} h^s_{\theta_1}}{\mu_1^{0(l)} h^s_{\theta_1} + \sum_{j=1}^{i-1} a_{11}^{(l)} g_{11}^{(l)}(t_i^s - t_j^s)} \qquad (2)$$

$$p_{ij}^{s\,(l+1)} = \frac{\sum_{j=1}^{i-1} a_{11}^{(l)} g_{11}(t_i^s - t_j^s)}{\mu_1^{0(l)} h^s_{\theta_1} + \sum_{j=1}^{i-1} a_{11}^{(l)} g_{11}^{(l)}(t_i^s - t_j^s)} \qquad (3)$$

Given the lead s, p_{ij}^s can be interpreted as the likelihood that the i-th event (u_i, t_i) is affected by the previous j-th event (u_j, t_j) for interaction sequence associated with s and p_{ii}^s is the likelihood that i-th event is sampled from the background intensity. Moreover, its advantage is the parameter μ_1 and a_{11} can be solved in closed forms, and the non-negativity constraint of μ_1^0 is automatically satisfied.

Zeroing the partial derivative $\frac{\partial \mathcal{L}}{\partial \mu_1^0}$ and $\frac{\partial \mathcal{L}}{\partial a_{11}}$ leads to:

$$\mu_1^{0\,(l+1)} = \frac{1}{\sum_s h^s_{\theta_1} T_s} \left(\sum_{s=1}^{m} \sum_{i=1}^{n_s-1} p_{ii}^{s\,(l+1)} \right) \qquad (4)$$

$$a_{11}^{(l+1)} = \frac{\sum_{s=1}^{m} \sum_{i=1}^{n_s-1} \sum_{j<i} p_{ij}^{s\,(l+1)}}{\sum_s \sum_{j=1}^{n_s-1} G_{11}^{(l)}(T_s - t_j^s)} \qquad (5)$$

Meanwhile, we solve the estimation of the exciting kernel scale parameter w_{11} in $g(t - t_j) = we^{-w(t-t_j)}$: note $e^{-w(T-t_i)} \approx 0$ when $wT \gg 1$ as suggested in (Lewis and Mohler 2011) which shows w can be approximated by:

$$w_{11}^{(l+1)} = \frac{\sum_{s=1}^{m} \sum_{i>j} p_{ij}^{s\,(l)}}{\sum_{s=1}^{m} \sum_{i>j} (t_i - t_j) p_{ij}^{s\,(l)}} \qquad (6)$$

Solving for $\boldsymbol{\theta}$ by fixing a, w and $\boldsymbol{\mu}^0$ Given the fixed exciting term parameters and the base intensity scaling factor, we adopt gradient descent to solve the sub-problem with respect to variable θ_1. More specifically, by dropping off the constant term $\sum_{j=1}^{n_s-1} a_{11} G_{11}(T_s - t_j^s)$ in $\mathcal{L}_1$, we obtain the following objective which is a function w.r.t. θ_1:

$$\sum_{s=1}^{m} \sum_{i=1}^{n_s-1} \log \left(\mu_1^0 h^s_{\theta_1} + C_i^s \right) - T_s \mu_1^0 h^s_{\theta_1} \qquad (7)$$

$$\text{where} \quad C_i^s = \sum_{t_j^s < t_i^s} a_{11} g_{11}(t_i^s - t_j^s)$$

For the constant encoded by θ_{10}, the partial derivative is:

$$\frac{\partial \mathcal{L}_1}{\partial \theta_{10}} = \sum_{s=1}^{m} \left(\sum_{i=1}^{n_s-1} \frac{\mu_1^0}{\mu_1^0 h_{\theta_1} + C_i^s} - T_s \mu_1^0 \right) \frac{\exp(-\theta_1^T \mathbf{x}^s)}{1 + \exp(-\theta_1^T \mathbf{x}^s)} \qquad (8)$$

For the other coefficients in θ_1, the partial derivative is:

$$\frac{\partial \mathcal{L}_1}{\partial \theta_{1k}} = \sum_{s=1}^{m} \left(\sum_{i=1}^{n_s-1} \frac{\mu_1^0}{\mu_1^0 h_{\theta_1} + C_i^s} - T_s \mu_1^0 \right) \frac{x_k^s \exp(-\theta_1^T \mathbf{x}^s)}{1 + \exp(-\theta_1^T \mathbf{x}^s)} \qquad (9)$$

Algorithm 1 Learning profile-specific decoupled two-dimensional Hawkes processes for lead win-propensity estimation

Input:
observed training samples i.e. leads $\{c_s\}, \sum_{s=1}^{m}$ where each lead is associated with an interaction event sequence $\{t_i\}, \sum_{i=1}^{n_s-1}$ which is tailed with the "won" time stamp t_{n_s} if lead c_s is won within a certain period e.g. a full quarter;
Profile attributes $\mathbf{x}^s = [x_1^s, x_2^s, ..., x_K^s]^T$ that is associated with lead c_s, as exemplified in Table 1;
Initial value for $\mu_1^0, \theta_1, a_{11}, w_{11}, \mu_2^0, \theta_2, a_{21}, w_{21}$, l=0;
Iteration stopping threshold L, gradient descent step-size α;
Output: Learned parameters $\mu_1^0, \theta_1, a_{11}, w_{11}$ for the self-exciting model, and $\mu_2^0, \theta_2, a_{21}, w_{21}$ for the affecting model.
Procedure:
for $l = 1 : L_{max}$ **do**
    // Solving for a_{11}, w_{11}, μ_1^0 by fixing $\boldsymbol{\theta}_1$
    Update $p_{ii}^{s\,(l+1)}$, $p_{ij}^{s\,(l+1)}$ by Eq. (2) and (3);
    Update $\mu_1^{0\,(l+1)}$, $a_{11}^{(l+1)}$, $w_{11}^{(l+1)}$ by Eq. (4), (5), (6);
    // Solving for θ_1 by fixing a_{11}, w_{11} and μ_1^0
    Update $\theta_{1k}^{(l+1)}$ by Eq. (10) by the gradients in Eq. (8), (9);
end for
Apply the same method for solving $\mu_2^0, \theta_2, a_{21}, w_{21}$.

Apply gradient descent to update θ_{1k}:

$$\theta_{1k}^{(l+1)} = \theta_{1k}^{(l)} - \alpha \frac{\partial \mathcal{L}_1}{\partial \theta_{1k}}, \quad k = 0, 1, \ldots, K \tag{10}$$

Similar iterative scheme can be performed for the the term $\mathcal{L}_2$. Thus we finally obtain the estimations of $\mu_1, \theta_1, a_{11}, w_{11}$ and $\mu_2, \theta_2, a_{21}, w_{21}$ separately. The overall optimization algorithm is summarized in Algorithm 1.

Related Work and Contribution

The Hawkes process dates back to (Hawkes 1971; Ogata 1988). The model partitions the rate of events occurring to background and self-excited components. The background events are statistically independent of one another, while the offspring events are triggered by prior events. Its applicability for time-series or event sequence data has stimulated attentions of diverse disciplines, e.g. seismology (Ogata 1988; 1998), finance (Weber and Chehrazi 2012), criminology (Lewis et al. 2010; Mohler et al. 2011) and asset management(Yan et al. 2013b; Ertekin, Rudin, and McCormick 2013) and the references therein. In contrast to the above work focusing on one-dimensional Hawkes process, this paper aims to seeking a comprehensive formulation and effective algorithm for *profile-specific multi-dimensional* Hawkes processes, which is a relatively new topic with several very recent literature (Liniger 2009; Zhou, Zha, and Song 2013a; 2013b; Li and Zha 2014; 2013; Li et al. 2014).

Technical-innovation Compared with the above mentioned work related to Hawkes process, all parameters in our model are assumed unknown and estimated by our proposed algorithm. However, (Zhou, Zha, and Song 2013a) assumes the bandwidth of the self(mutual)-exciting kernel w_{ij} is known, and the background intensity μ_u being a constant parameter for all samples, which ignores the heterogeneity of μ_u in real-world problems. This simplification is also used in (Zhou, Zha, and Song 2013b; Li and Zha 2014) and the latter work instead parameterizes the mutual influence a_{ij} via latent variables to reduce the model space induced by a large number of dimensions for their social infectivity analysis. In contrast, we address the inherent heterogeneity by parameterizing lead attributes in a tractable optimization scheme. Note that for a given lead with known profile attributes, our model assumes the background is a stationary point process equaling to Poisson process. This is because our practical problem is confined in a relatively short business period, e.g. one quarter, thus the secular trend rarely exists. Thus we do not need perform the background model fitting using different non-stationary assumptions as used in (Lewis et al. 2010).

Impact to real-world problems As far as we know, this is the first work to establish a *modern* machine learning paradigm, i.e. profile-specific two-dimensional Hawkes Processes and learning algorithm for applications to the sales pipeline prediction. Though there is a few precedent statistical methods (Zliobaite, Bakker, and Pechenizkiy ; Chen et al. 2010) for sales analytics, while these methods and applications differ significantly from ours in that the historical event sequences (sales interaction) are not captured. For instance, one straightforward way is collecting the basic statistics of events over a certain time window such as sum, variance etc. However, this aggregation would cause information loss which hurts the potential towards more advanced predictive modeling. Furthermore, our method can also be easily generalized to other practical problems. For instance, in asset management, given a sequence of different types of failure events associated with the asset, $\{a_{ij}, w_{ij}\}$ can model the mutual impact between different failure types, and $\mu_u(\mathbf{x})$ can model the background failure rate related to the asset profile $\mathbf{x}$ and failure type u. We have seen the early success of recent work on predictive maintenance to urban pipe network (Yan et al. 2013b) and grid (Ertekin, Rudin, and McCormick 2013), whereby only a one-dimensional Hawkes process is adopted with a constant background rate which ignores the type of failures and the diversity of each sample. The proposed model in this paper is more promising as it is more flexible to incorporate the rich types of failures (e.g. leak, burst for pipe failure), as well as to handle the heterogeneity of background rate with a parameterized profile model (e.g. consider the diversity of material type, diameter, age for each pipe). Other potential applications can also be found such as client purchase life-cycle analysis where each type of items can take one dimension and the background rate is personalized by the customer profile features.

Deployment and Evaluation

We perform our study on a Fortune 500 multinational technology company in the B2B market environment. Throughout this section, due to the sensitivity of the proprietary company-owned selling data, we de-identified the brand name and other profile information, only leave relative metrics such as AUC score. Our model was finished in the end of 2013Q2. To make an unbiased performance evaluation, the model was evaluated in 2013Q3 with blind testing data

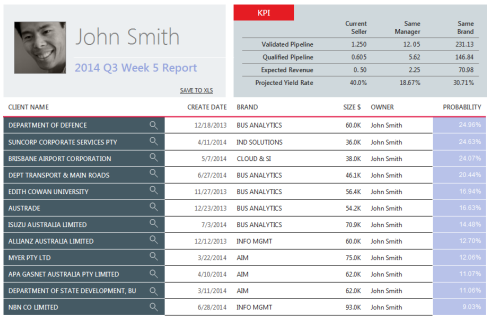

Figure 4: Web portal for 'next-two-week' win probability scoring. Sellers are able to view the latest scoring report.

in that quarter. In 2013Q4 it was released to sales team to impact sellers' decision, with the aim of transforming the selling ecosystem and methodology. For evaluation reported in this paper, due to business sensitivity, we randomly chose a subset lead set (100K-200K) for each quarter. The exact overall win rate is less than 20% and the win rate on the sampled set are disclosed in Table 2 and Table 3.

Application tools Before jumping into the detailed performance evaluation, we first present two downstream application tools derived by the proposed model. Fig.3 shows the pipeline quality and gap analysis given the quarterly quota target. This heating map, which covers various areas and product lines, by calculating the overall expected won revenue, is mainly used by sales leaders. Fig.4 illustrates another tool tailored for individual seller, especially for freshmen, who need some guidance to prioritize their workload.

Blind test As mentioned earlier, we use 2013Q2 data as the training set, and 2013Q3 as the testing set. In particular, in this case study, we chose two product lines across both mature market and emerging market. Apart from the baseline of sales subjective ratings, there are several machine learning methods being taken consideration in our evaluation: i) Logistic model, which extracts the sum and variance of past interaction events (whole time line so far and last five weeks) as additional input features besides profiles; ii) Cox point process model, where only profile information is used to model the hazard rate; iii) Constant background rate model similar to recent work Triggering Kernel Learning (TKL) (Zhou, Zha, and Song 2013a) that models the background rate using a constant parameter, and iv) our proposed background rate profile-specific Hawkes model.

Table 2 evaluates the AUC performances of ROC curve for these peer methods[1]. One can observe machine learning methods all outperform the subjective ratings, especially in emerging market due to the relatively young sales team there. The Logistic model and Cox model perform closely although the Cox model is assumed to be more suitable as it considers the observation window. In our analysis this is because Cox does not consider mutual exciting effect between interaction and win outcome. The simplified model by TKL

[1]In fact, we evaluate the model for each week by comparing the outcome in the next-two-week observation window. The average AUC over 11 weeks in that quarter are reported in this paper.

Table 2: AUC for win prediction on blind test data 2013Q3. The score for each lead is generated by integrating its win intensity λ_{win} over the next two weeks. 'BL' denotes 'Business Line', 'HW'('SW') for 'Hardware'('Software').

Market	BL	Lead #	Win%	Sales	Logit	Cox	TKL	Alg.1
New	HW	200K	15.7%	**.608**	.675	.678	.617	**.707**
New	SW	100K	12.5%	**.586**	.659	.661	.614	**.701**
Mature	HW	200K	19.5%	**.665**	.711	.718	.687	**.741**
Mature	SW	150K	14.8%	**.649**	.703	.709	.671	**.732**

(Zhou, Zha, and Song 2013a) that assumes the homogeneity of background rate for all leads, shows worse performance compared with the above two models, even it incorporates exciting effect. However, when it is combined by the base intensity personalization as solved in our model, it shows significant performance improvement. We argue that the relatively poor performance of TKL model further comes from the biased estimation of the background rate, which induces additional noise for learning the exciting parameters. Due to this limitation, the TKL model would also probably cause biased estimation in other applications such as (Yan et al. 2013b; Ertekin, Rudin, and McCormick 2013), which can be solved by our model by personalizing the background rate.

Interference test We release the scoring report to sales team in 2013Q4. Separate sales teams are receiving scoring reports generated by different models, and the performance is computed separately. For sales teams, they can make their resource allocation decision according to our predictions. The corresponding performance is reported in table 3. Compared with the blind test, our model still outperforms as it is likely that sellers' actions are influenced by prediction. They may invest more resource on the high-propensity leads judged by our model which induces regenerative effects between prediction and action. It is also worth noting that the sellers' estimation is improved compared with the blind testing data 2013Q3. Apart from the fluctuation across quarters due to other external factors, we do receive some feedback from sales team that some sellers would cross-check our report before enter their subjective ratings. This implies our report help sellers better evaluate their leads. We leave the analysis for how prediction and action influences each other to our long-term research agenda as it requires more data to calibrate other external factors.

Further discussion We believe we are still in the initial stage of advancing machine learning and AI in sales analytics which is a complex real-world problem yet relatively new to the computer science research community. Our model can further benefit from other data sources such as salesperson profile and activity, as well as marketing and promotion. New performance metrics beyond ROC AUC shall be studied. More comprehensive methodologies can be designed beyond the scope of this paper such as reinforcement learning (Kober and Peters 2012), or specifically Markov decision processes (MDPs) (White III and White 1989) and graph matching for resource allocation optimization (Tian et al. 2012; Yan et al. 2013a; 2014).

Table 3: AUC on interference test data 2013Q4.

Market	BL	Lead #	Win%	Sales	Logit	Cox	TKL	Alg.1
New	HW	200K	18.3%	**.628**	.681	.680	.612	**.729**
New	SW	100K	15.1%	**.618**	.672	.664	.604	**.715**
Mature	HW	200K	21.3%	**.689**	.727	.721	.693	**.751**
Mature	SW	150K	18.2%	**.680**	.731	.712	.689	**.749**

Conclusion

We have presented a modern machine learning method for sales pipeline win prediction, which has been deployed in a multinational Fortune 500 B2B-selling company. The proposed method is applicable to other real-world problems due to its generality and flexibility as discussed in the paper. We hope this paper can timely raise the wide attentions from industries as selling is essential to most business companies.

References

Chen, C.-Y.; Lee, W.-I.; Kuo, H.-M.; Chen, C.-W.; and Chen, K.-H. 2010. The study of a forecasting sales model for fresh food. *Expert Systems with Applications*.

Cox, D. R., and Oakes, D. 1984. *Analysis of survival data*, volume 21. CRC Press.

Daley, D. J., and Vere-Jones, D. 1988. *An introduction to the theory of point processes*, volume 2. Springer.

Ertekin, S.; Rudin, C.; and McCormick, T. H. 2013. Reactive point processes: A new approach to predicting power failures in underground electrical systems.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*.

Hunter, D. R., and Lange, K. 2004. A tutorial on mm algorithms. *The American Statistician* 58(1):30–37.

Kawas, B.; Squillante, M. S.; Subramanian, D.; and Varshney, K. R. 2013. Prescriptive analytics for allocating sales teams to opportunities. In *ICDM Workshop*.

Kober, J., and Peters, J. 2012. Reinforcement learning in robotics: A survey. In *Reinforcement Learning*. Springer. 579–610.

Lawrence, R.; Perlich, C.; Rosset, S.; Khabibrakhmanov, I.; Mahatma, S.; Weiss, S.; Callahan, M.; Collins, M.; Ershov, A.; and Kumar, S. 2010. Operations research improves sales force productivity at ibm. *Interface* 40(1):33–46.

Lewis, E., and Mohler, E. 2011. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*.

Lewis, E.; Mohler, G.; Brantingham, P. J.; and Bertozzi, A. 2010. Self-exciting point process models of insurgency in iraq. *UCLA CAM Reports 10* 38.

Li, L., and Zha, H. 2013. Dyadic event attribution in social networks with mixtures of hawkes processes. In *CIKM*, 1667–1672. ACM.

Li, L., and Zha, H. 2014. Learning parametric models for social infectivity in multi-dimensional hawkes processes. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Li, L.; Deng, H.; Dong, A.; Chang, Y.; and Zha, H. 2014. Identifying and labeling search tasks via query-based hawkes processes. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Liniger, T. J. 2009. Multivariate hawkes processes. *PhD thesis, Swiss Federal Institute Of Technology, Zurich*.

Linoff, G. S., and Berry, M. J. A. 2011. Data mining techniques: For marketing, sales, and customer relationship management. *Indianapolis, IN, USA: Wiley Publishing*.

Mohler, G. O.; Short, M. B.; Brantingham, P. J.; Schoenberg, F. P.; and Tita, G. E. 2011. Self-exciting point process modeling of crime. *Journal of the American Statistical Association* 106(493).

Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* 83(401):9–27.

Ogata, Y. 1998. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics* 50:379–402.

Ozaki, T. 1979. Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics* 31(1):145–155.

Rubin, I. 1972. Regular point processes and their detection. *Information Theory, IEEE Transactions on* 18(5):547–557.

Shivaswamy, P. K.; Chu, W.; and Jansche, M. 2007. A support vector approach to censored targets. In *ICDM*.

Tian, Y.; Yan, J.; Zhang, H.; Zhang, Y.; Yang, X.; and Zha, H. 2012. On the convergence of graph matching: Graduated assignment revisited. In *ECCV*.

Varshney, K. R., and Singh, M. 2013. Dose-response signal estimation and optimization for salesforce management. In *SOLI*.

Weber, T. A., and Chehrazi, N. 2012. Dynamic valuation of delinquent credit-card accounts. Technical report, EPFL-CDM-MTEI.

White III, C. C., and White, D. J. 1989. Markov decision processes. *European Journal of Operational Research* 39(1):1–16.

Yan, J.; Tian, Y.; Zha, H.; Yang, X.; Zhang, Y.; and Chu, S. 2013a. Joint optimization for consistent multiple graph matching. In *ICCV*.

Yan, J. C.; Wang, Y.; Zhou, K.; Huang, J.; Tian, C. H.; Zha, H. Y.; and Dong, W. S. 2013b. Towards effective prioritizing water pipe replacement and rehabilitation. In *IJCAI*.

Yan, J. C.; Li, Y.; Liu, W.; Zha, H. Y.; Yang, X. K.; and Chu, S. M. 2014. Graduated consistency-regularized optimization for multi-graph matching. In *ECCV*.

Zhou, K.; Zha, H.; and Song, L. 2013a. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*.

Zhou, K.; Zha, H.; and Song, L. 2013b. Learning triggering kernels for multi-dimensional hawkes processes. In *ICML*.

Zliobaite, I.; Bakker, J.; and Pechenizkiy, M. Towards context aware food sales prediction. In *ICDMW'09*.